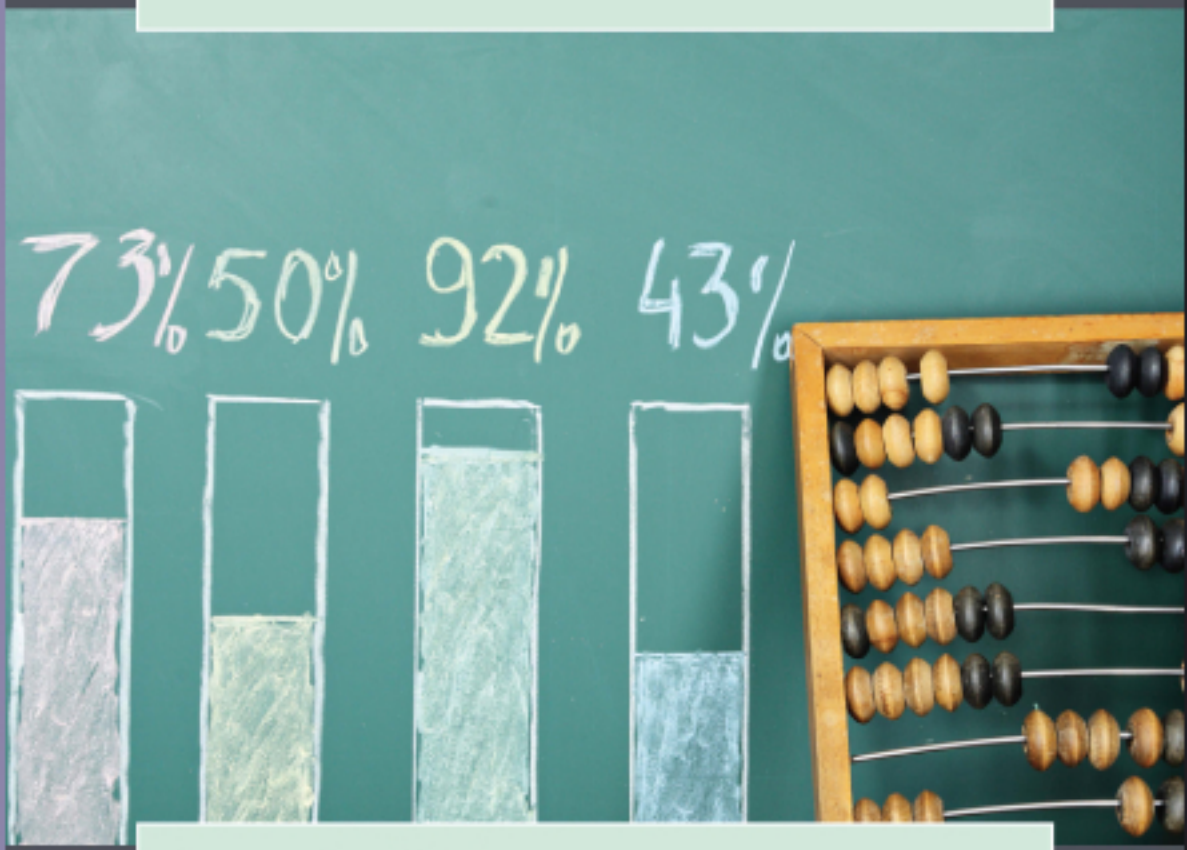
Statistics and Probability in High School

Carmen Batanero and
Manfred Borovcnik



SensePublishers

Statistics and Probability in High School

Statistics and Probability in High School

Carmen Batanero
Universidad de Granada, Spain

and

Manfred Borovcnik
University of Klagenfurt, Austria

SENSE PUBLISHERS
ROTTERDAM/BOSTON/TAIPEI

A C.I.P. record for this book is available from the Library of Congress.

ISBN: 978-94-6300-622-4 (paperback)
ISBN: 978-94-6300-623-1 (hardback)
ISBN: 978-94-6300-624-8 (e-book)

Published by: Sense Publishers,
P.O. Box 21858,
3001 AW Rotterdam,
The Netherlands
https://www.sensepublishers.com/

All chapters in this book have undergone peer review.

Printed on acid-free paper

TABLE OF CONTENTS

PREFACE

Research in statistics and probability education has produced a variety of results that would be useful for both secondary and high-school mathematics teachers and the educators of these teachers. Although there are many good textbooks in different countries that describe statistical ideas with a formalisation level adequate for students, usually these textbooks are written in a sequential way so that the different concepts and procedures are introduced in turn, with insufficient connections between them and limited attention to students' underlying intuitions.

There are, of course, excellent exceptions such as the books produced by the Schools Council (1980) in Statistical Education Project in the 1980's; yet, even, these textbooks do not include a detailed summary of research related to the teaching of the concepts, which started to get shape only after the first International Conference on Teaching Statistics in Sheffield in 1982.

In the later stages of our careers and, after collaborating and corresponding for many years in different projects, we decided to write a book directed to reinforce the mathematical and didactical knowledge of high-school teachers in statistics and probability. At the same time, we wish to offer examples of potential activities useful to introduce the main statistics and probability concepts and enhance the underlying ideas at this school level.

Consequently, in this book we provide examples of teaching situations, while at the same time we review research on adolescents' stochastical[1] reasoning and literacy, with the aim to provide recommendations and orientations for teaching these topics within high-school mathematics. The expression "high school" relates to different educational levels depending on the country; in this book, we will consider students from ages 14 to 18 (grades 9–12 in the United States of America curriculum). The book is organised in five chapters:

In the first chapter, we present some principles we use to select the content analysed in the book and the approach to teach this content. These principles emerge from:

a. Our own teaching and research experience;

b. An analysis of stochastic high-school curricula in several countries (e.g., ACARA, 2010; NCTM, 2000; CCSSI, 2010, MEC, 2007);

c. The synthesis of available research (as summarised, for example, in Biehler, Ben-Zvi, Bakker, & Makar, 2013; Chernoff and Sriraman, 2014; Garfield & Ben-Zvi, 2008; Jones, 2005; Jones, Langrall, & Money, 2007; Shaughnessy, 1992, 2007; Shaughnessy, Garfield, & Greer, 1996);

[1] In some countries the term stochastics is used to highlight the mutual dependence between probabilistic and statistical knowledge and reasoning. Throughout the book we occasionally use stochastics for statistics and probability to express our view that these fields are tightly interconnected and should be taught together.

d. Our own conceptions of statistical and probabilistic literacy, thinking, and reasoning; and

e. Our extensive experience with strategies that may help support student development in stochastic literacy, thinking, and reasoning.

The first chapter sets out key educational principles. Each of the following chapters (Chapters 2–5) has a focus on a group of related fundamental stochastic ideas, while taking into account that high-school stochastics should be built on basic ideas that students have encountered at primary and middle-school levels. These chapters are organized according to a common structure, including an introduction, with a short analysis of the main stochastic ideas in the particular topic and its place in the curriculum; some initial and more advanced specific examples that may serve to involve learners actively as they progress in their development of the concepts, a summary of what is known about difficulties students encounter with the related concepts, a synthesis of the main learning goals in the chapter, and finally, some additional resources that may help teachers and students. When possible, we make connections between the different chapters and include some historical notes that shed light on ways of thinking about the concepts.

We have tried to give a balanced view on probability and statistics, with a focus on the interrelated nature of the concepts, integrating probabilistic ideas at a level suitable for high school teaching, including the step from descriptive statistics to statistical inference. Where ever we could do it, we have also tried to integrate mathematical concepts and contexts so that the mathematics developed becomes meaningful for the learners. May our exposition contribute to an increase in statistical and probabilistic literacy in our societies.

We hope the book will be both useful for practising teachers, as well as for researchers in statistics education and practitioners in teacher educators (teacher trainers). The different chapters contain original materials, but build upon our extended set of publications, part of which is listed in the references.

We thank our colleagues and students who have commented several drafts of the chapters. Among them we want to name especially two who accompanied us in our research work now for decades: Juan D. Godino and, particularly, Ramesh Kapadia who was also helpful for improving the English. Finally, we would like to express our deepest gratitude to our families and friends for their encouragement and support over the years when we were writing the book.

May, 2016
Carmen Batanero and Manfred Borovcnik

CHAPTER 1

EDUCATIONAL PRINCIPLES FOR STATISTICS AND PROBABILITY

In this chapter, we describe those principles, which reflect our view of how stochastics should be taught at high-school level. Firstly, we suggest the need to focus on the most relevant ideas for the education of students. Secondly, we analyse the complementary nature of statistical and mathematical reasoning on the one side and of statistical and probabilistic reasoning on the other side. We then examine the potential and limits of technology in statistics education and reflect on the different levels of formalisation that may be helpful to meet the wide variety of students' previous knowledge, abilities, and learning types. Moreover, we analyse the ideas of probabilistic and statistical literacy, reasoning and sense making. Finally, we demonstrate how investigations and experiments provide promising teaching strategies to help high-school students to develop these capabilities.

1.1. INTRODUCTION

We are surrounded by uncertainty that affects our lives in personal, biological, social, economic, and political settings. This fact suggests that we need to understand random phenomena to make adequate decisions when confronted with uncertainty. We meet arguments based on data in everyday life; to critically evaluate these arguments, we need to understand the way in which the data is produced and how the conclusions are obtained.

Three widely accepted reasons for including stochastics at the school level are the essential role of statistics and probability in critical reasoning, its instrumental role in other disciplines, and its key role for planning and decision making in many professions (Franklin et al., 2007; Gal, 2002, 2005; Wild & Pfannkuch, 1999). These reasons have been recognised for some time and, hence, the teaching of statistics at secondary school has a history of about 30 years in many countries, for example, in Australia, Austria, France, Israel, Germany, Spain, and in the US. A recent innovation is the extension of statistics teaching to lower grades so that it is now included throughout the curriculum starting from the first year of primary school to graduate courses and postgraduate training at universities.

Statistical ideas originate from the natural sciences and demography and have transformed many fields of human activity over the past three centuries (Gigerenzer et al., 1989; Hacking, 1990). Today, statistics is pervasive; there is scarcely a political, scientific, or social issue without reference to statistical results. People encounter statistical information while shopping, watching TV, reading a newspaper, or surfing on the Internet. Furthermore, national statistics offices and international agencies such as the United Nations (UN) or the World Health Organisation (WHO) make their statistical studies available on the Internet.

Statistical methods are important not only in various disciplines in science but also for government and business systems, e.g., health records or retirement pension plans. It is essential to understand statistics and probability to critically evaluate how statistics are generated and to justify decisions, be they societal or personal (Hall, 2011). Accordingly, statistics educators, educational authorities, as well as statistical offices and societies call for a statistically literate society and support projects that help children and adults to acquire the competencies needed in the era of data information.

The aim of this chapter is to clarify what is meant by statistical literacy, statistical thinking, statistical reasoning, and sense making. These aspects of learning statistics have been widely discussed in the Statistical Reasoning, Thinking, and Literacy (SRTL) Research Forum, a series of conferences (srtl.fos.auckland.ac.nz/) starting in 1999 as well as in Ben-Zvi and Garfield (2004) and Garfield and Ben-Zvi (2008). We also suggest that students can acquire relevant competencies through statistical projects and investigations.

Teaching statistics and probability at high-school level is often embedded within mathematics. However, due to its peculiarities, statistics and probability require special attention on the part of teachers and curriculum designers in relation to the selection of content and the best way to make the statistical ideas accessible to the students. The goal of Chapter 1 is to present our overall perspective on the teaching of statistics and probability. This perspective is made more explicit in Chapters 2 to 5 that deal with the teaching of the main ideas of this subject at high-school level.

1.2. FUNDAMENTAL IDEAS IN STATISTICS AND PROBABILITY

Given that the time available for teaching is limited, it is important to select the key concepts that should be taught carefully. Several authors (e.g., Borovcnik & Kapadia 2014a; Borovcnik & Peard, 1996; Burrill & Biehler, 2011; Heitele, 1975) have investigated the history of statistics, the different epistemological approaches to probability and statistics, and the curricular recommendations in different countries, as well as the educational research. Based on their studies, they have proposed various lists of "fundamental" ideas. These ideas can be taught at different levels of formalisation depending on students' ages and previous knowledge (see, e.g., Borovcnik & Kapadia, 2011).

Our suggestions to teach the statistics content (Chapters 2 to 5) are organised around four main clusters of fundamental ideas. We consider these clusters as key foci around which activities can be organised by teachers in order to help students acquire the related key concepts. Each chapter analyses the essential content of one cluster using paradigmatic problems or situations that can be put forward to the students. When appropriate, we make connections across the content in the other chapters. In these chapters, we also inform teachers about the learning goals implicit in the activities, point out potential difficulties encountered by learners as described by research, and suggest promising teaching resources and situations that embed the ideas within instruction. A summary of the fundamental ideas included in each of these chapters follows.

1.2.1. Exploratory Data Analysis (Chapter 2)

Basically, statistics deals with collecting and analysing data, and making decisions based on data. The starting point in a statistical study is a real-world problem that leads to some statistical questions requiring data in order to be answered. To address the questions, such data may already exist; yet, often new data must be produced to provide sufficiently valid and reliable information to make a judgement or decision. Unlike mathematics, answers to statistical questions always involve an element of uncertainty. Furthermore, in contrast to mathematics, in statistics the context of the data is a critical component in the investigations.

For example, when studying the volume of a cylinder in mathematics, the same formula always applies no matter whether the cylinder is a juice can or part of a building. In statistics, however, the type of data and the context are essential for choosing the appropriate method of analysis and for interpreting the results. The challenge of statistical problem solving is the need to interpret the statistical concepts within a given context and choose the most suitable method from a variety of possible methods that may be applied for the problem. Although the computations in statistical procedures can often be completely outsourced to software, the decision about which procedures and techniques should be used and the interpretation of the results within the context of the data remain a big challenge when teaching statistics.

Today, there is a large amount of *data* accessible on the Internet on almost every topic that may interest students. This helps to facilitate working with real data, which can increase students' levels of motivation (see, e.g., Hall, 2011; Ridgway, 2015). Real data sets also help students investigate issues that are rarely mentioned in traditional problems in textbooks. For example, students can explore different ways of collecting data, design their own questionnaires or experiments, and gain a better understanding of different data types. When doing their own investigations, students can encounter additional problems in managing missing or incomplete data, come across data that are atypical (or even wrong due to a failure in the collection process). They have to refine the data set accordingly, or assess the reliability and validity of the investigated data.

Representations of data play a major role in statistics as a variety of graphs can be used to display and extract information that may be hidden in the raw data. The process of changing the representation of data in order to find further information relevant to the initial problem is called *transnumeration* and is considered to be an important process in statistical reasoning (Wild & Pfannkuch, 1999).

Variation and distribution are two complementary concepts that play a vital role in statistics. Although variables and variation also appear in many areas of mathematics, mathematics focuses on functional (deterministic) variation while statistics deals with random variation. Hence, a goal of statistics education is to enable students to reason about data in context under conditions of uncertainty, and to discriminate between statistical reasoning and mathematical reasoning. Wild and

Pfannkuch (1999) suggest that the perception of random variation is an essential component of statistical thinking. Moreover, statistics provides methods to identify, quantify, explain, control, and reduce random variation.

Distribution is a term that is specific to statistics and probability; it is a collection of properties of a data set as a whole, not of a particular value in the data set. A distribution consists of all the different values in the data including the frequencies (or probabilities) associated with each possible value. Variation and distribution are linked to other fundamental statistical ideas such as centre (as modelled by mean, median, or mode), *spread* (as modelled by standard deviation or variance), and *shape* (for example, bi-modal, uniform, symmetric, or L-shaped). Measures of *centre* summarise the information about a distribution while measures of spread summarise the variability within the data. Each value of a variable shows some deviation from the centre. In the context of measuring an unknown quantity (signal), this deviation may be interpreted as an error in measurement (noise). Metaphorically, the distribution embodies an overall "model" for potential errors or noise, while the centre can be seen as the signal (Konold & Pollatsek, 2002).

1.2.2. Modelling Information by Probabilities (Chapter 3)

Descriptive statistics, probability theory, and statistical inference complement each other and view the same information from different angles. Descriptive statistics investigates the information from *one* sample (data set) and summarises it using single numbers and graphical displays. Statistical inference goes beyond the present data set and tries to generalise the findings, that is, to transfer them to a wider population to which the data set belongs. Probability takes the role of a mediator as it supplies a justification for a generalisation beyond the initial data.

The information extracted from the data by using descriptive statistics can only be generalised by inferential methods that are established by probability models. At the same time some probability concepts are more easily understood as a generalisation of descriptive statistics (e.g., the idea of expectation can be understood as a generalisation of the idea of mean). These links have influenced and shaped school curricula. Yet, the study of statistics is incomplete if the reference to probability is missing in our teaching. The fundamental purpose of probability is twofold: to *measure* or to *evaluate* uncertainty. The first idea considers probability as a property of physical objects like length, while the second notion points towards a qualitative procedure of subjectively attaching numerical values to uncertain phenomena. In the book, we take into account these two ideas and the different meanings of probability.

In recent years, there has been a shift in the way probability is taught at school level, from a classical Laplacean (or "axiomatic") approach (common until the 1980s) towards a frequentist conception of probability, that is, an experimental approach where probabilities are estimated from long-range relative frequencies (Batanero, Chernoff, Engel, Lee, & Sánchez, 2016). Simulations and experiments are used to support students in understanding the Law of Large Numbers and

grasping the sophisticated interaction between the notion of *relative frequency* and the *frequentist conception of probability*. The *subjectivist view* of probability, which is widely used in applied statistics, has been developed hand in hand with the frequentist view so that the two complement each other (Hacking, 1975). Their interplay is relevant, especially for conditional probability. Bearing this in mind, we suggest a combination of both approaches in the teaching of probability.

The fundamental ideas of variable and distribution still apply for probabilistic modelling. However, probability distributions deal with the *potential* data of a "random experiment", which models how data will be generated, rather than with *actual* data that have been collected. A helpful idea is to think of repeated experiments that supply us with idealised frequencies. This metaphor helps us to transfer many concepts from descriptive statistics to probability and delivers a more concrete picture of what probability means.[1] A variety of representations of descriptive statistics can be transferred to probabilities. Furthermore, tree diagrams may be used to simplify the discussion of combined random experiments and the calculus of probabilities. All of these approaches are presented in Chapter 3.

1.2.3. Exploring and Modelling Association (Chapter 4)

In a statistical study, we often are interested in checking whether two (or more) variables are interrelated and whether some type of function may describe their interaction. Functions occupy a central place in mathematics and are used to describe connections between variables in many fields (such as economy or physics) in situations that have the following features. First, the function may be determined by general laws. Second, in practical experiments these functions may only be slightly blurred by small measurement errors so that the underlying functional relationship is still visible from graphs of the data. In physics, for example, the value of the independent variable *determines* a specific value of the dependent variable.

Independent (explanatory) and dependent (response) variables occur in mathematics and statistics but the link between them is not as strong in the statistical context as it is in mathematics. Data on variables such as heights of fathers and sons might show a similar tendency but its pattern is less clear than in data on variables that follows a relationship in physics; yet, a description of the interrelationship might be useful. When we study the relation between smoking and the occurrence of lung cancer, we deal with qualitative variables. Moreover, while the percentage of people with lung cancer is greater among those who have previously been smokers – which indicates that it is better not to smoke – this association cannot be interpreted directly in causal terms as it could be induced by third factors (hidden variables) that may be operating in the background. One example is that problems in handling stress can lead people to smoke and make them more prone to lung cancer.

[1] In Chapter 3, we also see that in many situations probability models are applied to one-off decisions.

Three different problems can be studied when modelling statistical relations between variables:[2]

1. Are two variables interrelated? This leads to the concepts of association and correlation.
2. Is there a mathematical function that may be used to describe the relationship and is it possible to justify the criteria in selecting such a function? This is called the regression problem. At high-school level, teaching is usually restricted to linear functions to describe the relations between the variables as they are easier to interpret.
3. How well does a specific function describe the relationship between the variables? This leads to the problem of finding a suitable measure of the goodness of fit. For general functions, the coefficient of determination, which coincides with the square of the correlation coefficient for linear functions, is used.

Regression becomes more complex if the influence of several variables on one (response) variable is analysed. Therefore, the mathematical techniques needed to find suitable functions to model multivariate relationships between the target variable and several independent variables are usually outsourced to statistical packages. Yet, people who perform the analysis need to understand the basic ideas behind these methods. They also should understand that data becomes more reliable if a special design to control for third factors is followed. As the topic is so relevant, elementary parts of it are contained in school curricula with just two variables under scrutiny, while supported by technology both for the required calculations and for drawing specific graphs. The interpretation of the results can be sophisticated and is often a challenge for teaching.

1.2.4. Sampling and Inference (Chapter 5)

The fundamental idea of statistical inference is to generalise information from data sets to wider entities. It is related to *inductive logic*, which deals with principles of generalising empirical information (Hacking, 1965; Popper, 1959). Although mathematics derives true statements by rules of ordinary logic, in statistical inference the generalisations are constructed on the basis of hypotheses that include probabilities. In spite of being a very young field of study,[3] statistical inference has paved the way for our evidence-based society.

Fundamental ideas of inference are *population* and *sample* and their interrelationships. We might observe a process of data generation for some time (e.g., the weight of fabricated units in a factory) or a sample of data from a subset of a finite population (e.g., the weight of some eight-year-old children in a

[2] Chapter 4 is restricted to the descriptive study of the topic although all of these problems can be generalised with inferential methods.

[3] Apart from rudimentary earlier traces, the methods were developed between 1920 and 1950. The successful axiomatisation of probability by Kolmogorov in 1933 increased the prestige of statistical methods, which are today applied in every area of human activity.

country). In both examples we might be interested in a number that describes the average weight of the population. In the first case the focus is on all future units produced, whereas in the second, the focus is all current eight-year-olds in the country.

An important idea is that of sample representativeness. If there are no biasing factors in how elements are selected, then the average weight of the sample data should be a reliable estimate of the average (expected) weight of the population. The statistical way to "guarantee" representativeness is to control the sampling process. More precisely, *random sampling* is a suitable method to obtain a representative sample, and it is possible to calculate the probability of obtaining a biased sample. The techniques for generalising the information from samples to the whole population are confidence intervals and tests of hypotheses.

Today, statistical inference has found its way into curricula all over the world with a variety of approaches that attempt to make the methods and the inherent notions more accessible to students. Specifically it is common to use simulation to facilitate parts of the computation and to visualise the sampling variability. More recently, resampling approaches have been used to simplify the probability models implicit in inference methods by focussing entirely on the data set that has to be analysed. All of these methods, as well as Bayes' rule to update information from empirical data, are presented in Chapter 5.

In conclusion, the most fundamental objective in probability and statistics is to offer models to understand and interpret real observations. A model does not completely represent reality; yet, it can be used for explorations that may lead to relevant results or insights about real data. A fundamental goal of teaching statistics is that students understand the *hypothetical character of probability and statistical models* and the possibility of applying these models in many different contexts.

1.3. COMPLEMENTARY VIEWS OF STATISTICS AND MATHEMATICS

Today, the teaching of statistics at university level is often separated from mathematics. For example, in countries like Spain or the US, distinct degrees or graduate programmes are offered in the training of mathematicians and statisticians. In other countries such as Austria or Germany, there are stronger connections between mathematics and statistics.

Research in statistics and probability (stochastics) is promoted by the International Statistical Institute and other organisations with specific conferences (e.g., the World Statistics Congress) and journals (*Annals of the Institute of Statistical Mathematics* or *Computational Statistics*). Statistics education research has received wide input from areas different from mathematics, for example, psychology, mathematics education, and general education (see Batanero, Chernoff, Engel, Lee, & Sánchez, 2016; Jones, 2005; Jones, Langrall, & Mooney, 2007; Shaughnessy, 1992, 2007; Shaughnessy, Garfield, & Greer, 1996; Vere-Jones, 1995, for a detailed description). As discussed in Section 1.6.3, more

recently the field of statistics education has grown to become a discipline in its own right.[4]

The specific character of statistics and probability is also reflected in the philosophical, ethical, procedural, and even political debates that are still ongoing within these areas and their applications, which does not often happen in mathematics.[5] Statistics and probability are closely related to other sciences such as demography, genetics, insurance, astronomy, and law, from which many statistical methods were developed. Furthermore, inferential statistics has formed the basis for a new scientific paradigm in social sciences, medicine, and other disciplines. These disciplines share the process of a scientific argument for *generalising empirical results* that leads beyond the subjectivity and the restrictions of a single experimental situation.

Subsequently, we describe the components of an empirical study, which starts with a contextual question (in Section 1.7.2). This question leads to an appropriate design with a corresponding statistical question as well as a plan on how to collect the data needed to address the question according to the chosen design. This careful planning is necessary in order to obtain useful information about the initial question and to keep the information free from confounding effects. The exploration and analysis of the data are followed by drawing some conclusion from the data (Wild & Pfannkuch, 1999). As we expose in Chapter 2, a crucial final step is the interpretation of the results in relation to the initial question given the context of the problem (see also the steps outlined in a statistical investigation in the GAISE project in Franklin et al., 2007).

The main interest in applying statistics in a real-world study concerns finding, describing, and confirming patterns in data that go beyond the information contained in the available data. Thus, statistics is often viewed as the science of data or as a tool for conducting quantitative investigations of phenomena, which requires a well-planned collection of observations (Gattuso & Ottaviani, 2011; Moore & Cobb, 2000). This feature of statistics explains why it is easy to establish connections between statistics and other school subjects and why it has sometimes been argued that statistics should be taught outside the mathematics classroom (Pereira-Mendoza, 1993).

Statistics is part of the mathematics curricula. As argued by Usiskin (2014), statistics involves a great deal of mathematical knowledge. Fitting statistics within mathematics teaching means that no additional time is needed for a separate subject in school, where time is limited. According to Usiskin, as well as to Scheaffer (2006), statistics and mathematics may support and complement one another in the school curriculum.

---

[4] For example, in the US, there are graduate programmes in statistics education and doctoral degrees have been awarded in statistics education.

[5] An example is the controversy around the use of statistical tests (Batanero, 2000; Borovcnik, 1986a; Hacking, 1965).

An example described by Usiskin is the natural extension from the straight line that goes exactly through two points to linear regression where the interest is in finding a line that fits to more than two data points in an optimal way. Finding the line of best fit is part of mathematical modelling; however, statistical modelling does not stop there. Lines are compared to other functions that may also model the data and the interpretation of the fitted function depends on the context. Typical questions are:

Do other functions fit better? What does it mean if we describe the relation between the two variables under scrutiny by a line? Can we predict the value of the dependent variable outside the range of data?

According to Usiskin, statistics education can also benefit from the strong movement towards modelling developed since the late 1980s especially promoted by the ICTMA, the International Community of Teachers of Mathematical Modelling and Applications,[6] and introduced only more recently in statistics education through the Exploratory Data Analysis (EDA) movement.[7] In this modelling approach, the motivation for a particular concept emerges from the context; the concepts are developed interactively by contextual considerations and mathematical principles. The teacher should find an appropriate real situation in which the new concept makes sense for the students. Throughout the book we present such initial problems that help students understand the related concepts, followed by more complex situations when space permits.

Statistics at high school may sometimes be a separate course (e.g., the advanced placement course in the US) or included in other subjects and taught when it is needed to understand the current topics. We agree with Usiskin that regardless of the placement, all students should experience substantial school work in statistics. In this way, they become competent to appreciate and criticise empirically-based arguments around them and empowered to make adequate and informed decisions.

Probability versus statistics. Some authors consider that probability is more strongly linked to mathematics than to statistics. Throughout the book, we try to make clear that statistics and probability complement each other and should not be completely separated. The frequentist view of probability serves to connect probability with a wide range of applications and probability cannot be understood without a connection to relative frequencies. Moreover, modelling probability as the limit of relative frequencies provides a first basis for introducing statistical inference from a frequentist approach.

The subjectivist view of probability extends the applications of probability to decision making in one-off situations where the frequentist view does not apply. Additionally, Bayes' rule establishes strong links between these two perspectives on probability and allows combining subjective probabilities *and* statistical data to

[6] The PPDAC cycle of Wild & Pfannkuch (1999) in Section 1.7.2 has close connections to the modelling cycle.

[7] Although EDA also started in the 1980s, it was then mainly a movement towards methods that were attractive to teaching because of their simplicity.

update prior probability judgements and make them "more objective". Finally, some probability ideas are needed to understand more informal approaches to inference when we use simulations or re-randomisation to generate empirical sampling distributions (see Chapter 5).

Thus, when teaching probability, one has to consider that probability is a theoretical concept – a virtual concept according to Spiegelhalter and Gage (2014) – and we speak about probability by using metaphors such as "propensity", "degree of belief", or "limit of frequencies", which convey only parts of this abstract concept. Even though the relationship between probability and relative frequencies is fundamental for the comprehension of probability and statistical methods, this relationship is not always well understood as some students confuse frequency with probability.[8] Moreover, the different representations of measures of uncertainty (as absolute numbers[9] versus probabilities) may also involve different levels of difficulty in understanding probability models. In Chapter 3, we discuss other difficulties encountered by the students when interpreting small probabilities, which usually occur in the case of risks of adverse events such as a maximum credible accident of a nuclear power station or dying from lung cancer.

1.4. THE ROLE OF TECHNOLOGY

Technology has revolutionised the applications of statistics and likewise statistics education. With software such as *Fathom* (Finzer, 2007) or *Tinkerplots* (Konold & Miller, 2011), specially designed to support the learning of statistics and probability, with a spreadsheet, or even with Internet applets, data analysis is no longer the exclusive domain of statisticians (Biehler, 1997; Biehler, Ben-Zvi, Bakker, & Makar, 2013; Pratt, Davies, & Connor, 2011). As demonstrated throughout Chapters 2 to 5, software can facilitate computations and the production of graphical representations of data. Thus, students can use methods such as fitting a variety of models to a scatter plot (see Chapter 4) that were not accessible to them a few years ago.

With modern technology, students can represent abstract interrelationships and operations, interact with the setting, and see the changes in the representation or in the results when varying some data or parameters.[10] Technology provides the possibility of dynamic visualisations where the impact of crucial parameters on a graph can be traced. This technique may serve, for example, to explore the waiting time for the first success in Bernoulli experiments (as done in Chapter 3).

Due to facility and speed of computations, the size of data sets is no longer a limitation so that it becomes easier to use real data collected by the students or taken from the Internet (e.g., from CensusAtSchool, n.d., or from several statistical

[8] The expression "empirical probability" is unfortunate in this regard as a probability is always theoretical; only the frequencies are empirical.

[9] In the sense of Gigerenzer (1994).

[10] Another possibility not studied in the book is multivariate dynamic visualisation that can be implemented with tools like those available from Gapminder (Rosling, 2009).

offices). In Chapter 2, we suggest that students collect and analyse physical measures of themselves such as height and weight, arm span, shoe size, etc. However, teachers should be careful about issues such as students being able to understand the data within the context and to connect the data set to the investigated problem. Otherwise students may lose interest (Hall, 2011).

Another advantage of technology is that it can help build micro worlds where students can explore conjectures and establish ideas from analysing what happens. Some examples are given in Chapter 3, where students can explore simulations of experiments to understand essential properties of a frequentist concept of probability and perceive that the variability of relative frequencies becomes smaller in longer series of trials. Working with such environments may help students to realise and change their probability misconceptions that would persist within a formal approach to probability (see Jones, 2005; Jones, Langrall, & Mooney, 2007).

Furthermore, technology offers the opportunity to students to learn about modelling since it enables students to build their own models to describe data or to explore probability problems. For example, a table-oriented method allows calculating the posterior from prior probabilities and given data, which is a method vital for a Bayesian decision-oriented approach towards inference (see Chapter 5). Above all, technological support may enhance students' understanding of the complementary role of probability and statistics; for this reason, it is particularly useful in the introduction of the frequentist view of probability (Chapter 3) and in obtaining empirical sampling distributions with simulation or re-randomisation to explore inference from data to populations (Chapter 5).

The use of technology, however, may also have drawbacks. It may hide the mathematical difficulties of a concept using artefacts and change hands-on activities on physical objects (e.g., spinners) into virtual activities ("spinners" run by software). For example, drawing a scatter plot requires quite a few steps including attributing the variables to the coordinates, deciding the scale of the axes, and plotting the points. These steps build an operative understanding of the final plot. If technology automatically supplies these operations, students may use the default options (e.g., an inadequate scale) in an uncritical way.

Moreover, technology may bias educational efforts. On the one hand, teachers may use technology to overemphasise computation neglecting the understanding of the statistical concepts. On the other hand, the learning goals may be reduced to the learning of simulation techniques not paying attention to the concepts or the reasoning behind the method applied. Another drawback of the overuse of simulation is the reduction of probability to a mere frequentist concept. To avoid such problems, we put special emphasis on analysing the learning goals underlying the activities proposed for the work with the students in the classroom and, throughout the book, we use technology as a complement to rather than a substitute for statistical reasoning.

1.5. ADAPTING THE LEVELS OF FORMALISATION TO THE DIVERSITY OF STUDENTS

Various factors suggest the need to be able to teach the same topic at different levels of formalisation. Among them, we list the diversity of curricular guidelines around the world, the different educational requirements of similar high-school grades (e.g., technical versus social-science strands), as well as the differing abilities and competencies of the students. As the book is intended to serve a broad international audience, we have tried to implement the principle of adapting to a diversity of students throughout the text.

Let us consider, for example, statistical inference, a quite sophisticated topic, where starting with an informal approach has been recommended to reduce the technicalities when introducing the topic (e.g., Makar, Bakker, & Ben-Zvi, 2011; Rossman, 2008). This informal approach could be used as an introduction for the majority of students. However, in some countries (e.g., in New Zealand, Germany, or Spain), a more formal approach to inference (including an exposition of confidence intervals) is required in the last grade of high school for particular strands.

We therefore start the exposition of basic inference methods like tests of significance and confidence intervals in Chapter 5 with an informal approach where the sampling distribution is estimated via simulation. We then later suggest that students with experience in probability rules or the binomial distribution may use this previous knowledge to compute the exact sampling distribution. We add further activities related to statistical tests as decision making or use of Bayes' theorem to update prior information about a parameter for those students with more advanced experience in probability.

Other resources considering the same curricular content at different standards of formalisation can be found in the GAISE guidelines (Franklin et al., 2007). They provide examples of how different levels of presentation require and reflect an increasing sophistication in understanding and applying stochastic concepts.

1.6. STATISTICAL AND PROBABILISTIC LITERACY

Statistics is embedded in a methodology that serves to generate evidence from data. Statistical knowledge is vital in research and in public discussion where it is used to empower arguments pro or contra some issue. Without statistical knowledge it is difficult to discern misuse from proper use of data. Statistical knowledge involves thinking in models, being able to apply proper models in specific situations, considering the impact of assumptions, deriving and checking the results, and interpreting them in the context.

1.6.1. Statistical Literacy

The relevance of statistical reasoning and knowledge to functioning effectively in the information society led to the introduction of the term *statistical literacy:*

> The ability to understand and critically evaluate statistical results that permeate daily life, coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions. (Wallman, 1993, p. 1)

Statistical literacy is emphasised in the GAISE report (Franklin et al., 2007) produced in collaboration between the American Statistical Association (ASA) and the National Council of Teachers of Mathematics (NCTM, 2000).

Literacy is defined as the ability to find, read, interpret, analyse, and evaluate written information, and to detect possible errors or biases within this information. Therefore, to be statistically literate, people need a basic understanding of statistics. This includes knowing what statistical terms and symbols mean; being able to read statistical graphs and other representations of data; understanding the basic logic of statistics; and understanding and critically evaluating statistical results that appear in everyday life. Statistical literacy should also enable people to question the thinking associated with a specific method, to understand certain methods and their limitations, or to ask crucial questions to experts and understand their answers. For example, to measure the success of some treatment as compared to alternative treatments, a doctor can use a variety of criteria such as life expectancy, 5-year survival rates, or the whole survival curve.

Several researchers have developed their own specific models of competencies to describe statistical and probability literacy. We describe some of these models subsequently.

1.6.2. Statistical Literacy Components

Watson (1997) described "statistical literacy" as a set of competencies that adults need to manage "life" in the information society, which include literacy, mathematical and statistical skills, as well as knowledge of context and motivation. She considered three levels of increasing complexity: a) a basic understanding of statistical terms and language; b) a more complete understanding of statistical language and the related concepts in the contexts where they are encountered; and c) a critical attitude and thinking ability to apply statistical ideas to analyse or debate statistical claims (see also Watson & Callingham, 2003).

Another widely accepted description of statistical literacy was proposed by Gal (2002) who identifies two interrelated components of the knowledge that an adult needs to become a competent "data consumer": a) the person's ability to interpret and critically evaluate statistical information, data-related arguments, or stochastic phenomena in diverse contexts; b) the related ability to discuss or communicate opinions on such information. These capabilities are based on statistical and mathematical knowledge, literacy skills, knowledge of the context, as well as specific attitudes such as a critical stance. Gal (2002, p. 47) suggests that the goal of statistical literate students is more easily reached when less emphasis is placed on the computational component of statistics teaching:

> Some schools […] teach statistics […] as part of mathematics, […] yet not in a way that necessarily emphasises the development of statistical literacy.

Gal's (2002) model includes both mathematical and statistical knowledge. Reading tables and graphs already requires mathematical competencies as well as taking a critical stance in regard to statistical arguments. Mathematical thinking is also included in modelling and understanding the definition of particular variables when collecting and analysing data. For example, unemployment is defined differently in various countries, poverty is defined relative to a population, and the proportion of poor does not change if all people get richer by a fixed amount of money.

When statisticians communicate results to the target audience, they do not always explain all of their assumptions; hence, the data may appear as if they were facts. Nevertheless the validity of results depends on these assumptions. Maintaining a balance between simplifying the communication of statistical results and preserving authentic statistical information is difficult: for in simplifying the information to reach more people, the statistician might disengage those who are really interested in obtaining more complete information.

Gal (2005) expanded these ideas to probability literacy. This extension includes the capability to interpret and critically evaluate probabilistic information and random phenomena in diverse contexts. Essential for such literacy are the abilities to understand the meaning and language of basic probability concepts and to use probability arguments properly in private or public discussions. Gal includes dispositional elements in his description of probability literacy, e.g., appropriate beliefs and attitudes, and control of personal feelings such as risk aversion without defensible reasons.

1.6.3. Actions and Resources Directed to Increase Statistical Literacy

Vere-Jones (1995) documented the early initiatives to promote statistical literacy. The *Education Committee* was founded in 1948 as joint initiative of the International Statistical Institute (ISI) and the UN. The Taskforce on Teaching Statistics at School Level, established by the ISI in 1971, organised a *Round Table*, which analysed the international status of teaching statistics at school level (Barnett, 1982).

To improve statistical education around the world, the ISI launched a series of conferences in 1982 under the title, *International Conference on Teaching Statistics* (ICOTS; for the latest, see Makar, Sousa, & Gould, 2014). Another milestone was the establishment of the *International Association for Statistical Education* (IASE) in 1991 as a separate section of the ISI (iase-web.org). The founding of specific journals[11] inspired and facilitated the exchange of ideas and consolidated the young research community.

[11] The main journals are *Statistics Education Research Journal* (iase-web.org/Publications.php?p=SERJ), *Journal of Statistics Education* (www.amstat.org/publications/jse/), *Teaching Statistics* (www.teachingstatistics.co.uk), *Statistique et Enseignement* (publications-sfds.math.cnrs.fr/index.php/StatEns), *Technology Innovations in Statistics Education* (escholarship.org/uc/uclastat_cts_tise) and *Stochastik in der Schule* (stochastik-in-der-schule.de/).

A major initiative towards statistical literacy for everybody is the *International Statistical Literacy Project* promoted by the ISI and the IASE. The activities include competitions for students, the World Statistics Day, the best cooperative project competition, a newsletter, and an electronic repository of teaching resources.

1.7. STATISTICAL AND PROBABILISTIC THINKING

In the previous sections, we discussed the specificity of statistical and probabilistic thinking. Stochastic concepts are derived by arguments with a strong mathematical component. Nowhere else in mathematics, has the overlap between the scientific and philosophical ideas been as strong, and these ideas were only separated relatively recently. Moreover, stochastics has shaped modern research methodology and is now an integral part of empirical research as shown in Popper (1959).

A source of difficulty is the mixture of theoretical ideas and personal conceptions when people apply stochastic reasoning. One example of this is the complex relationship between randomness and causality. In a causal situation, the outcome is completely determined by the conditions of an experiment; however, in a stochastic situation there is no explicit way to predict the outcome with certainty, even though the experiment is thought to be repeatable under the same conditions as in physics. The assumption is that the repetitions of a random experiment are "independent". Within probability theory, independence is reduced to the multiplicative rule. When this rule has to be applied to a situation with data in a real context, one has to motivate why the specific experiment can be modelled as an independent stochastic experiment. However, many textbooks refer to independence as a lack of causal influence, which only adds to the confusion.

When teaching these concepts, often only the mathematical definitions are developed while the intuitive side is neglected. Probabilistic modelling (described in Chapter 3), however, has to rely on some basic assumptions that all methods of statistical inference have in common; consequently, without a connection to probability, statistics can be deprived not only of its potential in applications but also of its sense-making.

Consistent with the ideas above, many authors have pointed to the unique nature of statistics and probability (e.g., Moore & Cobb, 2000; Scheaffer, 2006) whence it is wise to use a modelling approach for teaching. Modelling includes working with real data, searching for suitable variables to answer questions from a context (see Franklin et al., 2007), measuring objects on the scale of the variables, controlling for variation in data by the selection of samples, and reducing the effects of confounding factors by intelligent design of data production.

1.7.1. Statistical Thinking

Thinking is something human beings do all the time and it is influenced by their experiences and the theoretical frameworks they have acquired in their life. Mathematical thinking is thinking influenced by a mathematical background. It can be formal if the mathematical notions and the relationships are precisely used; however, according to Fischbein (1987) mathematicians also have an intuitive, yet mathematical way of thinking. In that case their thinking is guided by secondary intuitions that emerge from learning the mathematical concepts, replacing their primary intuitions that exist prior to formal education. Such secondary intuitions allow one to find short-cuts for a solution of a mathematical problem. Thinking mathematically comprises finding a suitable model to adequately represent the situation. Of course, in the modelling process the constituents are wider and also include – beyond mathematics – knowledge of the context, as well as criteria for assessing how well models match a situation.

Stochastic thinking has been described by different authors, starting from Heitele (1975) who described a list of fundamental ideas related to understanding probability. One influential example of a modelling framework for statistical thinking was developed by Wild and Pfannkuch (1999) who described the complex thought processes involved in solving real-world problems using statistics. They used the modelling cycle described by the modelling group within mathematics education (e.g., in Blum, 1985, or in Blum, Galbraith, Henn, & Niss, 2007) and filled its components with statistical ideas. They mainly focused on the process of empirical research with little reference to probability although it is possible to interpret their description to include probabilistic models. They developed a complex framework including the following components that we briefly summarise under the header *statistical investigation cycle*.

1.7.2. The Statistical Investigation Cycle

Solving statistical problems involves a complete research cycle of *Problem, Planning, Data, Analysis, and Conclusion* (PPDAC). A cycle (Figure 1.1) begins when a person conceptualises a statistical problem embedded in a larger real-world problem. Various examples are given throughout the book such as investigating the differences between the physical measures of boys and girls in a classroom (Chapter 2), deciding about the convenience of an insurance contract (Chapter 3), predicting life expectancy from other variables (Chapter 4), or fixing an acceptable proportion of defective items in quality control (Chapter 5). Most of these problems emerge from a practical need, where statistics is used to find a reasonable answer.

Problem. A solution to the real problem requires that the student understands the context and realises what is the relevant question to be answered. Often this original question is too wide and needs to be made more precise; for example, restricting the possible decisions in the insurance contract. The first step in the

modelling cycle is refining the problem in such a way that it can be posed in statistical terms; it is necessary to get familiar with the context and fix the specific goals that should be met in solving the problem.

Planning. Once the problem is manageable, students have to think ahead about how to develop a strategy to get a solution. In case the solution to the problem depends on some quantity that varies with the members of a specific population (shoe size, in Chapter 2), it is necessary to plan how to measure this quantity across the population members (boys and girls). Then there is the decision about which data should be collected (number of students, specific groups in the sample) and which strategy is followed to sample the population (systematic or random sample). Finally, a plan of the analysis is required (in this example, a descriptive analysis is appropriate).

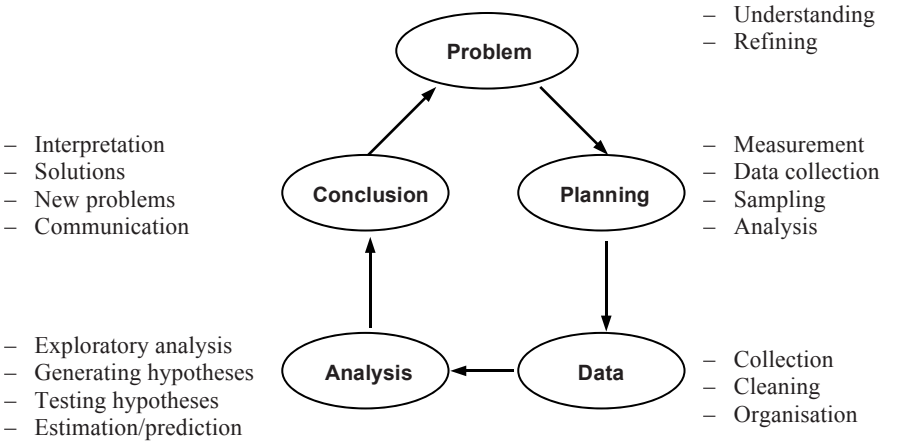


Figure 1.1. The PPDAC cycle (adapted from Wild & Pfannkuch, 1999, p. 226)

Data. The next step is collecting the data. Sometimes the students collect accessible data (e.g., from the Internet) or use data from previous questionnaires. Other times, the students design questionnaires or experiments to collect their own data. It has to be mentioned that their statistical understanding increases when they are involved in the design of a data collection instrument or experiment. An example where the students are involved in organising an experiment to collect data to test a hypothesis is given in Chapter 5. Once the data are collected, they are placed in a spreadsheet or other software that facilitates computations and visualisations. This sometimes requires reorganising the data, for example, in selecting the variables of interest and deciding the units to use for the analysis. Data should also be inspected for errors in measurement or recording.

Analysis. After the data are collected and cleaned, the analysis can start. If the students have no idea about what can be found in the data, they might carry out an exploratory data analysis (EDA). In this approach (see Chapters 2 and 4), the students summarise the characteristics of the data without testing pre-set hypotheses. Tukey (1977) was the main promoter of this approach, where multiple data representations and visualisations are used to discover hidden patterns in the data and to formulate new hypotheses. Other times the students perform an inferential analysis (see Chapter 5). In this case the interest is generalising a conclusion from the data at hand to a wider population from which the data are a random sample. Inference is used to make estimations or predictions about a larger population that the sample represents or to test hypotheses set prior to collecting the data.

Conclusion. The final step in the modelling cycle is the interpretation of the results from the analysis and relating this interpretation to the context in such a way that produces some answers to the original problem. For example, in the tea-tasting experiment to test a hypothesis in Chapter 5, we obtain data that are very unlikely if the hypothesis is true. We interpret the finding as evidence against the initial hypothesis and therefore reject it. In Chapter 2, we decide to repeat the analysis by discarding some atypical values. The modelling cycle can be repeated several times until a reasonable conclusion is reached.

1.7.3. Fundamental Types of Statistical Thinking

Carrying out a statistical research process requires that the person continually poses and solves new questions. Wild and Pfannkuch (1999) defined specific types of thinking processes that constantly appear during the PPDAC cycle. Pfannkuch and Wild (2004) defined these fundamental types of thinking as follows:

1. *Appreciation for the need for data.* As suggested in Section 1.2, data are essential in statistics; very few statistical investigations can be completed without properly collected and analysed data. Even if people have their own experience with the same type of situation they should not base a solution solely on their personal experience. Reliable data are essential to provide information for a solution and to reach an adequate decision or judgement about the situation. An important part of statistical thinking is being able to recognise the points in the process where new data are needed.

2. *Transnumeration.* By summarising and visualising data, patterns and deviations from these patterns (hidden in the raw data) might be discovered. Changing the data representation may enhance the understanding of the situation. The process of discovering ideas, tendencies, or structure from the data by changing the data representation is called transnumeration by Pfannkuch and Wild (2004).
   For example, by measuring, some characteristics of the situation are captured (e.g., different heights in a sample of students, different hair colours, etc.).

Visualising the data in a bar graph (if applicable) shows the mode (or modes if there is no unique peak) and the range of the variables. Changing to a box plot, the median, the quartiles, and the extreme values become visible (see Chapter 2). The change of representations might reveal new relevant information in the data.

3. *Perception of and attention to sources of variation*. Variation occurs in all areas of mathematics; however, random variation is specific to statistics. A particular type of statistical thinking is to differentiate statistical and non-statistical (deterministic) variation. It is also important to recognise the various sources of variation: natural variation in the population, error in measurement, or variation in sampling of data.
   A goal of statistics is to separate irreducible and reducible variation. Even natural variation in the population can be reduced; for example, in Chapter 2, the analysis of arm span should be separated between boys and girls. Statistics offers methods to control variation when the source of variability is known. Variation is also inherent in the conclusions; for example, a *p*-value or a confidence level indicate the quality of the used statistical argument based on samples from the population (see Chapter 5).

4. *Integration of statistics and context*. Contextual knowledge is vital in all steps of the modelling cycle. To highlight its relevance, Wild and Pfannkuch include the integration of statistics and context as a specific type of statistical thinking. The statistical model must be selected and exploited in such a way that the essential elements of the real situation are captured. At the same time it is necessary to be conscious that any model is different from reality; hence, some differences between the model and the investigated problem situation remain. The target is to generate data that contain adequate information needed to answer the initial problem; the summary report should be oriented to synthesise, understand and generalise the situation, when possible. Most importantly, integration with the context is essential in the conclusion phase where it is decided whether the solution is reasonable and applicable in the context.

5. *Using appropriate statistical models.* As in other areas of applied mathematics, modelling is essential in statistics. The opportunities for modelling are unbelievably wide in statistics (as seen from the examples in Tanur, 1989). Moreover, statistics has developed its own set of models that were specifically developed for the analysis of data. There is a wide range of statistical models, some of which are highly sophisticated. Examples include the normal distribution, regression models, or statistical tests, which can be generalised to complex situations. However, it is possible to use simple versions of these models at high-school level (see Chapters 3 to 5). A feature of statistical models described by Wild and Pfannkuch is that they help us to think in terms of distributions (aggregates) instead of concentrating on individuals.

1.7.4. Components of Probabilistic Thinking

Borovcnik (2011) described probabilistic thinking by elaborating on the following components: a) the ability to balance between psychological and formal elements of probability when using a personal scale of probability; b) the understanding that there are no direct success criteria in random situations;[12] c) the ability to discriminate randomness and causality; and d) the understanding of the difference between reflecting on a problem and making a decision. Other components of probabilistic thinking include:

1. *Influence of prior probability judgement*. Realising that many probabilities are dependent on other (prior) probabilities and for this reason these probabilities should be related to the proper subgroup. For example, the probability of a woman with a positive mammogram having breast cancer depends on the prior probability of having breast cancer, which describes the risk of "her" subgroup (Chapter 3).

2. *Asymmetry of conditional probabilities*. Understanding that conditional probabilities establish a non-symmetric relation between events is a key to dealing with probabilities and interpreting them properly. For example, if the probability of a positive mammogram is high given that the woman has breast cancer that does not imply that the reverse conditional probability is also high. It is vital for probabilistic thinking to be able to relate the reversed conditional probability to the prior.

3. *Theoretical character of independence*. Applying independence is often an inherent requirement of probabilistic models but is hard to check whether it is really appropriate. For example, independence usually cannot be applied when two pieces of circumstantial evidence are combined at court. Despite its *abstract* meaning, independence is a main concept in probability as frequencies make sense only if experiments are "independently" performed (Chapter 3).

4. *The problem of small probabilities*. Interpreting small probabilities appropriately is extremely difficult. Small (conditional) probabilities of observed results are used to reject the hypothesis in inference (see Chapter 5) but this does not establish a contradiction to the hypothesis (in the sense of logic). It is difficult but essential to understand the logic of a significance test as it weakens mathematical proof to empirical evidence. Small probabilities also occur in the evaluation of risks and are difficult to handle because the researcher usually does not have enough data. For example, Dubben and Beck-Bornholdt

[12] The person should understand that a correct strategy does not always assure a success.

(2010) describe an epidemiological study where *all* 331 BSE[13] positive cases could be false positives as there is no proper estimate of BSE prevalence.

5. *Correlation as probabilistic dependence*. Understanding that correlation is based on probabilities and conceptualises a much weaker form of relationship than functional dependence. Furthermore, it is important to accept that correlation can be increased, generated, or changed by other variables or artefacts[14] which are often neglected (see also Chapter 4). A proper interpretation of correlation marks a great step towards thinking probabilistically.

1.8. MAKING SENSE OF STATISTICS AND PROBABILITY

Often, teachers present statistical information such as definitions of new concepts or examples of procedures for solving statistical problems, and then give exercises to the students to practice what they learnt. The consequence is routine learning where students apply the formulas without any deeper understanding of the underlying concepts.

In order to improve the situation, statistics educators recommend refocusing the teaching of statistics on reasoning and sense making (e.g., Shaughnessy, Chance, & Kranendonk, 2009). According to the Common Core State Standards Initiative (CCSSI, 2010), to make sense of the problems posed to them, the students should first understand the goals and constraints, and conjecture a possible solution path before starting to solve the tasks posed. They may consider similar problems that have been solved before or solve a simpler form of the original problem (e.g., reduce the sample size). The teacher may provide support when needed; for example, suggesting the use of a particular type of graph to discover patterns and relationships in the data. Another strategy is organising the students in groups, where more advanced students help their classmates.

Throughout the book, we attempt to make sense of the different concepts and methods by using contexts where these ideas can be meaningful for the students. For example, in Chapter 3, we use contexts familiar to students to introduce three different views of probability: a) probability as a value to which the relative frequency tends in a large number of experiments; b) probability as a ratio of favourable to possible cases when the elementary events are equally likely; and c) probability as a personal degree of belief. In the same chapter, conditional probability is linked to circumstantial evidence. In Chapters 3 and 5, Bayes' rule is introduced as a method to learn from experience. The correlation coefficient is related to both the error in prediction and the spread of scatter plots in Chapter 4.

[13] Bovine spongiform encephalopathy or mad cow disease, a fatal neurodegenerative disease in cattle.

[14] Regression of average (aggregate) values instead of using the original data can increase the correlation considerably; by the Simpson paradox a positive correlation in all subgroups may be changed to a negative correlation in the whole group investigated.

We also try to build on students' previous knowledge to develop a deeper *understanding* of statistical concepts. For this reason, each concept is introduced via tasks that reveal how statistics and probability concepts may help to solve a given problem.

1.9. STATISTICAL INVESTIGATIONS AND EXPERIMENTS

A traditional teaching approach based on structured lessons and simple exercises can miss the goals of developing statistical and probability literacy and fail to make sense from concepts and procedures for students. Alternatively, experiments and projects have been recommended as substantial working modes for teaching statistics and probability (e.g., McGillivray & Pereira-Mendoza, 2011). Such activities may be linked to real-world problems and real or simulated data. Students can work – alone or in small groups – on their own projects and experiments. The focus may be put on practical work according to the cycles of statistical modelling in empirical research as suggested by Wild and Pfannkuch (1999).

For example, in Chapter 4, the context of life expectancy is used to introduce scatter plots, the correlation coefficient, regression lines, and the coefficient of determination. Within such a project-oriented approach, students can a) develop and reinforce their capabilities to formulate questions; b) design and implement a data collection plan for statistical studies including observational studies, sample surveys, and comparative experiments; and c) justify conclusions that are based on data (see, e.g., Australian Curriculum, Assessment and Reporting Authority [ACARA], 2013; NCTM, 2000).

In Chapter 5, an experiment about testing the possibility of merely guessing the order tea and milk are added in a cup of tea serves as a framework to introduce the relevant inferential concepts included in the high-school curricula. Other examples of projects and investigations are included throughout this book. We attempt to show how projects and investigations are a catalyst for student engagement, for learning to solve problems in a given context, for developing statistical reasoning competencies, and for synthesising learning. Above all, projects facilitate that students can make sense of statistical work.

Each of the content chapters (Chapters 2 to 5) is devoted to the elaboration of a set of related fundamental ideas that are grounded in the educational principles of the first chapter. We introduce concepts and procedures via investigations or experiments to assist students, classroom teachers, and teacher educators to understand the need for the stochastic ideas that should be learned.

1.10. FINAL THOUGHTS

In this chapter, we have described the educational assumptions inherent in the exposition of the ideas throughout the book. Teachers need specific support to facilitate students acquire the related statistical and probabilistic ideas, to appreciate and acknowledge the complementary features of mathematical and

statistical knowledge (and reasoning), to learn to make the best use of technology, and to orientate their teaching to students of differing abilities.

The aim of this book is to sustain teacher educators and teachers as well as to increase their interest, competencies, and knowledge in stochastic education. We hope to see that education researchers are encouraged to explore innovative ways and tools for educating teachers and students in statistics and probability using the ideas suggested in this book.

CHAPTER 2

EXPLORATORY DATA ANALYSIS

In this chapter, the concepts required to represent and analyse data and the techniques of exploratory data analysis will be developed within a teaching situation where students need to make sense of statistical concepts and procedures. After describing how these ideas are incorporated in high-school curricula, we present some teaching situations where students need to collect, summarise, and compare data sets to investigate and provide potential solutions to a research question. The summary and representation of the data can lead to the idea of distribution and to properties such as centre and location, spread, and shape that are useful when carrying out comparisons of distributions. Information about students' common difficulties with understanding data, empirical distributions, centre and spread, as well as the potential of technology to facilitate students' work with data will be part of our exposition.

2.1. INTRODUCTION

In a statistical study, the first steps are devoted to explore the data and to summarise their main characteristics using graphs and statistical summaries. Instead of focussing only on a few pre-determined hypotheses and apply statistical methods to test these hypotheses, Tukey (1977) introduced exploratory data analysis (EDA) as a method aimed to explore the data in order to find general patterns and to formulate hypotheses that lead to new experiments or the collection of further data.

Tukey promoted a series of graphs and suggested to summarise the data distribution using five numerical summaries: maximum, minimum, median, and quartiles because, unlike mean and standard deviation, these measures are robust; i.e., their values only change a little when atypical data or outliers are present. Typical graphs used in EDA include box plots, histograms, scatter plots, and stem-and-leaf plots. It is important to remark that Tukey did not simply replace old techniques by new methods, but he promoted a specific philosophy of exploratory data analysis. According to this philosophy, he suggested exploring data to find inherent patterns and detect deviations from these patterns, in order to search for generalisable findings from the data rather than mere testing of given hypotheses. In this exploratory approach, the main focus is on the questions that arise from the context. These questions are used to explore data for new insights. Consequently, the preliminary findings have to be interpreted within the context and this interpretation determines the next steps of an interactive analysis (see Borovcnik & Ossimitz, 1987).

Usually, EDA focuses only on the data set that is analysed since it is not assumed that these data are a sample from some larger population so that any

generalisation is tentative (Biehler, 1994). The interaction of the analyst with the data is determined by the context of the data and the knowledge about this context. The central role of the context (discussed extensively in Chapter 1) and the questions within the context makes it clear that EDA becomes more relevant for teaching when the students are familiar with the context. Thus, they can interpret the results and find answers to the questions from which the analysis started.

Today, exploratory data analysis is included in middle and high-school curricula. For example, the CCSSI (2010) recommends that grade 6 students should be able to display numerical data in different types of graphs including dot plots, histograms, and box plots. Students should relate the measures of centre and spread to the shape of the data distribution and to the context, in which the data were collected. In the GAISE project (Franklin et al., 2007), students are encouraged to identify appropriate ways to summarise numerical and categorical data using tables, graphical displays, and numerical summary statistics.

When entering high school, students should be able to use techniques for collecting data including census studies, sampling from a population, and observations; they should appreciate the impact that the quality of the data has on the conclusions of a statistical study (Ministry of Education, 2007; MECD, 2015). Students should also identify appropriate ways to summarise discrete numerical or categorical data using tables, simple graphical displays (e.g., bar graphs, dot plots), and numerical summary statistics (e.g., mean, median, mode, range, and interquartile range).

At high school, students' understanding of empirical distributions can be reinforced. Their previous knowledge should be expanded to other graphical representations of data (e.g., stem-and-leaf plot, histogram, and box plot), relative frequency tables for numerical data, and summaries such as quartiles, interquartile range, and mean absolute deviation, or standard deviation.

2.2. A TEACHING SITUATION TO INTRODUCE ELEMENTARY STATISTICAL CONCEPTS AND PROCEDURES

There are many possible situations to engage students to work with statistics (as discussed in McGillivray & Pereira-Mendoza, 2011). In this section, we suggest to work with data collected by the students themselves. The activity can start by collecting physical measures of students in the classroom such as height and weight, arm span, shoe size, waist perimeter, etc. The data used were collected from two classes of students in a high school in Spain (n = 60 in total, age 14–15).

2.2.1. Starting Questions

Exploratory data analysis, as well as any statistical investigation, always starts with questions that arise from a context. If the students are familiar with the context, the methods to investigate these questions make sense for them so that they can easily follow the process of interactively dealing with these questions. We might start with a discussion about characterising the students in the class and whether boys

and girls are different with respect to certain variables. Other possible topics, in which the students might be interested, are the distance to the school, the amount of pocket money, the use of mobile phones, or various apps. For the while, we focus on physical measurements like shoe size, body length, arm span, etc.

Table 2.1. Data from the whole class on several variables

ID-Nr.	Gender	Shoe size	Weight	Height	Arm span	Euros	Sport
1	1	35	65	164	158	28.0	2
2	1	40	64	166	171	12.4	1
3	1	36	65	179	171	3.6	2
4	1	42	68	170	172	218.0	1
…	…	…	…	…	…	…	…
21	2	36	68	158	150	16.0	1
22	2	35	70	156	152	7.0	1
23	2	36	50	159	153	20.0	2
…	…	…	…	…	…	…	…

The full data set is contained in the appendix (Table 2.10). The aim of the analysis is to find patterns in the data that might not only describe the class but may also apply for students of similar age and context and to quantify differences in subgroups. Instead of using these data, other students could collect their own measures and compare their results with ours. The methods will be introduced when they are needed; this will automatically show that they are useful. The interpretation of the results becomes the key factor in the analysis and this will enhance the used concepts.

2.2.2. Exploring Qualitative Variables

To reinforce students' previous knowledge of statistics, the teacher can ask them to produce frequency tables or simple graphs to represent the number of boys and girls in the classroom.

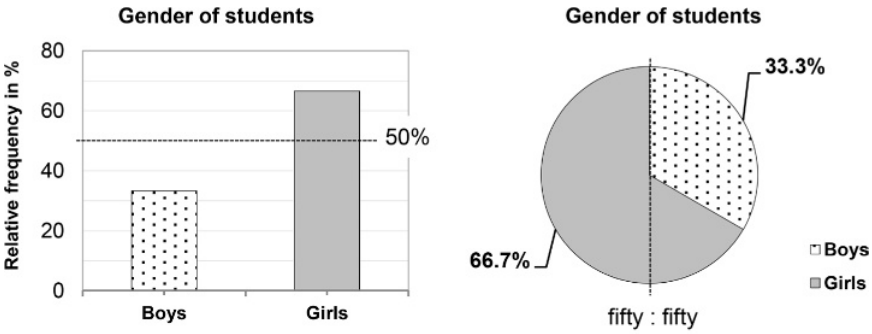
> *Task 2.1.* Summarise the individual data of the class on the variable gender by a frequency table that shows the absolute and relative frequencies of each gender. Represent the data by a suitable graph. Introduce a horizontal line at the level of 50% into the bar graph. What can you say about the percentages of boys and girls in comparison to such a benchmark at 50%? Can you calculate any statistical measures for the variable gender from the data?

The variable under scrutiny (gender) is qualitative; therefore, all we can do is to compute the absolute and relative frequencies (also in percentages; see Table 2.2), and represent the distribution using, e.g., a bar chart or a pie chart (Figure 2.1). Sometimes it enhances the data if we compare it to a benchmark; here the

50%-line. This reveals that our class has remarkably more girls than boys (two thirds as compared to 50% for an equal "share").

Table 2.2. Distribution of students by gender

Gender	Absolute frequency	Relative frequency	Percentage
Boys	20	0.333	33.3
Girls	40	0.667	66.7
Total	60	1.000	100.0



a. Bar chart *b. Pie chart*

Figure 2.1. Different graphs representing the gender of students (n = 60)

2.2.3. Exploring Numerical Variables

The analysis can continue with the exploration of the variable shoe size in the whole class (see Table 2.3). In Spain as everywhere else, the shoe size is measured by the foot length and is used to buy appropriate shoes; the values range from 17 (a baby's shoe size) to 52 for very tall men; in our classroom, the values ranged from 35 to 46.

> *Task 2.2.* Condense the 60 single data for the shoe size so that they can be analysed more easily. Build a table of (absolute and relative) frequencies of the different measurements; finally, draw a bar graph to represent the distribution of the shoe size. What can you see? Try to interpret patterns you find within the context. Compute the cumulative relative frequencies; can you explain the relatively large percentage of students with shoe size up to 38?

In addition to absolute and relative frequencies, now it is possible to compute relative cumulative frequencies that inform us of the frequency of students with shoe size less than or equal to a given value.

From Table 2.3, we see that the relative cumulative frequency of students with shoe size less or equal to 38 is 0.600 (i.e., 60% of students). The explanation for this high percentage is that there are more girls than boys in the group and girls tend to have smaller feet. This may be seen from Figure 2.2, where two groups are clearly visible as the distribution has two peaks (modes); one with a typical shoe size of 37 (the group of girls) and another at a value of 42 (the group of boys).

Table 2.3. Distribution of students by shoe size

Shoe size	Absolute frequency	Relative frequency	Relative cumulative frequency
35	4	0.067	0.067
36	8	0.133	0.200
37	14	0.233	0.433
38	10	0.167	0.600
39	2	0.033	0.633
40	4	0.067	0.700
41	5	0.083	0.783
42	6	0.100	0.883
43	4	0.067	0.950
44	1	0.017	0.967
45	1	0.017	0.983
46	1	0.017	1.000
Total	60	1.000	

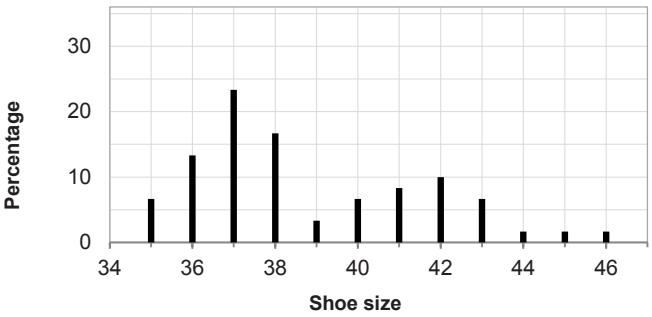


Figure 2.2. Distribution of the shoe size in the whole group of students

The distribution can be summarised further using the measures of location, centre, and spread (that students have learnt previously) that are displayed in Table 2.4. These measures are easily computed using a calculator or Excel, so that the main focus in the classroom can be put on the interpretation of the measures instead of their computation.

Task 2.3. Quantify the distribution of the shoe size by introducing the following measures and a new graphical representation:

a. Determine where the data on the variable shoe size is located; i.e., calculate the minimum and the maximum shoe size.

b. Use the mean, median, and mode to quantify the centre of the data. Can you interpret these measures within the context? Is it meaningful here to speak of a central tendency of the data? Which measure is easier to interpret? What does the median tell about the data?

c. Determine the range, the interquartile range, the variance, and the standard deviation of the shoe size data. Compare these values to each other and to your knowledge about the context. What can you learn from these values about the spread of the data?

d. Use the following five measures minimum, lower quartile, median, upper quartile, and maximum of the data to draw a box plot. Interpret the information in the box plot and compare this representation to the bar graph (from Task 2.2). Can you find advantages for either of these two representations?

The *measures of location* computed in a. help to see where the data are located. The shoe sizes range from the minimum to the maximum (from 35 to 46); however, later we will see that roughly half of the data range from the lower to the upper quartile (from 37 to 41).

In part b., the students compute the *measures of centre (central tendency)*: mean, median, and mode. Note that the mean of 38.8 in this data set does not correspond to any original value in the data set as no student has such a shoe size. While many sets are closed under arithmetic operations (such as the integers are closed under the operation of sums, i.e., the result belongs to the integers) the set of integer values is not closed under the operation of mean (the mean of integers such as the shoe size need not be an integer). Moreover, the mean is not well-suited to represent this data set as it is close to 39 (see Table 2.4), a value of shoe size that is very scarce in the data set. The median (38) is interpreted as half of the students have a shoe size below 38 and the other half has a size over 38,[1] but, for the same reason it is not representative of the data set. In fact – as we have already detected – there are two groups mixed with different modes, and each mode represents the typical shoe size of one of these groups (of boys and girls).

In part c., various *measures of spread* are calculated: The standard deviation (2.767) and variance (7.654) are the principal measures to quantify the spread of data but they lack an intuitive interpretation (see Table 2.4, also for other measures). However, they will make more sense to the students in connection to Chebyshev's inequality (Chapter 3) and the study of correlation and regression

[1] To be precise, one has to ask that at least half of the data are greater or equal to the median besides that at least half of the data are less or equal to the median. This complication is a due consequence of the fact that a finite number of values have to be split into two groups of the same size, one built of the smaller and the other of the larger values. This is not possible, for example, with 5 values. If one has 6 values, then any value between the third and fourth largest can be used as dividing point. Thus, while the idea of the median is intuitive, it requires an awkward definition to make it precise and complicated algorithms to compute it.

(Chapter 4). Simpler measures of spread are the range, i.e., the difference between maximum and minimum (46–35 = 11), and the interquartile range (IQR), which is the difference between upper and lower quartile (41–37 = 4). We can see that the "inner" 50% of the data have a much smaller range than all data (4 against 11). When compared to the mean, the measures of spread are small here (full variation of 11 as compared to a median of 38); this means that the spread in the data set is not too large.

Table 2.4. Summary statistics for the distribution of shoe size (whole class)

Measure	Values
Location	
Minimum	35
Maximum	46
Lower quartile	37
Upper quartile	41
Centre	
Mean	38.8
Median	38
Mode	37
Spread	
Variance	7.654
Standard deviation	2.767
Interquartile range (IQR)	4
Range	11

Part d. deals with the *five-number summary and the boxplot*: In addition to the median, it is interesting to find the quartiles; below the first quartile there is 25% of the distribution and above the upper quartile another 25%. This means that half of the data on shoe size of the students lies between 37 and 41. These five numbers (minimum, maximum, quartiles, and median) are displayed in the box plot (see Figure 2.3), a graph introduced by Tukey (1977) to summarise distributions of metric variables. In the intuitive way, "half" of the data are inside – half are outside the box; half of the data in the box are in the lower, half are in the upper part of the box; half of the data outside the box are below – half are above the box.

The five statistical summaries are directly signified by graphical elements. The box plot gives an indication of location of the data (shoe size between 35 and 46). It shows the centre (in the form of the median, which is 38) and it displays spread (IQR as well as range). It also displays whether the distribution is skewed (shoe size is highly skewed; in Figure 2.3, the lower whisker and the lower half of the box are much shorter than the corresponding items in the upper part). However, it misses to show that the distribution is bimodal, which is easily visible in the bar graph. Often box plots are drawn vertically from lower to upper values rather than from left to right. We will follow this practice subsequently and show the rotated box plot to illustrate its different appearance (Figure 2.4).

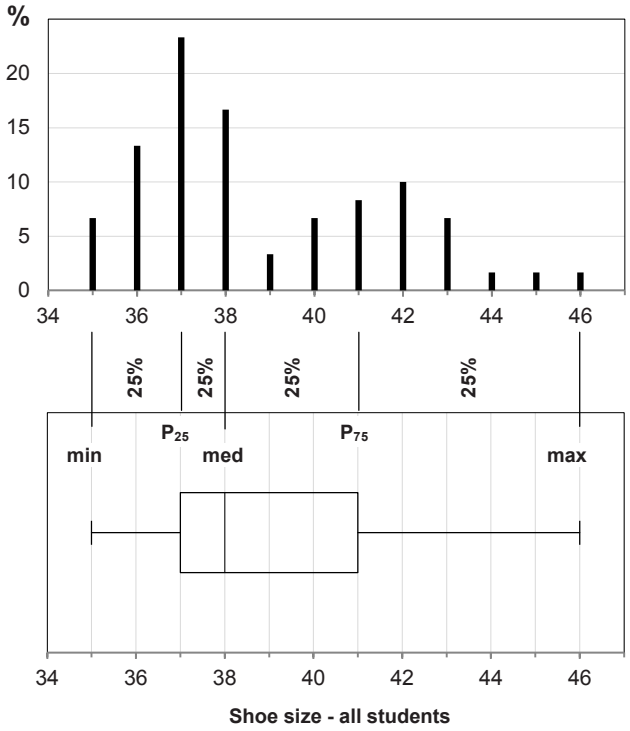


Figure 2.3. Box plot for the shoe-size data in the whole group as compared to the bar graph

Task 2.3 is continued by a discussion of the shape of the distribution of the shoe size, by searching for atypical values in the distribution, and by focussing on patterns in the distribution and deviation of the pattern.

Task 2.4. Using the various quantitative measures of location, central tendency and spread, as well as the box plot and the bar graph:

a. Discuss the shape of the distribution of the shoe size and explain its essential properties.

b. Discuss whether there are students with atypical values of the shoe size as compared to the general pattern of the distribution. What can you do if there are atypical values? How can we represent atypical values in the box plot?

c. Describe the distribution of the shoe size by taking into account its general pattern and deviations of single data points from this pattern.

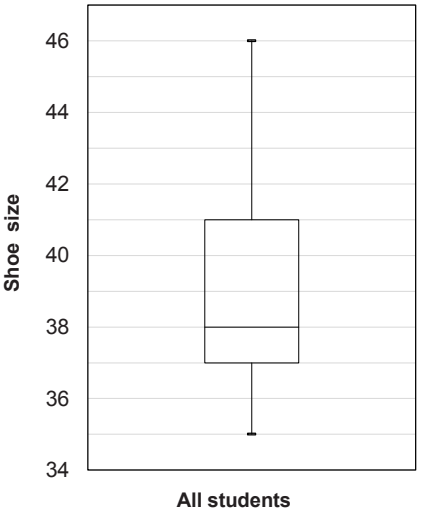


Figure 2.4. Vertical box plot for all data of the shoe size

Shape of the distribution: Once we detect that the distribution of the investigated variable has a *bimodal* shape (two peaks), any measure of centre gets cumbersome (as is the case in Figure 2.3). Two modes are an indication of the need to split the data for another variable to make it uni-modal and repeat the analysis within the subgroups. In this example, it is convenient to split the data by gender. This will also give rise to the question, by how much boys and girls differ in shoe size. Another factor that influences the interpretation of measures of central tendency and spread is whether a distribution is symmetric or skewed. For skewed distributions, the median becomes more relevant to describe the centre as it preserves the intuitive interpretation that it halves the data into a lower and an upper part while the mean is shifted out of the centre. However, the median does not represent a centre either (though it is not shifted so much out of the centre as the mean).

Atypical values and outliers: Notice that no atypical values are displayed in the box plot in Figure 2.3; this means that the smaller and larger shoe sizes are not too far away from the other values to be considered as atypical. In case of atypical data, the question arises whether to regard them as outliers and to eliminate them or not (be considered as mistakes or as not belonging to the analysed group). To help detect atypical values, the whiskers of the boxplot are no more drawn from the box to the extreme values. To mark the end of the whiskers, there exist several rules.[2]

[2] Tukey (1977) suggests determining the *preliminary* length of the whiskers as 1.5 times the IQR; however, the final whiskers would be reduced to that place where data are located so that at the very

Values outside the whiskers are then labelled to identify the statistical units behind these data so that it is easier to explain why these data are extreme and whether they should be eliminated as outliers or should be kept as substantial information about the group.

Pattern and deviation: A central idea of EDA is to split the data into a pattern that fits the major part of the data and deviations from that pattern. There are several ways to find the pattern. One is to focus on the central box, another is to describe the general pattern of the data by all data that lie within the whiskers and refer to the rest of the data as deviating data (atypical or even outliers).

2.2.4. Comparing Groups

Since Figure 2.2 suggests that two groups with different distributions for the shoe size were mixed, the teacher can ask the students to repeat the previous analysis separately for each group (boys and girls) and compare the results.

Task 2.5. Compare the distribution of the shoe size in the two groups according to gender. Condense the data (in Table 2.10, Appendix) by frequency tables, one for boys and one for girls. Represent the gender differences for the variable shoe size by aligned separate bar graphs. Discuss the different shapes of the distribution between boys and girls and try to describe the general pattern of these distributions.

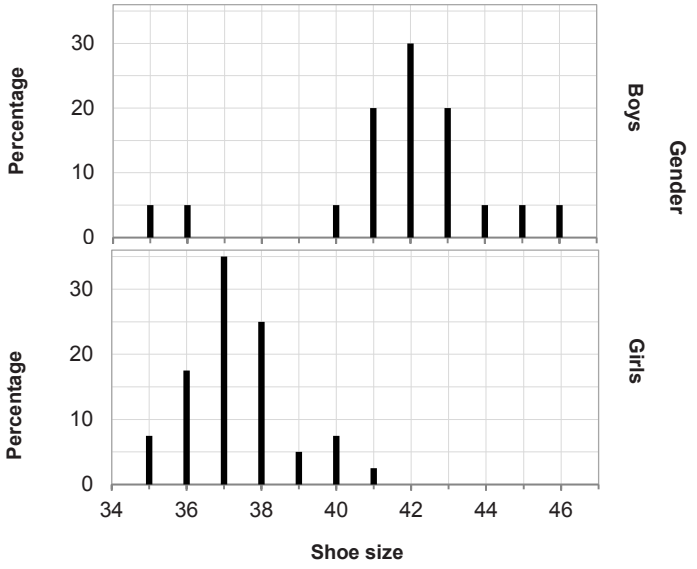


Figure 2.5. Separate bar graphs for the distribution of the shoe size for boys and girls

end of the *drawn* whiskers there are data in the data set. He considers data as atypical if they are beyond the *preliminary* end of the whiskers.

We will not repeat the frequency table for the shoe size for boys and girls but focus on the graphical representation of the distribution. In Figure 2.5, the bar graphs for both groups are displayed in the same diagram. The differences are clearly visible: the boys' shoe size is larger than that of girls with a few exceptions. Students can analyse what happens in these *atypical cases*. Given the age of students (14-16), some boys might still grow while usually girls have already matured at that age.

Task 2.6. Quantify the difference in the distributions of the shoe size with suitable measures of location, central tendency, and spread. How big are the differences in central tendency? Are there differences in spread of the variable shoe size between boys and girls?

The statistical summaries for these distributions are displayed in Table 2.5. In this table, the difference in central values is clearly visible since there is a difference of shoe size of about 5 between boys and girls (it does not matter whether the centre is measured by mean, median, or mode). The spread is also much larger in boys' feet, which is visible in the range (11 for boys and 6 for girls) as well as in the variance and the standard deviation.

Table 2.5 Summary statistics for the distribution of shoe size for boys and girls – all data

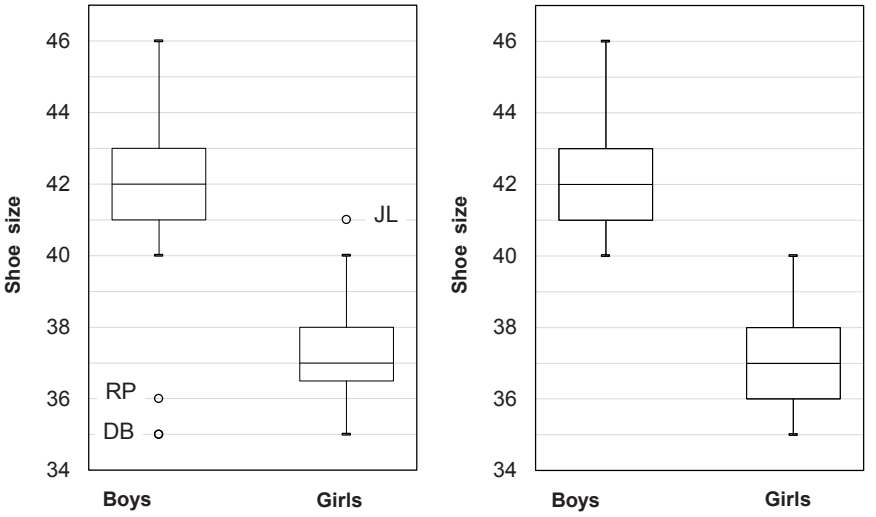
Measure	Boys	Girls
Location		
Minimum	35	35
Maximum	46	41
Lower quartile	41	36.5
Upper quartile	43	38
Centre		
Mean	41.7	37.4
Median	42.0	37.0
Mode	42	37
Spread		
Variance	6.537	1.926
Standard deviation	2.557	1.388
IQR	2	1.5
Range	11	6

Task 2.7. Draw two aligned box plots for the distribution of shoe size and describe the general pattern and the differences between boys and girls. Discuss the different shapes of the distributions; describe the general pattern and differences for boys and girls and whether you can detect atypical values within the two groups. Use whiskers that do not extend to minimum and maximum so that extreme values of the distribution are shown separately. Are there atypical values? Can these atypical values be considered as measurement errors (or other types of outliers) and discarded or do they signify

substantial phenomena from the context? Compare the distributions with and without atypical values. What can you see?

In Figure 2.6a, we summarise the distribution of the shoe size for boys and girls using two parallel box plots, where the atypical values are additionally represented and marked by labels that enable the analyst to identify the statistical units. Notice that these atypical values did not appear in the box plot for the entire group (Figure 2.4). The reason is that being atypical is not an absolute property but depends on the group. Thus, to have a shoe size of 35 is atypical for boys but neither for girls nor for the entire group.

In these box plots, it is clearly visible that the central values and spread in the boys' shoe size are larger than the related values of the girls. But the spread is not as much larger as one would expect since there are two atypical values, i.e., boys with very small feet as compared to the group of boys; at the same time, there is one girl with too big feet as compared to the other girls.



a. All data – whiskers with Tukey rule to detect atypical values

b. Atypical values eliminated – whiskers to min and max

Figure 2.6. Box plots for the distribution of the shoe size for boys and girls separately

Task 2.8. Calculate the measures for location, centre, and spread without the detected atypical values and discuss the effect from eliminating atypical values. Give a final description of the gender differences for the variable shoe size in the sense of the general pattern and individual deviations from it.

In Table 2.6, we see how removing the atypical values affects the different statistical summaries. While there is no change in the modes and medians, the mean of the shoe size of boys is now larger while that of girls is smaller (Figure 2.6b shows this effect graphically). This result is due to the fact that the mean is influenced by atypical values; a very high data value increases the mean and a very small one decreases it; on the contrary, the median is robust, i.e., it is less affected by atypical values. Consequently, when the distribution has atypical values, the median is preferable to represent the centre of the distribution.

Measures of spread have also been affected when removing the atypical values and the spread is now smaller in both groups, which is visible in both plots and in the statistical summaries. Moreover the difference in spread is nearly the same if these atypical values are eliminated. The initial difference in measures of spread does not indicate large differences in spread but are a sign of inhomogeneity within the groups of boys and girls.

Table 2.6. Summary statistics for the distribution of the shoe size for boys and girls – atypical values excluded

Measure	Boys	Girls
Location		
Minimum	40	35
Maximum	46	40
Lower quartile	41	36
Upper quartile	43	38
Centre		
Mean	42.39	37.26
Median	42.0	37.0
Mode	42.0	37.0
Spread		
Variance	2.252	1.617
Standard deviation	1.501	1.272
IQR	2	2
Range	6	5

The final description of the groups of boys and girls without atypical data yields the general pattern of the data: The shape of the distribution of shoe size is symmetric in both groups with respect to the inner core of the data (the box) and slightly skewed to higher values (for boys a little more than for girls). The spread is nearly the same in both groups. The pattern of the data may be described as a shift of 5 in the centre of the distribution with nearly the same shape (Figure 2.6b).

Apart from this pattern, there are deviations in both groups. In a way, boys and girls fall apart each in two further subgroups. A small group of boys has an extremely small shoe size while an even smaller group of girls has a very large shoe size. In fact, these values are not outliers; they are valid data. The explanation

for the boys (at age 14) is that some boys simply lag behind physically as their development is retarded.

2.3. ADDITIONAL ACTIVITIES

Although shoe size is a numerical variable, it only takes integer values and the number of different values is small. In many situations, however, variables are measured with decimal numbers since they represent continuous magnitudes, which theoretically can take any value in a given interval[3]. In the following section we are dealing with simple statistical methods to explore continuous variables.

2.3.1. Exploring Continuous Variables

Once the students get acquainted with statistical graphs and summaries adequate for numerical variables with integer values, the next step is to analyse continuous variables such as arm span.

Task 2.9. Represent the distribution of the arm span data using a stem-and-leaf diagram. Read from this diagram the minimum and maximum of the data and the percentage of data less than 160. Describe the general pattern of the distribution.

In Table 2.10, we reproduce also the data for the arm span of the 60 students in cm. Although we work with integer numbers because we have rounded the data to cm, in fact it would be theoretically possible to obtain any given number in each interval (for example, 187.317 cm). There are too many different values for the variable to build a frequency table or a bar graph. A different strategy is to simplify the data in grouping the values into intervals. A first step to do this grouping is to produce a steam-and-leaf plot, which has been introduced by Tukey (Figure 2.7).

```
Frequency   Stem |Leaf
    13        15 |0234555578889
    22        16 |00000001122344555566 88
    14        17 |11112225788999
    11        18 |00001255578
```

Figure 2.7. Stem-and-leaf plot for arm span in cm (60 data)

To build this plot, the data are split into two components: the first digits form the stem while the last digit is the leaf. For example, the length of 154 is decomposed in a stem of 15 (dm) and a leaf of 4 (cm). The data are arranged in

[3] Continuous variables may theoretically take all values in an interval, i.e., all real numbers that are contained in this interval; this is, for example, the case when we deal with magnitudes that are modelled by a normal distribution (see Chapter 3 for a description of this type of distributions).

rows, where every row corresponds to data with the same stem. Each stem is also displayed once followed by a vertical line; on the right-hand side of this line we display the leaves for all the data with the same stem in numerical order.

In a stem-and-leaf display, the absolute frequency of the intervals related to the stem may be displayed left of the stem label. It is easy to show to the students that the diagram in Figure 2.7 is basically a histogram with intervals of width 10. In the first row, all the data of arm span are displayed that fulfil $150 \le x_i < 160$.

The minimum of the data is 150 and the maximum is 188; the data is located between 150 and 188. As 13 values of the 60 data are below 160, 21.7% (roughly one fifth) of the data is below 160. The shape is skewed to the right as the two upper lines have nearly the same number of data as compared to one line below the modal line, which has a markedly higher number of data (22).

Task 2.10. Repeat the analysis of Task 2.9 using a stem-and-leaf diagram with groups (lines) of length 5. Discuss the shape of the distribution, which can be seen from this graphical representation.

The stem-and-leaf in Figure 2.7 does not give a refined picture of the distribution. That is why we present another diagram where one line from before is split into two lines so that we represent the data in groups of length 5 beginning with $150 \le x_i < 155$, $155 \le x_i < 160$.

```
Frequency     Stem |Leaf
    4           15 |0234
    9            * |555578889
   14           16 |00000001122344
    8            * |55556688
    7           17 |1111222
    7            * |5788999
    6           18 |000012
    5            * |55578
```

Figure 2.8. Expanded stem-and-leaf display for arm span in cm (60 data)

In Figure 2.8, a stem-and-leaf plot with intervals of width 5 is displayed. The shape of the distribution is now markedly skewed to the right. The modal line is 16, which represents the group $160 \le x_i < 165$. Note that the intermediate groups are always labelled by an asterisk; the asterisk below 16 designates the group $165 \le x_i < 170$. The modal line has now restricted to a shorter range of values. The shape is more clearly visible. We could continue to make the groups (lines) shorter. However, we would see that the single lines would tend to clutter and give no more a pattern of the distribution.

Task 2.11. Condense the data into groups (intervals) of length 10, determine absolute and relative frequencies of these intervals (use the same intervals as for the stem-and-leaf diagram) and display the distribution by a histogram. Approximate the mean from the condensed data of the frequency table.

Table 2.7. Distribution of students by arm span

Arm span (interval)	Mid-point	Absolute frequency	Relative frequency	Relative cumulative frequency
[150, 160)	155	13	0.217	0.217
[160, 170)	165	22	0.367	0.583
[170, 180)	175	14	0.233	0.817
[180, 190)	185	11	0.183	1.000
Total		60	1.000	

The histogram related to this frequency table (Figure 2.9) looks similar to the stem-and-leaf plot in Figure 2.7. The vertical stripes of the histogram have the same area as was covered by the numbers in a line earlier.

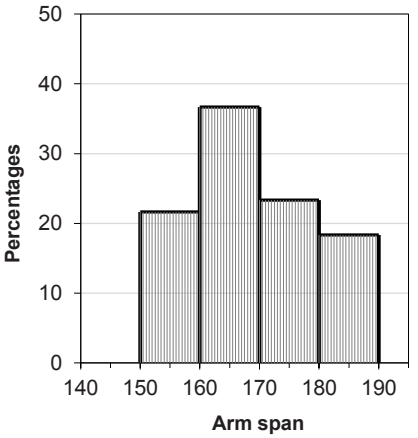


Figure 2.9. Histogram for arm span in cm (60 data)

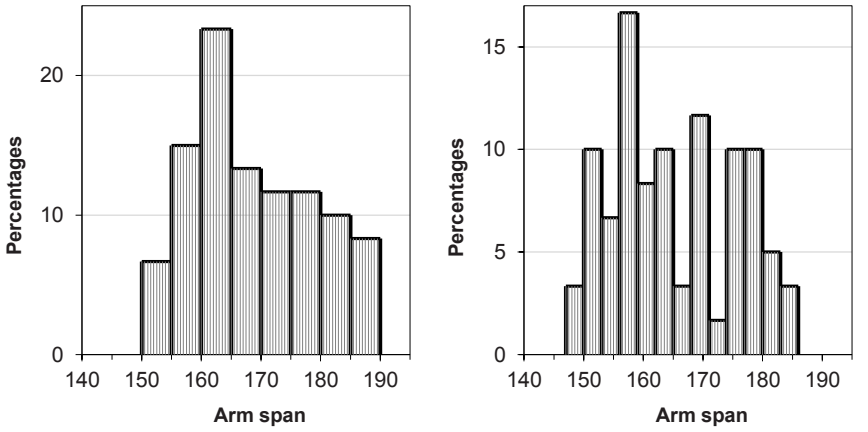
From the frequency table (Table 2.7) we cannot determine the single values of the data as we could still do from the stem-and-leaf plot (Figure 2.7). Thus, we cannot calculate the mean of the arm span data exactly. However, if we use the midpoints of the intervals instead of the exact data, we can re-establish the sum of all data approximately by

$$155 \cdot 13 + 165 \cdot 22 + 175 \cdot 14 + 185 \cdot 11 = 10130 \text{ and } \bar{x} \approx \frac{10130}{60} = 168.8,$$

which is quite close to $\bar{x} = 167.7$ (from all data in Table 2.10), especially if we consider that by replacing all data from the interval $150 \le x_i < 160$ by 155 the errors could be up to 5 each.

Task 2.12. Repeat the analysis of Task 2.11 using intervals of length 5 and 3. Determine the frequency table and draw the histogram. Comment on the shape of the distribution as it is shown by the histograms.

We omit the frequency tables and show the histograms in Figure 2.10. The left histogram with intervals of length 5 shows exactly the same shape as the stem-and-leaf diagram with lines of length 5. This gives a clear indication of a distribution that is skewed to the right. The histogram with intervals of length 3, however, is cluttered. The information shown here is too refined to enhance the general pattern of the distribution.



a. Intervals of width 5 *b. Intervals of width 3*

Figure 2.10. Histogram for arm span in cm (60 data)

An important remark is that the modal intervals vary with the width of intervals; therefore, for grouped data, special care should be taken for the interpretation of modal intervals. Thus, in the different histograms displayed for the arm span we get different impressions. While in Figure 2.9, the modal class is 160-170 cm, in Figure 2.10a, it is located around 162.5 (the class ranges from 160 to 165), and in Figure 2.10b, we observe two different groups with different "modes" each: one is located roughly around 160 cm (girls) and another between 174 and 180 cm (boys).[4]

[4] While the mode is easy to detect in histograms, it is not visible in box plots. The mode is highly dependent on the grouping process of the data as may be seen here. For the ungrouped raw data, the

From either diagram (histogram or stem-and-leaf plot), it is possible to build a frequency table (see Table 2.7). This time, each line (group) represents the different frequencies corresponding to an interval of values. What is important is to have a clear rule of where to count each observation.

In approximate computations, we often work with the interval midpoints (as in computing the mean above); therefore, the midpoints are usually added to frequency tables for grouped intervals. It is noteworthy to mention that the width of intervals is a matter of convention in both the stem-and-leaf diagram and the histogram. In our analysis, we decided to use intervals where the lower boundary is included while the upper boundary is not included; however, this is also a matter of convention.[5]

An advantage of the stem-and-leaf display over the histogram is that we can read off the exact original values for each data (at least exact to two digits) while these values are lost in the histogram representation. On the other hand, stem-and-leaf diagrams display only the absolute frequencies, so that – while they are perfectly suitable to investigate the shape of one distribution – they are not equally appropriate for the comparison of two distributions as this requires to base the comparison on relative frequencies (at least if the groups are different in size).

Task 2.13. Quantify the differences in the distribution of the arm span for the groups of boys and girls. Compare these two groups. In which respect do the distributions differ? Are there atypical values? Describe the general pattern of the distribution of the arm span for boys and girls.

Like in the case of the shoe size, we can separate boys and girls to compute the different summary statistics (Table 2.8). Students can interpret the differences in the groups from either of these statistics or from the box plots in Figure 2.11. They may also repeat their analysis by excluding the atypical values and investigate their effect on the statistics and the diagrams.

In the previous activities, the quartiles proved useful. The lower quartile corresponds to the dividing line for 25% of the lower values in the distribution and is also called the 25^{th} percentile P_{25}; the upper quartile corresponds to the dividing line for 75% of the lower values in the distribution and is also called the 75^{th} percentile P_{75}, which marks the boundary of the top 25%. Note that the percentiles are calculated from ordered lists of data.

---

mode may be useless as it need not be a typical value at all. The mode is more useful for detecting a bimodal shape of the distribution than to designate a centre of the distribution. In the case of bimodality, the analyst would try to split the group and analyse the subgroups separately.

[5] In some applications and in EDA, the boundaries are treated like here; however, in traditional statistics, the boundaries are usually treated just the other way round, i.e., the lower boundary is not included while the upper boundary is included.

Table 2.8. Summary statistics for the distribution of arm span for boys and girls

Measure	Boys	Girls
Location		
Minimum	158	150
Maximum	188	180
Lower quartile	172	158
Upper quartile	183.5	165.5
Centre		
Mean	178.1	162.5
Median	179.5	161.0
Mode[6]	–	–
Spread		
Variance	51.253	51.795
Standard deviation	7.159	7.197
IQR	11.5	7.5
Range	30	30

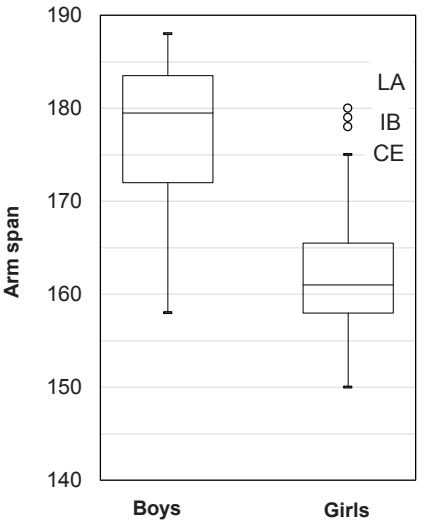


Figure 2.11. Box plot of the arm span – separate diagrams for boys and girls

[6] The mode is left out here. It does not make sense to compute it for continuous data (apart from referring to a modal class). It is too sensitive to changes in the data and needs not reflect patterns in the data. For the arm span of boys, the values of 172, 180, and 185 each had a frequency of 3 and could all be considered as modes.

Task 2.14. Draw the curve that corresponds to the cumulative relative frequencies for all students in the class and for boys and girls separately. Read off this curve the 20th percentile. What percentage of all students is less equal to 170 cm? Does it make a difference whether we refer only to boys or only to girls?

If the data are ordered, then the cumulative curve increases at each value of the arm span by 1/60 (1/n with n the number of data, Figure 2.12). In fact, from this curve it is possible to read off other percentiles such as P_{20} = 157. Alternatively, from this graph, it is possible to read, which percentile corresponds to a given value of the variable; if a student's arm span is 170 cm, the person corresponds to $P_{60;}$ or put in a different way, 60% of the data on the arm span in this group is less or equal to this student's value.

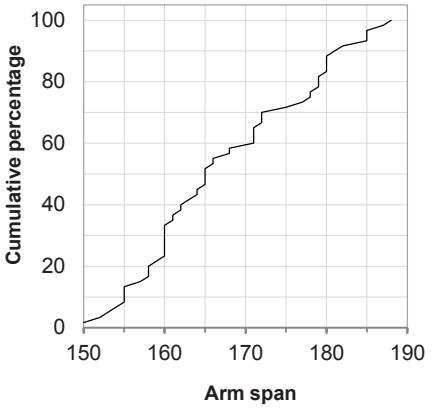


Figure 2.12. Cumulative frequency distribution for arm span (60 data)

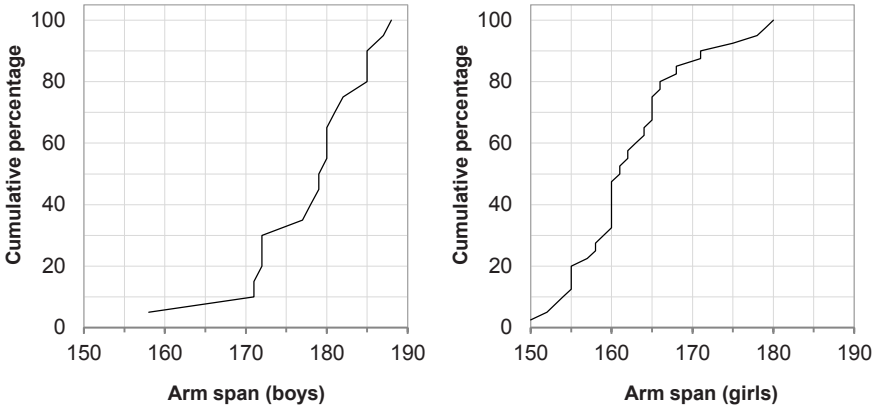


Figure 2.13. Cumulative frequency distribution for arm span in cm – separate curves for boys and girls

As the arm span is quite different for boys and girls, it is preferable to use the specific cumulative graph for their own group in computing the percentiles. In Figure 2.13, it is clearly visible that, while only 18% of boys have an arm span less than or equal to 170 cm, this value corresponds to the percentile P_{88} for girls.

The activities above may be repeated with the analysis of other variables such as pulse rate at rest, weight, or height. A different analysis consists of separating the students into three groups according to their grades of practicing sports and compare these groups with respect to the variables just mentioned.

All these activities are restricted to exploratory analyses with no intention to generalise the conclusions to other students beyond the class. Yet, we try to find general patterns in the data that help us to insights about the group under scrutiny. We discussed the shape, atypical values, and how boys and girls differ, and tried to explain the patterns and the deviations by referring to the context. This may generate hypotheses about other, similar groups of students.

The previous analyses may be complemented with inferential investigations if we collect the data at random from a larger population. For example, in case the 60 students were a random sample of all students in the school, we could try to generalise the findings to all the students in that school. As we have seen that boys and girls are different in this class, we might also expect differences by age so that we should select the students in a restricted way that ensures that all ages as well as boys and girls are well represented. More information about statistical inferential methods is included in Chapter 5.

2.3.2. Exploring Bivariate Relationships

The final activity here is a short introduction to the topic of correlation and regression, which is expanded in Chapter 4. This data set may also serve to explore linear relationships between variables as many physical measurements are approximately linearly related.[7]

When reaching to this point of analysis, many students would have observed that the arm span is very close to the height of a person. Moreover, the taller a person is, the larger is his or her arm span. Would it be possible to find a line that can be used as model to predict (estimate) the arm span of a person if the height is known?

> *Task 2.15.* Draw scatter plots of arm span versus height for boys and girls separately. Describe the general pattern (tendency) of the increase in arm span by an increase in height for boys. Are girls different from boys in this respect?

In Figure 2.14, the variables are represented in separate scatter plots for girls and boys and a line fitting the data is added to each plot. Since the sample size is small, these lines are not very stable and would probably change when more data are added to the sample. Anyway, these lines obtained with Excel can be used to

[7] In fact, as it is exposed in Chapter 4, correlation and regression methods emerged from the study of physical measurements.

remind the students of elementary properties of straight lines. For example, the slope in these lines indicates that the arm span for girls is increased by 1.25 cm if the height increases by 1 cm; in the case of boys, the additional gain in arm span equals 0.82 cm if the height increases by 1 cm. That does not imply that boys' arm span is smaller as the additional additive constant for the line describing the boys is 33.01 while for the girls it is –41.67.[8]

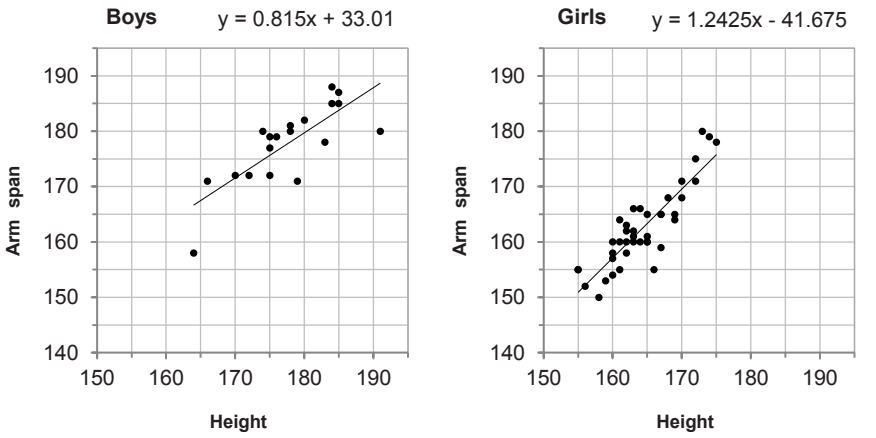


Figure 2.14. Scatter plot of arm span by height in cm – separate plots for boys and girls

The teacher can discuss with the students what is the aim of fitting a line to the data and also discuss with them about the differences between an association (between height and arm span as shown in the scatter plots) and functional dependency between two variables. The students can also find out the difference between direct and inverse relationships and search for other examples of this type of association.

2.4. SYNTHESIS OF LEARNING GOALS

There are a number of concepts, procedures and properties that underlie the activities described. They are all related to exploratory data analysis and part of them is previously known to the students.

An important goal at high-school level is, to make students aware of the main differences between *qualitative* (such as gender) and *numerical* (such as weight) variables, as well as of the different summary statistics and graphs that are adequate to represent them. EDA does not only offer more flexible tools and representations of data than traditional descriptive statistics, it is also characterised by a special approach towards data analysis (Biehler, 1986; Borovcnik, 1995). In

[8] This latter constant cannot be interpreted as a prediction for the arm span as it is the value for arm span if the height is 0.

the spirit of EDA, data analysis starts with questions from the context and returns to the initial questions, trying to make sense of the intermediary results. EDA explores data in multiple ways in order to detect inherent patterns and to find empirically-based arguments to the initial questions.

2.4.1. Distribution and Different Types of Frequencies

The first step in analysing a variable is forming its *distribution*, which includes all values obtained and their frequencies. The distribution describes the data set as a whole and therefore, every single data point influences the distribution. In Table 2.2, we displayed the distribution of gender using absolute and relative frequencies (also in percentages).

Let $a_1, a_2, \ldots, a_k$ denote the values of a qualitative variable, then:

- The absolute frequency f_i of a_i is the count of this value and describes how often a_i occurs in the data; note that the sum of all the absolute frequencies equals n, the size of the data set.
- The relative frequency of a_i is defined as $h_i = \frac{f_i}{n}$. It represents the fraction of the data that takes on the value a_i. The sum of all relative frequencies of all possible values of the variable equals to 1.
- When relative frequencies are multiplied by 100, we obtain percentages, which sum up to 100.
- In case of numerical variables, we can order the values (now denoted by x_i) by size and additionally obtain cumulative frequencies (see Table 2.3). The cumulative relative frequencies $H_i = \sum_{j=1}^{i} h_j$ are the relative frequencies of all data less or equal to x_i.

The empirical distribution of a variable with observed values (a_i or x_i) serves to investigate questions that arise from the context. At the same time it may also prepare the concept of probability distribution. There is a tight connection between these two types of distributions, starting from a connection between relative frequencies and probabilities of a specific value, as well as a connection between the mean value (or the median) of data and the expected value (median) of a probability distribution (see Chapter 3 for the interrelations between these concepts).

With data, these concepts have a material meaning so that the more complicated probability terms may be facilitated. To support the teaching of the concept of distribution, a variety of graphs may be used that mediate between statistical analyses and the context. Such graphs facilitate the modelling process by developing the concept of distribution; such graphs also enhance the interactive data exploration, which combines statistics and context.

2.4.2. Simple Univariate Graphs

There are many different statistical graphs suitable for exploratory analyses of data; some of these diagrams were specifically designed for this purpose (e.g., steam-and-leaf diagram or box plot) within a strand of applied statistics, which has been called EDA – exploratory data analysis, which also has developed a philosophy of data analysis that differs from traditional descriptive statistics (see Tukey, 1977, for other graphs created within EDA). Other plots like the histogram are non-specific to EDA but can be used within the EDA philosophy.

Bar graph and pie chart. Both types of diagrams are suitable to represent either qualitative or numerical variables holding few different values (such as the shoe size). While bar graphs use length to represent frequencies of a distribution and thus emphasise the part-to-part comparison in a data set, pie charts use areas and thus emphasise a part-to-whole relationship in the data set. It is important to notice that for qualitative variables the horizontal axis represents non-ordered values and the bar width plays no specific role (see Figure 2.1a) while for numerical variables the horizontal axis represents the number line. Thus, the values of the variable are ordered and the bars should be restricted to lines (with no width) as in Figure 2.2.

Histogram. This graph is adequate to represent the distribution of numerical data with many different values (usually related to a continuous variable). To construct a histogram, the data are grouped into intervals by dividing the entire range of the variable and counting the frequency of values that fall into each interval. A rectangle is drawn with an area[9] proportional to this frequency and a width equal to the interval. The rectangles of a histogram "touch" each other to indicate that the underlying variable is continuous.

Stem-and-leaf diagram. This graph is an intermediate step to building a histogram, which is a useful tool to get a first overview on the *distribution* of the data. To build this graph, the data are arranged in rows and each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit). An advantage (as regards histograms) is that the original data value can be recovered from the plot, at least to two significant digits. This data is ordered, which helps to calculate order statistics such as the *median* and the *quartiles*. A disadvantage is that they are less flexible than histograms since differing widths of intervals cannot be handled. Furthermore, stem-and-leaf displays only show absolute frequencies

[9] The height in a histogram represents density and corresponds to the theoretical density function of the variable, which is relative frequency divided by interval length. The concept of density is more difficult to interpret; it is comparable to the concept of population density. In many cases we can use intervals with equal width so that for this particular case the height of the histogram is proportional to the frequency in that interval. That is why in high school we often avoid calculating the density. However, discussing the idea of density when introducing histograms offers an opportunity to prepare the concept of density for continuous probability models such as the normal distribution.

and are, therefore, not useful for the comparison of two groups if the groups are considerably different in size.

Cumulative frequency graph. A graphical representation of the cumulative frequencies is very useful to determine percentiles and percentile ranks. Before we can draw this graph, we need to calculate the cumulative frequencies by summing up the frequencies from lower to higher values of the data. By drawing horizontal lines to represent 1/4, 1/2, and 3/4 of the total frequency, we can find the lower quartile, median, and upper quartile from the horizontal axis in a cumulative frequency graph.

2.4.3. Simple Summary Statistics

Although graphical representations are the main tools for data visualisation and exploration, summary statistics are often helpful to make more precise judgements about the context. For example, the boys have an 18.5 cm longer arm span than the girls (as measured by the difference in the median, see Table 2.8). This value quantifies the difference by a suitable and easy-to-interpret measure and thus specifies what we mean by longer. Relative to the median of arm span, this difference amounts to roughly 10% so that we can consider the difference between boys and girls as large. The two box plots (Figure 2.11) show that not all girls have a smaller arm span and that there is a great variation of arm span for boys as well as for girls.

Thus, two complementing conclusions emerge from the exploratory data analysis: one about the size of the difference (i.e., the general pattern for the majority of the students) and one about the spread of the variable (i.e., that individuals may deviate from their groups substantially). The difference in arm span is not only a shift in the centre of the distribution, but shape and spread are also slightly different in the two groups.

Students know most of these statistics from their previous schooling when they enter high school. However, at this educational level they can reinforce their knowledge of the same concepts, compute these statistics from more complex data using calculators or software, interpret their meaning in various contexts, and see how they support the exploration of questions emerging from this context. The following summary statistics are relevant at high-school level:

Measures of location: The data range from the minimum to the maximum; this gives a clear idea about where the data are located. If one considers only the core of the data – the 50% of data in the "middle", then these core data range from the lower to the upper quartile. For a quick judgement of the spread of the core data, a qualitative look at the data may be sufficient so that it is not necessary to calculate the quartiles at that first inspection of the data.

Measures of central tendency or centre: mean, median and mode. While the mode can also be used with qualitative variables, mean and median can only be used with

numerical variables. The mean changes quickly with atypical data so that in distributions with such atypical cases the median is preferable. The mode gives easily accessible information about the distribution if the distribution is clearly peaked. It also helps to identify whether the data originate from a homogenous group or from differing groups (as in the case of our class, which splits into the group of girls and boys). However, the mode is more suitable to represent a typical value if the variable is qualitative but gives less information about a continuous variable as it also depends on the way how intervals are introduced to draw a histogram. For asymmetric distributions, the median gives a clearer interpretation of the centre than the mean as the mean might be far out from the centre so that it does not represent a "centre" of the distribution at all. If the difference between mean and mode is larger, this is a clear indication of a rather skewed distribution.

Measures of spread: Standard deviation and variance are the most important measures of spread – also for embedding the results of the data analysis into the framework of statistical inference. However, range and interquartile range give a more intuitive access to the spread of the data as these measures are calculated only on the basis of two numbers each. The standard deviation is based on the average squared deviation of any single data from the mean over all data of the investigated "group". There is not an easy interpretation of the standard deviation.[10] The most intuitive way to interpret standard deviation is to calculate mean ± 2 (or 3) times the standard deviation; the interval that can be built by these two numbers covers the range of nearly all data, especially if the distribution is symmetric. In Chapter 4, Chebyshev's inequality will expand a related property to random variables.

Five-number summary: The minimum, maximum, quartiles and median can later be generalised to the idea of percentile and percentile rank. A percentile P_r is defined as a value of data that is higher than r% of the other data values.[11] Special examples of percentiles are the median (P_{50}) and the quartiles (P_{25} and P_{75}). These statistics may be introduced intuitively by splitting the ordered data into an upper and lower half and (recursively) split the two piles again with the result of dividing the data into four groups, from the lowest to the highest with 25% of data in each group.

---

[10] Carl Friedrich Gauss was one of the first to deal with the standard deviation, which was not yet officially introduced. He defined the precision of measurements (of the data) as $1/(\sqrt{2}\cdot\sigma)$ and used a quantity of $0.4769\cdot\sqrt{2}\cdot\sigma$ as probable error; the constant is derived from the condition that the probability of an error larger than that equals ½. Half of the errors are below and halve are above this "typical error", which corresponds to the same idea as used in the box plot where half of the data are inside and half are outside the box.

[11] This intuitive definition shapes the idea of quantiles, but for discrete data often there is no r% part of the data so that the precise definition seems cumbersome. Although the idea in the background remains simple, it is a cause of misunderstanding for the students as they tend to focus on the algorithm to calculate the percentiles rather than on the idea.

It is amazing that these five numbers give a clear overview on where the data are located, where the centre of the data lies, how the data are spread out, and even about the shape of the distribution. If one of these groups of the data extends over a smaller range, then the data are more densely distributed there; the shape is skewed if the ranges of the four groups are markedly different. To belong to the top group, the value has to be larger than P_{75} so that the upper quartile marks where the top 25% of the data "begin".

It is important that the students realise that these statistical summaries can take on values in a numerical set different from the original numerical set where the data take on their values. They are not *closed* operations in the original numerical set of the data in an algebraic sense. All these graphs and summary statistics should be introduced when needed in a project or an investigation. As in the example from Section 2.2, the interpretation of these statistics facilitates that students reinforce their statistical reasoning while at the same time the utility of the concepts and procedures is shown to them.

2.4.4. Spirit of Exploratory Data Analysis (EDA)

Sometimes, EDA is seen as a change of techniques as compared to traditional descriptive statistics; e.g., using the median rather than the mean. as it is more robust and not so much dependent on further assumptions on the data. However, EDA is more and can be seen as a change of the paradigm of data analysis. Thus, EDA has a specific spirit or philosophy, which differs from other approaches to data analysis (Borovcnik, 1995, p. 7). In fact, EDA tools intentionally support the interaction.

- Simple concepts enhance a direct interpretation of intermediate results.
- Flexible models allow nearly all types of questions to ask from the models.
- Visual representations of the data facilitate the direct (model-free) detection of inherent patterns.

In this style, two components are investigated in the data: the general pattern (model) or tendency and the residual, which may be written as a visual (heuristic) equation:

$$\text{Data} = \text{Model} + \text{Residual}.$$

The exploratory approach may be characterised as an interactive search for a model (or for several models) and an explanation of the residuals from the context. For example, a box plot summarises three different models (or three stages of modelling): a) the median, which marks the 50:50 dividing point; b) the box, which represents the pattern of the inner 50% of the data; and c) the box with the whiskers, which shows which data can be considered as regular and which should be considered as atypical.

It might pay the effort to label the atypical values in order to identify the units behind these data. While the bulk of the data that fits the general pattern is analysed anonymously as usual in statistics, the atypical values should be

explained from the context and are identified therefore. Sometimes, these outliers are eliminated while other times the analysis is done in two ways, one with these atypical values included, the other excluding these values (as we did in the example of shoe size). The main difference to traditional data analysis is that the data and its context become the centre of considerations and modelling, while in traditional statistics the data are carefully planned and then plays only the role of inserting it into the methods. One aim of EDA is generating hypotheses while inferential statistics resumes the role of testing hypotheses. However, the two different approaches can complement each other ideally as EDA techniques allow to find plausible hypotheses from the context, which should then be tested more formally.[12]

2.4.5. Basic Strategies in Data Exploration

In Section 1.2, we briefly described the fundamental ideas in probability and statistics that have been introduced through Chapters 2–5. In relation to exploratory data analysis, we emphasised the ideas of data, representation of data, variation, distribution, and centre. All these ideas were introduced in this chapter, together with those of spread, location, and shape. These ideas are realised by a series of basic steps in EDA, which were analysed by Borovcnik (1986b; 1987). Below we describe these steps.

Comparing a specific data to a framework. A first basic method in data analysis is to judge data in a specific context. We cannot state that a shoe size of 36 is low for a boy of 14 years unless we see the distribution of other boys of this age for reference. When we judge the road traffic accidents of today, we can use the same day of the year in the past years for comparison, or days with similar weather, or density of traffic (weekend), etc. We can use weekly or monthly data instead. The results will differ by the framework we use for such comparisons.

Condensing information. We condense information in order to simplify the data and gain some knowledge not visible in the data set. From the multivariate data set of the students (Table 2.10), we extract only the data on one variable (e.g., the shoe size), order the data by size and build a frequency table (Table 2.3). Or, we order the data on arm span by a stem-and-leaf diagram. From here, we can easily see where the data are located (minimum, maximum), see its range, etc. This idea is behind transnumeration as described by Wild and Pfannkuch (1999).

Describing the general pattern of the distribution of the data. We have drawn a bar graph for the data on shoe size (Figure 2.2) and we could recognise that the distribution is bimodal so that a split of the students into the subgroups of boys and girls helped to get a more compact distribution.

[12] Other characteristics of EDA are described in Behrens (1997), Biehler (1994), and Borovcnik (1995).

Quantifying the information. To describe the distribution of the shoe size of boys, we might use only a few statistical summaries that give precise information about the distribution. To state that the mean is 41.7 and the standard deviation is 2.56, we might think of the interval of plus or minus two times 2.56 around the mean (36.6, 46.8), which comes close to the whole range of the data. The five number summary also gives quite a few benchmarks for judging the distribution. Measures of location, central tendency and spread serve to summarise the general features of the information inherent in data. Moreover, they allow to judge whether individual data fit into the general pattern or they should be considered as atypical.

Comparing groups. In line with the first idea – to compare single data within a framework – we usually do not only describe the distribution of one group but we do compare it to another group. Thus, we compared the group of boys and girls with respect to the distribution of shoe size and arm span. Using the principles of condensing information, describing the general pattern, and quantifying the key elements of the distributions, we can describe the differences between boys and girls quite well. In a way, we could find a general pattern of a difference of 5 between both distributions while the distributions were quite similar in shape (nearly symmetric with the same spread) so that one can imagine that the distribution of the boys is just shifted by 5 shoe numbers downwards to represent the distribution of the girls then.

Looking for patterns and for deviations (a typical values or outliers). The main idea of data analysis is not just to describe a distribution but to search for generalisable patterns (model) in it and to split the data into the components of model plus residual. And, most important, finding deviations from the model might provide even better explanations for the investigated phenomena than the pattern itself. By using quantitative measures for the spread of the distribution, we could identify atypical values for the shoe size, both in the distribution for boys and for girls. This leads to more insight from the context (there are simply some boys lagging behind at the age of 14; a few girls have extraordinary physical measurements) so that the group of boys and girls fall apart into further subgroups: those who fit to the general pattern and those who are extremely far out in this pattern.

In-depth interpretation of the concepts. As a part of EDA, the intermediary results have to be interpreted to decide about the further steps to follow. In such an interpretation, it is essential to comprehend what the concepts used in the particular example mean more widely. This meta-knowledge about the concepts used is also a key for interpreting the final results in the context. In addition to be a genuine didactic aim, the theoretical interpretation of concepts becomes more important in the modelling view of applied mathematics and especially in EDA.

There are quite a few ways to interpret measures of data such as the mean or the standard deviation in a way that leads beyond their mathematical setting and optimising properties. For example, we could use the idea of a centre of gravity without details of physics. Another interpretation for the mean is due to its

equalising property: if all persons have the same value, their data coincide all with the mean value (and then the mean is interpreted as a fair share). Middle-point and representativeness ideas could also be discussed, in paying special attention to students' common pitfalls.

Standard deviation can be related to the idea of a typical deviation of single data. For quite many, fairly symmetric data sets, roughly two thirds of the data lie between mean ± one standard deviation, which corresponds in some sense to the width of the box in a box plot that contains nearly exactly half of the data. For the median, the idea of splitting the data into two halves might help even if this is not possible and algorithms to operationalise this split are cumbersome. Such ways to interpret statistical measures help to apply these measures in context and to find adequate interpretations of the data in context.

2.5. STUDENTS' REASONING AND POTENTIAL DIFFICULTIES

Even if most EDA methods involve only calculating and interpreting percentages or simple statistical measures as well as drawing and reading various graphs, research suggests that students have problems in understanding the concepts, and in relating these concepts to the context in a meaningful way. One reason for these difficulties is that teaching often focuses on the application of methods instead of the interpretation of results in a given context. Another reason is that even simple mathematical concepts are not as easy as often believed. Below, we present a summary of this research that is useful for teachers to help students to overcome these difficulties (wider summaries are included in Batanero, Godino, Vallecillos, Green, & Holmes, 1994; Shaughnessy, 2007, and Shaughnessy, Garfield, & Greer, 1996).

2.5.1. Graphical Competencies and Communication Skills

Graphical representations of statistical data are a relevant part of EDA. Visualisations are even more important when it comes to communicating scientific results to the public or when students learn about statistical concepts and their interpretation in the context of statistical investigations. Consistently, graphical representations play a central role in teaching statistics at high-school level. Despite this relevance, educational research alerts us that the level of competencies related to statistical graphs that is achieved by the students is not sufficiently high (for a survey, see Arteaga, Batanero, Contreras, & Cañadas, 2012).

Friel, Curcio, and Bright (2001) suggest that graph comprehension develops gradually, through the repeated construction and use of a variety of graphs in contexts that require the student to make sense of data. These authors identify the following elements in a graph:

Background of the graph. This includes colours, grid lines, and images, which could have been imposed on the graph, as well as title and labels, which inform

about the graph content, the represented variables, and the question from the context that is answered by the graph.

Framework of the graph. This includes axes, scales, grid lines, and reference marks, which provide additional information on the size and units of measurements, and where to find the various pieces of information in the graph.

Graph specifiers. Each graph includes specific elements that are used to represent data such as rectangles (in histograms) or points (in scatter plots). Not all specifiers are equally easy to understand. Friel et al. suggest the following increasing order of complexity: a) Position on a linear scale (line graphs, bar graphs, some pictograms), b) position on a non-linear scale (polar graphs, bivariate graphs), c) length (star graphs without reference axes, trees), d) angle (pie charts), e) area (pictograms, histograms), f) volume (cubes, some statistical maps), and g) colours (colour-coded statistical maps).

Bertin (1967) suggests that a graph is a complex semiotic object. The graph itself and every component of it are constructed from signs that require a semiotic activity by those who interpret them. According to Bertin, a reader has to perform three successive operations to read a graph:

External identification. Understanding the conceptual and real-world referents that relate to the information contained in the graph by an analysis of the graph's labels.

Internal identification. Identifying the relevant dimensions of variation in the graph's pictorial content and identifying the correspondence between visual and conceptual dimensions and scales.

Relating the graph to the represented reality. Establishing the correspondence between the visual and the conceptual dimensions in order to draw conclusions from the graph about the real context.

In addition to the competencies above, different authors analysed the depth of the activities involved in reading graphs (see Table 2.9). Bertin (1967) describes different levels of difficulty that were termed by Curcio (1989) in a slightly different way. Friel, Curcio, and Bright (2001) expanded the above classifications by defining a further upper level (marked by an asterisk).

Although students easily achieve the first two levels of reading graphs, the upper levels are hardly achieved even by prospective teacher students. For example, Burgess (2002) analysed prospective teachers' reports about a multivariate data set, and found that many participants were not able to make generalisations about the presented data. Similar results were obtained by Batanero, Arteaga, and Ruíz (2010) in a study with 93 Spanish pre-service primary teachers; although most participants were able to construct adequate statistical graphs as part of a statistical project, only one third of the participants were able to reach a conclusion regarding the underlying research question.

Table 2.9. Various classifications of levels of difficulty in reading graphs

Bertin's levels of difficulty	Curcio's levels *extended by Friel et al.
Extracting data Direct reading of the data in the graph.	*Reading between the data* Literal reading of the graph without interpreting the information behind it.
For example, reading the frequency of a specific value in a bar graph.	
Extracting trend Perceiving a relationship between two data *subsets* that can be defined a priori or are visible in the graph.	*Reading within the data* Interpreting and integrating the data in the graph by comparing single data to each other or to the whole graph.
For example, determining the mode of a distribution in a bar graph visually. This requires isolating the subset of data around the peak of the distribution (which is visible from the graph) and the remaining data and compare the frequencies in these subsets of data.	
Analysing the structure of the data Comparing trends or clusters and making *predictions* from data in the graph.	*Reading beyond the data* Making predictions and inferences from the data to information that is not directly reflected in the graph.
For example, judging visually the differences in modes and range of two distributions in an attached bar graph or estimating the *Y* value for a given *X* value in a scatter plot for a point not included in the graph.	
	Reading behind the data[*] Judging the method of data collection, the validity and reliability of the data.
For example; discussing factors blurring the findings, which have not been controlled by the data collection method.	

2.5.2. Errors in Producing Graphs

Other researchers discuss common errors students make when they produce statistical graphs. The first step in building a graph is selecting a type of graph that is adequate to represent the scale of the variable and the problem under study. Yet, students often fail to find an adequate graph. Li and Shen (1992) analysed the statistical graphs produced by their students when working within a statistical project. They found students who displayed the data of a qualitative variable by a frequency polygon, who used attached bar graphs to represent bivariate data, or represented unrelated variables in the same diagram. Similar errors were obtained by Batanero, Arteaga, and Ruiz (2010) in a study with prospective teachers.

Li and Shen also found out that the following problems are associated to the scales of the variables in a graph: a) choosing an inappropriate scale for the intended purpose (e.g., not covering the entire range of variation of the variable represented); b) ignoring the scales in one or both axes; c) not specifying the origin of coordinates; and d) not providing a suitable division for the used scale (their division was often not fine enough).

Wu (2004) categorises the errors of high-school students when building statistical graphs: 1) errors related to scales; 2) errors in specifiers, title, or labels; 3) confusion between similar graphs (e.g., between histogram and bar graph); 4) confusion between variable and variable value.

Other authors focussed on specific graphs. For example, Pereira-Mendoza and Mellor (1991) found that students make simple reading mistakes (e.g., using a horizontal bar graph instead of a vertical bar graph). Lee and Meletiou-Mavrotheris (2003) describe four main errors related to constructing and interpreting histograms:

– Perceiving histograms as a representation of isolated data assuming that each rectangle refers to a particular observation rather than to a range of values.
– Focussing on the vertical axis and compare the differences in heights of the bars instead of comparing areas (when comparing the variation of two histograms).
– Interpreting a histogram as a bivariate graph (e.g., as a scatter plot).

In a study with pre-service primary teachers in Spain, Bruno and Espinel (2009) found the following difficulties with the construction of histograms and frequency polygons: separating histogram rectangles (i.e., inserting a gap between adjacent rectangles), inadequate labelling of the axes (they might have problems with the number line), and ignoring zero-frequency intervals. All the above difficulties were replicated in a study with 207 prospective primary school teachers by Arteaga, Batanero, Contreras, and Cañadas (2016). The authors attribute these errors to *semiotic conflicts* or to failures in the semiotic activity involved in producing or in interpreting graphs.

2.5.3. Understanding Measures of Central Tendency or Centre

Although the idea of mean of a data set is apparently simple, a wide research, part of which is summarised by Jacobbe and Carvalho (2011), shows a variety of difficulties in students' and teachers' understanding of the concept of centre. In an early research piece, Pollatsek, Lima, and Well (1981) described that many students combined the means from two different samples as if they were simple data; i.e., they combined two means without weighting the values according to the relative size of the samples where they come from. The same mistake was observed by Li and Shen (1992) in students working with data grouped in intervals; many students ignored the frequency of the single intervals when computing the mean from this type of table.

In a research directed to evaluate conceptual understanding, Mevarech (1983) suggested that students mistakenly assume that a set of numbers together with the operation of arithmetic mean constitutes a mathematical group satisfying the

properties of operational closure (the mean of two means of samples is the mean of the combined sample), associativity, identity, and inverse element. Students may unconsciously use these properties in this context where they do not apply as they know that the arithmetical operations on which mean and variance are based fulfil these laws.[13]

Continuing the previously mentioned research, Strauss and Bichler (1988) studied the development of children's understanding of the following properties of the average (arithmetic mean):

a. The average is located between the extreme values.
b. The sum of the deviations from the data to the average is zero.
c. The average is influenced by all the values in the data set.
d. The average does not necessarily equal one of the single data values.
e. A value of zero must be taken into account in calculating the average.
f. The average is representative of the values that are averaged.

Their results also suggest that children's understanding of the mean changes with age and that properties a., c., and d. are easier than the others. Since the mean is often described as a "representative" value of the distribution, Campbell (1974) suggested that there is a tendency to situate the mean in the centre of the range of the distribution. This property, however, is only fulfilled when the distribution is symmetric; in case of atypical values or for skewed distributions, the median might not only be a better descriptive representation of the data[14] but also a statistic that is easier to understand.

Understanding the idea of "representative value" implies, according to Russell and Mokros (1991), three different competencies. The authors asked 12 year-old students to solve problems that involved the construction of data sets with a specific value for the mean. In other tasks, the students were asked to describe the information in the data provided by the various measures of central tendency (mean, median, and mode). The authors suggest that understanding these measures involves the following competencies (see also Mokros & Russel, 1995):

a. Selecting the best representative central value for a given data set.
b. Building a data set, which has a given central value (e.g., a given mode).
c. Understanding the effect that a change in part of the data has on the various statistics for the centre (mean, median, and mode).

Goodchild (1988) provided 12 year-old students with matchboxes with a label "average content 35 matches" on them and asked them to build a suitable distribution for the content of 100 boxes. The results showed a poor understanding of mean and variability in random settings since the distributions had little spread.

[13] There seems to be a strong tendency to transfer (unconsciously) familiar rules of calculation with numbers to other operations as here with the operation of averaging data. That does not imply that the students openly think in structures of mathematical groups.

[14] There is no general rule, which value to use as a representative of the data. It simply depends on the purpose. See our previous comments in Task 2.3 and in Section 2.4.3.

García and Garret (2006) replicated these findings when working with 17 year-old students.

Russell and Mokros (1991) classified students' misconceptions about central values into four categories:

a. The "most frequent value" or mode, when students assume any central value is the most frequent in a distribution.
b. The "most reasonable value", when it is assumed that any central value belongs to the data set.
c. The "midpoint", where many students assume that the mean or median is the geometric middle point in the distribution.
d. The "algorithm leading to the measure", where students remember the computation of the central values, but are unable to explain their meaning.

Of course, the statement in a. is not generally true even if the distribution is symmetric (it could have two symmetric peaks outside), and it does clearly not stand for the mean and the median of skewed distributions. The property in b. cannot hold generally as we have seen in the project on shoe size. It is amazing how strong the drive is to "force" the representative value to be one value that the investigated variable can attain. It might be induced by the wording "representative value". The misconception in c. about the midpoint might originate from the fact that, geometrically, a midpoint is in the middle of the range of the variable. The statement d. does not surprise as the algorithm to calculate is the only piece many students "understand" after teaching because interpretative activities are rarely used in the classroom. To prevent all these misconceptions, the teacher should clarify properties of the median and the difference of it to the mean in teaching.

All the above difficulties and misconceptions show that many students only acquire a superficial understanding of the mean and of the computational algorithm. Unfortunately, they do not attain a profound understanding of the various statistics for the central value. Parts of these difficulties are replicated in a study with prospective primary school teachers (n = 273) carried out by Navas, Batanero, and Godino (1997); they found problems in understanding the relationships between mean, median, and mode, and a quarter of the participants were influenced heavily in estimating the mean when atypical values occurred in the data.

Similar results were found by Groth and Bergner (2006) who investigated 46 pre-service elementary and middle school teachers. Other research (e.g., Watson & Moritz, 2000) is directed to describe students' developmental stages supported by neo-Piagetian frameworks such as Biggs and Collis (1982). Responses by students in different tasks are used to classify them in discrete stages along a uni-dimensional (continuous) scale.

A different approach is taken by Batanero, Cobo, and Díaz (2003) who perform an epistemological analysis of the concepts of central value, including their different properties, representations, as well as tasks and problems related to them. Based on the epistemological results, they constructed much more systematically varied items than those used in previous research and aligned them to a

comprehensive questionnaire to assess students' understanding of all these mathematical objects. Their results with a sample of 14 year-olds (n = 168) and 16 year-olds (n = 144) suggest a more complex non-linear structure of student's understanding and its development. In the older group, they also found a positive effect of instruction on understanding the properties of the mean, but no effect could be established for the median.

2.5.4. Understanding Spread

In addition to measures of the centre, measures of spread complement the description of a distribution since two different data sets with the same average may have different degrees of variability. Konold and Pollatsek (2002) emphasised the importance of jointly considering variability (noise) and centre (signal) because both ideas are needed to find meaning when analysing data. However, people tend to ignore the spread of data as Campbell (1974) has pointed out in an early research piece.

The standard deviation measures how strongly data depart from a measure of central tendency. Nevertheless, Loosen, Lioen, and Lacante (1985) noticed that many textbooks put a stronger emphasis on the heterogeneity among the observations than on their deviations from a statistics for the central tendency.[15] As Loosen et al. note, the words variation, dispersion, or spread can be interpreted in different ways, depending on whether they refer to a relative dispersion (in relation to a central value) or an absolute diversity. Students tend to confuse these types of variation. This confusion was also shown in Bakker's (2004) study, where students used expressions such as "average", "range", and "spread" in a non-normative way.

Recently, the task of comparing groups (with data often presented in graphical format) has been used to investigate students' understanding of measures of centre and spread. Using this type of task, Watson, Kelly, Callingham, and Shaughnessy (2003) describe students' levels of reasoning about variation on a one-dimensional scale, and find that only few of them reach the highest level.

In addition to conceptual difficulties, students also produce arithmetic errors. In university students, Mevarech (1983) found difficulties for the variance similar to those for the mean. Also here, the author explains the errors by a hypothesis that the students tend to assume a group structure for the operation of variance on data sets. To meet difficulties in understanding the standard deviation, delMas and Liu (2005) suggest that students need to understand the concepts of distribution, mean, and deviation from the mean first in order to build the notion of standard deviation.

[15] The interrelations are quite sophisticated. If we consider variance or standard deviation (and not other measures of spread), then the variation defined as mean of squared deviations between *all pairs* of single data equals two times the mean of the squared deviations from the mean value. For the related standard deviations a factor of √2 applies.

2.5.5. Understanding Order Statistics

The study of order statistics presents computational as well as conceptual difficulties. First of all, the computation of median, quantiles, and percentiles is taught with a different algorithm for data grouped in intervals and for non-grouped data; even for raw data, the definition does not always yield a unique measure, and software packages all use different algorithms that lead to different values for the statistics especially for smaller data sets.

Schuyten (1991) states that even university students find it difficult to accept two different algorithms for the same concept, and, moreover, different values for the same parameter depending on the chosen algorithm or on the width of the intervals. She also points out the large distance between the conceptual knowledge of the median and the algorithm employed to obtain its value. In going from the definition of the median as "middle value of the distribution" to its calculation, there are many steps, which are neither sufficiently stated nor well understood.

The final algorithm consists in solving two inequalities that are formulated by the empirical cumulative distribution function $F_n(x)$; one is derived from the condition that the relative frequency of data less or equal x has to be greater or equal ½; and – symmetrically to it – the relative frequency of data greater or equal to x is greater or equal to ½.[16] If F_n were continuous, one had just to find a value x with $F_n(x) = ½$. However, the difficulties arise from the fact that F_n is not continuous and is given only by means of a table of numerical values. If the raw data are available, the two inequalities are uniquely solvable if we additionally take the smallest value with such properties (or the midpoint of these values). If the data are grouped in intervals, then the graph of F_n is not known within the intervals. We know F_n only at the endpoints of the intervals and thus have to interpolate these values on the boundaries of the median interval in order to approximate the median. A similar procedure delivers the calculation of quantiles in general.

From the description of the mathematical subtleties of calculating the median, one can guess that the students must have severe problems to understand the algorithm and do not know why the mathematicians make it so complicated. As the conceptual task of dividing the data into two groups of equal counts (one containing the larger values and another including the smaller values) and obtaining a dividing point (the median) is easy to understand, the students may be puzzled why it is so difficult to calculate this dividing point. Furthermore, different software uses different algorithms[17] to estimate the quantiles from sample data, which contributes to the troubles of the students who find distinct numerical values from different software.

Barr (1980) noticed the lack of understanding of the median in a study with students aged 17 to 21 years. In Mayén and Díaz's (2010) research with 14 and 17

[16] The first condition may easily be formulated in terms of F_n as $F_n(x) \geq ½$. The second condition, however, is more complicated to denote as $1–F_n(x^-) \geq ½$ where x^- means a limit to x from below.

[17] The different algorithms used – each with its own merits – are determined by optimising the properties of *estimating* the related quantiles of probability distributions. There are various criteria to optimise and different software packages use different criteria.

year-olds, most of the students had grasped the idea that the median is a central value but made different interpretations of this centre. Some of them interpreted the median as the middle point[18] of the figures in the frequencies column, or as the middle point of the values of the variable column, or even as the middle point in the list of numbers before they have been ordered. While it is intuitive to talk about the lower half or the lowest quarter of the data, it takes a sophisticated definition to make this intuitive concept precise and to use this definition to calculate the median or the lower quartile. However, as the students are so focussed on algorithmic understanding, the algorithm is explained to them in detail rather than the concept. And here, the algorithm is the difficulty, not the concept.

2.6. ADDITIONAL RESOURCES

The analysis in the previous sections suggests that neither teaching nor learning statistics is easy. Fortunately, we can build on many educational resources that might help teachers in their educational tasks. A few of them are described below.

2.6.1. Journals and Books

First, there is a wide variety of papers that describe research on students' misconceptions and difficulties as regards different concepts or procedures, and statistical reasoning in general. This research also provides assessment items, which teachers may use in their own classroom either to evaluate the students or to discuss the concepts with the students. Many other papers describe successful teaching experiences or suggest innovative teaching resources.

Journals[19] like Journal of Statistics Education (JSE), Statistique et Enseignement (SeE), Statistics Education Research Journal (SERJ), Stochastik in der Schule (SiS), Teaching Statistics (TS), or Technology Innovations in Statistics Education (TISE) pursue the improvement of statistics education as their prime goal; many of them are freely available on the Internet. Some are more class-oriented and also supply material useful for teaching; others publish research papers in statistics education.

Moreover, some associations and research groups make their resources freely available on the Internet; for example, the *International Association for Statistical Education* (iase-web.org/) includes an online version of the proceedings of the ICOTS (International Conferences on Teaching Statistics) and IASE Round tables, and informs about other statistics education conferences.

---

[18] Many of these misconceptions are simply a wrong application of the middle-point idea. The students have heard that the median is something like a middle point – the point in the middle. Students have to learn that the procedure can only be applied to the ordered values of the variable, not to the unordered, and not to the frequencies.

[19] For convenience, we insert links to these journals here: JSE: www.amstat.org/publications/jse/; SeE: publications-sfds.math.cnrs.fr/index.php/StatEns; SERJ: iase-web.org/Publications.php?p=SERJ; SiS: stochastik-in-der-schule.de/; TS: www.teachingstatistics.co.uk; TISE: escholarship.org/uc/uclastat_cts_tise.

There are also excellent books devoted to the teaching of statistics. Some examples are the series *Statistics in your world* (Schools Council Project on Statistical Education, 1980), *Data Driven Mathematics* (in particular, the books by Landwehr, & Watkins, 1995, and Landwehr, Watkins, & Swift, 1987). Some specific books in different languages also discuss pedagogical issues in statistics education (Batanero, 2001; Borovcnik, 1992, Hawkins, Jolliffe, & Glickman, 1992, or Watson, 2006).

2.6.2. Data Sets

As we suggested in Chapter 1, interesting data sets are widely accessible on the Internet, which can increase students' interest in the study of statistics (Hall, 2011). When working with real data, students can better understand the different types of variables, make sense of the used modelling, and appreciate the value of statistics for clarifying questions within a context.

An example is the United Nations web server where the data used to establish the human development reports (hdr.undp.org/en/data) are available together with visualisation tools. We use such data in Chapter 4 to explore the ideas of correlation and regression. Other international organisations such as the World Health Organisation (www.who.int/) and many statistical offices around the world also offer free access to their databases. Some of these statistical offices include a section of their websites devoted to statistics education with a variety of resources. Some examples are Statistics New Zealand or the statistical bureaus of Portugal (provides material also in English) and Spain.[20]

In addition to general data bases, we find also specific data sets directly useful for statistics teaching. An example is the website of the Journal of Statistics Education that hosts a variety of data sets, partly linked to papers in the journal. The Data and Story Library (lib.stat.cmu.edu/DASL/) offers various data files and related descriptions that are classified by theme or by statistical methods useful for implementing specific methods in teaching or discussing contextual questions in the statistics class.

2.6.3. Internet Resources

Many teachers and researchers host specific web servers devoted to statistics education that contain varied resources. In general, these collections include various applets on descriptive statistics, probability, and inference. Some examples are listed below; the links are given in a footnote[21].

---

[20] New Zealand: www.stats.govt.nz/tools_and_services/schools_corner.aspx, Portugal: www.alea.pt/; Spain: www.ine.es/explica/explica.htm.

[21] CAUSE: www.causeweb.org/; CensusAtSchool: www.censusatschool.org.uk/international-projects; Gapminder: www.gapminder.org/; History of Statistics: www.york.ac.uk/depts/maths/histstat/; Rossman/Chance: www.rossmanchance.com/applets/; Statlet: www.math.usu.edu/~schneit/CTIS/; Understanding Uncertainty: understandinguncertainty.org/; VESTAC: lstat.kuleuven.be/java/version2.0/; XLstat: wwwg.uni-klu.ac.at/stochastik.schule/Boro/index_inhalt.

CAUSE. This webpage includes information and resources arising from a strategic initiative of the American Statistical Association. The aim is improving the teaching of statistics at undergraduate level. It includes teaching resources. The organisation also promotes workshops for statistics teachers and research.

CensusAtSchool. This is an international project aimed to encourage children to get involved with statistics and to provide data and resources for teaching. The data are collected by children from different countries taking part in the project.

Gapminder. A Swedish foundation that tries to promote sustainable global development and achievement by increased use and understanding of statistics. It includes the software gapminder that serves to visualise multivariate data along time and a variety of data sets as well as interesting videos.

Materials for the history of statistics. A collection of materials that includes bibliographies, biographies, portraits of statisticians, historical texts, e-resources on the history of the subject, and quotations.

Rossman/Chance applet collection. For descriptive statistics, there are applets showing a histogram and box plot simultaneously to learn how these representations shed a different light on the distribution, another applet for calibrating the intuitive estimation of correlation coefficients for data, an applet for linear models fitting to a scatter plot. Many other applets are related to the investigation of sampling distributions (for Chapter 3) and statistical inference (Chapter 5).

Statlet. Various applets to train an intuitive comprehension of mean, median, standard deviation from a bar graph of the data, and correlation from scatter plots. Further applets deal with random phenomena and methods of statistical inference.

Understanding uncertainty. This project is coordinated by David Spiegelhalter with the aim to make people more competent in dealing with probability and especially the concept of risk, which is based on probability. It provides insightful papers, videos, animations, and an interesting weblog where topical problems of (understanding) risk are discussed.

VESTAC. Produced at the University of Leuven within a European project on Visualization of Experimentation with STAtistical Concepts. Graphical representations of simulated data are used to illustrate probabilistic phenomena and statistical methods.

XLstat. It offers templates for simple techniques of data analysis that are not directly available in Excel; it also provides didactical animations for essential concepts of data analysis (and of probability).

In addition, many other web servers offer other types of statistics education resources. This includes interactive applets, research papers, assessment materials, and various kinds of data sets. We finish with a reference to a very beautiful resource, namely "The basic laboratory for randomness, probability, and combinatorics" (Laboratorio básico de azar, probabilidad y combinatorial[22]) but it is in Spanish.

APPENDIX: DATA

Table 2.10. Data sheet on the data of the class of students

ID-Nr.	Gender	Shoe size	Weight	Height	Arm span	Euros	Sport
1	1	35	65	164	158	28.0	2
2	1	40	64	166	171	12.4	1
3	1	36	65	179	171	3.6	2
4	1	42	68	170	172	218.0	1
5	1	42	68	175	172	19.0	3
6	1	41	69	172	172	0.0	2
7	1	41	66	175	177	0.0	2
8	1	45	74	183	178	64.0	3
9	1	42	69	176	179	13.0	2
10	1	43	74	175	179	29.6	2
11	1	41	82	174	180	11.0	3
12	1	42	68	178	180	7.0	2
13	1	46	86	191	180	2.4	3
14	1	41	62	178	181	15.4	1
15	1	42	74	180	182	71.4	2
16	1	43	72	184	185	300.0	3
17	1	44	70	*	185	25.0	1
18	1	42	68	185	185	5.0	3
19	1	43	71	185	187	28.0	3
20	1	43	81	184	188	204.8	2
21	2	36	68	158	150	16.0	1
22	2	35	70	156	152	7.0	1
23	2	36	50	159	153	20.0	2
24	2	37	55	160	154	16.0	3
25	2	38	64	166	155	0.0	3
26	2	37	57	161	155	18.0	1
27	2	37	50	155	155	20.0	3
28	2	35	50	155	155	20.0	2

(Continued)

[22] The link is: ntic.educacion.es/w3/eos/MaterialesEducativos/mem2010/labazar/.

Table 2.10. (Continued)

ID-Nr.	Gender	Shoe size	Weight	Height	Arm span	Euros	Sport
29	2	35	52	160	157	80.6	2
30	2	36	56	162	158	10.0	2
31	2	37	46	160	158	205.4	1
32	2	38	62	167	159	48.0	2
33	2	37	59	161	160	30.8	2
34	2	36	55	163	160	28.0	3
35	2	37	53	164	160	35.0	1
36	2	37	51	165	160	4.0	2
37	2	37	58	160	160	8.6	2
38	2	36	60	165	160	10.2	1
39	2	36	58	162	160	8.0	2
40	2	38	60	165	161	176.0	1
41	2	36	50	163	161	19.2	2
42	2	37	53	162	162	40.0	2
43	2	38	58	163	162	51.6	1
44	2	37	58	162	163	9.0	2
45	2	38	60	161	164	20.0	2
46	2	38	57	169	164	44.6	3
47	2	37	62	169	165	104.0	3
48	2	37	60	167	165	4.8	2
49	2	37	50	167	165	28.0	2
50	2	40	65	165	165	24.2	2
51	2	38	58	163	166	11.4	2
52	2	38	58	164	166	5.0	2
53	2	38	60	170	168	8.0	2
54	2	39	66	168	168	5.0	1
55	2	38	52	170	171	20.0	1
56	2	41	63	172	171	80.0	2
57	2	37	64	172	175	27.6	1
58	2	40	74	175	178	22.4	2
59	2	40	62	174	179	0.0	2
60	2	39	63	173	180	120.4	2

* missing value

CHAPTER 3

MODELLING INFORMATION BY PROBABILITIES

Probability embraces a cluster of ideas that help us to make predictions and judgements by modelling random situations suitably. Ideas such as experimental data, weight of uncertainty, and equiprobability contribute towards the concept of probability. The concept of independence is a basic prerequisite for the frequentist interpretation, whilst conditional probabilities are essential to adapt personal weights in view of new information. In this chapter, we explore such ideas around two main themes: modelling situations with the aim to achieve better decisions, and generalising information from (random) samples to populations. We explain central theorems that determine both areas of applications and describe the main teaching goals as well as students' difficulties and wrong intuitions concerning to that topic. Resources that facilitate a wider understanding of probability supplement the chapter.

3.1. INTRODUCTION

Once the students are familiar with exploratory data analysis, and before starting inference, they need some basic knowledge of probability. Thus, modelling competencies of the students are reinforced that will hopefully be useful for their personal and professional lives. Although the idea of chance is as old as civilisation (Bennett, 1999), the first theoretical developments in probability are quite recent and the concept has received different interpretations that still coexist and are included in the curriculum (Batanero, Henry, & Parzysz, 2005; Borovcnik & Kapadia, 2014a; Hacking, 1975).

Ancient times were signified by a deistic approach to probability; probabilistic rituals were used for divination and people did not think of analysing probability systematically as this seemed to contradict their purpose. Furthermore, mathematical progress in geometry, for example, fitted to an idealistic Pythagorean philosophy, whereas the systematic analysis of relative frequencies did not seem scientific though data from omnipresent games of chance have been available and the devices such as dice did show already a nearly perfect symmetry (see, e.g., Batanero, 2016; Bennet, 1999; Borovcnik & Kapadia, 2014a).

The poem *De vetula* from the 13th century may be seen as a first theoretical progress as – for the first time – a systematic evaluation of chances has been attributed to combinatorial multiplicity (Bellhouse, 2000). It took further centuries until the famous exchange of letters between Fermat and Pascal (1654/1962) for the next step of conceptual progress when a hypothetical continuation of a game was modelled by equal probabilities to decide the fair division of stakes (see Batanero, Henry, & Parzysz, 2005; Borovcnik & Kapadia, 2014a).

Bernoulli (1713/1987) is the next milestone of conceptual progress insofar as a theorem was proved that "relative frequencies" converge to the underlying probability. A bit further in time, with Bernoulli's theorem not yet known everywhere, Bayes (1763/1970) proved another theorem, namely the convergence of a judgement about an unknown probability of an event towards the relative frequencies of this event – it has become known as the inverse probability and establishes an early link between probability and statistical inference while at the same it openly combines the aspects of probability as a personal judgement and probability as relative frequencies.

Conceptually, the next great leap in progress was marked by the introduction of the normal distribution by Gauss (1809) and Laplace (1810), which was accompanied by an early proof of a simple version of the Central Limit Theorem.

After that, a decline was to be seen as probability was used carelessly in various fields among them also in law (by Poisson, e.g.). Yet, the rising physics needed the models of probability to formulate the new laws in thermodynamics so that at the turn to the twentieth century an urgent need was felt to provide a sound foundation for probability. Hilbert (1900) sets an agenda to axiomatise physics, which meant in fact, to axiomatise probability. However, though the community of mathematicians met Hilbert's challenge, it took one further unsuccessful attempt of von Mises (1919) to define axioms to characterise probability directly as a limit of relative frequencies and three more decades until Kolmogorov (1933/1956) solved the problem.

The axiomatisation was strongly linked to the frequentist interpretation of probability though it did only mimic the properties of relative frequencies.[1] It is interesting to note that de Finetti (1937/1992) derived another axiomatic solution for the concept of probability that was based on a preference system and characterised a non-empirical personal probability; consequently, the scientific debate reached the same point where it had already been with Bayes. Accordingly, in the 1950's, a fierce debate on the foundations was revived (see Borovcnik, 1986a, Hacking, 1975, 1990), which finally led an inconclusive situation.

Nowadays, the scientific community acknowledges that the conflict between objectivistic (frequentist) and subjectivist (personal) conceptions of probability cannot be resolved so that in applications, researchers use those methods that are most suitable for the purpose in question regardless of the connotation of probability. For teaching, the best choice is to enable a pluralistic conception of probability in the learners so that they can develop intuitions on the full complexity of the concept of probability.

In high school, probability is often presented as limit of long-term relative frequencies of an experiment that is repeated independently under the same conditions, which is the core of the frequentist approach and also connects to inference in Chapter 5. There are several reasons to discuss other interpretations of probability as well. Expectations for high-school level include (e.g., CCSSI, 2010; MEC, 2007; NCTM, 2000) the understanding of concepts such as sample space,

[1] The relative frequencies were not part of the axioms, however.

probability distribution, expected value of random variables in simple cases, conditional probability and independent events as well as computation of the probability of compound events, total probability, and Bayes' rule. In the GAISE document (Franklin et al., p. 84), an introduction to the normal distribution as a model for sampling distributions, basic ideas of expected value, and random variation are also mentioned. In the Spanish curriculum for the social sciences branch of high school (MEC, 2007; MECD, 2015), besides the binomial and normal distributions the Central Limit Theorem and its implications for approximating the binomial by the normal distribution is also included.

The aim of this chapter is, to summarise these ideas about probability from various perspectives that are relevant for students at this level and include meaningful applications. We emphasise the idea of modelling where the modeller perceives probability as a virtual entity rather than as property of reality, which is vital in applications beyond games of chance. After describing some basic ideas, we introduce several probability models that provide connections to inference (Chapter 5) such as the binomial and normal distributions and summarise the main learning goals. In Section 3.6, we discuss some persistent misconceptions of probability and provide research-based explanations for them. Additional teaching resources and ideas supplement the chapter.

3.2. TEACHING SITUATIONS TO CHARACTERISE PROBABILITY

Probability is a multi-faceted concept; therefore, we present a range of teaching activities related to frequentist, subjectivist, and Laplacean approaches. The mathematical concept unites all these interpretations; it is vital to interlink the seemingly differing conceptions so that a wider comprehension of probability emerges in the learners.

3.2.1. Frequentist Probability: Investigating Coin Tossing

A first way to introduce probability is to link it to relative frequencies, using the intuitive observation that relative frequencies "converge" to a value (the underlying probability) if the experiment is repeated many times. The following activity from Freudenthal (1972) is used to build up a frequentist interpretation; at the same time we clarify one key idea of probability and statistics: When estimating an unknown probability (or parameter), the variability of estimates gets smaller when the sample size increases. This activity also provides a starting point for the study of inference (Chapter 5).

Task 3.1. Each student tosses a coin 100 times recording 1 for heads and 0 for tails. Then the students should perform the following analyses:

a. Divide the series of 100 results into twenty samples of size 5 each and determine the number of heads (0, 1, ..., 5) in each sample (see for example for the first 20 results in left side of Table 3.1). Calculate the proportion (relative frequency) of heads in each

sample of size 5 and use it as an estimate for the "unknown" probability p of heads. Draw a bar graph of the distribution of the 20 estimates.

b. Combine pairs of samples of size 5 to form samples of size 10 and estimate the unknown probability p of heads (10 estimates per student).

c. Combine pairs of samples of size 10 and draw a bar graph of the estimates of p based on samples of size 20 (5 estimates per student).

d. Compare the graphs of the estimates among single students and discuss whether there is a general pattern to see. Then combine the results of all students and draw the graphs of the empirical distributions for the estimates from each sample size (Figure 3.1). Finally, try to recognise a pattern and discuss whether we can identify differences in the precision of the estimates for the probability of heads with various sample sizes. Is the quality of estimates for p influenced by the sample size?

When working with this activity, the patterns of small samples (size 5) show great variation between individual students. No general pattern is visible; obviously, an estimate for the probability of heads based on five tosses is unreliable and more data is needed.

Table 3.1. Results of the experiment of 10 students

a. First twenty tosses of one student

Toss N	Result	Number of heads in samples of size 5	10	20
1	0			
2	0			
3	1			
4	0			
5	1	2		
6	1			
7	1			
8	1			
9	0			
10	1	4	6	
11	1			
12	0			
13	0			
14	1			
15	1	3		
16	0			
17	0			
18	1			
19	0			
20	0	1	4	10

b. Frequency table of the data of ten students

Estimates of p from samples of size 5			
Number of heads	Sample proportion	Frequency absolute	relative
0	0.00	6	0.030
1	0.20	22	0.110
2	0.40	77	0.385
3	0.60	56	0.280
4	0.80	32	0.160
5	1.00	7	0.035
All		200	1.000

In Table 3.1, we present data from ten students that joined such an experiment. We only show the first 20 tosses of one student to give an idea how the analysis is performed. We use the relative frequencies in each sample of size 5 to estimate the probability p of heads; our repeated estimates vary greatly as shown by the first four samples: 2/5 = 0.40, 4/5 = 0.80, 3/5 = 0.60, and 1/5 = 0.20. The right side of Table 3.1 shows the distribution of the estimates from all 200 samples of size 5 that results from a statistical analysis of the data lists from all ten students.

Comment on the design of the task

Each student should toss a coin and insert the results into a prepared form with the number of the toss in the first column, the result of the toss in the second; in the third column (with a header 5) the number of heads in each five tosses is listed. Combining pairs of five tosses yields the result of 10 tosses, and finally combining pairs of ten yields the number of heads in 20 tosses (columns 4 and 5 with the headers 10 and 20). In this way, the protocol is easy to handle and it takes little time to establish it.

There are some advantages of the suggested approach. First, the calculation is done step by step and the numbers of heads in subsequent samples show the variability of the estimates of the unknown probability p (if divided by the sample size), which highlights the repeated estimation of the unknown probability. Second, coin tossing is done practically and not simulated on the computer where mathematical algorithms yield random numbers; something that establishes an inherent paradox and which may lead (at least) to confusion if not to intuitive objection on the learner's side. The disadvantage of the summarising procedure is that each student has only 20 samples of size 5 (100 tosses), 10 of size 10, and only 5 of size 20. The small data sets introduce a further element of variation; however, if there are ten students in the class then the combined protocol has enough data to discover a clear pattern.

Due to the reasons above, we prefer the manual approach in contrast to machine simulation of coin tossing where we could simulate many more samples of 5, 10, and 20 each to study the patterns.

We supplement the analysis for the larger sample sizes. The frequency distributions of the estimates (Figure 3.1) show a clear pattern: the larger the sample on which the estimates are based, the closer they are grouped around the value ½. The interval from 0.30 to 0.70 has been lightly shaded to highlight that the proportion of samples that fall within that interval increases with growing size of the sample.[2]

Students can simulate more tosses and analyse larger samples or just *extrapolate* the behaviour of estimates to larger sample sizes: The estimates of the unknown probability p improve if the size of samples increases and the *distribution* of relative frequencies tends towards this unknown probability.[3]

---

[2] Figure 3.1 illustrates the *Weak Law of Large Numbers* that states that the interval may be arbitrarily small around the "true" value of p and yet, with increasing sample size, the relative frequency will fall into that interval with a limiting probability of 1.

[3] Contrary to some students' beliefs, *patterns*, in which the results occur, play no role whatsoever.

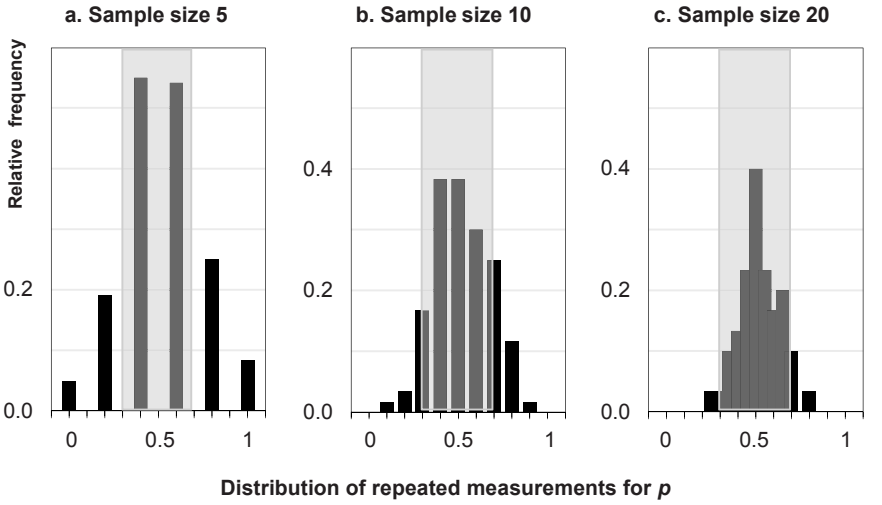


Figure 3.1. Precision of estimates of the probability of heads from different sample sizes

3.2.2. Subjectivist Probability: The Insurance Contract

Another key idea is using odds to calibrate subjective probabilities, which weight our confidence in uncertain situations. Betting on an event E, the odds for E are the ratio p:q of the probabilities that E occurs and E fails, i.e., $P(E):P(\bar{E})$. For example, odds for a six on a die are 1:5. From odds, we can compute the probability by $P(E)=\frac{p}{p+q}$, which yields $\frac{1}{1+5}$ for the six.

This view of probability is introduced in the following teaching situation (Borovcnik, 2006a). In this activity, the students investigate the convenience of taking out a full-coverage insurance policy for a new car and how both the car owner and the insurance company can benefit from the contract.

Task 3.2. A person is indecisive about taking out a car insurance or not. We simplify matters and consider only two cases of damage: total wreckage (with a cost of € 30,000) and no accident. This person then has two possible decisions: d_1 taking the car insurance with a cost of € 1,000; and d_2 having no car insurance with costs depending on future outcomes (see Table 3.2).

a. If the odds for total wreckage are 1:9, determine a value for each of the two decisions and find out, which decision is the better one.

b. Find odds for total wreckage, for which the value of both decisions is the same so that both decisions are equally good.

Table 3.2. Car owner's cost of the various decisions under the prospective circumstances

Potential future	Cost [€] according to selected decision	
	d_1 = Insurance	d_2 = No insurance
t_1 = No accident	1,000	0
t_2 = Total wreckage	1,000	30,000

The car owner pays to avoid uncertainty (possible loss) while the insurance company is paid for assuring the owner certainty (no loss). The insurance company grounds its contract on estimates of the probability for the various damages (and related payments) by past frequencies of accidents. The car owner has to consider personal risks: driving skills, driving habits, as well as possible incidents caused by other road users.

For question a., we calculate a probability for the wreckage of 1/10 from the given odds of 1:9. For option d_2, the expected cost for the driver is $30{,}000 \cdot 0.1 + 0 \cdot 0.9 = 3{,}000$ Euros. Ideally, we translate the probabilities into "frequencies" and do as if from 10 contracts we have one total wreckage and average the amount to pay over all contracts. For option d_1, the cost is 1,000 whether there will be an accident or not. Therefore – in terms of expected cost (used intuitively here) – it is cheaper to take out the insurance policy.

We can reformulate question b. into the phrase: In case of which odds for a total wreckage should the driver change the decision from buying the insurance to no insurance? We have to find the relative weights for t_2 against t_1 for which the decisions d_1 and d_2 become equivalent. A short calculation (with variable odds of $1{:}q$) shows that for 1:29 the expected cost for the car owner will be 1,000 Euros for either decision. If the odds for the total wreckage are smaller[4] (e.g., 1:35), no insurance is the cheaper decision. If the driver expects smaller but more frequent accidents such as parking damages, he has more reasons to sign the contract, irrespective of total wreckage[5]. As we see, personal probabilities may differ from a frequentist estimate for all drivers.

3.2.3. Laplace (A Priori) Probability: Calibrating Weights of Evidence

Games of chance establish an important field of application for probability. The assumption of equal probabilities for all elementary events easily leads to the definition of probability of an event as the ratio of the number of cases favourable to this event to the number of all cases possible. For a common die, the probability of 6 is 1 over 6 (one of six equal cases). If such a game is repeated independently, combinatorics is used to count the cases.

[4] The odds for total wreckage are smaller, e.g., 1:35 instead of 1:29 as the chances for the opposite (not a total wreckage) have increased. Consistently, the probability gets smaller.

[5] We could also include *utility* of money instead of money to improve our model. For fundamental components of an analysis of risk, see Borovcnik (2015b).

Task 3.3. In throwing three dice, is it better to bet on the sum 9 or 10? What are the probabilities for these events?

There was a historic debate around this problem initiated by the Duke of Tuscany (Borovcnik, & Kapadia, 2014a) whether ordering of the results of the dice had an impact on computing the different possibilities. Not considering order, six different cases each obtain the sum 9 or 10:

$$9 = 6 + 2 + 1 = 5 + 3 + 1 = 5 + 2 + 2 = 4 + 4 + 1 = 4 + 3 + 2 = 3 + 3 + 3.$$

$$10 = 6 + 3 + 1 = 6 + 2 + 2 = 5 + 4 + 1 = 5 + 3 + 2 = 4 + 4 + 2 = 4 + 3 + 3.$$

Galilei, respecting the order, discriminated $6 \cdot 6 \cdot 6 = 216$ possibilities and counted 25 cases for 9 and 27 for 10; therefore 9 is less probable than 10. Since all 216 cases are equiprobable, the probability of each ordered triple equals 1/216 and it holds $1/216 = 1/6 \cdot 1/6 \cdot 1/6$, which means that the single dice are independent.[6]

We can use games of chance to embody random situations in a holistic way. Assume we investigate a process with a binary characteristic (with "values" A and not-A) like coin tossing; we can map the process to an urn with two kinds of balls (balls labelled by A with a proportion p, and balls labelled by non-A with the complementary proportion $1-p$) and draw with replacement. If p is known, the urn simulates a process with independent experiments with the same probability p of success. If the probability p is unknown, the results might be used to estimate p (like in the coin-tossing activity).

In the subjectivist view of probability, comparing urns with specific proportions might help persons to assign a suitable value to their personal probability for a specific statement by trial and error.

3.3. TEACHING SITUATIONS INTRODUCING CONDITIONAL PROBABILITY

Probability is intricately bound to the twin concepts of independence and conditional probability. In the frequentist approach, we assume that random experiments are repeatable under the same conditions and independent from each other. Observed results in games of chance support the independence hypothesis, which, in general, is not open to proof. Independence is linked to conditional probability, which is useful in many situations where there is new information; this concept and Bayes' formula play an essential role to revise a probability taking into account further evidence.

[6] More precisely, the Laplace probability on the combined experiment with three dice is equivalent to equiprobability of three single experiments with one die, which are performed independently.

3.3.1. Conditional Probability and Circumstantial Evidence

Conditional probability is the key to evaluate and improve subjective probabilities. It links different conceptions of probability and also shifts the focus from probability as property of an object to probability as a judgement about the object. Basically, all probabilities are conditional to the status of information, which may change. Another vital reason to study conditional probability is that this concept is widely used in inference (see Chapter 5). Medical diagnosis is a natural context where we consider subgroups having different conditional probabilities concerning some health issue.

Task 3.4. In a study, we investigate whether a certain biometric value (such as mammography) is associated with a specific disease *A* (e.g., carcinoma of the female breast[7]) or not. We classify the patients by two characteristics: status of disease, where *A* denotes disease confirmed and not-*A* "no disease"; and a biometric variable with values *B* (positive, pointing towards the disease), or not-*B* (negative, indicating no *disease*). Inspecting 1000 patients, assume that we obtained the data in Table 3.3. Can we state that a positive biometric test is associated with the presence of the disease?

Table 3.3. Data on the association between a biometric test and the presence of a disease

Biometric test	Status of disease		
	A (disease)	Not-*A* (no disease)	All (sample)
B: Positive	9	99	108
Not-*B*: Negative	1	891	892
All	10	990	1000

We can compare the frequency of positive tests in the subgroups of people with *A* (90% = 9/10) and with *not-A* (10% = 99/990) and state that the difference of the occurrence of positive tests in these subgroups is large as 80% more have a positive biometric test for people with the disease as compared to those without the disease. Or, we compare the frequency of those who have the disease (category *A*) in the subgroups of people with positive tests *B* (8.3% = 9/108) and negative tests *not-B* (0.1% = 1/892). Again, we state a large difference (even though the percentages are small). This large difference in the subgroups indicates that the involved variables are related but that does not answer the question. Only with the methods in Chapter 4, we can calculate a (descriptive) measure for the strength of the association between the biometric variable and the presence of the disease.

Here, we change the perspective from frequencies to probabilities to introduce the concept of conditional probability and discuss its properties in the given context. If we *draw randomly* from the 1000 persons, we can compute different probabilities of interest in this scenario.

[7] Screening programmes are focussed on the more frequent types of cancer for women.

Task 3.5. Assume that we select persons at random from all persons from the study (Table 3.3) and investigate their status of health and their biometric test result. Calculate the probabilities of the following statements.

a. To draw a person with a positive test.

b. To draw a person with a positive test who has also disease *A*.

c. To find a positive test in the group of people with disease *A*; compare this probability to that finding a positive test if the person does not have disease *A*.

d. To select a person with disease *A* if the test for that person was positive; compare this probability to that finding a person with disease *A* when the test was negative.

In the following, we will describe the statements a. to d. by the related events and use the notation for conditional probability when we refer to the probability of a specified subgroup in an intuitive way. We will always relate to Table 3.3; the events are located either in single cells, in the margin of the table, or in specified rows or columns.

a. The probability to draw a person with positive test is: $P(B) = \frac{108}{1000} = 0.1080$. This is the sum (on the right margin) of the first row divided by the number of people in the scenario.

b. The probability to get a person with a positive test *and* disease *A* is: $P(B \cap A) = \frac{9}{1000} = 0.0090$. This is the number in the upper left cell divided by the number of people.

c. We compare the probability that the test is positive given the selected person *has disease A* to the probability of a positive result if the person *does not have disease A*, that is, we compare $P(B \mid A) = \frac{9}{10} = 0.9000$ to $P(B \mid not\text{-}A) = \frac{99}{990} = 0.1000$ and find that a positive result is 9 times more probable given that the person has disease *A*. To compute these probabilities we restrict our random selection to the related *column* and ask for the probability to get a person in the upper cell.

d. We compare the probability that the person has disease *A given* the test was positive (first row) to the probability of *A given* the test was negative (second row). That is, we compare $P(A \mid B) = \frac{9}{108} = 0.0833$ to $P(A \mid not\text{-}B) = \frac{1}{892} = 0.0011$ and find that it is roughly 80 times higher to have the disease given

a positive test result as compared to a negative one. The random selection here is restricted to the related *rows*.

In the previous considerations, we calculated a *conditional* probability that refers to a subgroup, which is specified by additional information; e.g., if we have someone with disease *A*, we can imagine that this person was randomly selected from the first column; there, 9 persons tested positive (*B*) and 1 person was negative (*not-B*) – whence the conditional probability of testing positive if *A* is actually present equals 9/10.

In our example, the conditional probability for a positive test is much higher in the group of people *who have* the disease than in the group with no disease. If we compare by rows, we note that the conditional probability for *B* in group *A* is much higher than in group *not-A*.

We finally highlight a basic identity for conditional probabilities:

$$P(A \mid B) = \frac{9}{108} = \frac{9/1000}{108/1000} = \frac{P(A \cap B)}{P(B)},$$

i.e., we can determine conditional probabilities from a quotient of (unconditional) probabilities, which will help to define the notion.

The context of medical diagnosis is also suitable for comparing various odds. We have two points in time: before and after we know the result of the biometric test. Consistently, we have prior and posterior odds and we can express the "value" of the biometric test result in the form of odds. It is interesting to compare these odds against each other. We relate to the data in Table 3.3 in the following task.

Task 3.6. Prior to the medical test, the odds for and against having the disease are $10:990 = 1:99$. As it is 9 times more frequent to have a positive test among persons with disease *A* than without *A* (9/10 to 1/10), a positive test has "odds" of $9:1$ for the disease. For people with a positive test result, the odds for disease *A* are 1:11. Find a mathematical equation that combines all these odds.

We switch in the representation of odds to fractions, i.e., we write 1:99 as $\frac{1}{99}$. In the new representation, we have the following odds: Prior odds: $\frac{1}{99}$; "odds" for disease by a positive biometric test $\frac{9}{1}$, and posterior odds (after a positive test): $\frac{1}{11}$. With a little imagination, we might see that the following equation holds:

$$\frac{1}{99} \cdot \frac{9}{1} = \frac{1}{11}.$$

That is, we can combine these two pieces of information by multiplying the related odds to obtain the *posterior* odds[8] for the disease, given a positive test. In a way, we may handle odds like scaling factors when copying a sheet of paper repeatedly: First we make it smaller by a factor of 1:99 (99 cm are rescaled to 1 cm[9]); then we enlarge it by 9:1 (1 cm is rescaled to 9). By the end, an object on this sheet is scaled by a factor that is obtained by multiplying the two single factors (odds not probabilities represent the scaling).

To finish the introductory activities around conditional probability, the teacher might discuss how new information changes the probability in situations like these: a) In tossing dice, one of the dice is not balanced and does not roll well. b) The age of the client (30 or 65) when an insurance company offers a life insurance for the period of ten years. c) Considering roulette, after 13 rounds of red, many players insist that black is overdue.

3.3.2. Conditional Probability and Compound Probability

In the context of medical diagnosis, we encountered various situations where adding information (such as belonging to a subgroup) affects the probability of an event. Using the basic identity for *conditional probability*, we define the probability *of A given B*[10] as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) > 0 .$$

This expression is the basis for deriving compound probabilities, that is, the probability that two events happen, which is given by the multiplication rule:

$$P(A \cap B) = P(A \mid B) \cdot P(B) \text{ if } P(B) > 0 .$$

In terms of proportions, the proportion of people who have both characteristics A and B is equal to the proportion of people with A multiplied by the proportion of people having B within the group of people with A. In the same way, it holds $P(A \cap B) = P(B \mid A) \cdot P(A)$ if $P(A) > 0$ (as A and B interchange roles).

The multiplication rule extends to more than two events: For example, for three events, one (of six possible) expressions of the multiplication rule is:

$$P(A \cap B \cap C) = P(A \mid B \cap C) \cdot P(B \mid C) \cdot P(C) \text{ if } P(B \cap C) > 0 .$$

---

[8] Revised values of odds after we observe the result in the biometric test.

[9] Of course, with uncertainty, we have not cm as unit.

[10] The verbalisation of conditional probability has a great impact on how people perceive the situation and recognise that it deals with a *conditional* probability. The exact wording is very sensitive to changes in the perception of people. The probability of an event A given that B happens, the probability of A if B, etc. are only two further useful variations among many ways to express it.

Students can use the multiplication rule to compute probabilities in compound experiments, that is, experiments that consist of two or more steps, each of which is defined by a random experiment. For example, flipping three coins at the same time can be considered as a three-step experiment, each of which consists in flipping a coin. Combined experiments may be analysed by means of tree diagrams where the edges are labelled by the related conditional probabilities and the probability of paths is calculated by the multiplication rule (see also Section 3.7.3).

A particular case of the multiplication rule happens when the events are independent, that is, the occurrence of one event does not add any information about the occurrence of the other event such as in flipping two or more coins where the result of each coin does not affect the results of the other coins. In this case, if one of the following relations holds, the others are valid, too (if only both events have a probability greater than 0):

$$P(A \mid B) = P(A),\ P(B \mid A) = P(B), \text{ and } P(A \cap B) = P(A) \cdot P(B).$$

While the first two relations are directed, the third shows that independence in fact is a symmetric relation between events.

In Chapter 4, we connect independence to association and present examples where conditional and total probability can be used to judge association in a contingency table. Though the idea of independence is basic, it is difficult to explain it without the reference to formulas. When should independence apply to two events, or when are two experiments performed independently so that the probabilities of the single events can just be multiplied? Motivations by a lack of causal influence confuse more than they can clarify. Often independence is just an assumption. When we toss a coin repeatedly, we assume that the single steps of the repeated experiment have nothing in common so that we can use the independence assumption.

3.4. ADDITIONAL TEACHING ACTIVITIES

In this section, we build bridges between probability and descriptive statistics and describe further ideas essential for understanding probability and its applications. We present standard situations and illustrate the central theorems of probability. The considerations prepare the concept of sampling distributions, which forms a further key to statistical inference (Chapter 5). For mathematical details, we refer the reader to Stirzaker (2003).

3.4.1. Random Variables

Sometimes we are interested in all the possible values that a variable resulting from a random experiment can take instead of focussing on a particular value of that variable. Consider, for example, the following activity.

Task 3.7. Fair game. Carmen and Daniel play a game with the following rules: They throw two dice and compute the difference between the higher and lower values on the dice. If the difference is 0, 1, or 2, then Carmen wins 1 Euro. If the difference is 3, 4, or 5, then Daniel wins 1 Euro.

a. Do Carmen and Daniel have the same benefit of the game? Would Carmen be willing to change her role with Daniel?

b. How much should each player win so that the rules treat them fairly, i.e., give both of them the same benefit?

After playing the game for a while, the students can see that the roles are not fairly attributed since Carmen wins about 2/3 of times (24 out of 36) the dice are thrown while Daniel only wins 1/3 of times. This is easily seen from Table 3.4 where we list the different outcomes of two dice and classify them according to the difference. Since there is a total of 36 different possibilities, a simple application of Laplace's rule leads to the probabilities for the values of the variable "difference between the upper and lower values".

Table 3.4. Probability distribution for the difference of two dice

Difference	Possible outcomes	Number of outcomes	Probability	Winner
0	11, 22, 33, 44, 55, 66	6	6/36	Carmen
1	12, 21, 23, 32, 34, 43, 45, 54, 56, 65	10	10/36	Carmen
2	13, 24, 31, 35, 42, 46, 53, 64	8	8/36	Carmen
3	14, 25, 36, 41, 52, 63	6	6/36	Daniel
4	15, 26, 51, 62	4	4/36	Daniel
5	16, 61	2	2/36	Daniel
Total		36		

As the difference between two dice depends on the result of a random experiment, we say we are dealing with a *random variable.* Each experiment can be linked to more than one random variable. In the example, we could have considered the sum or the product of the points of the dice. The set of values a random variable can take associated with the related probabilities is called a probability distribution. We can use the probability distribution in the example to compute the probabilities that Carmen or Daniel wins. While Carmen wins with a probability of 24/36, Daniel wins with a probability of 12/36 so that the game is not fair. We could give Daniel 2 Euros each time he wins so that in the long run the expected amount of money would be the same for both players.[11]

Random variables are vital for modelling random phenomena and are used in a variety of situations. A probability distribution connects the potential values with their probabilities; its shape is described by its expected value and variance. Several examples are presented in the next sections and in Chapter 5.

[11] Yet, the role of Daniel seems to be more risky as he wins more but with less probability.

3.4.2. Additivity of Expected Value and Variance for Repeated Experiments

In statistics, we find many random variables that are composed of the sum of simple random variables, for example, the sample mean or the sample proportion. In these applications we are interested in the expected value and variance of the sum. An important result is that the expected value for a sum of random variables equals the sum of expected values for each single random variable. In games of chance, this is backed up by the fact that playing a game two times costs twice as much as one game.

Additivity of expected value is counter-intuitive for dependent random variables as illustrated in the following teaching activity (taken from Borovcnik, & Kapadia, 2014b).[12] There are two ways to handle these relations in teaching: one is to illustrate that such laws should hold by simulating the assumptions of a situation, and the other is to arrive at insights by investigating a simpler case where the law applies. We will pursue both.

> *Task 3.8.* We throw three dice and the amount to win is the sum of the faces up. Can we find the distribution of this sum, which is a random variable with potential values between 3 and 18? Determine expected value and variance of single dice and of the sum. Can you confirm the additivity that was referred to above?

We simulate throwing three dice many times to investigate this distribution (see Table 3.5) and to examine whether the additivity of expected values can be seen from the data. For one dice, the expected value is 7/2 = 3.5 and the variance equals 35/12 = 2.9167.[13] In Figure 3.2 we compare the bar graph of the data to the exact probability distribution (which can be derived by combinatorial argument).

Table 3.5. Data from a simulation of 1000 throws of three dice[14]

Sum	3	4	5	6	7	8	9	10
Count	0	18	27	36	76	106	125	104
Sum	18	17	16	15	14	13	12	11
Count	11	16	32	42	76	100	115	116

[12] Thus, additivity is *not a stochastic property* of expected value (it is comparable to linearity in linear algebra) whereas it is a genuine stochastic property for the variance as it changes whether the random variables are independent or not.

[13] These parameters can be calculated by using formulas for the uniform distribution of one die or else estimated by simulation.

[14] Note: Simulation leaves us with an extra source of variation. We get better approximations if we generate 5000 experiments. The bar graph in Figure 3.2 (left side) shows some deviation from the perfect symmetric probabilities (right side). Even the relative frequency is slightly higher for 9 than for 10. If the difference of probabilities is small then simulated data may reverse the order if not enough data are generated.

From the data, we calculate a mean of 10.60 and a variance of 9.1632. Thus, adding the values of one dice three times yields roughly the estimated expected value and variance of three dice. This indicates that a law of additivity should apply for both the mean and the variance (yet, it cannot confirm it as the data would always deviate from the exact relation).

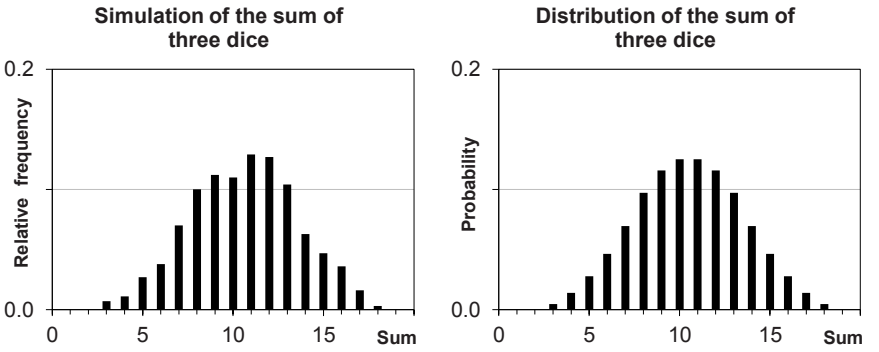


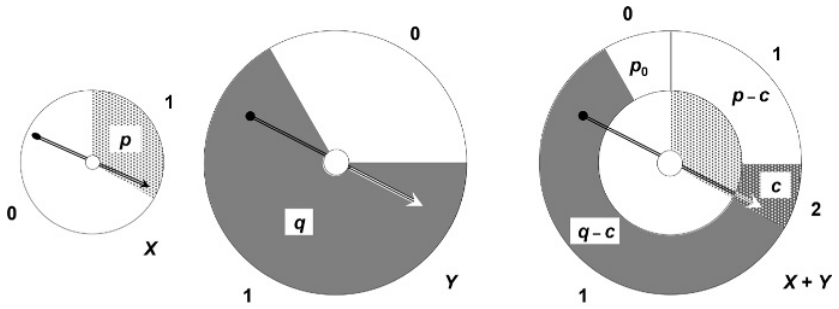
Figure 3.2. Sum of three dice: 1000 simulations and exact distribution

Task 3.9. Dependent and independent spinners. Imagine we have two binary spinners (Figure 3.3a); if the arrow points to the shaded area, € 1 is paid to the player; if it points to the white sector no payment is made. Let the winning probabilities be $P(X = 1) = p$ for the small and $P(Y = 1) = q$ for the big spinner. Determine the fair bet in each game?

The expected payments are $E(X) = p$ and $E(Y) = q$. If played independently one after the other, the fair bet equals $E(X + Y) = p + q$. People are surprised that the additivity for expectation still holds when we put the spinners one over the other to play the two games in one turn (Figure 3.3b); from Borovcnik, & Kapadia, 2014b, p.47).

$$E(X+Y) = 0 \cdot p_0 + 1 \cdot (p - c) + 2 \cdot c + 1 \cdot (q - c) = p + q = E(X) + E(Y).$$

A little more algebra shows that the variance follows an additive relation only if $c = P(X = 1, Y = 1) = P(X = 1) \cdot P(Y = 1) = p \cdot q$, that is, if the two wheels are independent. The variance is additive only for independent random variables, which is a basic relation of eminent relevance (we will come back to it in Sections 3.4.4 and 3.5.6). In this simple case, we can confirm the law and get an insight into the role of independence.



a. Independent *b. Dependent with overlap c*

Figure 3.3. Two independent and two dependent spinners

3.4.3. Distribution Models for Standard Situations

In all conceptions of probability, parts of the available information about an uncertain event originate from data. This raises two issues if we connect probability to data (inference): a) What is meant by a random sample and can we describe this concept in standard situations? b) What are central properties of relative frequencies and averages if taken from random samples?

Generally there are two ways to generate data, one is counting and the other is measuring. The binomial and normal distributions are a key for modelling these situations. The related mathematical theorems can be illustrated by simulation or simple examples. We present activities linked to these models; in the following sections, we explain central theorems and their implications for the concept of probability and its applications.

The Binomial model: Independent repeated binary experiments with a constant "success" probability throughout

The data generating process that fulfils these assumptions is also called Bernoulli series. In fact, the assumptions of this model form the basic random experiment in probability theory and may be illustrated by spinning a wheel of chance as in Figure 3.3a. Typical examples are games of chance, which are described by a random variable X with values 0 (failure) and 1 (success). There are also many other situations that can be modelled by the binomial distribution: defectives in a production process, passing a given examination or not, getting an illness or not, etc.

Consider a game. We denote the ith trial by X_i; the condition $P(X_i = 1) = p$ holds throughout and the games are independent, i.e., we can multiply single probabilities to get joint probabilities such as

$$P(X_1 = 1, X_3 = 1, X_7 = 0) = \; P(X_1 = 1) \cdot P(X_3 = 1) \cdot P(X_7 = 0) = \; p^2 \cdot (1 - p) \; .$$

Thus, the random variable *absolute frequency* S of success in a series of n independently played games, i.e., $S = X_1 + X_2 + ... + X_n$, is binomially distributed with parameters n and p. For example, the number of heads of 20 tosses of an "ideal" coin is binomially distributed with $n = 20$ and $p = 1/2$. From the equation defining S we calculate its expected value and variance (the latter because of the independence of the single trials) as:

$$E(S) = E(X_1) + E(X_2) + ... + E(X_n) = n \cdot p \;\text{ and}$$

$$\text{var}(S) = \text{var}(X_1) + \text{var}(X_2) + ... + \text{var}(X_n) = n \cdot p \cdot (1 - p) \; .$$

If we can model the data generating process by repeatedly spinning a wheel of chance, we call the process and the result a random sample and we can use the value of the random variable S/n (the relative frequency of successes in n trials) to estimate the (usually) unknown p (this was investigated in Task 3.1).[15]

Task 3.10. Binomial distribution. A simple multiple-choice test contains 12 items with 3 options each, of which exactly one is correct.

a. A candidate passes the test if 5 or more items are answered correctly. Determine the probability to pass the test if the candidate answers the items completely randomly.

b. Analogously, we have a multiple-choice test with 24 questions and the minimal requirement to pass is 10 or more.

c. How do the probabilities to pass change if we demand more than 8 correct answers in the short and more than 17 in the long test?

d. Are the assumptions of independent Bernoulli trials appropriate for the situation? Are multiple-choice tests suitable for any exam?

The students are expected to discover that if a person simply guesses, we can accept the assumptions of equal success probability of $p = 1/3$ and recognise the independence of achievement in different items. Therefore, we may model the number X of correct answers by a binomial distribution with n (number of items) and p. The following calculations on the basis of this distribution may also be cross-checked by simulation.

a. For the short test with $n = 12$ questions, we get $P(X \geq 5) = 0.3685$.

b. For $n = 24$, we obtain $P(X \geq 10) = 0.2538$. The longer test gives a person who simply guesses less chance to pass the exam.

[15] Alternatively we can draw n balls from an urn; there are balls with a sign 1 (with proportion p) and others with a 0; if we replace the drawn balls we get the same results as with the spinner. If balls are not replaced, we get a hypergeometric distribution for S.

c. The stricter conditions reduce the probabilities to pass to 0.0039 for the short and to 0.000036 for the long test.

d. While the assumptions of equal success probability and independence are fine for simple guessing, the general situation is more complicated: a student who has learnt may have skipped a chapter; or, when certain items are linked, the choice of answers might not be independent.

The normal model: sums of elementary "errors"

The normal distribution describes the distribution of variables that may be thought of as the result of adding random variables (so-called elementary errors). The speculation that biometric characteristics are the result of Nature adding such elementary errors to a target value[16] led Galton (1889) to the world-wide programme of measuring body characteristics. Surprisingly, the data reasonably matched the normal distribution, which founded the myth of the normal curve as a law of Nature. In the next activity, students will become familiar with some properties of the normal distribution.

Task 3.11. Consider the intelligence quotient (IQ), a score derived from standardised tests designed to measure intelligence. The IQ can be modelled by a normal distribution $N(100, 15)$.[17]

a. Simulate 600 values of IQ and plot a histogram. What features do you observe in the distribution?

b. Repeat the simulation with other parameters; use first a mean of 120 and the same standard deviation as for the IQ values; then use the same mean 100 as for the IQ values and double the value for the standard deviation to 30. Describe the common properties and the differences observed in the three distributions.

The students simulate the distribution of the IQs and plot a histogram (Figure 3.4) where they can observe how the values are grouped around the theoretical mean $\mu = 100$ as well as the approximate symmetry of the distribution, which is exactly symmetric in the model plotted over the histogram. This means that the probability to observe an IQ above or below 100 is exactly the same.

The students can observe the shape of the distribution (called bell curve) and estimate[18] the probability of obtaining values within a distance of 1, 2, or 3 standard deviations from the mean. Since 99.7% of cases fall within a distance of 3 standard deviations from the mean, IQ values below 55 or over 145 have a low probability and are very rarely observed in practice. None of our simulated values were so far out.

[16] Galton termed this target value *l'homme moyen* (Borovcnik, 2006a).

[17] The raw scores are transformed to fit to these parameters. If the average intelligence would increase in the future, the transformation would be adapted so that again the mean value equals 100.

[18] Rather than estimating these probabilities from simulated data, one might also use the normal distribution and *calculate* these probabilities.

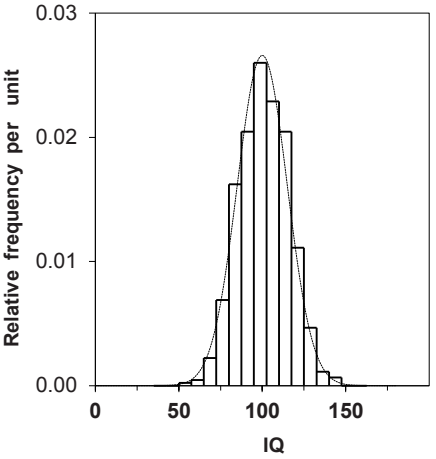
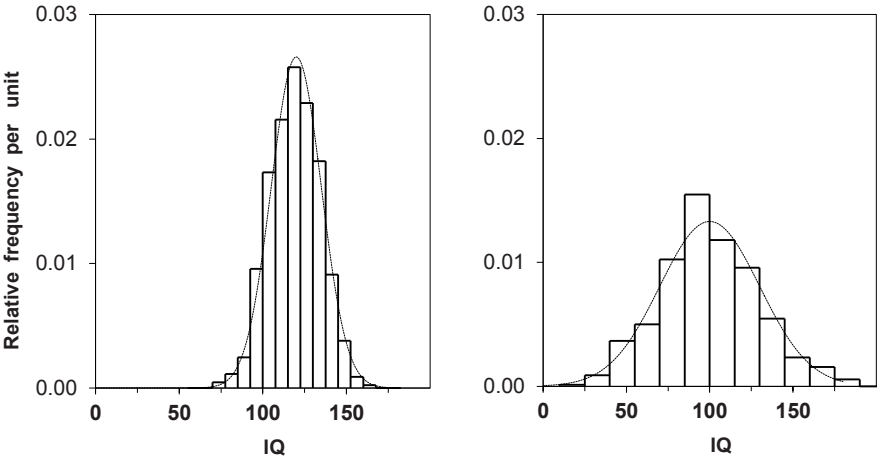


Figure 3.4. Simulation of 600 IQ values with a mean of 100 and a standard deviation of 15

The meaning of the parameters of the normal distribution μ and σ can be illustrated when comparing various normal distributions with different means or standard deviations (Figure 3.5). While a change in the mean only shifts the distribution, a change in the standard deviation also affects its shape (it makes it steeper or flatter).



a. Shift in mean to 120, standard deviation 15 *b. Same mean 100, standard deviation 30*

Figure 3.5. Simulation of 600 values of two different normal distributions

The normal distribution is very important in the theory of elementary errors in physics, which was introduced to justify this distribution as a model for measurement errors where each measurement is thought to be the result of "adding" a random error to the true measurement value. Many measuring processes can ideally be modelled by a normal distribution. The idea behind this approach is that the actual error is composed by a sum of latent (non-observable) elementary "pieces".

This method of modelling is justified by the Central Limit Theorem (see Section 3.5.6). In the next teaching activity, we use the normal distribution with results that do not match the empirical data or knowledge about the context. As a consequence, this sheds doubts on the validity of the model assumptions and thus yields some insight into the context of the activity.

Task 3.12. Height in married couples. Apply a normal model both for men and women. Estimate the parameters from demographic statistics and calculate the following probabilities. Assume that they are randomly selected and the selections of couples are independent from each other.

a. The man is taller by more than 10 cm than his wife.

b. The man is shorter than his wife.

c. Discuss the result recalling your knowledge about couples and height.

Let X describe the height of men and Y the height of women (all values are denoted in cm). From demographic data of adults in Germany, men have a mean of 178, women of 165; the standard deviations are (roughly) 7 for men and 6 for women. By the additivity, the difference $X-Y$ has an expected value of $\mu_X - \mu_Y = 178 - 165 = 13$ and a variance of $\sigma_X^2 + \sigma_Y^2 = 7^2 + 6^2 = 85$ (the latter because of independence).

Linear combinations of normally distributed random variables are also normally distributed so that we may "expect" that the normal distribution approximates also the distribution of the difference in heights. We can simulate the selection of couples many times and estimate the probabilities of the events by the corresponding relative frequencies; or, we calculate the probabilities with software. We standardise the variables and constraints so that we can use the standard normal distribution. Denoting its cumulative distribution function by $\Phi(z)$, we obtain:

a. $P(X-Y>10) = P\left(\frac{X-Y-13}{\sqrt{85}} > \frac{10-13}{9.2195}\right) = 1-\Phi(-0.3254) = 0.6267$ and

b. $P(X-Y<0) = P\left(\frac{X-Y-13}{\sqrt{85}} < \frac{0-13}{9.2195}\right) = \Phi(-1.4100) = 0.0793$.

c. Hence, in nearly two thirds of couples the man would be taller than the woman by at least 10 cm and in 8% the woman is taller than the man. As these probabilities seem too high compared to personal observation, the results imply that people do not "select" each other randomly.

3.4.4. Central theorems

In this section, we introduce simple versions of two important theorems of probability: the Law or Large Numbers and the Central Limit Theorem.

Bernoullian or Weak Law of Large Numbers
In Task 3.1, we illustrated that the proportion of heads in a sample has a smaller variability, the larger the sample size is (Figure 3.1). As a thought experiment one might expect that the distribution "converges" to one point, which is the (unknown) probability for heads if we increase the sample size beyond any limit. That is actually the Law of Large Numbers, which is discussed here in more detail. A key idea for statistical inference and the simulation method is that the relative frequencies weakly converge to the underlying success probability.

Task 3.13. Consider the standard situation of repeated binary experiments with a probability of success p. Describe the absolute frequency S by a binomial distribution:

a. Compute the mean and variance of this random variable.

b. Use Chebyshev's inequality to compute a bound for the probability that the difference between the probability p and the relative frequency S/n exceeds a specific positive real number ε in n trials.

c. What do you observe?

a. Applying properties of expected value and variance to the random variable "relative frequency", that is S/n, yields:

$$E\left(\frac{S}{n}\right) = \frac{1}{n}\cdot n\cdot p = p \text{ and } \operatorname{var}\left(\frac{S}{n}\right) = \frac{1}{n^2}\cdot n\cdot p\cdot(1-p) = \frac{p\cdot(1-p)}{n}.$$

b. We apply Chebyshev's inequality (see Section 3.5.6) to the random variable S/n and obtain the following upper bound for the probability of "large deviations":

$$P\left(\left|\frac{S}{n}-p\right| \geq \varepsilon\right) \leq \frac{\operatorname{var}(\frac{S}{n})}{\varepsilon^2} = \frac{p\cdot(1-p)}{n\cdot\varepsilon^2}.$$

c. As the upper bound on the right side converges to zero when n increases, the distribution of the relative frequency S/n will finally fall completely within an interval $(p-\varepsilon, p+\varepsilon)$ around the "unknown" probability p. Since a random sample is described as the independent repetition of the same binary random variable, this law is the key to estimate probabilities by relative frequencies from *random sample data.*[19]

As a consequence of this convergence, we can estimate an unknown probability from random data, which attain the value 1 independently with probability p.

[19] The theorem can be generalised to independent random variables (with finite variance): the mean value of a random sample converges to the expected value of the population.

Central Limit Theorem

We illustrate the fit of the normal approximation in two cases; one is the binomial distribution, the other deals with the distribution of the average of dice from samples. At the same time we develop ideas about the concept of random samples.

Task 3.14. Voters. Let the proportion p of voters of party A be 0.40. Determine the probabilities that in a random sample of n = 500 voters either of the following events occurs:

a. To have more than 300 A-voters,

b. To get between 180 and 220 A-voters in the sample. Compare the results to the normal approximation with mean $\mu = n \cdot p$ and variance $\sigma^2 = n \cdot p \cdot (1-p)$.

As a rule of thumb, the approximation is sufficient if the variance is greater 10 (and it is improved by the continuity correction; see Ross, 2010a).

a. Let F denote the cumulative distribution function of X (the number of A-voters), we calculate: $1 - F(300) < 10^{-15}$ (n = 500, p = 0.40, with software). Standardising of $k = 300$ yields $\frac{k - np}{\sqrt{n \cdot p \cdot (1-p)}} = \frac{300 - 500 \cdot 0.4}{\sqrt{500 \cdot 0.4 \cdot 0.6}} = 9.1287$, which shows that 300 is more than 9 standard units above the expected value. Using the 3-sigma rule, we see without further calculation that this has an extremely small probability.

b. $P(180 \le X \le 220) = F(220) - F(179) = 0.9689 - 0.0301 = 0.9388$. The normal approximation Y yields: $P(180 \le X \le 220) = P(Y \le 220.5) - P(Y \le 179.5)$ $= 0.9694 - 0.0306 = 0.9387$ (with continuity correction).

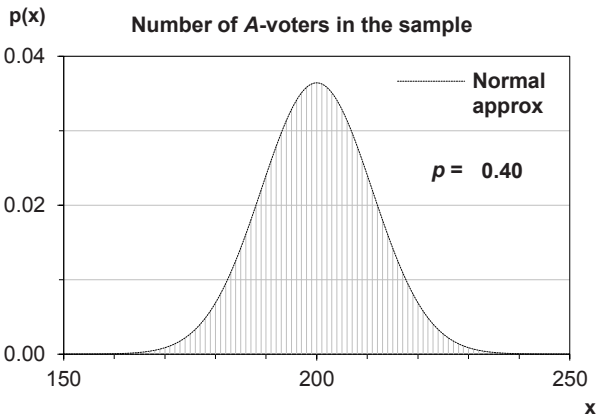


Figure 3.6. Number of A-voters in a random sample of 500: bar graph of the binomial distribution and the approximating normal curve

The activity illustrates that a binomial distribution may be approximated by a suitably chosen normal distribution. The reason is that the binomial variable can be split into a sum of zeros and ones; in this way the binomial variable is decomposed into a sum of other variables, which is similar to the elementary error hypothesis that was used to motivate the normal distribution as appropriate model for many random variables.

The mathematical justification for this normal approximation is the Central Limit Theorem. Apart from the fact that the binomial is a discrete distribution with gaps between the bars, the fit of the normal distribution (in Figure 3.6) is nearly perfect. In this way, the activity illustrates the Central Limit Theorem.

The activity can also be used to illustrate the concept of a random sample. The population is modelled by a wheel of chance: if a proportion p of the population has a characteristic (e.g., votes party A) we represent it by a wheel with a corresponding sector with 1 while the other sector is signed by a 0. Selecting one person randomly from the population is represented by spinning the wheel once. To select a random sample means to spin the wheel (independently) n times. The random variable S, here it is the number of persons voting A (that have been attributed a 1), is the absolute frequency calculated from the sample and has a binomial distribution. S/n is the relative frequency (as a random variable), which is used to estimate the (usually unknown) proportion p.

We have calculated that a random sample delivers a probability of 93% for a number of A-voters within 180 and 220, which corresponds to a fraction of 0.36 to 0.44 of A-voters in the sample if p were 0.40. The "if"-clause is a typical way to use a probability model. It shows that the percentage of A-voters in the sample is close to the proportion of A-voters in the population (at least with a high probability).

If there is a finite population and each of its members has a specific value for a characteristic that is measured, then we can determine the frequency distribution of this characteristic and reserve sectors on a wheel marked by that value with an angle that corresponds to that relative frequency. If we spin the wheel, we generate the random selection of one person and note its value. We generate a random variable by spinning the wheel. A random sample is generated by spinning the wheel (independently) n times.

Task 3.15. Investigating the distribution of the average. The students should simulate a sample of 5 dice and calculate the average number shown. Then they repeat this experiment to obtain 5000 samples of size 5, calculating the mean value in each sample. After that, they should investigate the obtained empirical distribution, draw a bar graph, and calculate its mean value[20] and variance. Finally, the students should repeat the analysis for 20 dice and compare the results to those with 5 dice.

[20] We have to be careful not to confuse the various mean values involved. The first sample has a mean value as have the following samples. All mean values provide the data to be analysed: they have a mean value again: the empirical mean value of the averages of 5 throws (we called them averages here to increase clarity of meaning in text). Furthermore, the population has a mean value, which

We simulated such data and present its frequency distribution on the left side of Figure 3.7. The average values in the sample have a mean value of 3.5050 and are scattered around 3.5, which is the exact mean of the population. If we repeat the whole process of simulating samples with $n = 20$ throws, we get the distribution on the right side of Figure 3.7. Again the averages of samples are scattered around 3.5 (they have a mean value of 3.4956 in our simulation). However, the variability of the averages with 20 data in the sample is much smaller. We learn that the expected value of one dice of 3.5 is reflected by the mean values of all samples, which tend to be closer to the population mean if the sample is larger. This is the Law of Large Numbers for averages.

In fact, in our simulated data we can observe that the variance of the simulated averages is close to the variance of one dice (35/12 = 2.9167) divided by n (0.5672 against the theoretical value of 0.5833 for $n = 5$ and 0.1482 against 0.1458 for $n = 20$). This is a stable relation between the variance of sample means and the variance of the population as can be seen by repeating the whole simulation scenario several times. It can be proven by the additivity of variances for sums (the average is the sum of single variables divided by the number n of data in the sample), see also Table 3.8.

In many cases we observe a random variable X by observing a process instead of selecting one person from a finite population, e.g., we measure the weight of packages of sugar coming from an automated filling line. Still we can think of a wheel of chance that produces the value of X; in this analogy, spinning the wheel (independently) several times means generating a random sample of X.

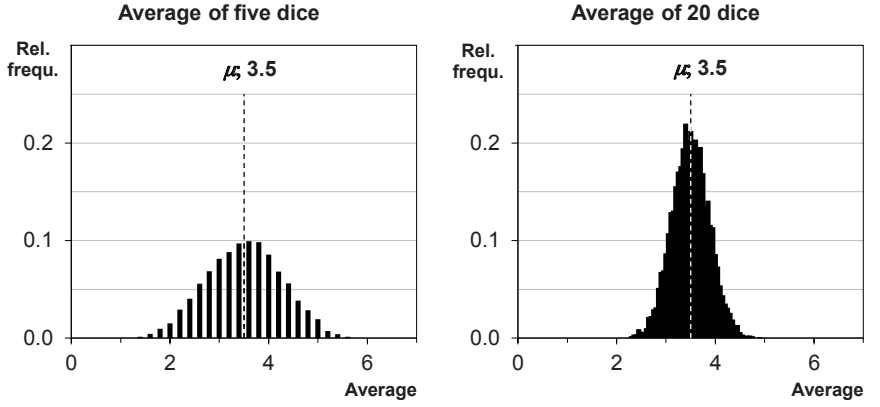


Figure 3.7. Distribution of averages of 5 (left) and 20 dice (right) from 5000 simulated samples of throwing 5 and 20 dice

becomes the expected value in the context of random sampling. This "population mean" is estimated by the sample mean – of the latter, we usually have only one value as we have only one sample.

Task 3.16. Text sums. Take any text of your choice. Attribute a code number from 1 to 1000 to each of the possible signs (your choice). Separate the signs into blocks ("samples") of 20.

a. Calculate the sum of the first 20-sample, calculate the sum of the signs of the second sample. Repeat the procedure until you have 1000 samples (the text has to have at least 20000 signs) so that you have 1000 data for the sum.

b. Calculate the mean and the standard deviation of these sample sums. After that standardise the data, i.e., subtract the overall mean from each sample sum and divide the result by the standard deviation.

c. Present the distribution of the 1000 standardised sums of the 20-blocks by a histogram or a frequency polygon.

The data will be within the limits of –5 and 5 and the distribution will be close to the standard normal curve. We can change the coding system, we can increase the sample size, or we can reorder the signs in the sample randomly. The latter will even improve the fit by the standard normal curve. The result seems surprising but can be explained by the Central Limit Theorem (for details, see Borovcnik, 2015a).

3.5. SYNTHESIS OF LEARNING GOALS

The main goal of the activities described in Sections 3.2 to 3.4 is to help students to get a clearer picture of how different situations can be modelled by probabilities, and how the various connotations of the concept influence the model and its interpretation. Here we summarise these ideas. Since teachers need knowledge wider than that of students, we also point out to how the axiomatic method is used to clarify the notion of probability.

3.5.1. Concepts to Model and Investigate Uncertain Situations

It is important that students perceive that the scope of probability involves two types of uncertain situations:

a. Situations that are signified by uncertainty, such as the *state lottery, insurance contract, and the price of a share*: Should one buy a lottery ticket? Should one take out full-coverage insurance for the new car? How does the price of shares change in the stock market?

b. Situations that can be handled in a better way if uncertainty is – intentionally – introduced. An example is *generalising findings from subgroups*. Suppose we get data from students of a course and find that the proportion of females is 66% and the body mass index is 23.2 on average. Can we and how can we generalise the results to all students in the country?

In these situations *there is no way to "derive" the exact outcome.* Probability provides concepts and models, which may be used to make adequate decisions in

uncertain situations.[21] Two basic steps of probabilistic modelling are to identify all (relevant) possible outcomes first (sample space) and then to weigh these outcomes by probabilities. The weights p_i are standardised *like* relative frequencies so that we have $0 \le p_i \le 1$, with 0 for "impossible" and 1 for absolutely certain events, ½ for "fifty-fifty", etc. There are several ways *how to obtain concrete values for the probabilities.*

1. Using assumptions about the symmetry of the underlying random experiment, e.g., in games of chance.
2. Using information about past events: relative frequencies of already performed random experiments or from data sets, which can be perceived as if they were generated by some random procedure.
3. Quantifying qualitative knowledge: e.g., from various studies on the effect of a health measurement like screening for certain diseases.
4. Calibrating personal knowledge: e.g., personal probability of damages in an insurance contract.

Situations are modelled by probabilities and the results are interpreted within a given context. Computed probabilities reveal insights about the phenomenon modelled. When calculated probabilities seem higher or lower than expected, this sheds doubt on the validity of the assumptions. In our height activity, it seems inappropriate to surmise that people start up a relationship independently of each other's characteristics. In the multiple-choice activity, the results help to improve the design of exam papers.

3.5.2. Different Connotations of Probability

In the philosophical debate, various conceptions of probability are discussed (Batanero, Henry, & Parzysz, 2005). Three different connotations appropriate and necessary for high-school level have been developed in mutual interaction and *none* of these can be separated from the others: a priori theory (APT), frequentist (FQT), and subjectivist (SJT) theories (see Borovcnik, & Kapadia, 2014a; b, where these abbreviations were introduced in order to highlight the multifaceted and inter-related conceptualisations of probability; see also Çınlar, 2011).

A priori theory (APT) – Probability as favourable to possible cases
It is remarkable that probability was used for centuries in quite complex situations but a formal definition was not attempted before Laplace. Laplace (1812/1951) defined the probability of an event A by the proportion of the number n_A of favourable cases to the number n_S of all possible cases for A, i.e., by

[21] Case studies illustrating the formative power of probabilistic modelling in decision making are found in Borovcnik, & Kapadia (2011).

$$P(A) = \frac{n_A}{n_S}.$$

Implicitly all elementary events have the same probability of $1/n_S$. To assure equal likelihood of all "cases", Laplace introduced the "principle of insufficient reason" according to which we may assume outcomes are equally likely if we have no reason to believe that one or another outcome is more likely to arise. A slightly better argument is based on the symmetry of an experiment (if relabelling of the individual outcomes leads to the same probabilities) or if there is no preference for any outcome. In practice, if we suspect that the physical symmetry is violated, we would monitor the frequencies of repeated trials.[22] Typical Laplace experiments are coin tossing and throwing dice.

Problems of the approach: The scope of this view of probability is restricted because in many random situations there are no obvious equiprobable elementary events. Even in games of chance historically, there was a long debate to come to the insight that the set of all (ordered) tuples from the sample space for a single experiment forms the natural sample space for the repeated experiment. Paradoxically, Feller (1957, p.39) notes that this "natural" sample space (Maxwell-Boltzmann statistics) for repeated experiments is not suitable for many phenomena in physics where equal probabilities hold for non-ordered tuples (Bose-Einstein statistics) even if identical components are excluded (Fermi-Dirac).

Frequentist theory (FQT) – Probability as limit of relative frequencies
An experiment is repeatedly performed "independently" under "exactly the same conditions". For an event A of the sample space S, the probability $P(A)$ is defined as the limit of the relative frequencies where $n(A)$ denotes the number of times that A occurred in the first n repetitions, i.e., by

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

This definition goes back to von Mises (1919) who attempted an axiomatic foundation of probability based on this convergence of relative frequencies, which proved inconsistent first and then too complex.

There are also drawbacks of this approach: How can we check that the sequence finally converges? How can we deduce a value for the probability from the relative frequency from a finite number of repetitions? How can we check for the *independence assumption*? Moreover, popular perceptions related to relative frequencies might mislead. For example, people tend to focus on patterns of the outcomes. However, it is important to note that all *patterns* of previous trials are *irrelevant* for probability and are useless for predictions. Such patterns only serve

[22] In casinos, such recording is done as a preventive measurement so that it can be investigated in case some doubt is expressed.

for checking the assumptions if the basic hypothesis of randomness already is in doubt (Borovcnik, & Kapadia, 2014b). Yet, the frequentist interpretation links probability to essential applications in the real world (e.g., mortality calculations based on casualties, or demographic data) and to statistical inference.

Subjectivist theory (SJT) – Probability as degree of belief
In this approach, any probability statement is based on personal judgement, which is conditional on the current status of knowledge. Orthodox Bayesians such as de Finetti (1974) do not recognise any rational constraints on subjective probabilities beyond conformity to probability calculus. From axioms for personal preferences, Bayes' rule, and the requirement of never accepting a bet that yields a loss whatever will happen, it is possible to derive a representation of personal preferences in the form of a probability measure P.

In any one-off decision, the interpretation of probability as a degree of belief is a helpful modelling perspective. The personal judgement, too, has to follow rationality criteria (e.g., the sum of probabilities of two exclusive statements can never be greater than 1) but is open to subjective views. This approach has been criticised as it makes probability appear to be some arbitrary judgement. However, the Bayesian school of inference that follows this approach is growing and argues that despite such problems there is no reason to exclude personal connotations of probability from consideration.

Axiomatic approach
Hilbert (1900) put the axiomatic foundation of physics on his famous agenda of open mathematical problems and with this point, he referred to the problem of finding a satisfying axiomatic foundation for the concept of probability. It was not until 1933 that Kolmogorov solved this problem satisfactorily. His solution was immediately acknowledged in the mathematical community and led to a fast development of probability and stochastic processes in the following decades.

According to Kolmogorov (1933/1956), probability is a *function* that attributes a real number to subsets of a sample space. Like relative frequencies, probabilities have to be a) non-negative, b) the probability for the sample space is normalised to 1, and c) probability is *additive*:

if events A and B have no element in common then $P(A \cup B) = P(A) + P(B)$.[23]

From the axioms and the concept of independence,[24] a rich theory has been built up (that contains the Law of Large Numbers, etc.). The axioms provide basic

[23] Probability has to fulfil this additivity property also for countably infinite events, which are mutually exclusive. More precisely, for an infinite sample space S and a countable collection of pairwise disjoint sets A_i, $i = 1, 2, \ldots$, we have to postulate $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ in the additivity axiom.

[24] Not to all subsets of the real numbers can a probability be attributed so that the axiomatic approach has to consider the system of the so-called Borel sets, which forms the domain of probability. The

mathematical rules for handling probabilities, while the values for probabilities are derived only by additional assumptions (based on the frequentist, classical, or subjective approach); this leaves much freedom for modelling the situations under scrutiny.

Although the axiomatic approach is not studied in depth in high school, it is desirable that teachers acquire basic ideas of it (for more details, see Borovcnik, & Kapadia, 2014b). Moreover, it is important that teachers understand the aforementioned controversies since they still exist and the interpretation of probability determines the methods used for statistical inference.

Complementarity of the different approaches

Shortly after Kolmogorov has published his axiomatic approach to probability that mimics the properties of relative frequencies, de Finetti (1937/1992) found an axiomatic approach for probability based on axioms of (personal) preference.

Several authors highlight the negative impact of reducing the teaching of probability to only one view (Batanero, Henry, & Parzysz, 2005; Carranza, & Kuzniak, 2008, Vancsó, 2009). The Laplace theory lacks wider applicability; furthermore, the equiprobability basis cannot be extended to experiments with infinite series of trials. The frequentist view solves the important problem of a probability concept on sequences of trials; however, it poses new problems such as how to check the assumption of independence or how to interpret the limiting behaviour.

Furthermore, many problems are one-off and non-repeatable so that we need the subjectivist theory (de Finetti, 1937/1992) where probability is derived as a numerical representation of a specific preference system (of a person). The axiomatic approach is intended to be free of any specific interpretation; however, it is mainly orientated to the frequentist interpretation and, therefore, biases the conception of probability towards this interpretation.

One major difference between personal and frequentist probability is the location of probability: in the first it is a property of some person (a judgement), in the second it is a feature of reality (like a physical constant, which has to be measured). Another distinction is that the first is a subjective while the second is an objective feature. It comes as no surprise that the axiomatic approach, too, misses constituent parts of probability. Borovcnik (2012a) argues that (conditional) probability enfolds its full potential only if both subjectivist and objectivist conceptions are jointly taken into consideration. This complementarity means that the concept of probability would fall apart if some connotations were neglected. This is a key idea and needs to be reinforced in the teaching of probability.

---

axiomatic approach only rules out the structure of probability but leaves the probability values open, which allows probability to become a flexible instrument of modelling. The independence concept discriminates probability from the concept of measures and, therefore, contributes substantially to the properties of the concept even though independence is not listed in the axioms but shifted to a definition.

3.5.3. Circumstantial Evidence and Bayes' Formula

In the activities on diagnosis, we used the data about the combined occurrence of a biometric test and the presence of a disease. We interpreted these data as probabilities, which was justified by a random selection of one of the 1000 persons. We rewrite Table 3.3 into Table 3.6 and fill in the combined events in addition to the counts.

Table 3.6. Probabilities on the biometric test and the presence of a disease

Biometric test	Status of disease		
	Disease	No disease	All
Positive	$B \cap A_1$ 9	$B \cap A_2$ 99	B 108
Negative	*not-B* $\cap A_1$ 1	*not-B* $\cap A_2$ 891	*not-B* 892
All	A_1 10	A_2 990	S 1000

We can determine the total number of positive results (B) in the test by adding all numbers in the first row; by dividing by the number of all persons, we obtain the *total* probability of B as $P(B) = P(B \cap A_1) + P(B \cap A_2)$. We have written the events as $A_1 = A$ and $A_2 = not\text{-}A$ and denote that events that define different columns are disjoint. The total probability law is simple if we interpret it as the sum of the probabilities in each row (or in each column).

Using partitions in the sample space to derive complex probabilities
When computing a probability, a useful strategy is to decompose the event of interest in other simpler events whose probabilities are known. Often we can find a partition of the sample space by two or more events in such a way that:

$$A_1 \cap A_2 = \varnothing \text{ and } A_1 \cup A_2 = S .$$

For example, a population can be divided into persons who have and who do not have a disease as we did in Table 3.6. In this case, a strategy to compute the *total probability* for an event B (such as a positive test) is given by:

$$P(B) = P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2) .$$

Note that one has to know the conditional probabilities of B for each of the events in the partition (the person has the disease or does not have it). This formula represents a short-cut for calculating the probability of interest $P(B)$ from the given conditional probabilities and the probability of having the disease.

Bayes' formula

We already have seen that our information relates to the conditional probabilities for a positive test given the status of the disease, that is $P(B \mid A_i)$. After a positive test *B*, we are interested in the other direction of the conditional probability, namely the probability to have the disease given a positive test, that is $P(A_1 \mid B)$.

The famous Bayes' formula links the two possible directions of conditional probabilities and allows calculating the new probability for the disease given a positive test. Regarding the definition, $P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$, the task is to replace the probability of *B* in the denominator by the total probability formula:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2)}.$$

If Table 3.6 has more than two columns, e.g., A_1, A_2, …, A_n, for *n* mutually exclusive diseases, the formula may be extended to *n* disjoint events that form a partition of *S* with a total probability of $P(B) = \sum_{j=1}^{n} P(B \mid A_j) \cdot P(A_j) = k$ so that Bayes' formula reads as

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{k}.$$ [25]

The situation may be represented using a tree diagram (see Section 3.7.3) where we enter nodes for the various states A_i on the first level and label the paths from the root to these nodes by $P(A_i)$; then we add a second level of nodes (*B* and not-*B*) and continue the paths to the event *B* labelling the edges with the conditional probabilities $P(B \mid A_i)$. The total probability of *B* is then the sum of the probabilities of all paths that end at nodes with a label *B*.

Bayes' formula serves to update prior probabilities $P(A_i)$ about A_i by further information *B* into posterior probabilities $P(A_i \mid B)$ and for this reason it is widely used in medicine, at court, etc., where prior judgements on a statement (of being ill, guilty, etc.) are revised by circumstantial evidence. The Bayesian approach to inference has been developed on the basis of this formula and forms one of the great schools towards inference (an example is shown in Chapter 5). Often people try to find an intuitive estimate of the posterior probability, e.g., for being actually ill given some data. Usually they fail (see also Section 3.6). Another difficulty is when the various states A_i are erroneously attributed equal probabilities.

[25] Kolmogorov interpreted A_i as *hypotheses*; this introduces an inherent link of conditional probabilities to subjectivist probabilities.

3.5.4. Random Variables and Expectation

In order to introduce the idea of a random variable, we use an analogy to *descriptive statistics* and games of chance to enhance this cluster of ideas. From the Law of Large Numbers, we know that relative frequencies tend to be closer to the (usually) unknown probability, the larger the sample is (see Task 3.1). That is, in any probability formula, we can replace the probabilities by *idealised* frequencies; we can also estimate the probabilities by the relative frequencies from real as well as from simulated experiments where the assumptions underlying the used model are implemented. This way, we can illustrate probability laws by simulation: a law is approximately fulfilled by the simulated data.

This interrelation can also be used in the other direction, that is, from probabilities to data: given the model probabilities, we can predict the behaviour of samples. For example, given the expected value of a random variable, we can predict the average value (the mean) of data.

Table 3.7. Correspondence between probabilistic and descriptive notions

Probability terms	Descriptive statistics
Prediction	
Potential values of a random variable $x_1, x_2, ..., x_k$	The different observed values of a data list (some occur repeatedly)
The probabilities p_i of x_i	The relative frequencies f_i of x_i.
Expected value $\mu = E(X)$ $= x_1 \cdot p_1 + x_2 \cdot p_2 + ... + x_k \cdot p_k$	Mean of data $\bar{x}$ $= x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_k \cdot f_k$.
Variance $\sigma^2 = \mathrm{var}(X)$ with p_i	The empirical variance s^2 with f_i
$(x_1 - \mu)^2 \cdot \frac{p_1}{f_1} + (x_2 - \mu)^2 \cdot \frac{p_2}{f_2} + ... + (x_k - \mu)^2 \cdot \frac{p_k}{f_k}$	
Estimation	

An overview of the correspondence between terms is presented in Table 3.7. The terms on either side share the same structure; replacing relative frequencies f_i by probabilities p_i (and vice versa) yields a link between them. The twin relations serve in two directions: From the probability at the left side we get predictions for the corresponding descriptive values at the right side and the descriptive values

conversely provide estimates for the probabilities, provided the data originates from a random sample.

This analogy between relative frequencies and probabilities leads to a general pattern of translating between descriptive and probabilistic terms, which forms the core of the frequentist interpretation of probability. For example, in the formula for the expected value, we can simply replace all probabilities and get the mean value of data with the related relative frequencies and vice versa.

Additivity of means and variances
Let $X_1, X_2, ..., X_n$ be n random variables, then (if the variances are finite):

$$E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n) .$$

This additive property is essential for statistical inference; in Chapter 5, we estimate the mean (expected value) of the population (described by a random variable X) by the mean of data in a sample of repetitions X_i of X, i.e., by the *random* variable $\frac{1}{n}S = \frac{1}{n}(X_1 + X_2 + ... + X_n)$ (see Table 3.8). For the variance, the additivity holds *only* if the games are *independent*:

$$\text{var}(X_1 + X_2 + ... + X_n) = \text{var}(X_1) + \text{var}(X_2) + ... + \text{var}(X_n) .$$

Note that the additivity of means is preserved in dependent experiments.

3.5.5. Standard Models of Distributions

In many applications of probability, there is no need to derive the probability distribution for random variables under scrutiny. We often find situations with similar characteristics that can be described by the same mathematical model. Two of the most useful models are the binomial and the normal distributions.

Binomial distribution
The binomial distribution describes the repetition of a series of independent trials in all of which we consider the same random variable with values 0 (failure) and 1 (success); the focus of interest is on counting the number of successes in n trials. For example, a political party could be interested in the number of people voting that party in a sample of n voters. This variable can take values between 0 and n. The probability distribution is given by:

$$P(X = k) = \binom{n}{k} p^k \cdot (1 - p)^{n-k} \text{ for } 0 \le k \le n .$$

For the mean and variance of the binomial distribution, we have:

$$E(X) = n \cdot p \text{ and } \operatorname{var}(X) = n \cdot p \cdot (1-p).$$

Normal distribution
The normal distribution describes various biological or psychological variables that are measured on a metric (ratio) scale. Its probability density is given by

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Although in theory, this variable can take any value, in practice, those values that are "far" from the centre of the distribution are unlikely. The density function is symmetric around the mean value μ so that values above and below the mean each represent 50% probability. Since the distribution is symmetric, mean, median, and mode coincide and values around the mean are the most likely values. The standard deviation is given by σ. Often, it is of interest to know the probabilities of symmetric intervals around the mean; the probabilities can be calculated from software:

– Values within one standard deviation from the mean have a probability of 68%: $P(\mu - \sigma \le X \le \mu + \sigma) = 0.6827$.
– Values within two standard deviations from the mean have a probability of roughly 95%: $P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.9544$.
– Values within three standard deviations from the mean have a probability of 99.7%: $P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.9973$.

Another important property is that the sums or differences of normal distributions also follow a normal distribution, whose mean and variance can be computed using the additivity properties for mean and variance (the latter only if the single variables are independent).

3.5.6. Law of Large Numbers and Central Limit Theorem

In Task 3.13, we introduced the Weak Law of Large numbers to highlight a key idea, namely that the relative frequencies weakly converge to the underlying probability of success in a series of independent binary experiments with the same success probability. In this section, we will deal with the Chebyshev inequality (that allows proving this convergence) and the Central Limit Theorem.

Tail probabilities, Chebyshev's inequality, and the Law of Large Numbers
Using Chebyshev's inequality, we find bounds for the probability of the tails of a distribution in terms of expected value μ and standard deviation σ; this explains why we use these parameters (and not others). As in descriptive statistics, the expected value signifies the central location and the standard deviation quantifies the width of a probability distribution.

For the normal distribution, almost all probability (0.9973) is included in the 3-sigma interval. As a rule of thumb, this holds as a rough approximation for virtually any (peaked) distribution. For our simulation with the three dice we have a mean of 10.60 and a standard deviation of 3.0271 with $\bar{x} - 3s = 1.51$ and $\bar{x} + 3s = 19.68$ so that in this case all data lie within this interval.

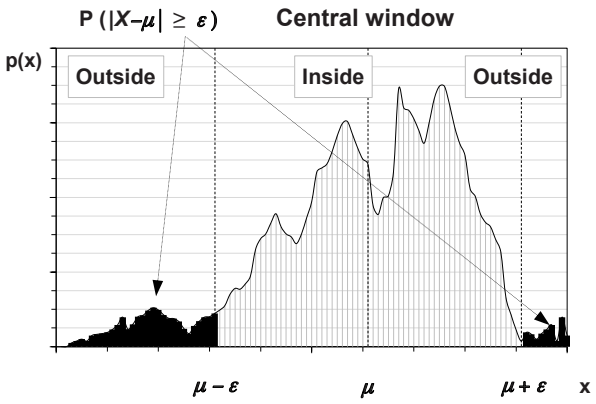


Figure 3.8. Tail probability of an arbitrary probability distribution

Chebyshev's inequality. In fact, we can find a general bound for the tail probability highlighted in black in Figure 3.8 in a simple way (see, e.g., Ross, 2010a). Let X be an arbitrary random variable with finite variance σ^2 and ε be a positive real number. Then $\mu = E(X)$ is a finite value and the probability that X will fall outside an interval $(\mu - \varepsilon, \mu + \varepsilon)$ is bounded by the following inequality:

$$P\left(\left| X - \mu \right| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}.$$

This bound for the tail probability of a distribution has the following intuitive properties: if the interval is smaller (as measured by ε), then the boundary for the probability of falling outside the interval gets greater; if the variance is smaller, the boundary shrinks. Of course, practically, the inequality is useless as, e.g., it yields an upper bound of 1/9 for the probability to fall outside the interval from above while for any normal distribution this value is 0.0027.

However, this inequality justifies the use of the variance or standard deviation as measures of variability. And, most importantly, it allows a simple proof of the Law of Large Numbers: the distribution of the relative frequencies (as a random variable) shrinks to the point of the unknown probability if the number of trials is increased beyond limit (see 3.4.4). This theorem justifies interpreting probability as

relative frequencies in independent trials and also permits to estimate unknown probabilities by the relative frequency of a random sample.

Central Limit Theorem

The Central Limit Theorem admits the approximation of the distribution of a sum of (independent) random variables by a normal distribution.[26] Due to this theorem, the idea about elementary errors that add up was developed to explain why the normal distribution is a good model and to justify it in specific cases. As the absolute frequencies are built by adding elementary units (0 or 1), we can approximate any binomial by a normal distribution provided the number of trials is long enough.

Although the mathematical proof is complex, we can illustrate the implications of the theorem by using simulation.[27] The focus in the classroom should be on an investigation how changing the distribution of the basic random variables (that build the sum) influences the speed of convergence to the normal distribution (see the related teaching activity in Chapter 5). We used this approach in Tasks 3.14–16.

Table 3.8. Relationships involved in the Central Limit Theorem

	Finite parent population – or data generating process	Mean of a random sample
Random variable	X	$\frac{S}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$
Expected value	μ	$E(\frac{S}{n}) = \mu$
Standard deviation	σ	$\sigma(\frac{S}{n}) = \frac{\sigma}{\sqrt{n}}$
Type of distribution	Arbitrary	Gets symmetric and bell-shaped

In Table 3.8, we present relationships that appear in the Central Limit Theorem. The more symmetric the distribution of the population is, the faster is the speed of convergence of the distribution of standardised means of random samples to the standard normal curve. This theorem is a key for statistical inference (see Chapter 5) as it implies that sums and averages get closer to a bell-shaped distribution, the more components are involved.

[26] What is surprising is the fact that the single summands can have an *arbitrary* distribution if only the variance of the distribution is finite.

[27] Already Laplace proved the convergence of standardised binomial distributions towards the standard normal distribution (see, e.g., Borovcnik, & Kapadia, 2014b).

3.6. STUDENTS' REASONING AND POTENTIAL DIFFICULTIES

Probability has been taught in school for several decades; yet, many problems still occur frequently. One reason is that probability is a *virtual* concept; the feedback from reality is indirect at best as most data fits to more than one hypothesised model (Borovcnik, 2011).

The plurality of objects described in the previous sections as well as the different interpretations of probability may explain the many misconceptions students have that deviate from a mathematical view on probability. Spiegelhalter (2014) stated "I often get asked why people find probability so unintuitive and difficult. After years of research, I have concluded it's because probability really is unintuitive and difficult." Below we describe some examples of these difficulties (see also Batanero, Chernoff, Engel, Lee, & Sánchez, 2016; Borovcnik, & Kapadia, 2014b; Chernoff, & Sriraman, 2014; Jones, 2005; Jones, Langrall, & Money, 2006; or Shaughnessy, 1992).

3.6.1. Misconceptions and Heuristics (Strategies) in Probability Situations

We refer to milestones in research on misconceptions and classify misconceptions intending to relate them to archetypical general strategies of human beings.

Seminal works

- Piaget and Inhelder (1951) is the earliest study on children's *understanding* of probability. Piaget conceived randomness as *opposed* to causality. He developed a theory of stages of cognitive development; according to him, randomness is an *irreversible* process so that adequate handling of probabilities is not possible before reaching the stage of formal operations (age 12+).
- Fischbein (1975, 1987) introduced the notion of primary *intuitions*, which emerge without formal education, and secondary intuitions, which develop from learning. Fischbein noted that children's probabilistic understanding sometimes decreases with age and referred this phenomenon back to the focus of education on deterministic relations such as causality in science.
- Green (1983) pioneered a large-scale empirical study of children's understanding of probability including the language of probability. Among other findings, Green accumulated evidence that these children (11-16 year old) progress with age in probabilistic and combinatorial reasoning but are not good at discriminating random from non-random sequences.
- Kahneman and Tversky (1972) began to study adults' understanding of probability. Their work interpreted adults' behaviour in probabilistic situations as biases and formulated the key heuristics (intuitive strategies) behind them (Kahneman, Slovic, & Tversky, 1982; Tversky, & Kahneman, 1980;). Their research has also found that private conceptions are not always stable; people change their views; sometimes they ignore facts or reinterpret facts to fit to their view. Description and "explanation" of some of these strategies (heuristics) follow.

Availability
Probability is intuitively approximated to the ease of recalling relevant cases from memory. But memory is confused with fantasy; people remember events selectively (events may only be registered if they are adverse; recording may start only after a series of "events", etc.). The recall is also biased especially from selective memory (e.g., emotional incidences are much more likely to be recalled).

Equiprobability bias
Lecoutre (1992) described the *equiprobability bias* according to which people tend to judge cases as equally likely. An explanation is that the idea of fairness meets a deep desire in human nature (equity, equal chance) and equal probabilities have been used to express fairness. To reverse the argument, if randomness is "fair" it should allocate equal probabilities for all cases.

Equal probabilities are used for making fair decisions when it is hard to find a solution that is agreeable to all parties; the decision is handed over to randomness so that the person who should decide gets free of the responsibility for the decision. It may be compared to seeking the help of "higher" forces as in a divination of ancient times. In another example, in football, the toss of a (fair) coin decides who will kick off. Equal probabilities are an embodiment of fairness. It is no surprise that they are transferred to situations inadequately and that persons tend to judge given possible outcomes as equally likely, especially if only *two* are involved.

Control of the future
Probability deals with uncertain situations relating to the past and to the future. A "need" to predict the future seems to be an archetype of thinking that is met by models of causal connections as investigated in physics. Causal schemes are occasionally transferred from science to probabilistic situations where these schemes do not apply (Laplace, 1814/1995, viewed probability as a substitute for causal connections). Another alternative is to *foresee* (or positively influence) future outcomes by exploring God's will by divination. In line with the wish to control the future, Konold (1989) recognised the *outcome orientation*: a further tendency in persons' behaviour to reinterpret given probabilistic information as a direct tool to predict (with certainty) the exact outcome of the next experiment.

Representativeness
In the *representativeness* strategy, the probability of one single outcome is equated to the probability of the group of similar outcomes (of which the specific outcome is a representative member). One explanation is that group features, according to an archetypical human desire are often transferred to individuals that belong to that group. Fashion is one example; we like to dress according to fashion to belong to the innovative active group of successful people. Group features play an eminent role for human beings, which in fact let the individual member benefit from the properties of the group (e.g., a person might be much better in climbing in a group than as a single). This "group transfer" relates the representativeness heuristics to an archetypical strategy.

For the state lottery, some combinations of numbers have many similar combinations and "belong" to a large group; others have only a few and belong to a small group. In selecting one, some people believe that it increases their chances to win if they take a combination from a large group as if it would inherit the greater probability from the *group*, to which it belongs. As an example, compare the reactions of people to the question, which sequence of numbers they would prefer for the state lottery and why: 1, 7, 11, 21, 22, 36 or 1, 2, 3, 4, 5, 6? Many prefer the first; one explanation is that they think it belongs to a larger group of similar numbers compared to the second and thus has a greater probability. However, in doing so they ignore that they win only if their *specific combination* is drawn.

Anchoring

Probability judgements are influenced by additional (irrelevant) information given or highlighted recently. Anchoring is a primitive way for processing information without analysis, which helps a person to react fast. As it is known that people anchor their judgements around pre-set information (whether it is relevant or irrelevant), this tendency may also be used for manipulation.

For example, an airline proud of being on time may reinforce how much effort they do by a simple trick: announce a delay of 35 minutes and, soon after, tell that the plane will be on time. Now, by anchoring, the passengers recognise that it is not easy to be always on time. Anchoring may also be the basis of misperceived mathematical relations like the misinterpretation of the Law of Large Numbers (frequencies tend to the probability in large samples) into a law of small numbers (frequencies should balance in any sample). Mixed with representativeness (the sample should be similar to the population) and confronted with a series of 4 heads in five tosses, people may think tails is more likely in the fifth toss.

Patterns

People perceive rules such as the regularity of the cycle of the sun, geometric and arithmetic patterns, and generalise these patterns (even when they do not apply). It is not surprising that people focus so much on patterns of random sequences and draw wrong conclusions from them. Random sequence patterns are highly instable (fluctuate greatly by the random source) so that we can be misled by drawing conclusions from short series. The randomness hypothesis is that patterns do not occur. If this hypothesis gets doubtful as observed patterns seem awkward they have to be checked *formally* by statistical tests and cannot be judged intuitively (Batanero, Arteaga, Serrano, & Ruiz, 2014).

Personal experience and information

People like to speak about individual cases and draw upon their personal experience. One drawback is that we do not consider our *individual* case as one data point within a *general* distribution of data about a random variable. "According to the statistics there is only 1 out of 10000 with a complication of this operation (a risk of 0.0001) but my father died from it". Personal experience like this raises questions in people's minds about the value of probabilistic information.

3.6.2. Independence and Conditional Probability

Mathematical knowledge alone is not always helpful if it comes to handling conditional probabilities appropriately. Many misconceptions are described in research (e.g., Falk, & Konold, 1992). For example, people transfer properties from some puzzling situation to others where they do not apply; the overlap with causal connotations might be confusing in problems of conditional probabilities. Finally, the symmetric notation we use to represent the concept provokes a reversal of conditional and conditioned event.[28]

Task 3.17. Independence and causality – time-dependent thinking. Suppose we have an urn with two white and two black balls and draw twice without replacement. Consider these two tasks (Falk, & Konold, 1992).

a. What is the probability that the second ball is white given that the first ball is white?

b. What is the probability that the first ball is white if the second ball turns out to be white?

c. Finally, do these conditional probabilities change if we draw with replacement?

The counter-intuitive answer is that the probabilities are the same for both questions a. and b., irrespective of replacement. As in question a., there are many situations where the conditional probability is known from the circumstances of the experiment. After we see that the first ball is white, we have an urn with one white and two black balls (if we draw without replacement). We therefore have a probability of 1/3 for white at the second draw.

In other occasions (as in question b.), we have to *calculate* the conditional probability asked for from the given probabilities regarding the rules of probability. Causal thinking is deep-seated and time-bound. In question b., people are convinced that what happened later lacks any relevance for previous events. Consistently, they often assume independence and answer ½ while it is 1/3 as in a. (Borovcnik, 2011). Causal schemes seem to attract more trust than probabilistic rules, which are rather abstract.

Such "causal thinkers" also doubt a formal solution by establishing the sample space of the two-stage experiment as this method "hides" the time-order; solving the task using the multiplication rule does not appear more convincing as conditional probabilities are often "related" to time order. One strategy to challenge causal thinkers about the effect of time might be to draw subsequently, first with the left and then with the right hand and only then – after both balls have been drawn to show the ball in the left hand (question a.) or to show the ball in the right hand (question b.) and ask for the colour of the ball in the other hand.

---

[28] There is an intuitive difficulty between the two conditional probabilities $P(A \mid B)$ and $P(B \mid A)$ even though they are different. The seemingly symmetric notation $P(A \mid B)$ does not help; a different notation as $P_B(A)$ instead (the B-probability of A) may be better. It goes back to Kolmogorov (1933/1956) and shows that conditional probability P_B is a probability measure per se, which is justified as it fulfils the axioms but this notation has not found wide acceptance.

Reversing conditions and conditioned statements

For conditional probability statements $P(X \mid D)$, research has found that people react very sensitive to the type of the statements involved: D may be perceived as cause for X; or D is a diagnostic event if X is a potential cause for D (Tversky & Kahneman, 1980). In a class discussion on conditional probabilities, the teacher might investigate whether the students distinguish the two types of situations and how their probability judgements are influenced if the events have such a connotation from context.

For example, the probability $P(+ \mid D) = 0.99$ of a positive medical test given that a person has an illness D is easily perceived as causal information. How do they judge the reverse – *diagnostic* – probability $P(D \mid +)$? Will the students find their own idiosyncratic explanations for their judgements of this probability or follow the arguments found in research? Either this probability is ignored as diagnostic probabilities are not well-understood or it is simply reversed and also evaluated as 0.99. Of course, as shown in Section 3.3.1, the second probability is much lower and depends on the prevalence of the disease.

3.6.3. Taking into Account Students' Reasoning to Improve Teaching

The previous analysis refers different ways of reasoning that deviate from normative probability models back to archetypical strategies, which may explain their persistence in spite of teaching efforts. Kahneman and Tversky have been criticised because their strategies were not predictive of behaviour as answers can often be explained by different views. Yet, teachers should know this description of human behaviour in judgements under uncertainty because it can help learners to become aware of their own criteria and begin to understand, which of them are fallacious.

There are approaches that make intuitive ideas about probability explicit in class. Lysø (2008) suggests starting the elementary probability course with an empirical investigation where the students serve as participants. This way, we can build the bridge between their intuitions and the mathematical concepts. Borovcnik and Bentz (1991) describe a conflict between *action* and *reflection*: reflection might deliver equal probabilities for each outcome; however, this gives no advice about which of the outcomes to choose and therefore, the person can take any. Moreover, giving a reason for the *actual choice* might provoke completely different views in the individual person as the argument that led to *equal probabilities* cannot be used to justify the choice actually done.

In reaction to such problems, Borovcnik and Peard (1996) suggest face-to-face or small-group teaching in the form of a guided interview to meet the specific requirements of probability. Similar suggestions have been made in previous research described, for example, in Jones (2005), or in Chernoff and Sriraman (2014). In the next section, we suggest other resources and activities that enrich the repertoire of tools mediating between calculating and understanding probabilities.

3.7. ADDITIONAL RESOURCES AND IDEAS

There are many further resources that may enhance the ideas behind probability or help to calculate probabilities from the givens. We suggest exploring the strange intuitive thoughts of students about randomness and random behaviour and introduce a different representation of Bayes' formula, which highlights the updating of conditional probabilities from the givens. Furthermore, we discuss tools to enhance given and calculated probabilities and to simplify the calculations.

3.7.1. Investigating Randomness: Generating Coin Tossing from Memory

In Sections 3.4.1–3, we introduced random variables as key for modelling real situations. To become aware of the assumptions that such models impute on reality, it may be wise to let the learners invent data that fit to a specific model. Due to systematic biases in mimicking random sequences from memory, it is possible to recognise whether a person has invented data or has used a random generator.

Batanero Arteaga, Serrano, and Ruiz (2014) suggest asking the students to invent 20 coin tosses. They investigated a group of 200 student teachers and found that the students balanced the proportion of heads and tails quite well but alternated the results too often so that the number of runs (sequences of the same symbol) was too high and the longest run too short.

Interestingly, many people think that pure randomness alternates the outcomes often and patterns of longer runs of the same symbol appear to them as not random. They focus so much on patterns, which meets an archetype of thinking but it does not apply at all with randomness. While patterns are misleading within random situations, specific patterns might indicate that the series inspected is non-random.

3.7.2. Odds and Bayes' Formula – Revising Weights of Evidence

We present a variation of Bayes' formula, which is adapted to the process of revising prior probabilities in the light of new evidence. Odds provide advantages for reading Bayes' formula and may help to understand the situation and how it is captured by the model (Borovcnik, & Peard, 1996). We present only the special case of two disjoint and exhaustive hypotheses H_1 and H_2 with prior probabilities $P(H_i)$, which are under revision by way of new evidence E to posterior (conditional) probabilities $P(H_i \mid E)$.

The usual representation of Bayes' formula as $P(H_1 \mid E) = \frac{P(H_1) \cdot P(E \mid H_1)}{\ldots \ldots}$ (denominator omitted) gives a straightforward way to calculate the required conditional probability but no clue about the direction of change or an insight into the final probability value. We omitted the confusing term in the denominator by intent as it is neither needed to understand the formula nor is it needed to calculate the probability.

Instead, we transform the formula to odds, that is, we calculate the quotient $P(H_1 \mid E) : P(H_2 \mid E)$. In this way, we actually eliminate the denominators and obtain the new formula:

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(H_1)}{P(H_2)} \times \frac{P(E \mid H_1)}{P(E \mid H_2)}$$

New – posterior – odds	=	*Old – prior – odds*	×	*Likelihood ratio*
relative to empirical data		before collecting data		power of data

In this new representation, the posterior odds are the product of prior odds and a second quotient, which may be interpreted as the power of the data (the empirical evidence): it is the quotient of the probabilities of the evidence E under the two hypotheses under scrutiny. If this likelihood ratio is small, then E has a small probability under H_1 as compared to H_2; in this case, E is strong evidence against H_1 and should decrease the odds for H_1; if this ratio is large, then E is strong evidence for H_1 and should increase the posterior odds of H_1.

This formula reflects the *multiplicative structure* of the situation: If prior odds or likelihood ratios double, posterior odds double. We also can see that posterior odds depend on two influential parameters: prior odds *and* power (likelihood) of data to discriminate between hypotheses.

We remind the reader of the intuitive interpretation of these parts of Bayes' formula as scaling factors (as developed in Section 3.3.1); the scaling factors are established by odds (and not by probabilities) and have a straightforward meaning and a simple structure. The influence of both prior odds and data "odds" (likelihood) is open to the intuitive thought of "scaling" the uncertainty (the odds of uncertainty). At the first stage, uncertainty is "scaled" by prior odds. After new evidence has become available, the likelihood ratio re-scales uncertainty. The two scaling factors are multiplied and deliver the final scaling of uncertainty.

In this representation, the influence of both input parameters can easily be perceived. These ideas are important to explore with students to help develop their intuition.

3.7.3. Mediating Tools to Support Teaching

We illustrate tools for easier calculation of probabilities that also help to communicate the process of modelling situations by probabilities. We present the following techniques: the method of tree diagrams; the re-formulation of conditional probabilities by contingency tables, which implies the transformation of conditional probabilities to expected values (statistical villages method); and the technique of dynamic animations that allows to study the effect of input parameters to enhance their meaning.

Tree diagrams and sets of outcomes

Task 3.18. Drawing from urns repeatedly. An urn is filled with balls; three are blue, three red, and one is green. We mix the balls and draw two balls without replacement. Determine the following probabilities:

a. the balls drawn have the same colour;

b. first ball is red conditioned to the event that the colour of both drawn balls is the same.

We solve the problem with two methods: The technique with sets is inefficient if the sets become larger; using combinatorics puts severe demands on the learner. That is why tree diagrams are a popular alternative. This involves a change of representation that has to be made more explicit as we transform the description of sets to *statements* which link strongly to the subjectivist connotation of probability.

First approach using sets of outcomes

For drawing one ball, the sample space is represented by $S_1 = \{b_1, b_2, b_3, r_1, r_2, r_3, g\}$; for two repetitions of the experiment, we have the Cartesian product of all pairs of S_1, i.e., $S_1 \times S_1$. Then, we eliminate those pairs with two identical coordinates, which yields $7^2 - (3+3+1) = 49 - 7 = 42$ elements; the set may be denoted as $S^* = \{(x_1, x_2) \mid x_i \in S_1, x_1 \neq x_2\}$. There is a symmetry argument to justify that all cases are equally likely. For calculating the probabilities, we have to count the favourable cases.

a. We count 12 pairs favourable to the event *CS* "colour is the same" (we sum the possibilities for each colour $3^2 + 3^2 + 1^2$ and subtract the 7 pairs mentioned above). Laplace's rule yields $P(CS) = \frac{12}{42}$.

b. The event "red at the first draw" corresponds to the set: $R_1 = \{(x_1, x_2) \mid x_1 \in \{r_1, r_2, r_3\}, x_2 \in S_1, x_1 \neq x_2\}$. For the probability of red first *given* the *same colour*, we investigate the intersection $R_1 \cap CS = \{(x_1, x_2) \mid x_1, x_2 \in \{r_1, r_2, r_3\}, x_1 \neq x_2\}$ and count 6 cases whence $P(R_1 \cap CS) = \frac{6}{42}$ and, with the result of a., $P(R_1 \mid CS) = \frac{P(R_1 \cap CS)}{P(CS)} = \frac{6}{42} / \frac{12}{42} = \frac{6}{12} = \frac{1}{2}$.

Second approach using tree diagrams

As the ball drawn is *not* replaced, the composition of the urn and the *conditional* probabilities for the second draw change. Instead of mapping the complete tree, we simplify it by denoting only the colour at the nodes, R_1, B_1, G_1, and use an index 2 to designate the second stage. R_1 means "red at first draw" (not to be confused with the ball labelled r_1, which can also be drawn at second draw); R_2 denotes "red at second draw" (Figure 3.9). We know the conditional probabilities from the

experimental conditions, e.g., $P(R_2 \mid R_1) = \frac{2}{6}$ as one of the red balls is eliminated (it does not matter, which one) and the possible cases are reduced by 1.

The endpoints of a path correspond to *combined events*; the edges of the tree are labelled by the appropriate conditional probabilities. There are two rules with tree diagrams for calculating probabilities:

– Rule 1. Multiplication of the conditional probabilities alongside a path yields the conjunction probability that corresponds to the event that the path represents.
– Rule 2. If an event consists of more paths, we obtain the probability for the event by adding the probabilities of all corresponding paths.

It is important to note that tree diagrams shift the focus from single outcomes of a sample space to events, which represent the nodes; this way they reduce the complexity of the problem.

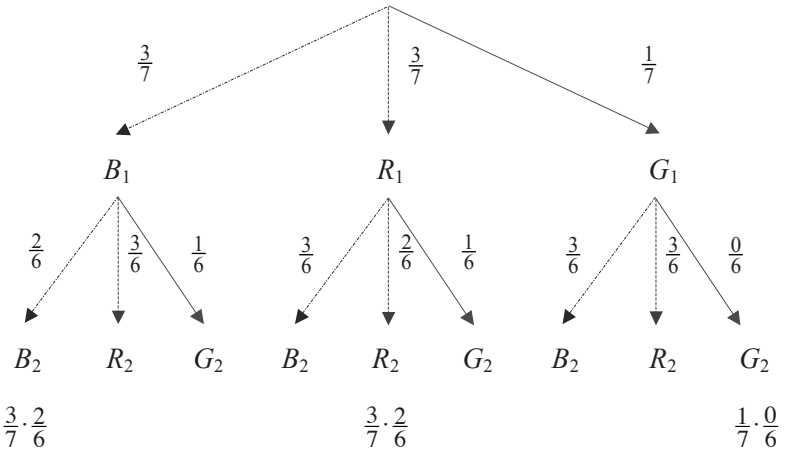


Figure 3.9. Tree diagram built up by the stages of the repeated experiment

a. From the tree diagram, we read off $P(CS) = \frac{3}{7} \cdot \frac{2}{6} + \frac{3}{7} \cdot \frac{2}{6} + \frac{1}{7} \cdot 0 = \frac{12}{42}$.

b. If we condition on the event *CS*, then only three paths remain. The only path that belongs to R_1 is the dashed one in the middle. Its *relative* weight yields the corresponding conditional probability $P(R_1 \mid CS) = \frac{6}{42} / \frac{12}{42} = \frac{6}{12} = \frac{1}{2}$.

Efficient use of tree diagrams

Many people are puzzled with the high probability of a coincidence of birthdays. What is the probability that of 30 persons no one has the same birthday ("no match"). Or, what is the probability that at least two of them have the same birthday ("at least one match")? People are usually surprised to find two persons in a class of 30 students who share the birthday.

The problem is usually solved by combinatorics; we refine the technique of tree diagrams (Figure 3.10). The sample space for each of the persons is $S_1 = \{1, 2, ..., 365\}$ for the days numbered subsequently. For *n* persons we have the *n*-fold Cartesian product.

Rather than label the nodes by numbers for the days, we label them by the statement whether the first person gives a date that already has been announced, and then the second person gives a date that leads to a match of birthdays (or not). We reduce the binary tree to the one path with no matches. We bring the persons in a sequence and ask them one after the other. The first has all days as favourable, for the second the favourable cases reduce by one. The multiplication rule yields 0.2937 for no match at all. Then, by the complementarity rule, the probability for at least one match is 0.7063 for 30 persons.

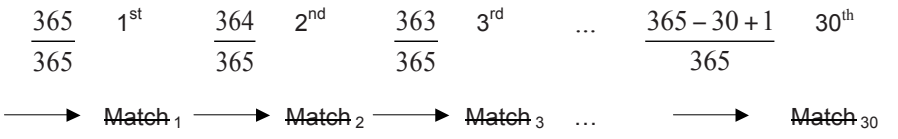


Figure 3.10. Tree encoded for statements related to the specific results

Teaching the use of contingency tables

We start with the data from the example on medical diagnosis (Section 3.3.1). From the contingency table (Table 3.3), the students should first determine the column percentages (right hand of Table 3.9), which show the distribution of positive and negative test results in the subgroups of people with and without the *disease* (columns); then they should determine the row percentages, i.e., the distribution of the disease in the subgroups of positive and negative test results (rows; left hand of Table 3.9).

It is vital for students' progress that they recognise that both distributions look at the same problem from different angles and that they learn to investigate how the distributions differ, and how they can be linked.

Usually, the column percentages are known (by other medical investigations like biopsy) and the row percentages are asked for (a positive or negative medical test should be interpreted). Some data read of Table 3.9 are: 90% of the group with disease *A* have a positive result; 8.3% of the group with positive result actually have *A*.

As relative frequencies are estimates of corresponding probabilities, we can interpret the row and column percentages as conditional probabilities. This is justified by the analogy between probability and descriptive statistics (see Table 3.7), which yields a general pattern for the interpretation of probability. In this activity, students can use this link to switch from data to probabilities. The row (column) percentages offer an easier access to what finally is a probability distribution related to the subgroup that is defined by the row (or column).

Table 3.9. Row and column percentages for comparing subgroups (data from Table 3.3)

Distribution of disease split by test result				Distribution of test results split by disease			
Biome-tric test	Status of disease *A*			Biome-tric test	Status of disease *A*		
	Disease	No disease	All		Group 1 Disease	Group 2 No disease	All
Group 1 Pos. +	8.3	91.7	100.0	Pos. +	90.0	10.0	10.8
Group 2 Neg. –	0.1	99.9	100.0	Neg. –	10.0	90.0	89.2
All	1.0	99.0	100.0	All	100.0	100.0	100.0

Statistical villages

This resource works the other way round; it starts with probabilities and conditional probabilities and establishes a table of natural (absolute) frequencies to illustrate and simplify the calculation of further (conditional) probabilities as has been suggested by Gigerenzer (2002)

Task 3.19. In medical diagnosis of disease *A*, a positive test result is taken as circumstantial evidence that the patient has *A*. The quality of such tests is characterised by their sensitivity and specificity. Sensitivity means the probability that the test result is positive if the person suffers from this disease – we assume 0.99; specifity of (assumed) 0.95 means that the test result is negative with probability 0.95 if the person does not suffer from *A*.

a. What is the probability that a person with a positive test result actually suffers from *A*?

b. In the population, *A* has a prevalence of 0.001 (1 of thousand). What is the probability of the person to have *A* if the test result is positive?

c. The person belongs to a risk group with a prevalence of 10% of *A*. What is the probability of the person to have *A* if the test result is positive?

Students should see that they cannot answer a. until they have some information about the prevalence, i.e., the percentage of the population with the disease. For question b., the students could use a tree diagram to obtain the conditional probability $P(A|+) = 0.0194$ given a positive test, which increases to 0.6875 in the risk group in c.

Rather than using tree diagrams, the students could apply a different strategy here, which may be described as investigating what happens if the given probabilities apply to a statistical village of 100 000 people. By this approach, all (conditional) probabilities are transformed to expected values (numbers), so that a (theoretical) frequency table for that village is reconstructed. If N persons are involved and p is the probability of the subgroup then Np are expected in the corresponding cell.

Table 3.10. Numbers of people of the various characteristics in the statistical village with prevalence of A of 0.001

Biometric test	Status of disease *A*		
	Dis-ease	No disease	All
Positive +	99		
Negative –		94 905	
All	100	99 900	100 000
Prevalence *P*(*A*):			0.001
Sensitivity *P*(+|*A*):			0.990
Specifity *P*(–|not-*A*):			0.950

Biometric test	Status of disease *A*		
	Dis-ease	No disease	All
Positive +	99	4 995	5 094
Negative –	1	94 905	94 906
All	100	99 900	100 000

Probabilities given test positive	
No disease:	0.9806
Disease:	0.0194

The numbers are filled in step by step; first use the prevalence to determine the column totals, second apply sensitivity and specificity to the column totals, and third complete the contingency table by row totals (as indicated by arrows in Table 3.10).

If a person is randomly selected from this village, the known probabilities are mirrored; e.g., $P(A) = 100/100{,}000 = 0.001$. Other probabilities of interest may also be read off this table. For the probability of A given a positive test, the person is selected from the subgroup that corresponds to the first row, which yields $P(A|+) = 99/5094 = 0.0194$.

Use of dynamic animations

We finally give an example of didactical animations that may help to recognise essential features of a specific random situation and of the model we use to describe it. We deal with the first success in a Bernoulli series.

To investigate waiting times[29] is a key idea of Bernoulli experiments with success probability p. Students could establish the distribution of the waiting time (a geometric distribution) for a reference value of $p = 0.25$ by simulation and ask for the waiting time that is exceeded with a probability of 0.10. To analyse the effect of changing the value of p on the calculated threshold might clarify misbeliefs about waiting times. For that purpose it is better to draw the exact geometric distribution and not to simulate data according to the assumptions of the model.

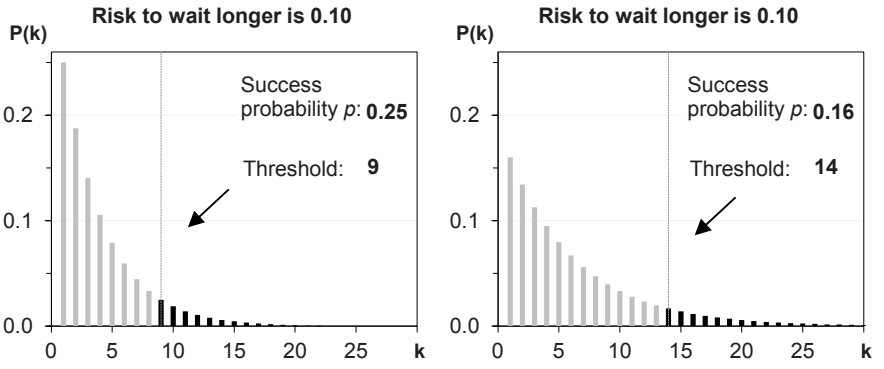


Figure 3.11. Dynamic animation with changing the success probability: waiting for the first success – current and initial situation

As may be seen from Figure 3.11, this threshold is 9: there is a risk of 10% of waiting longer than 8 (i.e., 9 and more) times. The students should investigate how this value changes with a different success probability; by changing the value of p with a slider gradually, they might pursue the effect on the calculated threshold in an "animated film". It is amazing how fast the distribution for the waiting times slips out to the right and by reaching the level of $p = 0.16$ (approximately the six on the ordinary die), waiting longer than 13 (14 and more) times has a risk of already 10%.

Thus, with a die it is quite usual to wait such a long time for the first six to occur. The animation might clarify a basic misconception on probabilities as waiting times heavily bias the perception and judgement of probabilities.

[29] While time is a continuous variable, our "waiting time" is discrete as it equals n if the event occurs at the nth trial. It might be better to use "number of the trial with the first success". Yet, it is suggestive to refer to it as waiting time here.

CHAPTER 4

EXPLORING AND MODELLING ASSOCIATION

In this chapter, we analyse the ideas of association, correlation and regression. We also discuss the underlying concepts that serve to model statistical relationships between variables and extend functional dependence to random situations. After describing how these ideas are introduced in high-school curricula, we present two teaching situations. The first introduces association in contingency tables, which are used when both variables are either qualitative or quantitative with only few different values. The second activity deals with correlation and regression, which apply when both variables are measured on a ratio scale. We restrict our presentation to exploratory methods and simple cases, which are taught in high school: 2×2 contingency tables and linear regression. A synthesis of the main learning goals, research on students' difficulties with these concepts, and teaching activities that may help students to develop their statistical reasoning complete our exposition.

4.1. INTRODUCTION

In a statistical study, researchers are often interested in finding relationships *between* variables and, when possible, in finding a mathematical model (for simplicity, often a straight line) that fits to the data and can be used to predict the values of a so-called *response* variable when some values of the *explanatory* variable are known. To solve these two problems, the ideas of association, correlation and regression were developed; they extend the concepts of functional dependence to random situations.[1]

These concepts form the foundation for many statistical procedures, including simple and multiple regression (with one or several explanatory variables), analysis of variance, and most multivariate methods (like factor analysis). According to Stanton (2001), Galton's work on inherited characteristics of sweet peas (e.g., the diameter; see Galton, 1889) led to the initial conceptualisation of linear regression. Bravais (1846) is credited for the initial mathematical formulas for correlation, while Pearson (1896) published the first rigorous treatment of correlation and regression. Subsequent joint efforts by Galton and Pearson brought about the more general technique of multiple regression.

As regards contingency tables, Stigler (2002) points out that this method of representing data has rarely been used before 1900. An exception are 2×2 tables that have a long history in logic (going back to Aristotle) and a long tradition in probability. In his study of fingerprints, Galton (1892) developed a formula to

[1] Association is the general term to denote a statistical relationship between variables. The concepts of contingency tables, association coefficients, and correlation and regression have been developed within statistics to study association.

obtain the expected frequency in a cell, from the total count, the marginal frequencies, and the assumption of independence of the two involved characteristics. In 1900, Karl Pearson introduced the chi-square test (Pearson, 1900).

These topics are included in the high-school curricula in many countries. For example, the Common Core State Standards Initiative (CCSSI, 2010) suggests that grade 8 students should discover patterns of association in 2×2 tables by comparing relative frequencies; they also should investigate scatter plots to evaluate and describe positive or negative correlation, as well as linear and nonlinear tendencies. Students should also fit a straight line to bivariate data, perceive this line as a model for the trend in the data, interpret slope and intercept of this line, and informally assess the fit of the model. For grades 9–12, students should use functions to describe the trend in bivariate data; if the data suggest a linear relationship, the strength and direction of the relationship should be expressed and interpreted by the correlation coefficient. Similar contents are included in the New Zealand curricula for grades 8–12 (Ministry of Education, 2007) and in Spain for grade 11 (MECD, 2015).

In the GAISE project it is recommended that, at intermediate level B, students quantify the strength of association and develop simple models to describe the association between two variables; this includes contingency tables for two categorical variables[2] and straight lines for modelling association between two metric variables. At the upper level C, students should be able to recognise when the relationship between two metric variables is reasonably linear, compute and interpret Pearson's correlation coefficient as a measure of the strength of the linear relationship, and understand the least-squares criterion for line fitting.

In this chapter, we introduce simple methods that help students to conceptualise and evaluate statistical dependence between two variables and to understand that this relationship differs from functional dependence (a concept, high-school students should be familiar with). Further learning goals are that students discriminate between statistical association and cause-effect relationships and recognise that statistical methods for the study of interrelations depend on the level of the scale on which the variables are measured.

The chapter starts with a teaching example from which we derive simple methods to explore the association in 2×2 tables: conditional probabilities, simple graphs, comparison of row or column proportions, or comparison of *observed* and *expected* frequencies in the table.

In a second activity with metric variables (i.e., the underlying scale allows all calculations with the numerical values), an informal approach to interpreting a scatter plot is a good start for any investigation of the ideas of regression and correlation. Students can judge intuitively whether two variables are reasonably

[2] At university level, the chi-square test or various coefficients of association may be used to study the association between two categorical variables. The observed value of this statistic may be used to test for significance, i.e., to decide whether the hypothesis of independence between the variables can be rejected or not.

linearly connected or whether another type of function better describes the "dependence" (the general trend) between the variables.

An introduction to Pearson's correlation coefficient is prepared by the quadrant-count ratio. Students then are introduced to ideas of curve fitting and regression lines. The aim is to apply the regression line (or curve) to model the general trend of "dependence" between the involved variables, to use the model function to make predictions about the dependent variable from the value of the independent variable, and to informally judge the adequacy of the model applied to fit the data.

In this chapter, we also discuss the teaching goals and students' difficulties in judging or estimating association from different representations of data or in understanding the properties of the concepts of correlation and regression. As in the other chapters, additional activities and resources that address such difficulties complete the discussion.

4.2. A TEACHING SITUATION TO EXPLORE CONTINGENCY TABLES

In Chapter 3, we introduced contingency tables as a useful representation of bivariate data. We used medical diagnosis as a natural context to introduce conditional probabilities. Here we adapt these tools and replace probabilities by (absolute and relative) frequencies in order to develop simple methods to assess association in 2×2 tables.

4.2.1. Exploring Association in 2×2 Contingency Tables

The lesson can start by posing to the students the following problem, which is adapted from Batanero, Cañadas, Díaz, and Gea (2015).

> *Task 4.1.* A psychologist was interested in studying whether the number of siblings of a child influences the circumstance that the child displays behavioural problems. The first variable is simplified in terms of two values: an "only child" and a "child with siblings". The second variable expresses whether the child displays behavioural problems or not. Can data provide empirical evidence about a relationship between these variables? The psychologist collected data, which are presented in Table 4.1. Looking at the data, do you think there is a relationship between being an only child and displaying behavioural problems? Explain your response.

After looking at the data in Table 4.1, some students may argue that the tendency to be problematic in only children in this sample is clear as there are more children with behavioural problems in the only-child group. Other students might even draw an attached bar chart for these data (Figure 4.1) to compare both groups and support their claim. The teacher should remark that the counts of the two groups cannot be compared directly as the groups are different in size and can suggest drawing a stacked bar chart instead, which uses relative frequencies (Figure 4.2). In this graph, each row in Table 4.1 is scaled to a row total of 100% and the percentage of children displaying problems (or not) relative to the row total is presented.

Table 4.1. Data on joint occurrence of behavioural problems and siblings

	Child displays behavioural problems		
Type of family	Yes	No	All
Only child	80	30	110
Child with siblings	40	100	140
All	120	130	250

With this new representation a clearer pattern evolves indicating a strong association between the variables. Students might object using percentages and still prefer their representation (Figure 4.1). A further example with a larger difference in size between the two groups should suffice to convince them.

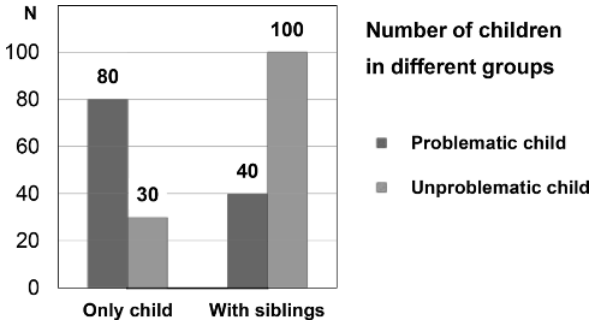


Figure 4.1. Attached bar chart for the child behaviour in children with and without siblings

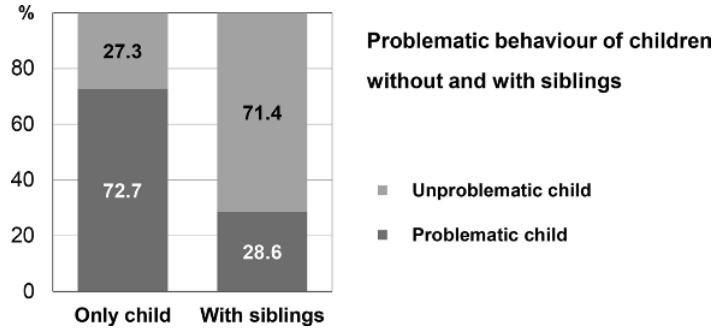


Figure 4.2. Stacked bar chart with percentage of children displaying behavioural problems in the subgroups of children with or without siblings

4.2.2. Different Distributions in 2×2 Contingency Tables

In this section, we introduce further concepts and procedures for evaluating the association in 2×2 tables. We work with percentages that are more intuitive for students; however, the same analyses may be carried out using relative frequencies when all the values are rescaled to the interval [0, 1]. Different from univariate distributions where we can compute only one percentage for each value of the variable, in contingency tables we can compute various *types* of percentages for the combination of values that correspond to each cell.

In Table 4.1, the cell in the first row and first column – with a count of 80 – represents the number of children in the sample that are "only child" *and* display behavioural problems. These 80 children can be referred to the whole sample of 250 children, which makes an overall percentage of 32%. They can be referred to all 110 children of the subgroup "one child" and we obtain that 72.7% of children without siblings display behavioural problems. We can also compare the 80 children in this cell to the first column; in this case, we compare them to the subgroup of all 120 children that display behavioural problems 80/120 = 33.3%.

In Figure 4.2, we represent the percentages of types of *children per row*, that is, when the row total used to compute these percentages instead of the overall total. These percentages are displayed in Table 4.2. In each row, the percentage of a cell is computed under the condition that the child belongs to this row (for the first row, the condition is "only child", for the second, it is "has siblings"). For this reason, we name these percentages conditional to the corresponding row (or row percentages); each row represents a subgroup and a univariate distribution for the variable "behavioural problems"; therefore, the row percentages add up to 100% in each row.

Table 4.2. Conditional distributions per row in Table 4.1 (row percentages)

	Child displays behavioural problems		
Type of family	Yes	No	Row total
Only child	72.7	27.3	100.0
Child with siblings	28.6	71.4	100.0

In the same way, in Table 4.3, the conditional distributions of the data in Table 4.1 are displayed per column; that is, these percentages are computed using the column total and now the sum of each column equals 100%. We name them percentages conditional to the corresponding column or simply column percentages.

We may instead be interested in the percentage each cell represents within the complete sample; these percentages (Table 4.4) represent the frequency of the occurrence of combined values as compared to the full sample. This is called the *joint distribution* of the two variables that build the contingency table. The joint distribution compares each cell to the full data set while the row and column

distributions compare the cell only to the corresponding row or column. All of them give a different – and differently useful – way to address various aspects of association of the involved variables.

Table 4.3. Conditional distribution per column of the data in Table 4.1 (column percentages)

	Child displays behavioural problems	
Type of family	Yes	No
Only child	66.7	23.1
Child with siblings	33.3	76.9
Column total	100.0	100.0

Table 4.4. Joint distribution of data in Table 4.1 (whole sample percentages)

	Child displays behavioural problems		
Type of family	Yes	No	All
Only child	32.0	12.0	44.0
Child with siblings	16.0	40.0	56.0
All	48.0	52.0	100.0

We switch back to the single variables. From the contingency table (Table 4.1), we can also determine the counts for the values of each variable. If we sum the counts in each row, we get the absolute frequencies of children from "only child" and from "with siblings" families. Thus, on the right margin, we have data of the type of family, a variable that defines the values (the subgroups) represented in each row. As this distribution stands on the (right) margin of the table, it is called the *marginal distribution* of the variable "type of family".

In the same way, if we only consider the variable "behaviour", we find the absolute frequencies on the bottom margin of the table (Table 4.1). Some students get confused, which marginal distribution is represented on the right and which one on the bottom. A look back at the table should clarify this issue.

Another representation of the joint distribution is the mosaic plot (Figure 4.3). In this diagram, we represent the different percentages of the cells as percentages from the whole sample using areas. We construct this graph in two steps, starting from a unit square:

a. The height of the unit square (situated on the vertical axis) is divided proportionally to the number of "only children" and "children with siblings". This way, two rectangles (horizontal stripes) with the same basis are constructed. Each of these rectangles represents one subgroup of children.

b. Each rectangle is then divided horizontally according to the proportion of children with and without problematic behaviour that applies to the corresponding group and is coloured accordingly.

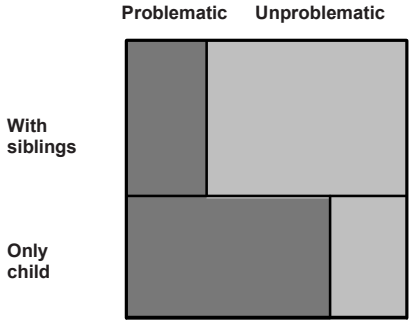


Figure 4.3. Mosaic plot of the children data

The mosaic plot displays three different distributions. The marginal distribution of the first variable (type of family) is represented by the two horizontal rectangles with the same base and heights proportional to the sample size in each group. The conditional row distribution is represented by the split of each horizontal block into two smaller rectangles with bases proportional to the number of problematic and unproblematic children (displayed in two different colours). The joint distribution is represented by the area of the final four rectangles.

The diagram is hierarchically constructed and one has to pay attention to which variable is used for the vertical axis; in case we want to represent the other conditional distribution of the type of family within the subgroups of problematic and unproblematic children, we need to build another mosaic plot using this variable in the vertical axis.

4.2.3. Simple Methods to Evaluate Association in 2×2 Tables

Once the students understand the various frequencies and distributions that can be deduced from the data of a contingency table, we can introduce elementary methods to study the association of the variables, which is "hidden" in the data. For this purpose, we use students' previous knowledge of probability. These methods are not well understood by students unless they are taught as shown in Batanero, Cañadas, Arteaga, and Gea (2013).

Comparing conditional probabilities

Students may use conditional probabilities, which were introduced in Chapter 3 to evaluate association in 2×2 tables. Assuming we randomly choose a child from our sample of 250 children (data in Table 4.1) and consider the events *Y:* "yes" (child

displays problematic behaviour), *N:* "no" (child does not display problematic behaviour), *O*: only child, *S:* child has siblings, then we obtain the following probabilities:

– Probability that the child is problematic: $P(Y) = \frac{120}{250} = 0.480$.

– Conditional probability of problematic behaviour given we deal with an only child: $P(Y \mid O) = \frac{80}{110} = 0.727$; conditional probability that the child is problematic given that the child has siblings: $P(Y \mid S) = \frac{40}{140} = 0.286$.

The conditional probabilities for "yes" (problematic behaviour) differ enormously in the subgroups with respect to the type of family. They are also different from the marginal probability of being a problematic child. These differences reflect that the data provide empirical evidence for an association between the type of family and the occurrence of behavioural problems.

Comparing relative risks

In our example, the probability of displaying problematic behaviour is remarkably higher for only children than for children with siblings. We can compute the *relative risk* of problematic behaviour for "only children" compared to "children with siblings", i.e., we divide the corresponding conditional probabilities in these two groups. In our sample, it is 2.55 times more likely for an only child to have behavioural problems than for children who have siblings:

$$\frac{P(Y \mid O)}{P(Y \mid S)} = \frac{0.727}{0.286} = 2.55 .$$

4.2.4. Expected Frequencies for Independent Variables

In order to introduce the idea of independence, the teacher may ask the students to guess, which would be the *expected frequency* in each cell in Table 4.1 in case that both variables were independent (i.e., no relationship between the two variables under scrutiny) if we know only the absolute numbers of the marginal distributions. Students could be asked to fill the empty cells in Table 4.5 by trial and error.

Table 4.5. Estimating expected frequencies of cells in case of independence

	Child displays behavioural problems		
Type of family	Yes	No	Row total
Only child			110
Child with siblings			140
Column total	120	130	250

Since the overall rate of problematic children is $\frac{120}{250} = 0.48$ and since no dependence between the variables is presupposed, this rate should apply to both types of family (only child or child with siblings). Then, for the group of only children, the expected number of children with problematic behaviour should equal $0.48 \cdot 110 = \frac{120 \cdot 110}{250} = 52.8$; rounding to integers, we expect 53 children in this group. In the same way, the students can compute the expected number of children in the other cells of Table 4.5 (results are presented in Table 4.6) and recognise the following relationship between margin cells and cells:

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Sample total}}.$$

Students may also notice that in each row we need to compute only one expected frequency since the row total is fixed. Furthermore, the second row can be deduced from the first row. Comparing the observed to expected frequencies (Tables 4.1 and 4.6), we find that there are more children than expected with problematic behaviour in the group of only-child families (80 vs. 53) and less in the group with siblings (40 vs. 67). To facilitate the comparison between observed and expected frequencies in the various cells of the contingency table, one should display both tables side by side.

Table 4.6. Expected frequencies in case of independence (rounded)

	Child displays behavioural problems		
Type of family	Yes	No	Row total
Only child	53	57	110
Child with siblings	67	73	140
Column total	120	130	250

These differences are also visible in the mosaic plot as the coloured rectangles corresponding to problematic children have distinct base lengths for both groups that represent the type of family (Figure 4.3). As for independence these lengths should be equal, the data provides evidence of an *association* between the variables.

We can compare the differences in the various cells qualitatively – as we have done above – to see whether the data are reasonable or whether these differences are too large when we accept the independence hypothesis. Another idea is to detect those cells that have the largest differences so that we can describe, in which way the observed data deviates from expected data assuming independence. These differences do not only establish empirical evidence of an association but also deliver potential explanations why the variables are associated.

Rather than calculating simple differences between the entries of corresponding cells (in Tables 4.1 and 4.6) or absolute differences, the squared differences are divided by the expected frequency of the cell. These "differences" form the basis for the chi-square statistics (Table 4.7; see also Section 4.7.1) and related measures for the strength of the association between the two variables. We will use only the table of these deviations to judge the degree of association qualitatively here. The cell for problematic *and* only-child children marks the largest deviation from independence; there are *far more* children with behavioural problems in the group of "only children" compared to the number that is expected under the independence hypothesis.

Table 4.7. Chi-square "differences" between observed and expected frequencies

	Child displays behavioural problems	
Type of family	Yes	No
Only child	14.0	12.9
Child with siblings	11.0	10.2

4.3. LIFE EXPECTANCY: A TEACHING SITUATION TO EXPLORE CORRELATION AND REGRESSION

In this activity, taken from Batanero, Gea, Díaz, and Cañadas (2014),[3] the students explore open data. An enormous amount of data on a wide range of variables is collected by international institutions and processed by the United Nations (UN) in their human development reports. These data are freely available and the official website (hdr.undp.org/en/data) also provides graphical facilities to explore the data. The activity focuses on investigating possible factors (variables) that might be related to life expectancy at birth in a given country. Students will be interested in analysing the variables that affect life expectancy as well as in observing the general increase in life expectancy over the past decades and the differences between the countries.[4]

[3] Many of the tasks analysed in this section have been used within a workshop directed to prospective teachers.

[4] The analysis of factors influencing the capacity of memory is another project to discuss potential relations between variables. Borovcnik and Kapadia (2012) used this context in teacher in-service courses. The factors investigated were the kind of words (strange or trivial words), the sequence, in which the words were presented, the place of a word (when it was presented), and gender, all of which played a key role for the success of retrieving the words from memory a short time after the words have been presented to the experimental group. Besides regression and correlation, the authors also use flexible diagrams to investigate these relations.

4.3.1. Exploring and Explaining Correlation

When working with contextual data, it is important that, first of all, the meaning of the variables involved is clarified. After that, we can continue by investigating scatter plots.

Understanding the variables
To start the activity, the teacher provides the students with a spreadsheet that contains data on a series of variables.[5]

> *Task 4.2.* Read the data for your country and compare it to neighbouring countries and to countries of your interest. Find the descriptions of the various variables, their range of variation and make sure you understand the meaning of these data. With respect to which variables is your country "better" than the countries you compared it to? Can you interpret the size of the differences?

In Figure 4.4, the variables used in this chapter are displayed; each student could add and explore other variables related to life expectancy or else perform the analyses with other dependent variables. After the teacher explains the source and relevance of the data, the discussion with the students should clarify the meaning of each variable and describe the way the data were collected.

	Life expectancy	Human developent index (HDI)	Gross national income per capita	Adolescent fertility rate	Infant mortality rate	Expenditure on health, public (% of GDP)	Educat-ion index	Total population (mio)	% Population urban
Afghanistan	47.9	0.387	1.200	121.3	199	1.8	0.359	30.58	22.3
Albania	76.6	0.734	7.449	14.2	15	2.9	0.719	3.19	50.9
Algeria	72.7	0.691	7.421	7.3	32	3.6	0.646	34.95	65.9
Andorra	80.7				4	5.3		0.08	88.4
Angola	50.3	0.481	5.278	123.7	161	2.0	0.422	18.56	57.6
Antigua a. Barbuda	72.4		17.052		12	3.2		0.09	30.3
Argentina	75.6	0.788	13.202	56.9	14	5.1	0.802	40.06	92.2
Armenia	74.0	0.712	4.794	35.7	22	2.1	0.760	3.09	64.1
Australia	81.7	0.926	34.259	14.9	5	6.0	0.980	21.90	88.9
Austria	80.5	0.879	34.673	12.8	4	7.7	0.851	8.37	67.3
Azerbaijan	70.4		8.752	33.8	34	1.0		9.07	51.8
Bahamas	75.1	0.769		53.0	12	3.7	0.671	0.34	83.9
Bahrain	74.8	0.805		16.7	12	2.6	0.744	1.17	88.6
Bangladesh	68.3	0.491	1.286	71.6	52	1.1	0.410	147.03	27.6
Barbados	76.5	0.790		42.7	11	4.4	0.746	0.27	43.8
Belarus	69.8	0.746	11.841	21.3	12	4.9	0.776	9.64	74.2
Belgium	79.9	0.883	32.395	7.7	5	7.0	0.880	10.66	97.4
Belize	75.6	0.696	6.019	78.7	18	2.6	0.661	0.31	51.8
Benin	55.2	0.422	1.369	111.8	118	2.5	0.362	8.60	41.6
Bhutan	66.4		4.643	38.3	79	3.3		0.71	33.9
Bolivia	66.0	0.656	4.013	78.2	51	3.4	0.741	9.77	66.1

Figure 4.4. Data for the activity stored in an Excel file

[5] We downloaded the data from 2009 from the UN server. The data are aggregated over countries. Each country is one statistical unit regardless of the number of inhabitants.

These variables are described in the UN web site in the following way:

- *Life expectancy at birth (LEX)*: Number of years that a new born could expect to live if the prevailing patterns of age-specific mortality rates at the time of birth stay the same throughout the infant's life.
- *Human development index* (*HDI*): A composite index measuring average achievement in three basic dimensions of human development – a long and healthy life, knowledge, and a decent standard of living.
- *Gross national product per capita* (*GNP*): Gross value added by all resident producers in the economy plus any product taxes and minus all subsidies not included in the value of the products (expressed in international dollars using purchasing power parity rates) divided by total population during the period.
- *Adolescent fertility rate*: Number of births per 1000 women in the interval of age 15–19.
- *Infant mortality rate*: Number of children dying between birth and exactly age 5, expressed per 1000 live births.
- *Public expenditure on health*: Central and local public budgets on health, external borrowings and grants (including donations from international agencies and non-governmental organisations), and social or compulsory health insurance funds, expressed as percentage of GNP.
- *Education index*: an index varying between 0 and 1 that takes into account the years of schooling and other educational variables.
- *Population, total*: number of all people (in thousands).
- *Population, urban*: percentage of total population living in areas classified as urban according to the criteria used by each country.

Interpreting correlation from scatter plots

Once the students are familiar with the meaning of the variables, the class should discuss factors that could influence or be associated with life expectancy. The students might argue that the total population should not be related, while expenditure on health should influence the health status in the country and, as it is stable over years, it should indirectly influence life expectancy. On the other hand, the education index and life expectancy are both related to the standard of living (a so-called common third factor), which increases the education index as well as life expectancy. The discussion shows that there are different kinds of relations between these factors and life expectancy and that the strength of relation differs.

After this class discussion, the teacher provides the students with a series of scatter plots (Figures 4.5 and 4.6) that – for each country – represent life expectancy (*Y* axis; response or dependent variable) against other variables (*X* axis, explanatory or independent variable). The students are encouraged to work in pairs to complete the following tasks.

Task 4.3. Assign a score between 0 and 1 to each of the following scatter plots, which corresponds to your visual judgement of the strength of the relationship of the particular explanatory variable with respect to life expectancy, where 0 means there is no relationship at all and 1 is the maximum strength of relationship.

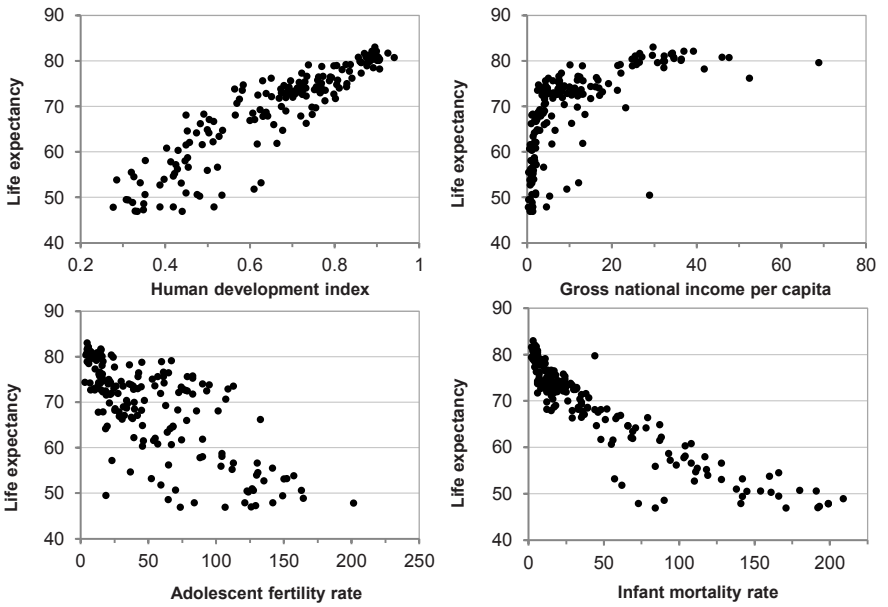


Figure 4.5. Scatter plots of life expectancy against various variables

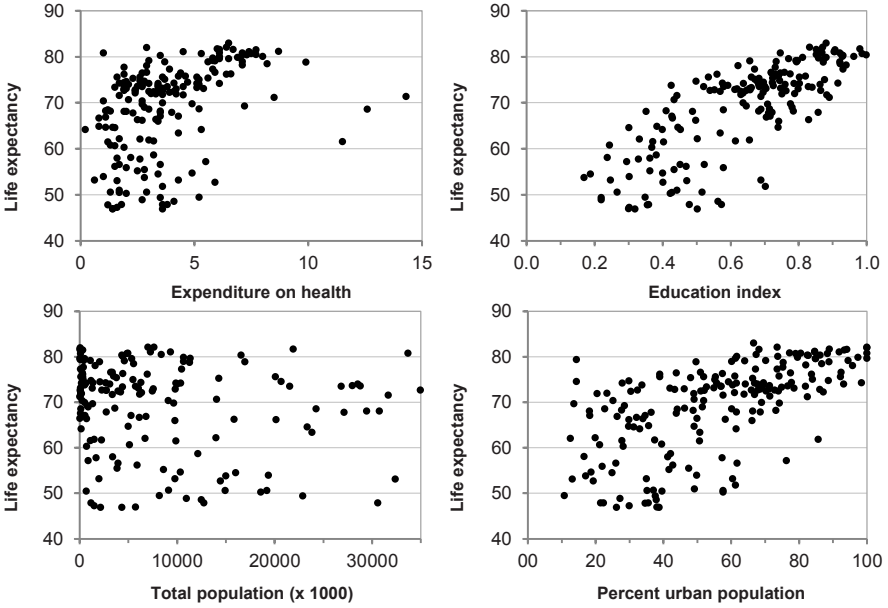


Figure 4.6. Scatter plots of life expectancy against other variables

Task 4.4. Assign a sign (+ or –) to each scatter plot depending on whether you think the relationship is direct (both variables vary in the same direction; when one increases the other also increases and when one decreases also does the other variable) or inverse (the variables vary in opposite directions; when one increases the other decreases and vice-versa).

In these scatter plots, the students will find examples of different types of relationships with life expectancy: direct or positive interrelations (human development index) as well as inverse or negative relationships (child mortality rate); strong (human development index), moderate (percent of urban population) and no association (total population). Furthermore, they will recognise patterns in the scatter-plot tendency (linear or otherwise). An intuitive idea is that the higher the spread in the scatter plot, the lower the strength of the relationship between the variables involved.

The next step is introducing the concept of covariance and the *Pearson correlation coefficient r*. To facilitate the introduction, in Figure 4.7, the data of only 16 countries are plotted and parallel lines to the axes are added crossing the centre of gravity of the distribution (i.e., the point with coordinates $(\overline{x}, \overline{y})$). These two lines form a new coordinate system that divides the plane into four quadrants (Figure 4.7).

As suggested in Figure 4.7, if there is a direct association, the points tend to appear in quadrants 1 and 3; if there is an inverse association (for example, life expectancy and child-mortality rate), the points tend to be located in quadrants 2 and 4. Holmes (2001) used this property to propose the *Quadrant-count ratio* (QCR) as an intuitive measure of association in metric variables. If n_i is the number of points in Quadrant i, then:

$$\text{QCR} = \frac{(n_1 + n_3) - (n_2 + n_4)}{n_1 + n_2 + n_3 + n_4}.$$

This measure is not used in the applications of statistics; yet it is recommended in the GAISE project (Franklin et al., 2007) as a preliminary step to introduce correlation because of its following intuitive properties:

– In case all the points are located in quadrants Q_1 and Q_3, QCR = 1;
– In case all the points are located in quadrants Q_2 and Q_4, QCR = –1;
– If more points are in Q_1 and Q_3, the relationship is direct and QCR > 0;
– If more points are in Q_2 and Q_4, the relationship is inverse and QCR < 0.

From Figure 4.7, we count 13 points for a positive and 3 for a negative association so that QCR = (13–3)/16 = 10/16 = 0.6250, which is an indication for an intermediate association between life expectancy and HDI. A high HDI in a country indicates a high local life expectancy. The measure, however, is crude insofar as it delivers the same value regardless how the points are located in the quadrants (close to a line or function or scattered all over).

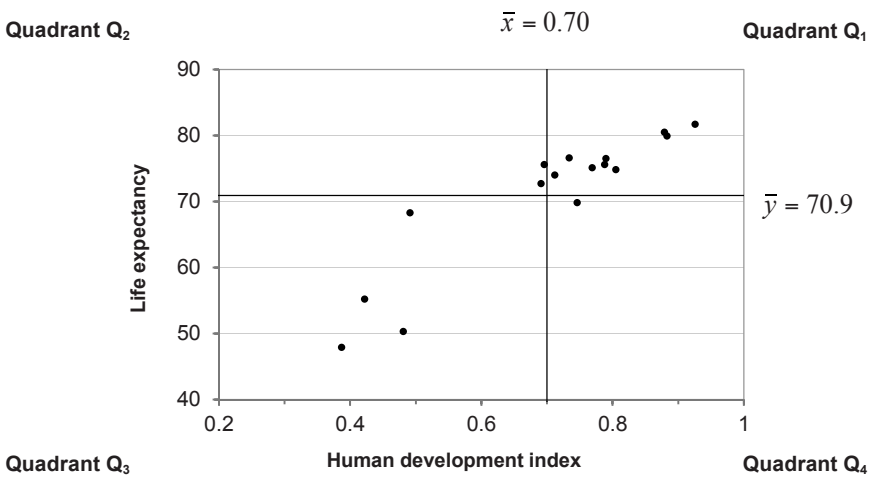


Figure 4.7. Scatter plot of a direct relationship

Covariance and pearson correlation coefficient

When using QCR, we count only the number of points in the four quadrants and neglect the distance of these points from the centre of gravity. Each point has the same weight regardless how well it reflects a direct or indirect relation between the variables. Intuitively, to check the *strength* of the relationship, it is important to take into account the deviation of each point from the centre of gravity. A statistic that considers these deviations is the covariance, which provides a measure of the direction (direct or inverse) *and* the strength of the association. It is defined as:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}).^{6}$$

The sign of the covariance indicates the direction of the relationship; if $s_{xy} = 0$, the variables are uncorrelated. As suggested in Figure 4.8, if there is a positive association, then the deviations from the mean $(x_i - \bar{x})$ and $(y_i - \bar{y})$ tend to have the same sign, because either both coordinates are above or below the average; hence the products of these deviations tend to be positive, producing a positive sum. Similarly for a negative association, the deviations from the mean tend to have opposite signs for *x* and *y* values producing a negative product leading to a negative sum.

[6] In inference, $n-1$ is used in the denominator in order to obtain unbiased estimators. In exploratory data analysis, *n* is preferred. For large *n,* the difference is negligible.

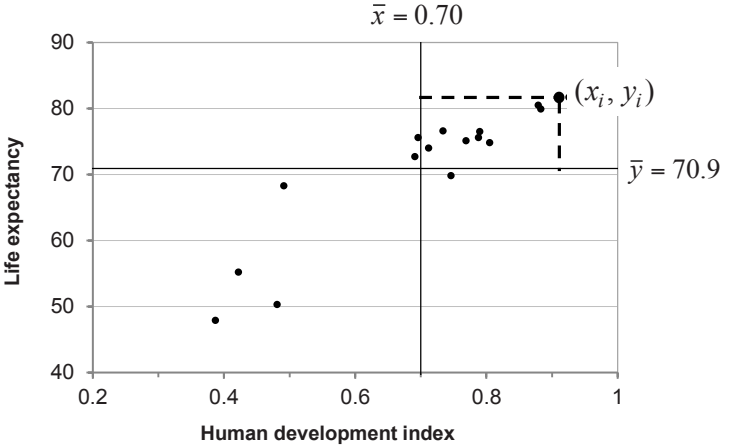


Figure 4.8. "Deviations" from a specific point to the centre of gravity (data from 16 countries)

From the covariance formula it is clear that the larger the deviations $(x_i - \bar{x})$ and $(y_i - \bar{y})$, the larger the value of covariance; hence, the covariance takes into account not only the number of points in each quadrant but also their (weighted) deviations to the centre of gravity.

A drawback of the covariance is that it is not bounded; therefore it is difficult to compare the covariance from different scatter plots to decide which association is stronger. This problem is solved by the *Pearson correlation coefficient r*, the most commonly used measure of the strength of linear relationships. This coefficient varies between −1 and +1 and takes the values +1 or −1 only if all the points have perfect linear dependence (i.e., they are all located on a straight line):[7]

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right).$$

The value of *r* has the same sign as the covariance since:

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

[7] The correlation coefficient is a standardised measure of association.

The values of s_{xy} and r are easily obtained by calculators or software; for example, spreadsheets may be used. The value of the Pearson coefficient r measures linear correlation; if the spread of the scatter plot is smaller, the relationship is stronger and r tends to −1 or +1; if the points are more widely scattered, r tends to 0. However, it is important to remark that it is possible find a small value of r in case of perfect association when the association is not linear (e.g., if the points lie on a parabola). In the following activities, students are to revise the values they previously assigned to each scatter plot in Figures 4.5 and 4.6.

Task 4.5. Identifying correlation values from scatter plots: In Table 4.8, some correlation coefficients are provided including all the values that correspond to the scatter plots presented in Figures 4.5 and 4.6. The students should find out, which coefficient corresponds to which scatter plot. The teacher can initiate discussions about possible strategies to solve the task:

a. Separate positive and negative correlations and independence;

b. Order the coefficients by size and the scatter plots by relative spread;

c. Finally use the spreadsheet to compute the correlation coefficient for each scatter plot and compare the obtained values to your earlier attributions.

Table 4.8. Correlation coefficients from different data sets

−0.40	1.00	0.78	0.91	−0.92	0.50
0.62	0.20	−0.73	0.00	0.38	0.61

Explaining the observed correlation

The previous explorations can be complemented with a discussion of the different situations that may lead to a correlation between two variables. The teacher can pose the following task to the students.

Task 4.6. Which variables in Figures 4.5 and 4.6 have a cause-effect relationship with life expectancy? Why? What other types of relationships different from cause and effect may lead to a correlation?

The initial discussion about various relationships is resumed here. The first approach was based only on students' previous expectations. Meanwhile, they have investigated the facts in form of scatter plots and have been introduced to various methods to measure the strength of co-relation. In the last task, they had to attribute various correlation coefficients to the corresponding scatter plots (without any calculations). After that quantitative analysis, the students are referred back to the context and should interpret the relations between the investigated variables to enhance their new knowledge.

It is essential that students realise that *a strong correlation does not yet imply causation* (i.e., one variable plays the role of the cause and the other that of the effect). Firstly, causation is an asymmetric relationship; we need to know the direction of the influence (which variable is the cause); while correlation is a *symmetric* relationship. For example, in Figure 4.5, we can think that a high infant mortality rate is a potential cause for low life expectancy as if many children die before reaching 5 years of age, the average time of life in this country automatically decreases.

Secondly, there are other types of relationships that explain correlation (see also Section 4.5.2). Students will recognise that some pairs of variables are interdependent, that means that one can think of both directions of influencing each other although the influence is not causal. Other pairs of variables co-vary due to third factors, which influence both. We have already mentioned the education index that shares the influence of a general standard of living with life expectancy. The students will also see that correlation is more useful for linear relations than for other functions that might describe the general trend of co-variation between two variables.

4.3.2. Fitting Models to Bivariate Data

Once a strong correlation is found, the interest is to fit a model to the data that serves to predict the *response* variable from the known values of the *explanatory* variable. We use simple functions for this purpose such as linear, exponential, or power functions. Of course, since association is only statistical, the *regression* model will not perfectly fit the data so that some differences will remain between the observed data and the prediction provided by the model.

Such a model is useful, for example, when trying to predict a future value given a (known) value of a variable, such as predicting the score of a student on a future test when the student's score in another test is known. Furthermore, the method of regression can be used to estimate the value of a variable, which is difficult to measure, by another variable that is easy to measure if a regression model is available that describes the statistical relationship between these variables. For example, assessing the state of an unborn baby from a physical measurement of the mother. The study of regression can start with the following task.

> *Task 4.7.* Look at each graph in Figures 4.5 and 4.6. Do you think it is possible to use a mathematical model (a function) to estimate the value of life expectancy given the value of the other (explanatory) variable? What type of function do you suggest?

The students should try to recognise the type of functions that would be useful models to fit to the data in the scatter plots represented in Figures 4.5 and 4.6. For the total population, no obvious trend is visible in the scatter plot and the correlation coefficient is nearly zero so that no function can be attached to describe the relationship. From the other scatter plots, the factors HDI, adolescent fertility rate, and education index have a distinct linear cloud of points and it is easy to fit a

straight line by eye even if the interpretation of a linear model for the fertility would raise some doubt from the context. We will resume the task of interpreting whether the model delivers an adequate contextual description of the relationship later.

To find other mathematical functions, the students have to recall their repertoire of functions like polynomial, exponential, or logarithmic functions. A spreadsheet like Excel offers a great help here as it lists various functions and draws the best-fitting function[8] (see below) of that type into the scatter plot so that the students can orientate themselves whether this delivers a good fit to the data or not.

Linear models: line of best fit

The first step in regression modelling is to decide about the *type* (family) of functions (e.g., linear functions) that should describe the relationship between the investigated variables. The second step is to select a specific function from that family that fits better than any other function from the chosen family. When the tendency in the scatter plot suggests that a straight line with the equation $Y = a + bX$ is an adequate model, the parameters *a* and *b* should be estimated from the data. In case a different function is preferable, for example, a parabola with the equation $Y = a + bX + cX^2$, three parameters *a, b,* and *c* have to be estimated from the data.[9]

The common method to estimate these parameters uses the *least-squares criterion*, which is based on making the sum of squares of the errors e_i of the predictions (usually called residuals) as small as possible. Note that the prediction error for the *i*th observation equals $e_i = y_i - (a + b\,x_i)$.[10] To obtain this line one has to find values *a* and *b* such that $\sum_{i=1}^{n} e_i^2$ is minimal. The solution is given by (see, e.g., Ross, 2010a):

$$b = \frac{s_{xy}}{s_x^2}, \quad a = \bar{y} - b\,\bar{x}\,.$$

The equation for the regression line can be transformed to the following formula (see below), from which an essential property is visible: the centre of gravity $(\bar{x}, \bar{y})$ lies on the regression line. The parameter *b* is called slope of the regression line and *a* is called *intercept*:

---

[8] A better model will use several explanatory variables to predict life expectancy; however, to introduce regression at high-school level, we use only one.

[9] Once we have decided the specific type of function to use, the regression method consists in finding the parameters for that type (family) of function.

[10] In fact, this optimisation problem can be solved for the linear model analytically. For other models there are no analytic solutions; only approximate solutions can be found with software.

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

If we replace the covariance in this formula by the correlation coefficient, we get another form of the equation in terms of standard scores:

$$\frac{y - \bar{y}}{s_y} \quad = \quad r \cdot \frac{x - \bar{x}}{s_x}.$$

This representation delivers a fascinating interpretation: The standard score of the y value of a statistical unit is predicted to be r times the standard score of its x value. If, e.g., $r = 0.5$ and the standard score of x equals 2, the predicted standard score of y is $r \cdot 2 = 1$.

In Figure 4.9, we add the best-fitting line and its equation to the scatter plot from Figure 4.7. This regression line implies that for each additional unit in the HDI, the increment in life expectancy is 57.8 years (rounded to one decimal). This line may be used to predict values, which are not in the plot: The predicted average life expectancy for a country, which HDI is 0.6 equals $y = 30.464 + 57.774 \cdot 0.6 = 65.1$.

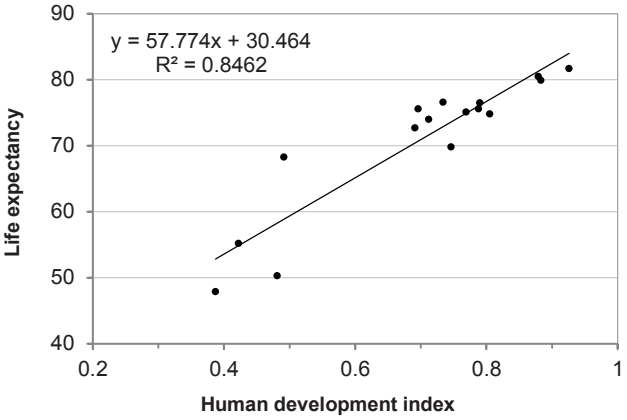


Figure 4.9. Regression of life expectancy versus HDI (16 countries)

Another remark is that the coefficients for the regression line change if more data are added and – as a result of considering more data – the model will be more reliable. Then, for the entire data set in Figure 4.10, the line is close (but different) to that one obtained with only 16 countries.

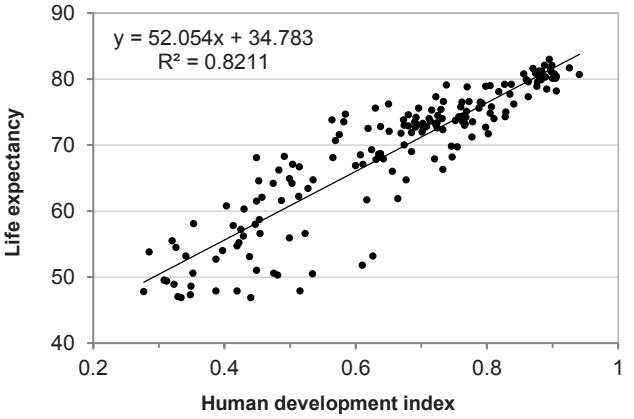


Figure 4.10. Regression of life expectancy versus HDI (entire data set)

Two regression lines

It is important to remark that the regression line expresses the mean values of Y for each fixed value of X. When the goal is to estimate the mean values of X for a fixed value of Y, then a different regression line has to be used (x and y interchange roles):

$$(x - \bar{x}) \;=\; (y - \bar{y})\; \frac{s_{xy}}{s_y^2}.$$

In general, both regression lines are different and coincide only in case of perfect linear relationship. In case of perfect independence, the lines are orthogonal and parallel to the axes. We can plot the data of life expectancy (LEX) and HDI, now with HDI as dependent variable on the y axis and determine the least-squares line. The result is displayed in Figure 4.11. The equation for the model is:

$$y = -0.3385 + 0.0146 \cdot x\,.$$

Note that we cannot get back the previous best line of predicting life expectancy from HDI. Neither by interchanging the variables to

$$x = -0.3385 + 0.0146 \cdot y$$

nor by solving the model equation for x (life expectancy):

$$x = \frac{y + 0.3385}{0.0146}.$$

This is the reason why some software denotes the model equation always with the variable names. The two best-fitting lines are then:

– For predicting LEX from HDI: $\text{LEX} = 30.464 + 57.774 \cdot \text{HDI}$.
– For predicting HDI from LEX: $\text{HDI} = -0.3385 + 0.0146 \cdot \text{LEX}$.

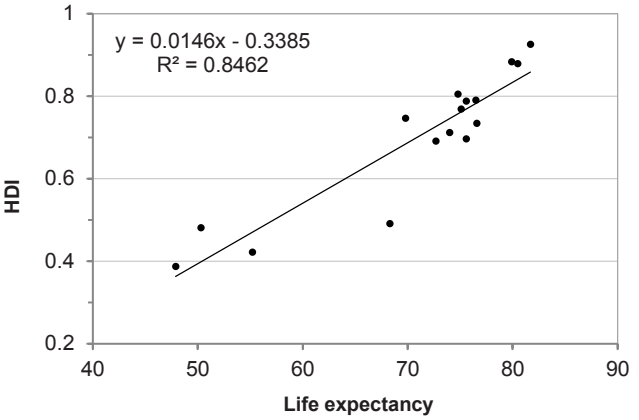


Figure 4.11. Regression of HDI versus life expectancy (16 countries)

Coefficient of determination

In Figures 4.9 and 4.10, the value of r^2 (r square)[11] is added to the equation of the regression line. The last step in the lesson is to analyse the meaning of this coefficient. We propose the following activity to the students.

> *Task 4.8.* Order the variables in the investigated scatter plots in Figures 4.5 and 4.6 according to your perception of their usefulness for predicting the value of life expectancy from least useful to most useful.

Intuitively, the smaller the spread in the scatter plot is around a trend, the more useful is the explanatory variable to predict life expectancy. For linear tendency it holds, the smaller the spread, the larger is the value of the correlation coefficient. To start the activity, the students can either consider the spread in the scatter plot or – if the trend is linear – the value of the correlation coefficient, which has been determined for each plot in the previous tasks. In fact, its square is called coefficient of determination:

[11] In the scatter plots, this is denoted as R^2 by the spreadsheet used.

$$D = r^2.$$

Given the properties of r, it is easy to understand that D varies between 0 (lack of linear correlation) and 1 (perfect linear correlation). Below, we will describe another meaning of D as percentage of explained variation that can give more meaning to the correlation coefficient and – at the same time – can also be applied to non-linear regression models.

The variance $s^2_{residual}$ of the residuals (i.e., the prediction errors) corresponds to the variation that remains after the regression model has been applied and thus reflects that variation that cannot be explained by the regression. The variance in Y reflects the total variation in the dependent variable. The difference $s_y^2 - s^2_{residual}$ (between initial and final variance) can be interpreted as reduction of variance by using regression, or variation explained by regression for which we denote $s^2_{\text{due to regression}}$ (Ross, 2010a). Graphically, this is mirrored by the smaller scatter of the residuals as compared to the general vertical width of the scatter plot. Thus, it is possible to decompose the variance of the response variable into two parts:[12]

$$s_y^2 = s^2_{\text{due to regression}} + s^2_{\text{residual}}.$$

We can divide this equation by the variance of Y and get:

$$1 = \frac{s^2_{\text{due to regression}}}{s_y^2} + \frac{s^2_{\text{residual}}}{s_y^2}.$$

That part of the total variance that is due to regression is called *coefficient of determination* and is usually expressed in percentages:

$$D = \frac{s^2_{\text{due to regression}}}{s_y^2}.$$

For example, the used regression model explains 60% of the total variation in the dependent variable if $D = 0.60$. An interpretation of the coefficient of determination as explained variance is enhanced by comparing a naïve procedure

[12] Statisticians use the corresponding sums of squares rather than "variances" as the sums of squares have to be divided by the so-called degrees of freedom rather than by n–1. It is a little algebra beyond high-school level to show that the variance due to regression equals the variance of the predicted values and can in fact be interpreted as variance. By (dynamic) numerical investigations in a spreadsheet, Borovcnik (2012b) shows that such a split of the total variance can be corroborated.

of estimating the value of the dependent variable (as the mean of the y values) and the procedure of estimating this value by a linear model (for details, see Borovcnik, 2006b and 2007b).

The coefficient of determination can be calculated for any regression function. For the linear model with predicted values of $a + b\,x_i$, it coincides with the square of the correlation coefficient so that it holds

$$\frac{s^2_{\text{due to regression}}}{s_y^2} = r^2 \text{ and } \frac{s^2_{\text{residual}}}{s_y^2} = 1 - r^2.$$

In the following two special cases, this is in fact true and "explains" the role of the two variance components:

a. For perfect linear relationship (i.e., $r^2 = 1$), all the variance in the data is explained by the regression model:

$$s_y^2 = s^2_{\text{due to regression}} \text{ and } s^2_{\text{residual}} = 0.$$

b. For perfect independence (i.e., $r^2 = 0$), all the variance in the data is due to randomness in the data (residual):

$$s_y^2 = s^2_{\text{residual}} \text{ and } s^2_{\text{due to regression}} = 0.$$

In the activity students order the variables according to the values of the coefficient of determination. The teacher discusses with them the meaning of this coefficient and lets them know how regression helps to decrease the unexplained variance in the data (since part of the variance is explained by the regression model). To improve the model (i.e., to decrease unexplained variance to an even larger extent), further factors should be incorporated in the regression equation. We mention multiple regression only as an appropriate method but it is not our goal to introduce it at high-school level.

4.4. ADDITIONAL ACTIVITIES

To complete the previous activities, other target variables can be investigated. Students could compare the set of influential factors to those we found for life expectancy. In this way, a network of statistically associated variables can be established that might describe general health, or the economic and environmental situation, and how they are interconnected. We will extend the previous analysis by more complicated functions to model the interrelations.

In Section 4.3, we only dealt with linear regression since only this type of regression is included in the middle and high-schools' official curricular documents. However, once the students are familiar with linear regression models, it is easy to intuitively introduce other types of regression, which are frequently used to describe social, economic, or biological phenomena. This extends the modelling repertoire and the relevance of the models as often linear models do not provide an adequate description of the interrelations between the variables involved.

Task 4.9. Try to improve the goodness of fit of the model for the relationship between gross national income per capita and life expectancy by choosing a function that fits better than the linear model (Figure 4.12). What can you observe? Judge the fit of the models by the coefficient of determination.

It is clear that there is a moderate relationship between these variables but the tendency in the scatter plot is clearly non-linear. Accordingly, the linear model explains only 36.9% of the variance of life expectancy.

If we change to a logarithmic model (Figure 4.13), the percentage of explained variance increases to 63.2%. The model equation $y = 6.2216 \cdot \ln x + 57.807$ is also easy to interpret: life expectancy equals a constant (57.807) plus 6.2216 multiplied by the logarithm of the gross national income per capita.

Numeric explorations (in a spread sheet) show an interesting pattern: if, e.g., the gross national income is doubled, the *additional* lifetime is always 4.31 years *regardless of the initial income*. This fact also can be used to discuss with students on the unequal distribution of wealth in different countries and clarify matters that wealth alone does not assure longer life or even improve the quality of life. Therefore, this statistical investigation may help reinforcing the social consciousness of students.

We could think of shifting a little part of the wealth from the richest countries (as measured by gross national income) to the poorest countries. The amount shifted marks a small percentage of the "donator" country but a great percentage in the "receiving" countries. This implies that the effect on life expectancy should be small in the donator country but high in the receiving countries. There, the money may be used to improve factors that influence life expectancy and – according to that relation (see the scatter plot with the logarithmic model) – it would have a large effect in these countries.

Now students may easily explore other types of simple mathematical functions they know from calculus to determine a model that fits the data better than a straight line. This exploration is facilitated by software; it can be automatically performed with a spreadsheet like Excel.

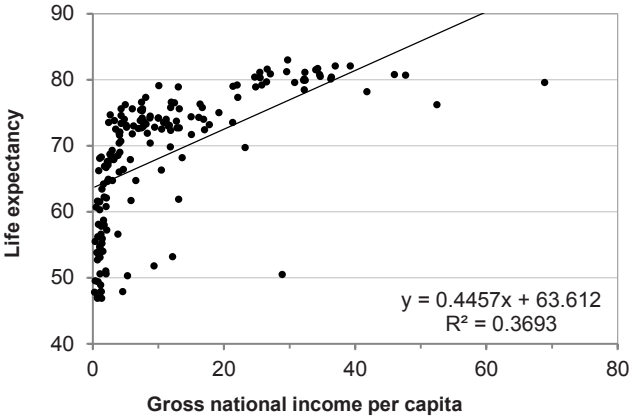


Figure 4.12. Life expectancy and gross national income–linear model

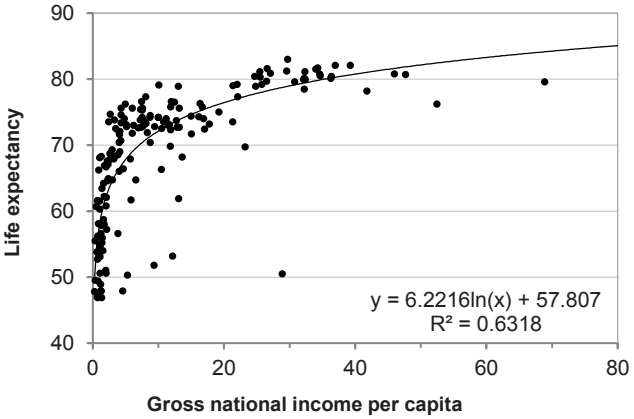


Figure 4.13. Life expectancy and gross national income–logarithmic model

4.5. SYNTHESIS OF LEARNING GOALS

The teaching activities described in Sections 4.2 to 4.4 may serve to introduce basic ideas about association in 2×2 contingency tables and correlation and regression in numerical variables at high school. As these two topics are so different, we synthesise our considerations about the learning goals separately. Below, we first summarise the purpose of the activities related to contingency tables and then expose the main goals involved in the activities linked to the study of correlation and regression.

4.5.1. Contingency Tables

A contingency table is used to summarise data when the statistical units are classified by two variables. We first intend the students to become familiar with the different frequencies that can be computed from the same data and simple methods to evaluate association between the variables represented in the rows and columns of such tables. In Table 4.9, a scheme is shown for the simplest case when both variables are binary.

Table 4.9. Scheme of a 2×2 table

	A	Not *A*	Total
B	*a*	*b*	*a+b*
Not *B*	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*a+b+c+d*

Different frequencies in a 2×2 table

In spite of its apparent simplicity, 2×2 tables are complex semiotic tools since there are many mathematical objects linked to these tables (Estrada, & Díaz, 2006). The absolute frequencies *a, b, c, d* in the cells equal the number of elements that fulfil a double condition (the specific values of the variables represented in that row and that column). Moreover, from a given cell, it is possible to compute different relative frequencies, depending on the total used in the computation; when this total is the sample size we obtain the joint relative frequency; if we used the row or column totals we obtain the row or column conditional relative frequencies. For example, from cell *a,* we can compute three different relative frequencies (or three different percentages if we multiply each of these relative frequencies by 100). We illustrate the various possibilities by the context and the data from Task 4.1 (Section 4.2.1).

– *Joint relative frequencies:* $\frac{a}{a+b+c+d}$; the joint relative frequency of siblings and being a problematic child is $\frac{40}{250}=0.160$.

– *Conditional relative frequencies as regards the row total*: $\frac{a}{a+b}$; the relative frequency of problematic children in families with siblings is $\frac{40}{140}=0.286$.

– *Conditional relative frequencies as regards the column total*: $\frac{a}{a+c}$; now, the relative frequency of children with siblings among all problematic children is $\frac{40}{120}=0.333$.

– *Marginal relative frequencies of rows and columns*: we can compute $\frac{a+b}{a+b+c+d}$ and $\frac{a+c}{a+b+c+d}$; that is, we have a relative frequency of children with siblings of $\frac{140}{250}=0.560$ and a relative frequency of problematic children of $\frac{120}{250}=0.480$ in the whole sample.

In order to make sense of these different frequencies, we introduced various diagrams such as the stacked bar chart and the mosaic plot.

Measuring association in a 2×2 table

In the activity included in this chapter, we used simple probabilistic reasoning to measure association such as comparison of probabilities, comparison of relative risks, and comparison of observed and expected frequencies in the table. For judging the association between the type of family and the occurrence of behavioural problems, we introduced the idea of expected numbers (of the cells) based on the additional assumption that the two variables are independent. The expected numbers (Table 4.6) and the observed numbers (the data in Table 4.1) were compared: as more problematic children were observed in single-child families (than expected) and less were observed in sibling families, this was interpreted as empirical evidence for a stronger association between the involved variables.

However, the strength of this association was not yet quantified as was done with metric variables by the correlation coefficient. For qualitative variables, analogue coefficients are available to measure the degree of association, which ideally vary between 0 and 1. Various methods of measuring association in these tables are discussed in Section 4.7.1.

4.5.2. Correlation and Regression

In Section 4.3, we used real data to help students to reason with evidence and multivariate data, where the relationships between the variables are not limited to linear regression. We also tried to provide them with the opportunity to observe different signs and strengths of correlation in real-world situations. We hope that this activity will help students to increase statistical literacy and support them to become informed citizens who would make better decisions in personal and public issues (Ridgway, 2015; Ridgway, Nicholson, & McCusker, 2006).

To identify the key factors that influence the difficulty of correlation tasks, the following criteria, suggested by Batanero, Gea, Díaz, and Cañadas (2014) and Estepa and Batanero (1996) were taken into account and investigated in the study (Table 4.10):

Table 4.10. Criteria used in selecting the independent variables

Variable / Index	*r* value	Fitting model	Explaining the relationship	Subjects' expectation
Human dev. index (HDI)	0.91	Linear	Inter dependence	Coincidence
GNP per capita	0.61	Logarithmic	Indirect dependence	Coincidence
Adolescent fertility rate	–0.73	Linear	Indirect dependence	No expectation
Under-five mortality	–0.92	Exponential	Cause-effect	Coincidence
Expenditure on health	0.38	Polynomial	Cause-effect	Weaker than expected
Education index	0.78	Linear	Indirect dependence	No expectation
Population total	0.00	Independence	Independence	No expectation
Population urban	0.62	No model	Indirect dependence	Contrary

- *Strength of correlation*: ranging from very strong correlation to independence. Correlation is more easily perceived in case of a strong positive association; however, in social sciences it is common to find weaker and also negative correlations.
- *Sign of correlation:* including both positive and negative correlations, which are harder to be perceived by the students according to Estepa and Batanero (1996).
- *Type of model that is fitted to the data:* including linear and non-linear models. Since high-school students have studied simple functions such as polynomial, exponential, or logarithmic functions, this activity also helps them to identify examples, in which these functions are used in a real-life context.
- *Explanation of correlation:* we used relationships that may be explained by cause-and effect, as well as interdependence or indirect dependence. The goal is that students learn to clearly separate between correlation and causation. They should also learn to deal with negative correlations and link these to decreasing linear trends.
- *Agreement between students' previous expectation and correlation in the data:* students should learn to read the facts without mingling them with their own expectations that are based on their previous knowledge, which may be anecdotal and thus not generalisable.

Starting from an intuitive activity where students are asked to informally evaluate the strength and the sign of the dependence between life expectancy and several independent variables, the goal is to introduce various concepts and procedures progressively as described in the following sections.

Intuitive interpretation of scatter plots

The study of the topic should start with students' exploration of the graphical representation of bivariate data by a scatter plot so that they can make sense of bivariate distributions. Using scatter plots, they can learn to interpret spread and tendency in the data, visualise differences between functional and random dependence (including strong or weak dependence and independence), and understand the idea of centre of gravity in bivariate distributions.

A first intuitive measure of association in numerical data is the Quadrant-count ratio (QCR), introduced by Holmes (2001); it is a summary statistics similar to the one proposed by Inhelder and Piaget (1955) to quantify association in 2×2 contingency tables (see Section 4.6.1). Like the measure proposed by Piaget, QCR is unreliable as it may provide the minimum or maximum values 0 and 1 in cases of imperfect independence or imperfect association. Moreover, with the same number of points in the different quadrants we obtain the same value for QCR, no matter what the spread and the shape in the scatter plot actually is; this means that QCR is not appropriate to measure the strength of the relationship.

This may also be seen from the fact that the counts in the quadrants correspond to an over-simplification of the values of the variables to the circumstance whether they are above or below the mean value of the related variable. That means that, instead of measuring the value x_i, we attribute a *u* if $x_i > \bar{x}$ and a *v* if $x_i \leq \bar{x}$. For the *y* variable, we rescale analogously. From this simple rescaling[13], we can expect that a measure of association derived from the transformed data can only reflect a rough picture of the co-relation between the involved variables.

Covariance and correlation

Once the students have informally explored sign and strength of dependence by the inspection of scatter plots and by using QCR, we introduce the covariance and the correlation coefficient. The students should be able to interpret the meaning of these statistics, their variation for different types of association (strong, weak, independence, direct, inverse), and the relation between covariance and correlation coefficient.

In the formula for the correlation coefficient, the differences of the *x* values from their mean are weighted by the inverse of the standard deviation s_x; the same procedure is applied to the *y* values. This method may appear "strange" for the student but it yields a statistic that varies exactly between −1 and 1, which makes it easier to interpret its values.

Furthermore, it is important that students realise that association does not automatically imply causation and that we can find other types of relationships that lead to a correlation of the involved variables (Batanero, Gea, Díaz, & Cañadas, 2014):

[13] The values *u* and *v* can be arbitrarily chosen if only $u < v$. That implies that we rescale the metric variable to a dichotomous one that is measured on an ordinal scale.

– Diagnostic relationships (see Chapter 3). If a specific symptom points into the direction of a disease, it is likely to have this disease in case that the symptom occurs; however, the disease is not caused by this symptom.
– The correlation between two variables X and Y may be generated by a common cause Z. For example, the percentage of urban population and life expectancy are positively correlated; these two variables co-vary due to underlying variables; e.g., better life conditions in urban than in rural locations will increase life expectancy in urban regions and thus countries with more urban population will have a higher life expectancy.
– There might be no connections between two variables and still we find a (spurious) correlation. There are many examples of this effect; one could look for any two variables changing in time. The two variables are then correlated – but without any reason.

Fitting models to bivariate data

When the association between the variables is strong, there is an interest in finding a model that serves to predict new data. Students are introduced to linear regression and to the computation and interpretation of the regression lines and their parameters. They should learn to identify a linear tendency in a scatter plot as well as identify other simple functions that fit the data when the tendency is not linear. With the help of calculators they can determine the equation of the regression line and interpret its slope and intercept.

It is essential that students understand the difference between the model (regression line) and the data (observations) and appreciate that models are useful for predicting new data when direct measurement is difficult. The concept of residuals helps to perceive this difference; at the same time it enhances understanding how the model explains part of the data's inherent variability.

For the statistical object i with data point (x_i, y_i), the actual value of the dependent variable y_i is compared to the prediction $a + b\,x_i$ based on the linear model. The model predicts points with coordinates $(x_i, a + b\,x_i)$ instead of the actual data. The error (residual) is the difference of these two expressions. Geometrically, these errors are the vertical differences between the data points and the regression line (see Figure 4.14).

Two different regression lines. Another goal of teaching regression is that students understand the asymmetric roles of the variables in the model. While for correlation both variables play a symmetric role, in regression, one of them is the response variable and the other is the predictor variable. Hence, there are two different regression lines that coincide only in case of perfect linear regression.

Goodness of fit. The coefficient of determination reinforces the meaning of residuals and serves to introduce the idea of goodness of a model. The difference between initial and final variance can be interpreted as reduction of variance by using regression, or variation explained by regression. This explained variation as percentage of the variance of the response variable gives the coefficient of

determination. In case of linear regression, the value of this coefficient coincides with the square of the correlation coefficient.

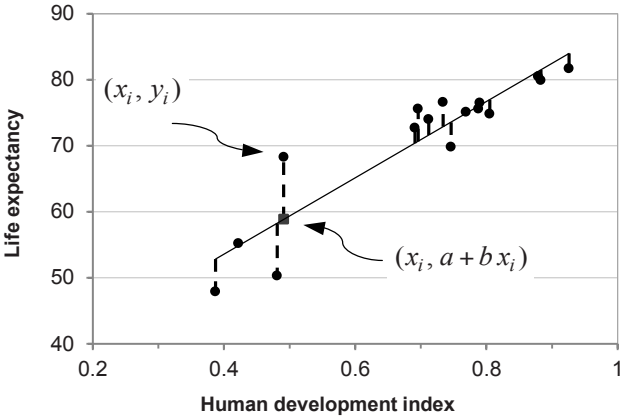


Figure 4.14. Residuals from each point to the regression line

4.6. SOME POTENTIAL DIFFICULTIES

Although there is little research carried out by mathematics educators into students' interpretation of contingency tables as well as correlation and regression, we find abundant related literature in the area of psychology, where it is connected to the problem of making decisions in uncertain environments (Scholz, 1983; 1991). The interest of these studies is to analyse the way in which human beings make decisions because of the profound implications in areas such as medical diagnosis, economy, and law. Below we first summarise research carried out about understanding contingency tables. We then give a brief review on research on correlation in numerical data.

4.6.1. Intuitive Strategies in Contingency Tables

In our analysis of research on dealing with contingency tables we use and expand some surveys included in papers such as Batanero, Cañadas, Díaz, and Gea (2015), Batanero, Estepa, Godino, and Green (1996), Engel and Sedlmeier (2011), or Pérez-Echeverría (1990).

The seminal work on this topic is due to Inhelder and Piaget (1955), who conceived the evolution of the concepts of association and probability as related and viewed association as the last step in the development of probabilistic reasoning. Consistently, they only investigated adolescents in their formal operation stages (13–14 year old) using interviews. The students had to judge the

association between eye and hair colour (simplified to blue and dark eyes and fair and brown hair) using data that were structurally equivalent to Table 4.9. According to these authors, to understand the association[14] between two variables in such a data format, it is essential to recognise that the cases in cells *a* and *d* support a direct (positive) association and the cases in cells *b* and *c* contribute to an indirect (negative) association.

Piaget and Inhelder found out that in a first stage of reasoning, students only analyse the relation between the positive cases (cell *a* in Table 4.9) and the total number in the sample. In Task 4.1, they would consider only the 80 cases with both characters (being only child and being problematic) and compare it to the total of 250 so that they would judge the association between behaviour and type of family as small.

In a second stage, adolescents compare *a* with *b* or *a* with *c*. In Task 4.1, they would "see" 50 more problematic than unproblematic children in the group of only children; or, 40 only children more than children with siblings in the group of problematic children. Thus, they would conclude that the variables are associated. However, this strategy works only in particular cases.

Inhelder and Piaget suggested that, although these adolescents can compute single probabilities, understanding association requires considering the frequencies $(a+d)$ as favourable to a positive association and the frequencies $(b+c)$ as favourable to a negative association and that students need to consider the relation:

$$R_P = \frac{(a+d)-(b+c)}{a+b+c+d}$$ [15]

where R_P represents the ratio of the difference between the number of cases confirming a positive and the cases confirming a negative association and the number of all cases involved. We observe that $R_P = 1$ when all the data fall into cells *a* and *d*; $R_P = -1$ when all the data fall into cells *b* and *c* and $R_P = 0$ when $(a+d)=(b+c)$ so that R_P is an intuitive measure of association in a 2×2 table. According to Piaget and Inhelder, recognition of these properties does not happen before 15 years of age.

After this pioneering study, many psychologists studied the perception of association in 2×2 tables with adults; this research showed poor reasoning with different tasks and all types of subjects. It is very common to base the association

[14] Inhelder and Piaget used the word "correlation". We denote relationships between qualitative variables as "association" and use correlation only for metric variables.

[15] Positive or negative association makes sense only for ordinal variables; yet, it is used in this context. How difficult it is to judge the strength of the association, may be seen from the complexity of formulas for coefficients of association. In terms of the data format in Table 4.9, the phi coefficient is given by $\varphi = (a\,d - b\,c)/\sqrt{(a+b)(c+d)(a+c)(b+d)}$.

judgment only on the frequency in cell *a* or only on a comparison of the frequencies in cells *a* and *b*. Jenkins and Ward (1965) pointed out that even the strategy of comparing the probabilities of the table diagonals $(a+d)$ and $(b+c)$ (computing R_P) considered correct by Piaget and Inhelder is not always valid; if the difference in the totals in the rows or columns in the table is large, this short-cut strategy can produce errors. Nevertheless, it is widely used by adults (see Allan & Jenkins, 1983).

Pérez-Echeverría (1990) classified the identified strategies in judging association by levels of complexity: In Levels 0 to 3, people only use data in 0 to 3 different cells; in Level 4, subjects base their judgment on additive comparisons of the data in all four cells. In Level 5, subjects base their judgment on multiplicative comparisons between the four cells. Level 5 strategies are the only correct strategies in general; although in some particular tables, the subjects can succeed with lower-level strategies.

Generally, subjects are not aware of the full complexity of a task and use naïve strategies instead. This might be sufficient in some problems as they lead to an approximate answer (e.g., in Batanero, Godino, & Estepa, 1998; Cañadas, Díaz, Batanero, & Estepa, 2013). From Section 4.2, we see that the correct strategy requires comparing conditional probabilities; for example, comparing the probability of *B* in both groups of people with and without *A* or comparing relative risks.

4.6.2. Linear Regression and Correlation

The study of relationships between two metric variables includes two different problems: correlation and regression. For correlation, the two variables play a symmetric role. The aim is to determine whether they co-vary or not, to find out whether this occurs in the same direction (the higher *x*, the higher *y*; positive correlation) or in the opposite direction (negative correlation), and to quantify the strength of this association.

Fitting a model to data

If a relatively strong association is perceived, the next task is to find a function $y = f(x)$ (in the simplest case a regression line) that may be used to predict the values of *y* from the values of *x* or describe the general shape of the relation between the variables. Since the dependence is not functional, only the mean value of *y* for each given *x* (the mean of the conditional distribution of *y* given *x*) is predicted. This problem does not have a unique solution and a sequence of choices needs to be made to approach it:

– *Choosing a model for the relation between the variables*. Which *family of functions* will be used (e.g., whether linear or exponential functions, etc. are appropriate) to select the regression function? This decision is based on previous knowledge about the phenomenon under study as well as on the result from the inspection of the scatter plot.

– *Selecting a decision criterion to find a best-fitting function*. After we decide about the type of model, for example, a linear model, different criteria are applicable to decide what is meant by best fit of a line, e.g., the least-squares criterion or regression to the median (Tukey line). It is important to assure that students understand the chosen criterion; this understanding is needed to interpret the final approximation provided by the regression method, as well as the fit of the used model to the data.[16]
– *Interpreting the best-fitting function from statistics and from context.* When the regression model has been determined, it is still possible to make mistakes in its interpretation or in its use to make predictions. The best-fitting function may have a small coefficient of determination and prediction based on it has a large error. The model may describe the relationship only within the range of data for the independent variable. It is vital to interpret the model within the context and understand its implications on further tasks within the context.

Some pitfalls

The systematic investigations of scatter plots can reveal many more pitfalls (see Borovcnik, 2012b, and Engel & Sedlmeier, 2011). Some are induced by the technique applied, which has hidden features that are not easy to recognise:

1. It is difficult to estimate (judge) the size of the correlation coefficient from scatter plots in some circumstances.[17] Rescaling of the axes can produce *any* impression about the correlation coefficient involved; therefore, this coefficient should always be computed in addition to displaying scatter plots. To calibrate the visual impression, it is advisable to scale the axes in such a way that the final cloud of points approximately lies within a square.

2. If individual statistical units are grouped (aggregated), then correlation increases solely because of the grouping process. For example, if data on a district level are grouped over the districts of a county, a large part of variation between the districts is lost and a model fits much better to the aggregated data on the county level than it fits to the local data.

3. The phenomenon of "regression towards the mean" is famous but is only an artefact, i.e., it is solely due to the method and not due to relations in the investigated context. It occurs when independent and dependent variables represent the same measurement at different times or for linked statistical units.

[16] A main issue in statistics is to develop criteria for the fit of models to data. This task requires deliberate consideration of the *statistical* properties of the resulting estimators.

[17] Sánchez-Cobo, Estepa, and Batanero (2000) described variables that affect this estimation such as the sign or strength of correlation. Estimations are relatively good in tasks that involve linear, strong, and direct correlation, and poorer if inverse, non-linear, or moderate correlation is involved. In case, people's expectations about the correlation do not coincide with the correlation in the data, the estimation is poor.

Historically, Galton and Pearson investigated the height of fathers and sons[18] (see Freedman, Pisani, & Purves, 1978). If the father is two standard deviations taller than average, then the son is predicted to be above average, too. Yet, the prediction for the son's height is, according to the usual linear regression, r times two standard deviations above mean, which is less extreme than the father. In fact, r coincides with the correlation coefficient in modern terminology; at that time it was named coefficient of reversion to the mean, or regression[19] coefficient. This historic confusion has given name to the procedure developed – the method of regression.

4. Another incidence of the phenomenon is when "measurements" are repeated independently. The second measurement tends to be less extreme than the first. If the measurements relate to sports, then undue interpretation of that tendency to fall back to the mean is often observed.

5. Some students do not distinguish the explanatory from the response variable and use the same regression line for prediction, no matter which variable should be predicted (Estepa, Sánchez-Cobo, & Batanero, 1999).

Interpreting the value of a correlation on a ratio scale

From early childhood on, students have learned to compare on proportional scales. No wonder that students tend to interpret the size of the correlation coefficient linearly, like fractions of unity.

Many students interpret a correlation coefficient of 0.40 as double the strength of a correlation of 0.20 although they cannot be compared in this way. A coefficient of 0.40 is larger than 0.20 but not double the value. As discussed in Section 4.3.2, a correlation coefficient of 0.40 gives rise to an explained variation of 16% as compared to 4% with $r = 0.20$. Explained variance in fact is a fraction, namely that part of the total variance that is explained by linear regression. An r^2 of 0.40 is double the value of an r^2 of 0.20 as 40% rather than 20% of the variation of the dependent variable is explained by the model. However, the connection between correlation coefficient and its square is intuitively not well controlled. High correlation coefficients do not mean high explained variation: A correlation coefficient of 0.80 only explains 64% of the total variation; a correlation coefficient of 0.50 (still much larger than many in empirical investigations) even much less, namely, 25% of the total variation.

The relation between correlation coefficients and the length of prediction intervals for the dependent variable is even more complicated. By linear regression,

[18] Galton and Pearson had a substantial interest to "prove" that intelligence is hereditary. As intelligence measurement had not yet been developed, they had to investigate variables that are easier to measure like height of people, or diameter of peas. The socially induced construction of the methods of regression and correlation is well-described in MacKenzie (1981).

[19] In the sense of the Latin verb *regredior*, which means to go back, to turn back, and to return: to return back to the mean by the factor *r*.

the prediction intervals are shorter by the factor $\sqrt{1-r^2}$, which is the square root of unexplained variation (see Section 4.3.2). People have no intuitive access to the fact that to halve the prediction intervals by regression (as compared to not using regression), it is necessary that the correlation coefficient has a size of 0.8660. To improve the length of prediction intervals by 10%, it still takes a correlation coefficient of 0.4359; to increase the improvement to 20% (doubling it) the correlation coefficient has to increase to 0.6000.

Such mathematical interrelations are not learned in the statistics class and it is no wonder that even researchers familiar with regression and correlation by far underestimate how big the correlation coefficient has to be so that some desirable effect is achieved, either in percentage of explained variation or in reducing the length of prediction intervals.

4.6.3. Misconceptions Related to both Association and Correlation

In addition to the above mentioned difficulties and incorrect strategies that are linked to either association or correlation, research has reported biases that affect subjects in similar ways when they deal with these problems.

Simpson's paradox

When interpreting an association or a correlation between variables, it is important to control other variables that may affect the relationship under study. Simpson (1951) described an effect, where a tendency that appears in different subgroups, may dissolve or even change the direction when the subgroups are combined.

> *Task 4.10.* Let us consider two hospitals A and B and a specific surgery that bears a great inherent risk to die from it. The survival data are in Table 4.11. Determine the risk of not surviving the surgery in the two hospitals. Which hospital would you prefer for that surgery? Consider also the survival statistic of the two hospitals, which discriminates the cases according to the health status of the patients (Table 4.12). What can you tell about the risks now?

Apparently, it is preferable to undergo the surgery in Hospital B since there the rate of failure is only 2% ($= \frac{16}{800}$) as compared to 3% ($= \frac{63}{2100}$) in Hospital A. If the patient has initially good health, however, Hospital A is preferable (only 1% $= \frac{6}{600}$ failures as compared to 1.3% $= \frac{8}{600}$ in Hospital B). The same happens when the person has initially poor health (3.8% $= \frac{57}{1500}$ failures in Hospital A and 4% $= \frac{8}{200}$ in Hospital B).

Initially, Hospital B is better. When patients are classified according to their initial state of health, however, the situation changes and the *tendency* is reversed and Hospital A is always better. The paradox is solved when we observe that the

majority of people with poor health have taken the surgery in Hospital A, which explains the worse overall failure rate in Hospital A.

Table 4.11. Survival of a surgery in two hospitals

Survival	Hospital A	Hospital B
Do not survive	63	16
Survive	2037	784
Total	2100	800

Table 4.12. Survival of a surgery according to hospital and initial health

	Good health		Poor health	
Surgery	Hospital A	Hospital B	Hospital A	Hospital B
Do not survive	6	8	57	8
Survive	594	592	1443	192
Total	600	600	1500	200

A Simpson effect is also possible for metric variables. An example is given by Krämer (2007). An investigation in the first income at workplace after graduation shows a strong relation between the length of studies (in semesters) and the income (see Figure 4.15): The scatter plot reveals a linear increase in income with longer studies ($r = 0.61$). Should one give the advice to study longer? A closer look at the data reveals that there is a third variable in the background: the data falls apart into three separate groups, namely in business and administration, physics, and chemistry graduates.

Within these studies, data suggests that those who study longer will have less chance for a good salary. However, chemistry is a longer study linked to higher initial incomes while business studies are shorter but linked to lower initial incomes. That both phenomena can go together seems paradoxical to many people. Factors that blur, change or even reverse the direction of a relation between variables are called third factors.[20] Most people find it very unintuitive that third variables can influence the relation of two variables under scrutiny. They think if two variables are linked they are linked in the same way under all circumstances.

[20] If no data on such third variables have been collected (due to failures in the early discussion on the problem), such an effect cannot be traced back and an assumption on possible effects has to remain a speculation.

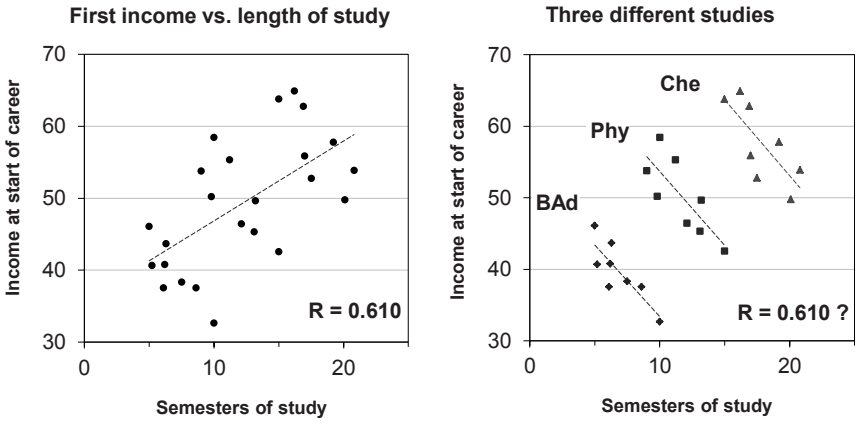


Figure 4.15. Left: A positive relation between first income and study length. Right: Within studies the relation is reversed (negative)

Illusory correlation

The most widely diffused bias was firstly described by Chapman and Chapman (1969, pp. 151) who used the term "illusory correlation":

> The report by observers of a correlation between two classes of events which, in reality, (a) are not correlated, (b) are correlated to a lesser extent than reported, or (c) are correlated in the opposite direction from that which is reported.

Many researchers have replicated this bias, which disturbs the estimates of association (e.g., Wright & Murphy, 1984; Meiser & Hewstone, 2006). It has been found with both qualitative and metric variables.

Other misconceptions

Batanero, Estepa, Godino, and Green (1996) analysed the performance of 213 17 year-old high-school students and their strategies in judging association. They observed a wide-spread lack of correspondence between strategy and association judgement. For example, a correct strategy was followed by an incorrect strategy in some tables. The authors analysed the arguments of the students and the frequencies in the table and suggested that these students showed different misconceptions of association.

In another study, Batanero, Estepa, and Godino (1997) observed that these misconceptions were partly resistant to instruction and that they also affect the judgement of correlation of metric variables. They discriminated the following idiosyncratic approaches:

– *Causal conception* according to which the subject considers association between variables only when it can be explained by the presence of a cause-effect relationship;
– *Unidirectional conception,* by which the student does not accept a negative association;
– *Local conception*, where the association is deduced from only a part of the data. This misconception underlies the incorrect strategies based on the use of only one cell or only one row when judging association.

In particular, the causal conception is related to illusory correlation and is induced by what people expect from the method or from the context of the data. Many persons over-interpret the size and the relevance of correlation coefficients if the relation is perceived as causal. Galton and Pearson are a good example of such wishful thinking as they found stable correlation coefficients around 0.50 (for similar variables as the height of fathers and sons) as "proof" for a hereditary (causal) law. On the other hand, an analysis of correlation and regression may be completely ignored if the relation seems far from being causal.

In the following example, we clearly know that "the stork does not bring the babies", which may – at first sight – be concluded from a correlation coefficient of 0.9832 (Figure 4.16; see Sies, 1988). Other times, researchers are erroneously induced to reflect about potential causal interrelations between the variables under scrutiny if a correlation is somewhat higher. Thus, while higher correlation coefficients are a good filter on potential relations between variables, there still remains the task of embedding the statistical result into the context and interpret its value accordingly.

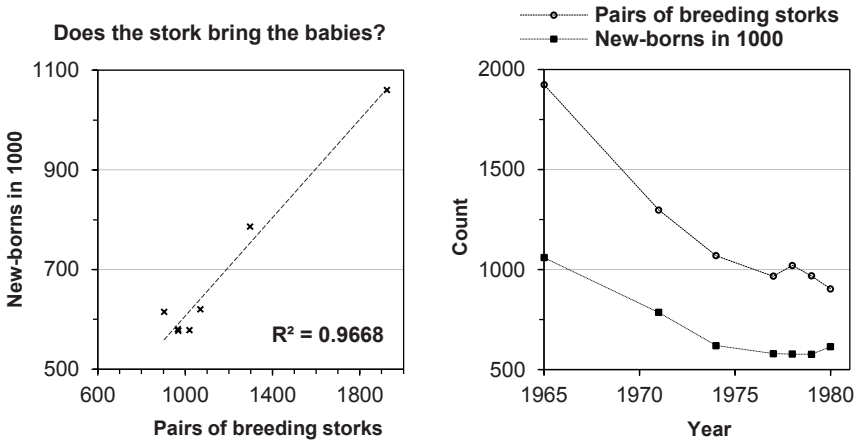


Figure 4.16. Left: A strong influence of storks to birth rates? Right: Two time-related variables are both decreasing

4.7. ADDITIONAL RESOURCES AND IDEAS

In this section, we complement the discussion about measuring the strength of association by introducing various statistics that quantify the degree of association also for qualitative variables. The measures follow the same idea: to have a normalised statistic (varying now from 0 to 1) to reflect the degree of association. Then we pursue the idea of studying more than one predictor variable by a visual approach rather than by multiple regression. This approach heavily draws on the new possibilities arising from data visualisation based on modern computer technology.

4.7.1. Measures of Association in Contingency Tables

In our study of contingency tables only simple probabilistic methods were used to informally evaluate the strength of association. At a more advanced level, formal association measures can be introduced.[21] The basic chi-square statistic can easily be explained as "deviation" between data and model by denoting the entries of the observed data as O_i and the corresponding entries in the table with expected numbers as E_i:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} .$$

The expected numbers E_i are calculated under the assumption of independence of the two variables investigated.[22] The problem with the resulting chi-square statistic is that it is not easy to interpret its value. There is no upper bound on the values and both the dimension of contingency tables and the total number n of statistical units tend to increase the chi-square statistic. Consistently, two different chi-square values can only be compared if the tables have the same number of rows and columns and if the sample size is equal.

Therefore, this statistic cannot serve to measure the strength of the association between the involved variables. However, the chi-square statistic can be used to test whether the observed data are reasonable given the underlying independence assumption, using a significance test (see Chapter 5). If for a 2×2 table the statistics

[21] In an introductory statistics course for psychology or business and administration students, various measures of association (and correlation) have to be introduced. It is essential that the students understand such measures and can apply them in context. These students are quite comparable to high-school students in their capacity and interest in mathematics. For ways to introduce such concepts and special strategies to make them meaningful to the students, see Borovcnik (2013a) and Cañadas (2012).

[22] Independence of variables can be investigated by checking whether all conditional row distributions are identical to the marginal distribution of the Y variable so that the entry of the row (i.e., the value of the X variable that defines the subgroup of that row) has no influence. Independence can also be checked probabilistically (see, e.g., Ross, 2010b).

is larger than 3.84, then we could "reject" the independence assumption (see, Ross, 2010b). In such a case, we could assume that the two variables are associated. Yet we would not know anything about the strength of the association.

Our intuitive comparison of single entries in the table of expected and observed frequencies, and further on, our intuitive judgement of the entries of the squared differences divided by the expected numbers (i.e., $\frac{(O_i-E_i)^2}{E_i}$) – as we have done in Section 4.2.4 – would yield hints about why the two variables are associated (because we can identify those cells with the largest entry) but still give no information about the strength of the overall association between the variables.

However, the chi-square statistic can be transformed to serve the purpose of measuring the strength of the association. To force the statistic into the desirable range between 0 and 1, the following coefficients are used. The phi coefficient is used for 2×2 tables (with n as the number of data):

$$\varphi = \sqrt{\frac{\chi^2}{n}} .$$

Cramer's contingency coefficient is used for general $r \times c$ tables (with r rows and c columns and k as the smaller number of r and c):

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} .$$

For the association between type of family and the occurrence of problematic behaviour, the phi coefficient (and V) yields 0.4387, which corresponds to a moderate association.

However, as discussed in Section 4.6.3 with the interpretation of the usual correlation coefficient, the new scale is *not linear* and one cannot state that an association coefficient of 0.40 indicates an association that is twice as "strong" as one with an association coefficient of 0.20.

In-line with the coefficient of determination (for correlation) where a correlation coefficient of 0.70 yields 49% explained variation, a rule of thumb states that association coefficients larger than 0.70 indicate a strong association. For reasons of symmetry, values smaller than 0.30 indicate a weak association.

4.7.2. Introduction to Multivariate Visualisation

Although in previous sections we only dealt with one influence factor (bivariate regression and correlation), it is clear that the context analysed in our examples implies the consideration of further influence variables and a more complex relationship. A current recommendation is, using the facilities provided by technology and the open-data movement, within which many organisations make

high-quality data available to citizens in general and to students and teachers in particular (Ridgway, 2015). An adequate analysis of these data would include multivariate statistical techniques that are out of the reach of high-school students.

It is possible, however, for the students to explore some of these complex relationships and grasp some idea of what multivariate relationships mean using visualisation resources. For example, in the UN Public Data Explorer facility (hdr.undp.org/en/data-explorer), the student can simultaneously represent four different variables in our file (or other statistical variables taken from the UN web server).[23]

> *Task 4.11.* Investigate the type of relations between fertility and life expectancy and, how this relation changed over time (1970–2013; see Figures 4.17 and 4.18). Check, whether specific countries (e.g., Israel, or India) follow the general trend or, in which direction they deviate from it. Do the various regions (represented by colour) follow the general pattern or do they show remarkably different patterns concerning the relation between fertility and life expectancy in this time interval? What about the additional influence of region and size of population on the fertility rate?

In Figure 4.17, for example, we represent the birth rate (Y axis) as a function of life expectancy (X axis). The tendency suggests a negative non-linear relationship: countries with larger life expectancy tend to have lower birth rates. Since the size of the bubbles represents the size of the population, we also see that the population size has some influence on the birth rate but not as high as for life expectancy. The bubble colour (region) suggests that Europe and Central Asia have lower birth rates.

The platform offers to change the time dynamically so that the graphical display of the scatter plot evolves like an animated film. This facilitates to study the interrelations between the variables along the time axis. We can also observe the changes in the variables in the whole distribution of countries or in specific countries. So, in comparing Figures 4.17 and 4.18, we observe a general improvement of life expectancy with time. In the specific case of Israel, the life expectancy increased by even 10 years since 1970. Fertility generally has dropped in this time period with the exception of Sub-Saharan Africa. The bubbles representing countries have moved to the right and downwards indicating that the trend was towards smaller fertility rates and higher life expectancy. The population (represented by the size of the bubbles) has increased nearly in each country, in all regions.

The relation between fertility and life expectancy shows a markedly concave decreasing curve in 1970 and has attained a shape close to a decreasing line (slightly convex) in 2013. The different type of function that describes the interrelation between fertility and life expectancy yields a picture of the joint development during the time span investigated.

[23] Similar resources are available from Gapminder World (www.gapminder.org/).

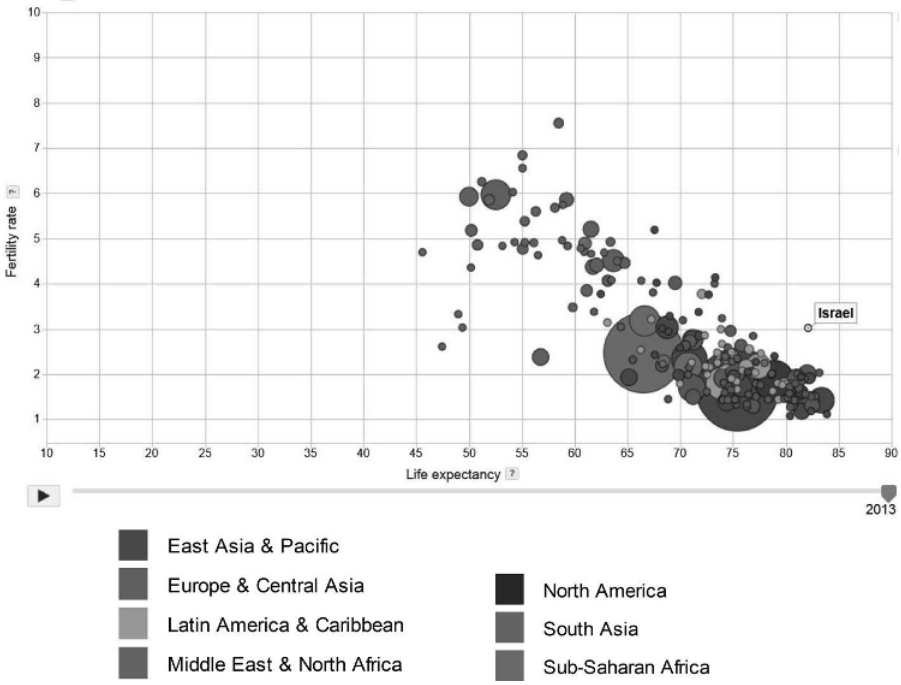


Figure 4.17. Birth rate as a "function" of life expectancy, region, and size of population for the year 2013

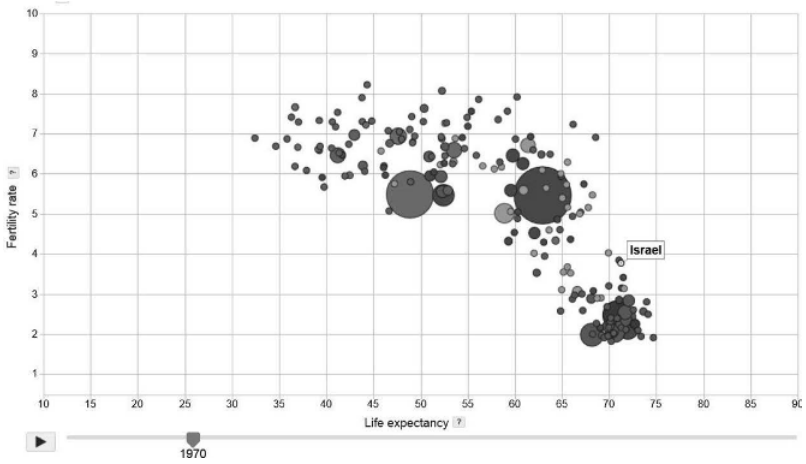


Figure 4.18. Birth rate as a "function" of life expectancy, region, and size of population for the year 1970

The 5 years increase in life expectancy is associated with a decrease of 1 in fertility rate in 2013; in 1970, however, there was nearly no decrease in the fertility rate in countries with smaller values of life expectancy whereas countries with higher expectancy (right part of the bubble plot) were associated with an even higher decrease of fertility. This relation differs hugely within the regions. For example, in Sub-Saharan Africa there is nearly no correlation between life expectancy and fertility whereas in East Asia and South-Asia higher life expectancy is markedly (negatively) correlated with the fertility rate.

One could also highlight other countries such as India to see that the increase of life expectancy is even larger; this increase, however, relates to a much lower state of average life in 1970 and thus reflects an increase that seems "easier to achieve" (e.g., by eliminating the causes for infant mortality). One could also insert a benchmark for life expectancy or for fertility; it enhances the development greatly if one inserts a line marking the average life time in both investigated years. It gets much more clearly visible how the different countries (or regions, which are marked by colour) develop in-line with the general trend of increase in average lifetime, or whether they are developing faster or slower. For that purpose it is also helpful to align the diagrams for the two years that are compared.

The same analysis applies also for the development of fertility. Once, both benchmarks are added to the graph, the co-development of both fertility and life expectancy can be studied in the time span investigated. This investigation may also be enriched by relating the co-variation to the size of population or the region (which is coloured). There is no limit to creatively using this tool and many questions from the context will arise for the students interactively. Some may have a clear answer, others will initiate further studies.

CHAPTER 5

SAMPLING AND INFERENCE

In this chapter, we analyse the main ideas of statistical inference that are relevant for high-school curricula. Since these ideas are sophisticated, we suggest elementary ways to introduce statistical inference in high school. We first present an example that serves to introduce different approaches to inference for a population proportion, including statistical tests and confidence intervals. We then suggest additional activities to explore the Central Limit Theorem as well as hypothesis testing as decision making, and include a synthesis of the main learning goals. Furthermore, we summarise the research on students' difficulties with sampling and finally describe additional teaching activities and ideas that may help students to develop inferential reasoning.

5.1. INTRODUCTION

Statistical inference is widely used in sciences, business, and many other branches of human activities. According to Hald (2008), Bayes and Laplace independently solved what is known today as the problem of inverse probability between 1764 and 1786. This problem consists of updating the (posterior) distribution for a parameter such as the population mean or proportion given a sample of observed data and the (prior) distribution of the parameter. Laplace also developed the first test of significance for the mean assuming that the observations come from a uniform distribution. He based his reasoning on the probability that the deviation from the expected value of the mean was as large as or larger than the value that was observed in the given data.

Between 1809 and 1828, Laplace and Gauss investigated the normal distribution and proved that it provides an approximation to the sampling distribution of the mean in large samples. In the 1920's to 1940's, statistical tests and confidence intervals were systematised by Fisher, Neyman, and Pearson.

Statistical inference is included in various curricula and scheduled for the last level of the high school (for 17–18 year olds). For example, the curricula in Spain and South Australia introduce statistical tests and confidence intervals for both means and proportions (Ministerio de Educación y Ciencia, MEC, 2007; Ministerio de Educación, Cultura y Deporte, MECD, 2015; Senior Secondary Board of South Australia, SSBSA, 2002). Students in New Zealand study confidence intervals, resampling and randomisation (Ministry of Education, 2007). In the last year of high school in France (*terminale,* 17 year olds), confidence intervals and an intuitive introduction to hypothesis testing are included (Raoult, 2013).

Other basic elements of inference are included in various secondary school curricula. For example, the National Council of Teachers of Mathematics, NCTM

(2000), and the Common Core State Standards Initiative, CCSSI (2010), recommend that students at grades 9–12 use simulations to explore the variability of sample statistics. Students are expected to understand how a sample statistic reflects the value of a population[1] parameter and to be able to use sampling distributions to perform informal inferences. For the last level of high school, the CCSSI (2010) also recommends that students make use of sample data and simulation models to estimate a population mean or proportion, develop a margin of error, and use data from randomised experiments to compare two treatments in order to decide whether the observed differences between them are significant.

In the same way, the American Statistical Association's Guidelines for Assessment and Instruction in Statistics Education, GAISE (Franklin et al., 2007), highlight the need for students to look beyond the data in making statistical interpretations when variability is present. They expect middle-grades students to recognise the feasibility of making statistical inferences and high-school students to learn how to perform inferences both with random sampling from a population as well as random assignment to experimental groups.

All these techniques are built upon the concepts of probability, distribution, centre, spread, uncertainty, and randomness that students should have acquired prior to inference (Liu & Thompson, 2009; Harradine, Batanero, & Rossman, 2011). Basic knowledge of widely used distributions such as the normal and binomial distribution is also included in various high-school curricula and the availability of software makes the related computations easy. Furthermore, it is possible to introduce some basic inferential concepts and the underlying reasoning with an informal approach, using simulation.

The majority of the ideas are introduced in the chapter using a specific paradigmatic situation that may serve to engage learners actively; later additional activities and inference methods are suggested in order to complement learning.

5.2. A TEACHING SITUATION: THE TEA-TASTING EXPERIMENT

In this section, we describe a teaching situation inspired by *the lady tasting tea*, a famous randomised experiment devised by Fisher (1935/1971).[2] Based on an episode in his life, Fisher developed the *test of significance,*[3] part of which will be used in the discussion below. We summarise the original situation and propose a slightly different experiment to introduce the main elements of significance tests, sampling distributions, and confidence intervals for a proportion in a simple way.

---

[1] The term population is used in a wide sense here. A process is an infinite population; its elements are the outcomes produced in an infinite series of repetitions of the process.

[2] Reproduced in Fisher (1956).

[3] Tests of significance were introduced by Fisher. Neyman and Pearson hold a different view of statistical tests as rules for deciding between *two* hypotheses (see Section 5.3.4).

5.2.1. The Lady Tasting Tea

The lesson can start by posing the following problem to the students:[4]

> *Task 5.1*. A lady claims that by tasting a cup of tea made with milk she is able to tell whether tea or milk was added first to the cup. How can we organise an experiment and collect data to check her assertion?

The teacher collects suggestions in the classroom. First, the students are asked whether the fact that the lady correctly classifies the order of milk and tea for *one cup* should be accepted as evidence that the lady has the ability to do so always. The teacher may also ask the students to provide personal examples of surprising coincidences (e.g., two students have the same birthday).

Historical note. Fisher's original experiment was designed to check the lady's claim properly. He prepared exactly 4 cups of tea where milk was added first and another 4 with tea added first, informed the lady of the setting, and presented the 8 cups in random order to her. One possible answer of the lady could have been *TMTT MMTM* (*T* means tea first; *M* milk first). Under the null hypothesis that the lady just guesses (i.e., classifies randomly), all 70 sequences have the same probability of 1/70 = 0.01429 in Fisher's calculations.[5] As the lady classified all 8 cups correctly, the result was considered as providing statistically significant evidence of her ability. Fisher suggested that the experimenter fixes the value of probability considered small enough to reject the null hypothesis before the experiment is performed. He used a 5% level of significance to decide whether the data provided support to reject the null hypothesis or not.

5.2.2. Using Experimental Data to Test a Hypothesis

In the classroom, a variation of Fisher's original experiment will serve to introduce various methods of inference for a population proportion. In our setting, each cup is prepared independently of the others so that the number of cups with tea added first can vary from 0 to 8.

---

[4] Parts of the activities included in this section have been adapted from a simpler version of the same project, which is described in Batanero (2001) and Batanero and Díaz (2011). Another teaching experiment with a natural embedding of the process of conceptualisation of the statistical notions is a memory experiment where the students may test whether they are better in memorising than a psychological "law" formulates. For details, see Borovcnik (2014).

[5] The lady's answer is a permutation of 4 *T*'s and 4 *M*'s (or a combination of 8 places taken 4 at a time); $C(8, 4) = \binom{8}{4} = \frac{8!}{4!4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = 7 \cdot 2 \cdot 5 = 70.$

The setting of the experiment
One student prepares eight cups of tea; another student flips a coin to decide the order; heads means milk is added first to the cup, tails means that tea is added first. Meanwhile, Peter, a third student who plays the role of the tea-tasting lady, is in a different room; he should taste the tea and classify the order in which milk and tea were added. After Peter correctly classified two of the first three cups, the teacher asks the following questions.

> *Task 5.2.* Is this result a proof that Peter can distinguish the order in which milk and tea were added to the cups? Or do you think that Peter makes his classifications at random? In case Peter predicts the order at random (simply guesses), what is the probability for Peter of getting 2 or 3 correct classifications out of *three* cups, just by chance?

To compute this probability, let *S* and *F* represent Peter's success and failure in predicting the order of tea and milk. As the order of ingredients was determined by flipping a coin, students should easily model the probability of success by ½, i.e. $P(S) = P(F) = \frac{1}{2}$ and assume that the consecutive tosses are independent (which does not apply for Fisher's original experiment).

Coin simulation[6]
If students are unfamiliar with elementary rules of probability, the experiment (tasting a cup of tea and milk) can be simulated by coin tossing. Success (*S*) is replaced by heads and failure (*F*) by tails. For three cups, each student flips three coins and counts the number of heads and tails representing now successes and failures of classifying the order of tea and milk in the tea-tasting experiment.

The number of successes in each sample of three cups is a variable because it varies in the different samples (substituted by the three-coin simulation). Its distribution is called *sampling distribution*. This distribution can be estimated by combining the simulated results from all students as shown in Figure 5.1.

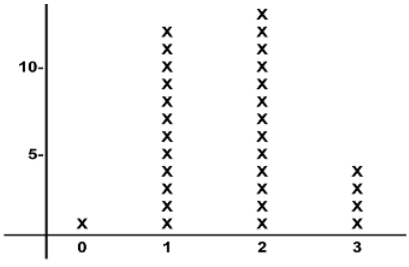


Figure 5.1. Number of heads in 30 simulations (three coins)

[6] The aim is to find out how rare an observation is, in case we assume a given hypothesis.

From this empirical sampling distribution we build Table 5.1 and *estimate* the probability of 2 or more correct classifications $P(S \geq 2)$ by the relative frequency of times we obtained two or three correct classifications:

$$P(S \geq 2) \approx h(S = 2) + h(S = 3) = \frac{13}{30} + \frac{4}{30} \approx 0.5667 .$$

Table 5.1. Empirical sampling distribution: three-coin simulation

Number of heads	Counts	Frequencies	Cumulative
0	1	0.0333	0.0333
1	12	0.4000	0.4333
2	13	0.4333	0.8667
3	4	0.1333	1.0000
Number of samples	30	1.0000	

Thus, 2 or 3 successes cannot be considered "rare" as they occurred in 57% of the experiments. As an intuitive cross-check, we might compare the result with the frequency of three-child families with 2 or 3 boys.

Then, the teacher suggests continuing the experiment and getting the results for eight cups of tea. Let us assume, for example, that Peter correctly classifies the order of milk and tea in 6 cups and fails in 2. Then, the teacher encourages further discussion by posing the following questions.

Task 5.3. Do you think that 6 out of 8 successes is sufficient evidence to reject the conjecture that Peter merely guesses the order in which tea and milk were added to each cup? What is, by pure chance, the probability of correctly classifying 6 or more cups? Which number of successes would you require as evidence to reject our initial conjecture, i.e., ascribing a special ability to Peter to identify the order of milk and tea?

The teacher asks the students to estimate the probability of getting this (6 successes) or a more extreme result (7 or 8 successes) if Peter predicts the order of milk and tea randomly. Again the students simulate the experiment by coin tossing. Each student flips eight coins and counts the number of heads and tails, which represent successes and failures in the tea experiment. The number of successes X for each student in the simulation can vary from 0 to 8, because not all of the students will get the same number of heads. The results for the whole class can be displayed in a dot plot to estimate the sampling distribution for this variable (Figure 5.2).

We are interested in the probability of getting 6 or more successes:

$$P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8) .$$

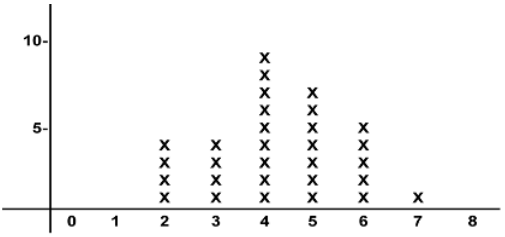


Figure 5.2. Number of heads in 30 simulations (eight-coins)

This probability can be estimated using the relative frequency in the simulation; as 6 of the 30 students in our scenario got a result of "at least 6", we estimate:

$$P(X \geq 6) \approx h(X \geq 6) = \frac{6}{30} = 0.20 .$$

Again, this event is not "rare" as we get it in roughly 1 out of 5 times when tossing eight fair coins. On the contrary, getting 7 or 8 heads (which represent 7 or 8 correct classifications in 8 cups of tea) would be considered as a *significant* result as it happens only in about 1 out of 30 times (3.3%).

The rationale behind significant results

Note that in order to reach a conclusion, we *assumed* that our initial conjecture (*hypothesis or model*) was true (no ability of classification) and accepted ½ for the probability of success. We organised an experiment and collected some data to test this conjecture. Consequently, we estimated the probability of obtaining the observed result (6 successes in 8 trials) or more extreme values (7 or 8 successes) given the conjecture (the probability of success is ½). This probability is a *conditional probability* and is known as *p-value*.[7] It measures how unusual the observed result is given that the conjecture were true.

We notice that the null hypothesis may be true and yet we get a "significant result" just by pure chance. Consequently, we do not logically falsify our conjecture; we only *reject* it because the data would be very unusual if the conjecture were true. Another remark is that when we use simulation (instead of directly computing the probabilities), we do not get the exact values for the probabilities of interest but only estimates for them. Teachers should be conscious of the different nature of frequencies (obtained from data; varying in different samples) and probability (the unknown constant that we try to estimate by the simulations).

[7] The term *p*-value was introduced by Karl Pearson (1900) and resumed by R. A. Fisher (1925/1954); it is a mere coincidence that the probability of a binomial model is often signified by the same letter *p* as is done here.

5.2.3. Different Approaches to Compute the p-value

Depending on students' previous knowledge, teachers may use various approaches to compute or estimate the *p*-value: a) computer simulation; b) elementary enumeration, c) probability rules, or d) the binomial distribution.

Computer simulation

Instead of using coins, we can simulate the experiment with computers. Any statistical software and even general purpose software such as Excel provides random number generation as well as graphs and statistical summaries that can be used to quickly analyse the experimental data. In Fathom, we collect data with 8 attributes (cup1 to cup8) in each of which we simulate random successes (1) and failures (0). Two new attributes contain the number and the proportion of successes in each sample of 8 trials. We then produce 200 replications of 8 cups in the tea-tasting experiment and display the empirical sampling distribution by a dot plot.

In the simulation study in Figure 5.3, there were 2 experiments with 8 successes, 4 experiments with 7 successes, and 34 with 6 successes from a total of 200 experiments. We use the relative frequencies to estimate the requested probability:

$$P(X \geq 6) \approx h(X = 6) + h(X = 7) + h(X = 8) = \frac{34}{200} + \frac{4}{200} + \frac{2}{200} = \frac{40}{200} = 0.20 .$$

Such an event that happened 20 out of 100 times cannot be considered as "rare". Therefore, we do not consider "6 or more successes" out of 8 cups as evidence to reject the conjecture that Peter makes his predictions randomly. But the event "7 or more successes" happened only in 3 out of 100 times:

$$P(X \geq 7) \approx h(X = 7) + h(X = 8) = \frac{4}{200} + \frac{2}{200} = \frac{6}{200} = 0.03 .$$

Consequently, we consider this result to be "statistically significant". That is, if Peter could correctly identify 7 or more of the cups, this "evidence" suggests that we should reject that Peter was just guessing at random.

Applying elementary probability rules or enumeration

Instead of estimating probabilities by simulation, we might compute the exact probability of 2 or 3 successes in the three-cup experiment given the conjecture is true (in this case $p = ½$) by applying the product rule. As the trials are independent from each other, it holds:

$$P(SSS) = P(S) \cdot P(S) \cdot P(S) = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8};$$
$$P(S = 2) = P(SSF) + P(SFS) + P(FSS) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8} .$$

Or, we establish the sample space E = {*SSS, SSF, SFS, FSS, SFF, FSF, FFS, FFF*}, enumerate the favourable cases, and determine that the probability of

getting 2 or 3 correct as 4/8 or 1/2 by Laplace's rule as the elements are equally likely (see Chapter 3).

Tea-testing-experiment: 8 cups

	cu...	cu...	c...	c...	c...	cu...	c...	c...	succe...	sampleprop..
1	0	0	0	0	1	1	0	0	2	0.25
2	0	1	0	1	1	0	1	0	4	0.5
3	0	0	0	0	0	1	1	1	3	0.375
4	0	1	0	0	1	0	0	1	3	0.375
5	0	1	1	0	0	0	0	1	3	0.375
6	0	0	1	1	1	1	1	1	6	0.75
7	1	0	1	0	1	1	1	0	5	0.625
8	0	0	0	0	0	1	1	1	3	0.375
9	1	0	1	0	1	1	1	1	6	0.75
10	0	1	1	1	0	1	0	0	4	0.5
11	0	1	0	0	1	1	0	1	4	0.5
12	1	0	0	0	1	1	1	0	4	0.5
13	0	1	1	1	0	1	0	1	5	0.625
14	1	1	0	1	0	1	0	0	4	0.5
15	1	1	0	1	0	1	1	0	5	0.625
16	0	1	1	1	0	1	1	0	5	0.625
17	0	0	1	0	0	0	0	0	1	0.125
18	0	0	1	1	0	0	1	0	3	0.375
19	0	0	1	0	0	1	1	1	4	0.5
20	1	0	1	0	0	0	1	0	3	0.375
21	0	1	1	1	1	1	1	0	6	0.75
22	1	1	0	1	0	0	0	1	4	0.5
23	1	1	1	1	1	0	0	1	6	0.75
24	1	1	0	1	0	1	1	0	5	0.625
25	0	0	0	0	0	1	1	1	3	0.375

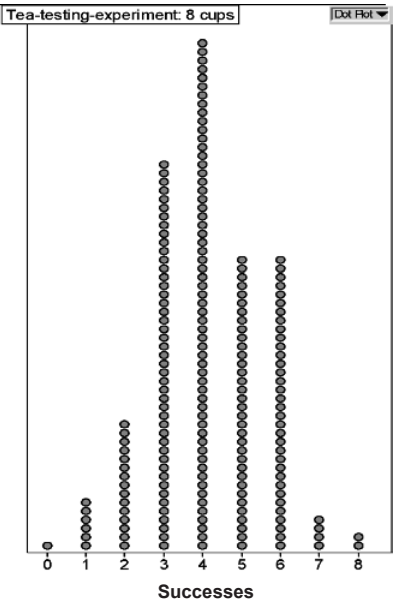


Figure 5.3. Simulation of 200 eight-cup experiments with Fathom

Direct computation using the binomial distribution

With 8 experiments, however, the enumeration method becomes difficult. Theoretically, the sampling distribution of the number of successes in the tea-tasting experiment is a binomial distribution $B(n, p)$ where n is the number of cups and p is Peter's capacity to classify a cup correctly. When students have studied this distribution (described in Chapter 3), they can use it (once the basic assumptions are clarified) to calculate the probability $P(X = k)$ for k successes in all n trials by:

$$P(X = k) = p_k = \binom{n}{k} p^k (1-p)^{n-k} = \binom{8}{k} 0.5^k (1-0.5)^{8-k}, \quad k = 0, 1, ..., n .$$

Alternatively, students may use a calculator, a spreadsheet, or software to compute the probability of obtaining 7 or more successes in 8 trials:

$$P(X \geq 7) = P(X = 7) + P(X = 8) = \binom{8}{7} 0.5^8 + \binom{8}{8} 0.5^8 = 0.0313 + 0.0039 = 0.0352.$$

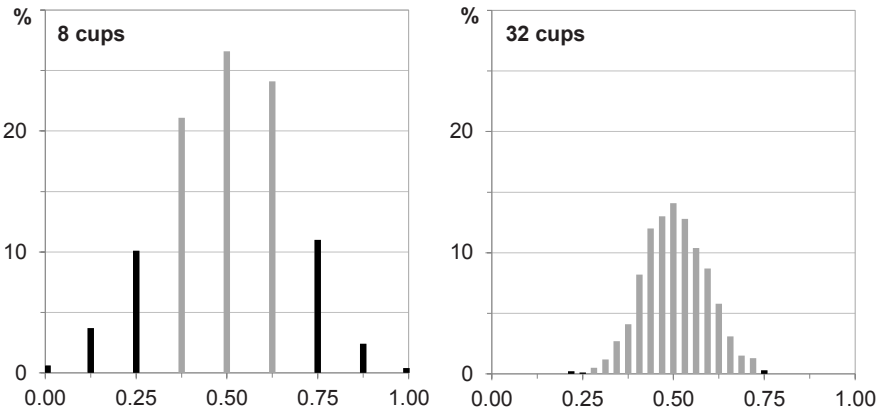
5.2.4. Sampling Distribution for the Proportion and the Effect of Sample Size

To estimate the proportion of successes (unknown capacity) p, we use the sample proportion $\hat{p}$, which is determined by dividing the number of successes by the sample size n. As stated in Chapter 3, the sample proportion converges to the unknown value p; in the sampling distribution for $\hat{p}$, the standard deviation gets smaller when we increase the sample size. This feature – that the variability decreases with the sample size – also holds for the sampling distribution of means. It is important to have students reflect on these patterns in sampling distributions (Chance, delMas, & Garfield, 2004), which is easily understood using simulation.

> *Task 5.4.* Compare the results of simulating 8-cups and 32-cups experiments on the assumption that Peter is merely guessing the order of tea and milk. Simulate 1000 experiments each and draw a bar graph of the sampling distribution. What do you notice?

To investigate the effect of the sample size, students compare the sampling distribution for the two experiments with 8 and 32 cups (Figure 5.4). Again they estimate the sampling distribution by simulating 1000 experiments each. The centre of both distributions is about the same (close to 0.5). However, the standard deviation of 0.0885 for $n = 32$ is about half of the value of 0.1776 for $n = 8$.

This observation reflects general properties of the theoretical sampling distribution, which is a binomial distribution (with n the sample size and $p = 0.5$ the success probability) divided by n. Therefore in both experiments $\mu = p = 0.5$ and as it holds $\sigma = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$, we obtain $\sigma = 0.1766$ for $n = 8$ and $\sigma = 0.0884$ for $n = 32$ as exact values for the standard deviation.



a. Samples with n = 8 *b. Samples with n = 32*

Figure 5.4. Sampling distribution for the proportion – 1000 samples with two sample sizes

In Figure 5.4, there are 289 samples (28.9%) with proportions equal to 0.75 or larger or equal to 0.25 or smaller in the 8-cup experiment; however, these values are rare in the 32-cup experiment as there are only 5 (0.5%). These results illustrate that the sample proportion is closer to the population proportion in the longer experiment. Students may explore such properties with other sample sizes and different initial values for the population proportion. Further activities related to sampling variability are described in Section 5.3.1.

5.2.5. Estimating the Population Proportion by a Confidence Interval

Once we reject the hypothesis that Peter was merely guessing the order of tea and milk (we model this merely guessing as randomly choosing the answer with equal probabilities), we assume that his probability of correct classifications is $p \neq 0.5$. To complete our study, we can estimate the true value for the population proportion p (Peter's theoretical proportion of correct classifications).

> *Task 5.5.* Assume that Peter correctly classified 75% of the cups in a 32-cups experiment (i.e., 24 successes out of 32 cups). Estimate Peter's capacity to correctly classify the order and discuss the precision of this estimate.

Our best estimate for Peter's proportion of success is $\hat{p} = 0.75$ since we expect the population proportion to be close to the sample proportion. However, as we expect some variability in the estimate, we can explore this variability with simulation and introduce the idea of *confidence intervals*.

We can produce 1000 simulations of the 32-cups experiment assuming that the true value of the population proportion equals $\hat{p} = 0.75$ and study the variability of the sample proportions in size-32 samples (see Figure 5.5). The most likely values range around 0.75 and the likelihood gets smaller when we deviate from this value. As the 5th and 95th percentiles of this distribution are $P_5 = 0.62$; $P_{95} = 0.88$, we can state that in 90% of the size-32 samples the sample proportion will vary between 0.62 and 0.88 given that the population proportion is 0.75.

The interval (0.62, 0.88) reflects the accuracy of our estimate. It is remarkable that this *intuitive interval* numerically coincides with what is called a 90% confidence interval if only the sample is moderately large.[8] The meaning of a *90% confidence interval* for p is that we expect that for 90% of all samples taken repeatedly from the population, the intervals that are built on these samples (using the proportion $\hat{p}$ in the sample) will capture the true value of the parameter.

If we want a confidence higher than 90%, we need a wider interval of estimation. Another way to increase the confidence level of the interval is to use a larger sample size. The teacher can ask the students to produce computer

[8] More information about confidence intervals can be seen, e.g., in Moore (2010).

simulations of the experiment for larger sample sizes and calculate the intuitive confidence intervals for the new situation with different degrees of confidence. In case the students have studied the normal approximation to the binomial distribution, they can use this knowledge to compute an approximation to the confidence interval.

As the theoretical sampling distribution is binomial (32, 0.75), we have $\mu = p = 0.75$ and $\sigma = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \approx 0.0765$. In the normal approximation $N(0.75, 0.0765)$, we obtain $P_5 = 0.6241$ and $P_{95} = 0.8759$. Consequently, the approximate 90% confidence interval in our example is (0.6241, 0.8759).

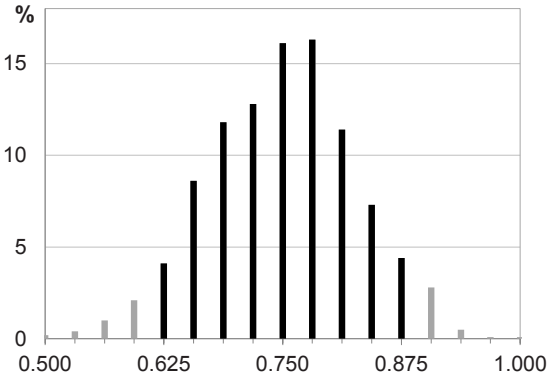


Figure 5.5. Variability of the sampling proportion when p = 0.75 and n = 32 – simulation of 1000 samples

5.3. ADDITIONAL ACTIVITIES

In this section, we suggest a few more classroom activities that complement the previous activities related to statistical tests: investigation of the sampling distribution by simulation; illustration of two types of inference by examples, and informal introduction of statistical tests as decision rules.

5.3.1. Exploring the Central Limit Theorem

From Chapter 3, we know that the normal distribution is an approximate model for the distribution of the sum of independent random variables. For this reason, it provides a suitable model for the sampling distribution of many statistics such as the mean or the proportion, which are computed from a random sample of data by adding values of single statistical units. A basic understanding of the Central Limit Theorem can be progressively built if we ask the students to simulate the sampling

distribution of the mean for different shapes of the population distribution and different samples sizes.

> *Task 5.6.* Investigate the sampling distribution of the mean of a sample from a distribution that describes the population of the values 1 to 6; use
>
> a. a uniform distribution, or,
> b. a J-shaped distribution on this set.
>
> In both cases, simulate 1000 samples of size 5, compute the mean for each sample, and draw a bar graph for the distribution of the simulated means. Compare the sampling to the parent distribution and compare the sampling distributions to each other. Repeat the investigation by generating samples of size 20. What can you see?

First, the students simulate the selection of 1000 samples of $n = 5$ elements from the uniform distribution for question a. and from a specific J-shaped population distribution for question b. (first row in Figure 5.6). Then, for each sample, they compute the mean and draw a bar graph of the empirical sampling distribution for the sample means (second row in Figure 5.6). Finally, the students repeat the whole process with $n = 20$ (third row) and investigate the influence of the sample size on the shape of the sampling distribution.

The first conclusion is that the centre of the distribution of the sample means is very close to the mean of the population for either population that is sampled. The second conclusion is that the distribution of the sample means is more compact as the sample size increases (Law of Large Numbers) and this effect is even more visible if we use a sample of larger size. The third conclusion is that the shape of the distributions becomes more symmetric and resembles a bell-shaped curve[9] the larger the sample size is even if the parent population is J-shaped.[10]

5.3.2. Inference for Proportions

In the tea-tasting project in Section 5.2, we dealt with a binary variable. This experiment is a particular example of problems with only two possible outcomes where the binomial distribution (introduced in Chapter 3) applies. We recall that a *binomial experiment* has to satisfy the following requirements:

---

[9] As the graphs will contract to one point (the mean of the population) with increasing size, one has to *standardise* the statistic to preserve a distribution that can still be seen.

[10] In Figure 5.6, both distributions restrict to one point if the sample size is increased beyond limits. Therefore, the shape of the distribution is stabilised by standardising the random variables so that rather than S/n, the standardised variables $\widetilde{S}_n = \dfrac{S/n - E(S/n)}{\sigma(S/n)}$ are investigated. The Central Limit Theorem states that asymptotically, the standardised variables $\widetilde{S}_n$ follow a standard normal distribution with parameters 0 and 1 if only the variables are independent and have a finite variance. The speed of normalisation is influenced by the shape of the parent distribution.

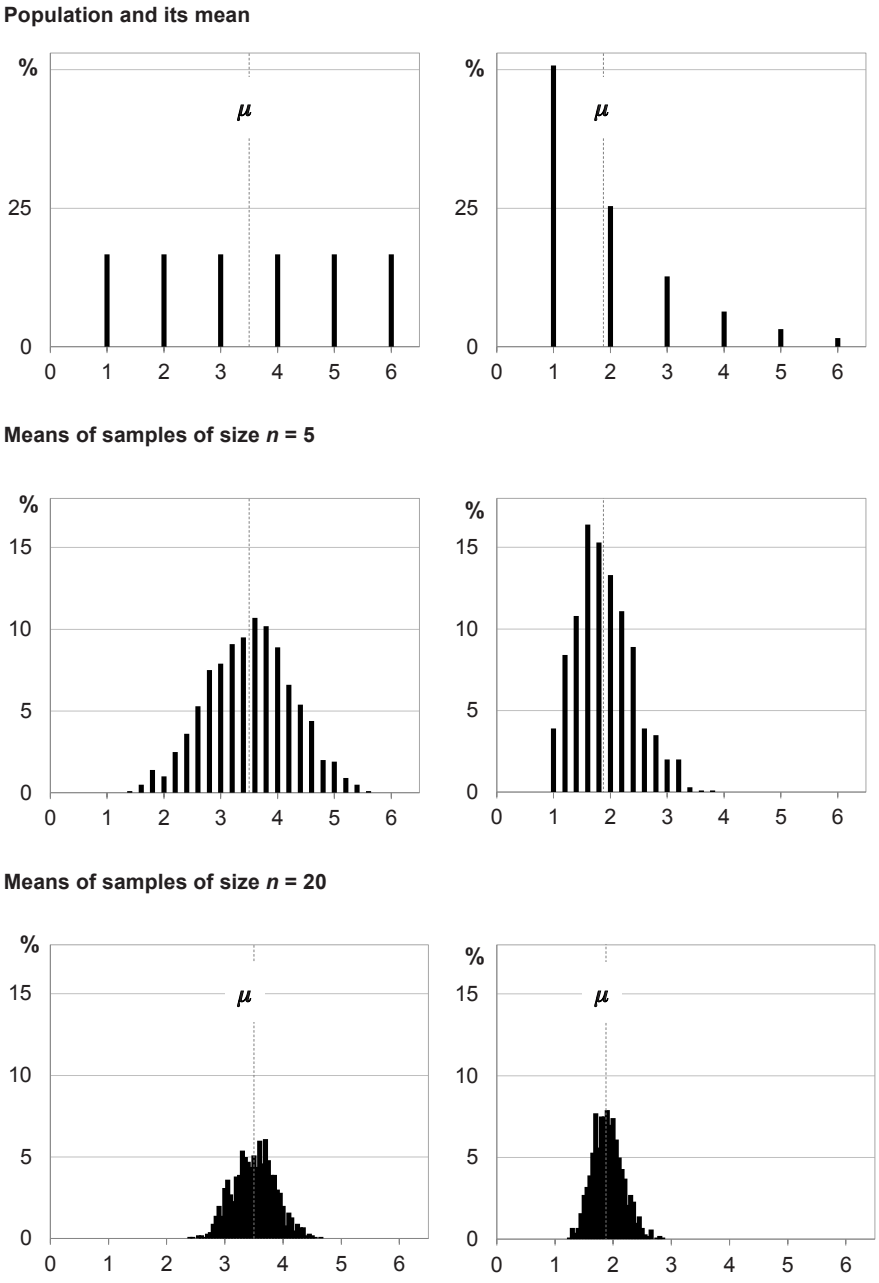


Figure 5.6. Convergence of sampling distribution of the mean to the normal curve – 1000 samples of size 5 and 20 for two parent populations simulated

1. Each trial can be reduced to two outcomes, which are considered as either success or failure; the probability of a success p remains the same for each trial.
2. The trials are performed *independently*.
3. The number of trials is fixed before the experiment.

The teacher may pose new inferential tasks where one may be interested in either estimating p or testing a hypothesis about the value of p; the parameter p may also be interpreted as the proportion of success in the population. Some examples follow.

– *Voting.* Different institutions publish surveys during election campaigns. Students may interpret reported confidence intervals for proportions of votes for a particular candidate or party. Such discussions may serve to clarify students' understanding of terms like stratified sample, estimate, margin of error, or 95% confidence. Questions about whether the proportion of votes will change with age, socio-economic background or country region will also serve to discuss bias in sampling.
– *Medical treatment.* Arthritis, a painful, chronic inflammation of the joints is a leading cause of disability for elderly people. Examples of proportions in which the health services are interested are the proportion of people who suffer from arthritis at a given age or the proportion of patients who experience side effects of pain relievers. Such investigations may relate to any other disease.
– *Normal approximation to the binomial distribution.* In Section 5.2.5, we used the normal approximation to the binomial distribution to compute an approximate confidence interval for a proportion. The normal distribution may be used to solve other inference problems involving the population proportion p. Students may also use software and plot the sampling distribution of the proportion to investigate this property for different values of the parameter p and increasing sample sizes.[11]

5.3.3. Inference for a Mean

A very common situation is estimating the mean μ or testing hypotheses about its value in a population, which is normally distributed $N(\mu,\sigma)$; e.g., the average height of boys in a given population and age. If the value of the standard deviation is known, the sampling distribution for the mean $\bar{x}$ in a sample of n elements from the population is also normal $N(\mu,\frac{\sigma}{\sqrt{n}})$. In practice this applies approximately if the population distribution is symmetric and single-peaked unless the sample is small (see Section 5.3.1).

[11] By a rule of thumb, the approximation is regarded as sufficient if $n \cdot p \cdot (1 - p) > 9$.

Task 5.7. Consider the intelligence quotient (IQ), a score derived from standardised tests designed to measure intelligence. The IQ can be modelled by a normal distribution N (100, 15). Simulate 1000 samples of 50 IQ values and compute the sample means. Then plot the empirical sampling distribution of these means and compare them to the theoretical distribution. What can you see?

The result of a simulation is shown in Figure 5.7. Note the small variability of repeated means, which lie between 95 and 105, which reflects that the theoretical standard deviation equals $\frac{15}{\sqrt{50}} = 2.1213$. Using this empirical sampling distribution, students can build confidence intervals or test hypotheses on the IQ for particular groups of the population (e.g., gifted students). If the students are familiar with the normal distribution, they can directly work with it.

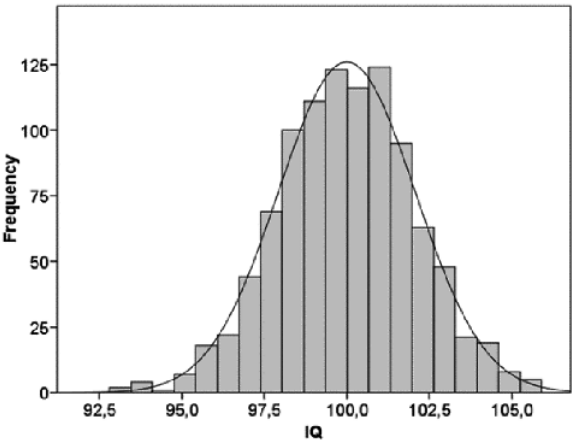


Figure 5.7. Empirical sampling distribution of means compared to the normal model: simulation of 1000 samples of 50 IQ scores each

5.3.4. Statistical Tests as Decision Rules

In Section 5.2, we used tests of significance to analyse whether given experimental data provided evidence to reject a previously assumed null hypothesis. In other situations, *statistical tests* are used to make a decision between two possible actions.[12] In this case, we consider two different hypotheses: the null and the alternative as in the following example.

[12] Neyman and Pearson viewed statistical tests as rules of inductive behaviour assuming some risks. Testing a statistical hypothesis is a tool to help making a rational choice between two courses of action (Neyman, 1950). For a general analysis of components of risk, see Borovcnik (2015b).

Quality control as a problem of decision

Manufacturing a product must meet the customers' specifications, while limiting cost. A key issue is controlling the proportion of defective items coming off the production line. As we cannot inspect the entire production, statistical sampling is used to decide whether the production is "under control". At regular intervals (hourly, daily, etc.) items are randomly selected and inspected. As it is impossible to completely eliminate defects, the producer will tolerate a reasonable proportion p_0 of defects and still assume that the process is under control; otherwise, measures should be taken to improve the production process. The decision depends on the competing costs of paying warranty for a proportion p_0 of defective items versus the cost to decrease the number of defects further. Let us assume $p_0 = 0.04$ and that each day a random sample of 100 items is inspected.

We are interested in the following questions: a) what is the limit for the number of defects we should allow in the sample to assume that the process is still under control? and b) what are the associated probabilities of errors in the decision? This is a typical situation where we should decide among two competing hypotheses:

– H_0: the process is under control; the proportion of defects is $p_0 \leq 0.04$. This is our *null hypothesis*, i.e., no change in the production (no effect).
– H_1: The process is out of control; the proportion of defects is $p_0 > 0.04$. This is our *alternative hypothesis*.

We start (as in the significance test for the tea-tasting experiment) by assuming that the null hypothesis is true.

Task 5.8. To simplify the situation from the quality control context above, we assume that $p = 0.04$.[13] Simulate situations with different probabilities p of defects (use $p = 0.10$ for example) and try to establish reasonable limits for deciding between the two hypotheses. Repeat the analysis for different sample sizes; first use 100 and then 400 items in the samples. Can you discuss the consequences of your decision rules?

In Table 5.2, we reproduce the empirical sampling distribution of the number of defects in 5000 simulated samples of n = 100 items, when p = 0.04 (null hypothesis). Using the cumulative frequencies, we estimate the probability of obtaining 7 or more defects when p = 0.04:

$$P(X \geq 7 \mid p = 0.04) = 1 - P(X \leq 6 \mid p = 0.04) \approx 1 - 0.8956 = 0.1044 .$$

If we fix 5% as the maximum probability of rejecting the null hypothesis when it is true (statistically significant result), the limit should be 8 defects in the sample to decide that the process is out of control since:

$$P(X \geq 8 \mid p = 0.04) = 1 - P(X \leq 7 \mid p = 0.04) \approx 1 - 0.9548 = 0.0452 .$$

[13] In fact, we are dealing with a composite null hypothesis. However, we only need to consider the case where p = 0.04, since whenever a result is significantly higher for this single hypothesis, it will also be significant for any value from $p \leq 0.04$.

Errors in decision rules

Rejecting the null hypothesis when it is true is the *Type I error*; in quality control it implies the risk of stopping the production process when in fact it is under control. In this context we are also interested in the risk of continuing the production when the process is out of control and we are producing too many defective items. The *Type II error* consists in not rejecting a null hypothesis when in fact it is false (that is, the alternative hypothesis is true).

We are interested in the probabilities of Type I and Type II errors, which are usually denoted by α and β; α coincides with the level of significance (from Fisher's test) and is pre-set right from the beginning. Computing the Type II error probability is difficult because we have different values for β depending on the unknown value of the parameter. In our example, it depends on the unknown value p of the proportion of defects.

In Table 5.2, we simulated another 5000 samples of n = 100 items when p = 0.10 to illustrate the outcome of a situation where the production is out of control.[14] If we decide to stop the production when $X \geq 8$, we do not reject the null hypothesis (process under control) if $X \leq 7$. Then the probability of Type II error is very high as $P(X \leq 7 \mid p = 0.10) \approx 0.2074$. We could diminish this probability if we change the decision rule; e.g., rejecting the null hypothesis already when $X \geq 7$. In this case, β equals 0.1202 but α increases now to 0.1044 (we have calculated this latter number above).

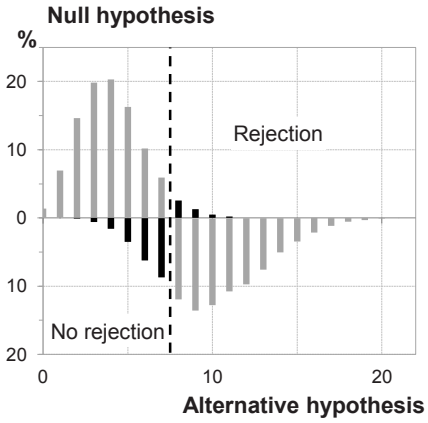
There is no rule about setting the values for these error probabilities because this will depend on the consequences, e.g., the cost of higher numbers of warranties if more defects are produced versus the cost of lost production time due to unduly stopping the production. From this example, we see that the probabilities of Type I and II errors are interrelated and that if we decrease one of them, the other increases. However, a larger sample size allows downgrading both errors to a tractable size.

In Figure 5.8, we reproduce the result of 5000 samples with n = 100 and with n = 400 items for the two hypotheses under study. The empirical distribution under the alternative hypothesis p = 0.10 is drawn upside-down in order to avoid an overlap with the distribution under the null hypothesis p = 0.04. The rejection rule is represented by a vertical bar: if the number of defects exceeds this point, the null hypothesis is rejected. Moving the bar illustrates our dilemma: moving it right, we diminish the Type I error and increase the Type II error, while moving it left, the effect is converse. However, while in the simulations with 100-item samples it is not possible to obtain small probabilities for both types of errors, for 400-item samples, X = 28 is a reasonable criterion to separate between rejection and non-rejection.

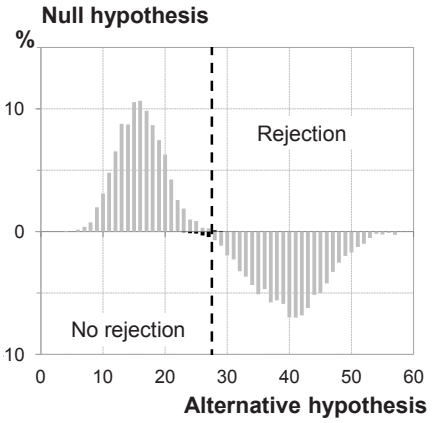
[14] The probability of a Type II error is smaller if $p > 0.10$.

Table 5.2. Simulation of 5000 samples of n = 100 items (p = 0.04 and p = 0.10)

N. of	Null hypothesis: $p = 0.04$			Alternative hypothesis: $p = 0.10$		
defects	Count	Frequency	Cum.Frequency	Count	Frequency	Cum.Frequency
0	68	0.0136	0.0136	0	0.0000	0.0000
1	348	0.0696	0.0832	1	0.0002	0.0002
2	732	0.1464	0.2296	5	0.0010	0.0012
3	991	0.1982	0.4278	30	0.0060	0.0072
4	1015	0.2030	0.6308	79	0.0158	0.0230
5	814	0.1628	0.7936	175	0.0350	0.0580
6	510	0.1020	0.8956	311	0.0622	0.1202
7	296	0.0592	0.9548	436	0.0872	0.2074
8	128	0.0256	0.9804	597	0.1194	0.3268
9	63	0.0126	0.9930	678	0.1356	0.4624
10	25	0.0050	0.9980	639	0.1278	0.5902
11	9	0.0018	0.9998	539	0.1078	0.6980
12	0	0.0000	0.9998	488	0.0976	0.7956
13	0	0.0000	0.9998	380	0.0760	0.8716
14	1	0.0002	1.0000	253	0.0506	0.9222
15				173	0.0346	0.9568
16				108	0.0216	0.9784
17				58	0.0116	0.9900
18				28	0.0056	0.9956
19				16	0.0032	0.9988
20				5	0.0010	0.9998
21				1	0.0002	1.0000



a. Samples of 100 items



b. Samples of 400 items

Figure 5.8. 1000 simulations of the number of defects with different hypotheses

5.4. SYNTHESIS OF LEARNING GOALS

The experiments described in Sections 5.2 and 5.3 may serve to introduce basic inference ideas, as well as the logic and procedures of hypothesis testing and confidence intervals. Below, the main learning goals underlying the activities are summarised.

5.4.1. Basic Inferential Concepts and Procedures

Hypothesis or model

In order to test the possibility that Peter was able to correctly classify the order of tea and milk, we started the tea-tasting activity from the hypothesis that Peter was merely guessing and predicted the order randomly (with equal probability for the two possible classifications).

A statistical inferential process starts by assuming a *model or hypothesis* of the situation. Because we face a random situation, we cannot deductively prove that the hypothesis is true; hence, we follow the typical procedure of statistical tests of significance.[15] We usually work with a *null hypothesis* that describes a model for a situation assuming chance is the only explanation. It states that the results observed in a study are no different from what is expected as a result of chance. In the tea-tasting experiment we assumed that Peter had no special ability to classify the order of tea and milk; his chance of being correct was not different from random guessing, therefore $p = 0.5$.

Experiment and observed data

In order to test a null hypothesis, we organise an experiment and collect some data so that it is possible to divide all results in two classes: results that are compatible with the null hypothesis and results that are not compatible with the null hypothesis (statistically significant results). In the classroom, we replicated the tea experiment: we considered that getting 7 or 8 correct classifications out of 8 cups was statistically significant whereas a smaller number of successes was not considered as significant.

Randomisation and control

Randomisation is a core idea that assures the validity of inference made from data. Variables affecting the results should be controlled in order to assure that the assumed hypothesis is the only possible explanation for the observed results. If perfect control proves difficult, randomisation assures the validity of inferences. It reduces the chances of other potential explanations for the results in the observed

[15] Tests of significance were introduced by Fisher who only considered the null hypothesis (of no effect). Later Neyman and Pearson viewed statistical tests as rules for deciding between two hypotheses (null and alternative hypotheses).

data, apart from the hypothesis being tested. For example, in the tea experiment, we might accidentally put different amounts of tea or milk in the different cups, mix the liquids more, or heat them to a different temperature. By randomly selecting the order in which the ingredients are mixed, the chance that any of these potential confounding factors may affect only one type of outcomes is reduced (e.g., only the cups where milk was added first).

Sample and sampling variability

Assume in the 8-cup experiment we obtained exactly the series *SSSF SFFS*, where the order of letters indicates the order of Peter's successes. It is important that students keep in mind that this data is just one of all possible series of eight results we could have obtained in the experiment. When collecting different samples from the same population, not all of them are identical (*sample variability*) even though the samples resemble the population (*sample representativeness*). Simulation can help students understand these two complementary ideas, especially when students can retrospectively examine and compare the different samples produced (see the activity in Section 5.2.3).

Parameter, statistic, and sampling distribution

We are usually interested in a statistical summary of the whole population, which is unknown (it is called parameter). For example, assume that we are interested in the exact proportion of correct classifications that Peter can make in the long run. We use the value of the statistical summary in the sample (i.e., the proportion of correct classifications in the 8-cup experiment) to estimate the value for the parameter (Peter's capacity).

The observed data constitutes just one of all possible samples that may have been drawn in the experiment. We observe that the proportion of correct classifications (sample statistics) varies in different samples. The distribution of the sample statistics in all the possible samples with a pre-determined size from the population describes the variability of these samples and is known as the *sampling distribution.* We obtained an approximation of it (called empirical sampling distribution) in our simulations of the experiment.

In other situations, we are interested in the population mean and use the sample mean as estimation and then estimate the sampling distribution of the mean. As suggested in Section 5.3.1 as well as in Chapter 3, the sampling distribution of means or proportions can be approximated by a normal distribution, also for samples of moderate size. Another important property is that the variability in the sampling distribution decreases when the sample size increases.

Probability of obtaining the observed or a more extreme result if the hypothesis is true (p-value)

In order to come to a conclusion, we want to assess the unusualness of the observed data or more extreme values in case the null hypothesis is true. In our example, we compute the probability of obtaining 6, 7 or 8 successes in 8 trials based on the assumption that $p = 0.5$. This probability is today known as *p*-value; it

is a conditional probability. We can compute its value using probability distributions or estimate it via simulation.

– *Exact computation* of *p*-values. In some cases, the null hypothesis leads to a known probability distribution such as the binomial or normal distribution. Students who have previously studied these distributions can directly compute the *p*-value.
– *Estimation* of *p*-values via simulation. In this method we simulate the conditions of the null hypothesis to generate one sample, compute the given statistics (the number of successes in the tea experiment), and then repeat the whole process to get an empirical distribution, which is an approximation of the sampling distribution of the statistics. From this empirical distribution we estimate the *p*-value.

Statistical significance

Any possible result of the experiment (e.g., any number of correct classifications of 8 cups) may happen despite its low likelihood. Consequently, the experimenter fixes a *level of significance* α (usually set to 0.05 or 0.01)[16] as a criterion to decide, which data is compatible with the null hypothesis. If a result happens and the probability of this or even more extreme results is lower than α given the null hypothesis is true, then the observed data is called *statistically significant* and leads to a rejection of the null hypothesis.

Rejecting the null hypothesis

In the tea experiment, we decided that obtaining 6 or more correct classifications for 8-cup experiments was not unlikely, because such a result happens by chance on average 1 in 5 times when we repeat the experiment. On the contrary, the probability of getting 7 or more correct classifications is only 0.03 and this result might be considered unusual, given the assumed hypothesis.

To determine whether a result is statistically significant, the researcher compares the *p*-value with the significance level α. The null hypothesis is rejected if the *p*-value is lower than the significance level α. In the example, we reject the null hypothesis and conclude that Peter does not merely guess the order randomly but has a special ability to identify the order (his probability to classify the order of milk and tea correctly is higher than 0.5). This is the usual reasoning in *significance tests,* a method used in statistics to provide evidence against a given null hypothesis.

Interval estimation, margin of error, and confidence level

Often, a statistical summary in the sample (e.g., the sample mean) provides an estimate for an unknown population parameter (the population mean). We are also

[16] Using 0.05 or 0.01 significance levels is a convention; consequently, the experimenter could select other values.

concerned about the *margin of error* of an estimate, as well as a measure of the "confidence" that our estimate is close to the parameter.

A *confidence interval* is an interval of possible values for the parameter, which a statistician regards as plausible after the data of a sample is known. Together with the interval, we provide a *confidence level* with the following property: if we take a random sample of fixed size from the population and calculate a confidence interval with confidence level 95%, then – on average – 95% of the intervals from *repeated* samples will contain the unknown parameter. As for the significance level, the value 95% is a matter of convention and can vary. A key property is that the width of the interval estimate increases with higher confidence levels and the width of the interval estimate with fixed confidence level decreases with larger samples.

5.4.2. Additional Advanced Learning Goals

Deciding between two statistical hypotheses

If more time is available, if we deal with talented students, or if curricular guidelines require more advanced learning, we can expand the ideas of inference and include situations like quality control where we are interested in deciding upon a null and an alternative hypothesis. The situations may be explored informally using simulation with no need of more formal statistics. In such contexts, it is vital that students understand that null and alternative hypotheses do not play symmetric roles as we assume that the *null hypothesis* is true to build the sampling distribution and to determine the decision rule (i.e., when to reject it).

Two opposing types of errors

Two kinds of errors are attached to decisions made due to the result of the test: A *Type I error* is the incorrect rejection of a true null hypothesis and a *Type II error* is the incorrect non-rejection of a false null hypothesis. The significance level α is the probability that is pre-set for the Type I error. Because there are many different possibilities for the value of the parameter under the alternative hypothesis, the probability of Type II errors is a function of the "true" alternative.

5.5. SOME POTENTIAL DIFFICULTIES

As it may be seen from the previous sections, statistical inference is based on many key concepts, as well as on a complex reasoning, and, due to its complexity, statistical inference is often misinterpreted or misused. Below we summarise the wide research on understanding inference and suggest some implications for teaching (complementary sources are Batanero, 2000; Borovcnik & Bentz, 1991; Castro Sotos, Vanhoof, Noortgate, Onghena, 2007; Harradine, Batanero, & Rossman, 2011; or, Shaughnessy, 2007).

5.5.1. Understanding Random Sampling and Sampling Distributions

There are quite a few components of understanding random samples starting from the demand that a sample should be representative, but should be random; at the same time, it is important to avoid any personal influence and bias including intuitive heuristics (that impede to apply mathematical models). We also discuss the different types of reasoners, the many kinds of distributions that are involved, and finally, the *distributional* view of a single value as opposed to an individual consideration of it.

The paradox of representativeness and variability of random samples

The key idea in inferential reasoning is that a sample provides some information about the population from which it is drawn even though this information is incomplete. When students understand this idea, they can balance two apparently contradictory properties of random samples (Batanero, Godino, Vallecillos, Green, & Holmes, 1994):

– *Sample representativeness.* When the sample is adequately selected, it will often reproduce some characteristics of the population; for example, the sample mean will be close to the population mean.
– *Sample variability.* Not all the samples are identical and not all the features in a sample reproduce exactly the values they take in the population; for example, one does not obtain the same mean in all the possible samples. According to Shaughnessy (2007), this second property is not always well-perceived by students as they are used to solve problems by unique values (e.g., a probability or a mean) instead of being asked for a range of values (e.g., building a data distribution with a given mean).

Judgemental heuristics

An important research programme in the context of decision making attributes these errors to certain judgemental heuristics or cognitive processes that reduce the complexity of a problem during the solution of a problem (see Kahneman, Slovic, & Tversky, 1982). One of these heuristics is the *representativeness heuristic*; people who unconsciously use this reasoning estimate the likelihood of a sample by only taking into account how well this sample represents the population. In consequence, they are insensitive to sampling variability and may neglect the effect of sample size on variability. These heuristics are pervasive in both students and teachers (see, e.g., Batanero, Serrano, & Garfield, 1996).

In extreme cases, some people are over-confident in reasoning from small samples because they trust the results, no matter how small the sample is. This tendency has been termed *"Law of Small Numbers"* because people incorrectly generalise from the Law of Large Numbers (that relative frequencies converge to the underlying probability) to a small number of repetitions of the experiment. Students with this misconception believe that random samples should be very similar to each other and accurately reflect the population regardless of how small the samples are.

Types of reasoners

Shaughnessy (2007) described a series of studies directed to explore students' understanding of sampling variability. A prototypical task asks the students to predict the number of items with specific properties when taking samples from a finite population and to explain their reasoning. A typical context is drawing from a container with 20 yellow, 50 red, and 40 blue candies. The students have to predict the number of red candies that will result in 5 consecutive samples of 10 (with replacement). Using variations of this task with students of different ages and countries, *a progression* in students' reasoning has been identified (Noll & Shaughnessy, 2012; Shaughnessy, Ciancetta, & Canada, 2004):

1. *Idiosyncratic* students base their prediction of sample variability on irrelevant aspects of the task (e.g., preference, physical appearance, etc.);
2. *Additive reasoning* students tend to predict the samples taking only absolute frequencies into account.
3. *Proportional reasoners* use relative frequencies and connect proportions in the sample to proportions in the population; they tend to predict samples that mirror the proportion of colours in the container.
4. *Distributional reasoning* involves connecting centres and spread when making the predictions of sampling variability. In addition to reproducing the proportion, their arguments also reflect random variation.

Different types of distributions involved in inference

Teachers should be conscious that inference involves three types of distributions that students have to discriminate carefully (Harradine, Batanero, & Rossman, 2011):

– The *population distribution.* We are interested in one *variable* in a specific population; for example, the weight of boys at a given age in a specified country. This variable is modelled by a theoretical probability distribution, e.g., by a normal distribution $N(\mu,\sigma)$, which is fully determined if we know the values of its mean μ and standard deviation σ. A typical inferential situation involves predictions about the value of μ when σ is known.
– The *observed distribution* of a single *data set.* In order to estimate the average weight μ of this population of boys, we collect a random sample from that population (e.g., we sample the weights of 100 boys). In this empirical distribution of data, the sample mean $\bar{x}$ is used to estimate the unknown value of μ since we know that the sample mean will be close to the population mean provided that the sample size and selection were adequate. In practice, this is the only sample we select and analyse.
– The *sampling distribution.* When we collect different samples from the same population, we get different values for each sample mean $\bar{x}$. The *theoretical* distribution of the sample mean in all the possible random samples of the same size from the population is called the *sampling distribution* and models the variability of the sample mean $\bar{x}$. For example, if the sample size is large

enough, the distribution of the sample mean is approximately normal $N(\mu, \frac{\sigma}{\sqrt{n}})$ where n represents the sample size. This sampling distribution is used to perform inferential processes such as testing hypotheses or building confidence intervals.

Consequently, for statistical inference it is necessary to distinguish among: a) the unknown theoretical mean in the population; b) the particular mean obtained in our sample; c) the different means that would be obtained with the all possible samples of size n from the population (a random variable); and d) the theoretical mean of this random variable, which coincides with the population mean if the sampling process is performed randomly. Conceptually, this is very difficult and it requires a long period of learning. However, with the current teaching of statistics across all levels of education and with the help of resources such as simulation and visualisation, teachers may help students to construct this knowledge progressively. Starting with sampling experiences, the students observe the variability of samples, solve simple inference problems, and finally, reflect and discuss about what was learnt in the classroom.

Singular versus distributional view of sampling

Saldanha and Thompson (2007) argue that students tend to focus on just the individual sample and its particular statistical summary instead of thinking of collections of samples and a distribution of the sample statistics. They recommend helping students construct this distributional view of sampling by guided simulation activities: Ask students to collect a sample from a population and then compute a statistic (e.g., take 5 balls from an urn containing balls of two colours and determine the proportion of balls with a given colour) and repeat the process. Then, encourage the students to observe the variation of the proportion in the different samples and think of these proportion values as each representing a specific sample. The last step is representing the distribution of the proportion values and using that distribution for making inferences about the proportion in the population.

5.5.2. Understanding Statistical Tests

In Section 5.2, we described tests for a simple hypothesis concerning the population proportion. The same method can be applied with little change to many other situations: testing the mean value of a population; testing the difference between two means, or the difference between two proportions, comparing several means (analysis of variance), etc. All these procedures share the underlying reasoning and the basic concepts described in Sections 5.2 and 5.3. Below, we describe the main misunderstandings of these elements (for more details see Batanero, 2000; Borovcnik & Bentz, 1991; Falk, 1986; Falk & Greenbaum, 1995, or Harlow, Mulaik, & Steiger, 1997).

Mathematical proof and statistical evidence

First, it is important that students perceive the difference between a deductive mathematical proof and a statistical test and do not interpret test results in a deterministic way. However, some students consider that, after obtaining a significant result, they have mathematically falsified the null hypothesis. It is important to help them understand that a statistically significant result may happen and yet the null hypothesis is true (Type I error) because unlikely results are not impossible and thus may happen by chance (Batanero, 2000).

Interpreting p-values and significance levels

Two frequently misinterpreted concepts are significance level and *p*-value. The *p*-value is defined as the "probability of obtaining the observed or more extreme data, given that the hypothesis is true".[17] In the tea experiment, the *p*-value is the probability of obtaining 6, 7 or 8 successes in 8 trials in case the probability of single successes is 0.5. A general definition of *p*-value is:

$$p\text{-}value = P(\text{Data-or-more-extreme-observations} \mid H_0 \text{ true}).$$

This expression is a conditional probability; however, many researchers (e.g., Falk, 1986; Falk & Greenbaum, 1995) have noticed that people either exchange the event and the condition or they disregard the condition. Consequently, different misinterpretations of the *p*-value[18] have been described:

– Misinterpreting the *p*-value as "the probability that the hypothesis or the model is true, after we have the observed data", that is,

$$p\text{-}value = P(H_0 \text{ true} \mid Data);$$

– Erroneously assuming that the *p*-value is "the probability that the hypothesis is true", i.e.,

$$p\text{-}value = P(H_0 \text{ true}).$$

Only in a Bayesian framework, it is possible to attribute a probability to a hypothesis, see Section 5.6.3.

Another common misinterpretation is to believe that the *p*-value is the probability that the result is due to chance. This is only true when there is a good experimental control that ensures that all the conditions (except the hypothesis) are controlled. Otherwise, a significant result might be due to other factors; in the tea-tasting experiment, if we do not control the experiment properly, the success in

[17] The significance level of a test $\alpha = P(\text{Rejecting } H_0 \mid H_0 \text{ true})$ is the probability of rejecting a null hypothesis given that it is true.

[18] Similar misinterpretations have been found for the significance level; for example, people change the order of the events in the definition of α and believe α is the *probability* that the null hypothesis is true after the decision to reject it has been taken.

classifying the order correctly may be due to, for example, adding different amounts of sugar to some cups. What we reject in a *statistical test is the null hypothesis,* and therefore we cannot infer the existence of a particular cause in an experiment from a significant result.

An important remark is that the overall significance level differs from the significance level in a single test when multiple tests are carried out on the same data set (Moses, 1992). It is common to apply many significance tests to one data set; for example, the scores in a test for a sample of students could be compared by gender, age, social class, etc.

A level of significance of 5% means that, given that the null hypothesis is true, we will reject it on average in 5 out of 100 times. Consequently, if we carry out 100 different tests on the same data set using a significance level of 0.05 in all of them, according to the definition we *expect* that 5 out of the 100 tests will be significant just by chance given that the null hypothesis is true in all of them. Thus, it becomes likely to get significant results by applying many tests and this is why the significance level has to be corrected in these circumstances (Batanero, 2000; Harradine, Batanero, & Rossman, 2011).

Statistical and practical significance

A significant result means that the experimental data provided evidence against the null hypothesis. However, it is important to distinguish between *statistical and practical significance*. For example, in comparing two means from large samples, the difference may be negligible for practical purposes and yet be significant (as we obtain a small *p*-value). Other times we find practical "significance" with no statistical significance when the sample is small. If a result is statistically significant, the researcher should still investigate whether the effect (the observed difference in the example) is practically relevant.

5.5.3. Understanding Confidence Intervals

Confidence intervals are – at first sight – easier to understand than hypothesis tests (Fidler & Cumming, 2005) although they involve complex ideas. A confidence interval is an interval of *plausible* values for an unknown parameter in the population (e.g., the mean) based on data from a random sample. The lower and upper end points of this interval are computed from the sample data and we have "some confidence" that the true (but unknown) population parameter lies within the calculated limits. A key concept is the *level of confidence*.

Coverage of the unknown parameter in the long run

The correct interpretation of a 95% confidence interval is that, when taking repeated random samples of the same size from the same population and computing the confidence interval for each of these samples, on average, 95% of these intervals will include the population parameter. After selecting a particular sample and computing the interval for that parameter, this interval may or may not cover the value of the parameter; it is a matter of uncertainty. However, some

people still believe that the parameter is *necessarily* included in the interval. Other students believe that a confidence level represents the probability that the parameter falls between the specific confidence limits.

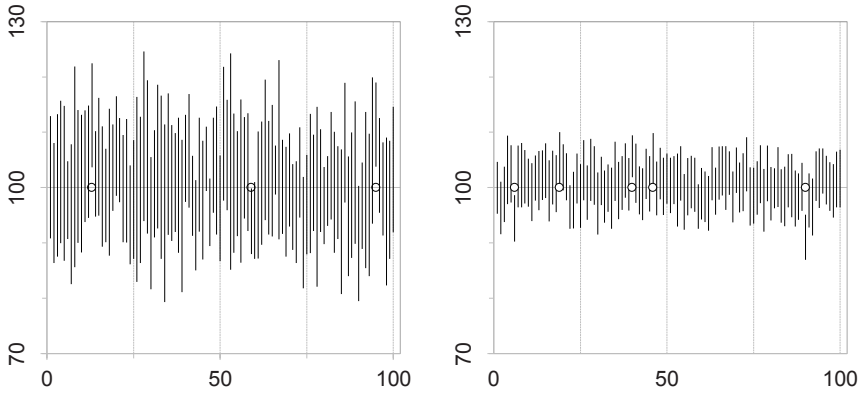
It is important that students understand that confidence intervals vary from sample to sample and analyse the factors that influence their width:

– When we fix the population and the confidence level, the width decreases as the sample size increases.
– When we fix the population and the sample size, the width increases as the confidence level increases; thus, a 99% confidence interval is wider than a 95% confidence interval.

Students may explore these effects using simulation and visualisation in the following task:

> *Task 5.9.* Simulate 100 samples of 10 IQ scores, calculate a 95% confidence interval and represent the intervals as vertical bars ordered by time (number of generated sample). Repeat the simulation with samples of 40 IQ scores. What can you observe?

The results are shown in Figure 5.9. For samples of size 40, students can observe that intervals are much shorter. The activity could be repeated with 90% and 99% intervals.



a. Sample size n = 10 *b. Sample size n = 40*

Figure 5.9. Confidence intervals from 100 random samples of n IQ scores (circles mark intervals that do not cover the population parameter)

Coverage of the mean of the first sample

Students should be warned of the conceptual difference between confidence level and probability of replication, since Fidler and Cumming (2005) found that even researchers confused these concepts and held the misconception that a specific *c*% confidence interval for the population mean will, on average, capture the first sample mean in about *c*% of the new samples when we replicate the experiment.

5.6. ADDITIONAL RESOURCES AND IDEAS

In this section, we first discuss the possibility of teaching inference with informal approaches. Then, we present resampling methods that may be used in the classroom to simplify the mathematics involved in statistical inference. Finally, we introduce elementary ideas related to Bayesian inference, where Bayes' formula is used to update prior probabilities for hypotheses.

5.6.1. Developing Informal Ideas of Inference

The suggested activities in the previous sections, particularly those based on simulation and visualisation are adequate for the last year of high school (17–18 year olds) in those countries where the curriculum explicitly includes ideas about parameter estimation, confidence intervals, or statistical tests.

To avoid the difficulties in understanding that were discussed in Section 5.5, many authors recommend that students are exposed to the conceptual bases in the simplest possible way, before we attempt to teach them formal inference (e.g., Harradine, Batanero, & Rossman, 2011; Wild, Pfannkuch, Regan, & Horton, 2011). With 14–16 year olds, we could use informal approaches to reinforce ideas of sampling, sample variability and representativeness, sample statistics as estimates of population parameters, effect of sample size on sampling variability, and gradually develop the idea of sampling distribution in the students.

These activities, like those based on simulation that were described in the previous sections are part of what is today termed as *informal inference*, a didactical approach useful when students start to study inference. For example, Rossman (2008) presents activities within this approach, which involve drawing conclusions beyond a given data set. In these activities, students are asked to generalise their findings from a descriptive analysis in a sample to a larger group. For Rossman, informal inference includes procedures aimed to eliminate chance as an explanation for the observed data, through an argument that employs no formal probability.

Zieffler, Garfield, delMas, and Reading (2008) structure their examples by three different problem types that can be investigated by an informal approach: a) estimating a parameter for a population using data from a sample like we did in the tea-tasting situation (Section 5.2); b) comparing two or more samples and generalising a difference in the data to the population from which the samples were collected; and c) deciding between two chance models, which is more likely to produce some given data as we did in the quality control problem (Section 5.3.4).

5.6.2. Resampling Methods

Newer procedures in statistical inference are entirely based on computer-intensive methods where the results are obtained merely by simulation. They require fewer assumptions on the population distribution and the philosophy behind this approach is that *all* the information needed is available in the given data.

Borovcnik (2007a) gave an early account for the simplicity of this approach, while Cobb (2007) was advocating it to replace traditional statistical inference completely.[19]

The key idea is drawing many samples with replacement from the original sample at hand (bootstrap) or sampling from all possible permutations of the original sample values (rerandomisation). A statistic is computed in each of the different re-samples to obtain an empirical distribution and to draw inferences about the population.

In Section 5.3.3, we derived the empirical sampling distribution for the sample mean of IQ scores by repeatedly computing the mean of samples taken from a hypothetical normal distribution $N(100, 15)$ (see Figure 5.7). In the bootstrap method, we also collect samples, compute the sample mean for each of the samples. The difference of bootstrap with the earlier simulation method described in Section 5.3.3 is that, instead of drawing the samples from a hypothetical normal population, we obtain samples with replacement from the original sample of IQ scores. The resulting distribution is similar to the sampling distribution but conceptually different from it; it is usually called *bootstrap distribution*. In the following task, students are given a simulated random sample of 50 IQ scores of the normal distribution with mean 100 and a standard deviation of 15 to work with resampling.

> *Task 5.10.* Take 200 samples (of size 50) with replacement from the initial data set. Calculate the mean IQ from each re-sample and present the distribution of the mean. What can you conclude about the mean of the population from which the initial data is a sample?

We use Fathom to apply this resampling method upon 50 IQ scores (our original data set) from which we try to generalise the value of the mean (statistic) to the mean of IQ scores (parameter) in the population from which we collected the first sample. Using these data, we form a new collection "sample of IQ scores" where we collect another sample of 50 IQ scores that is drawn with replacement from the

---

[19] The ongoing discussion about the relative educational merits of the various approaches may be seen in Borovcnik (2013b), or Borovcnik and Kapadia (2015) who criticise the narrow limitations of a pure resampling approach towards statistical inference. Their arguments remain valid also from an educational point of view and restrict the educational value of its simplicity considerably. The main advantage of informal approaches based on resampling may be seen as an *intermediate step* before students can learn more formal inference.

Even for this step, a serious drawback might be seen in the complete reduction of probability to a pure frequentist concept, which has to be contrasted to Spiegelhalter and Gage (2015) who state that probability is a *virtual* concept and we need metaphors to be able to speak about it. However, one should take into account that in the practice of statistics all the approaches described in the chapter are useful in different circumstances and have their own limitations and advantages. For controversies that arise from any attempt to reduce to one approach see, e.g., Batanero (2000), or Borovcnik (1986a).

initial sample and compute the sample mean. Now, we repeat the process many times. The size of these samples should be identical to the size of the original data set. In our example each new sample consists of 50 elements.

In Figure 5.10, we collected 200 different samples with replacement from the original collection of IQ scores and plotted the distribution of all the sample means, which is our bootstrap distribution. We can use this distribution to test hypotheses about the mean in the population of IQ scores or to compute bootstrap intervals: for example, in Figure 5.10, we used the lines at the 5^{th} and 95^{th} percentile of the bootstrap distribution of sample means to find a 90% bootstrap interval for the population mean. Such an interval approximates the usual confidence interval quite well, especially if the first sample is not too small. From the resampled data, we calculate 96.55 and 101.24 for the bootstrap interval for the mean of the population.

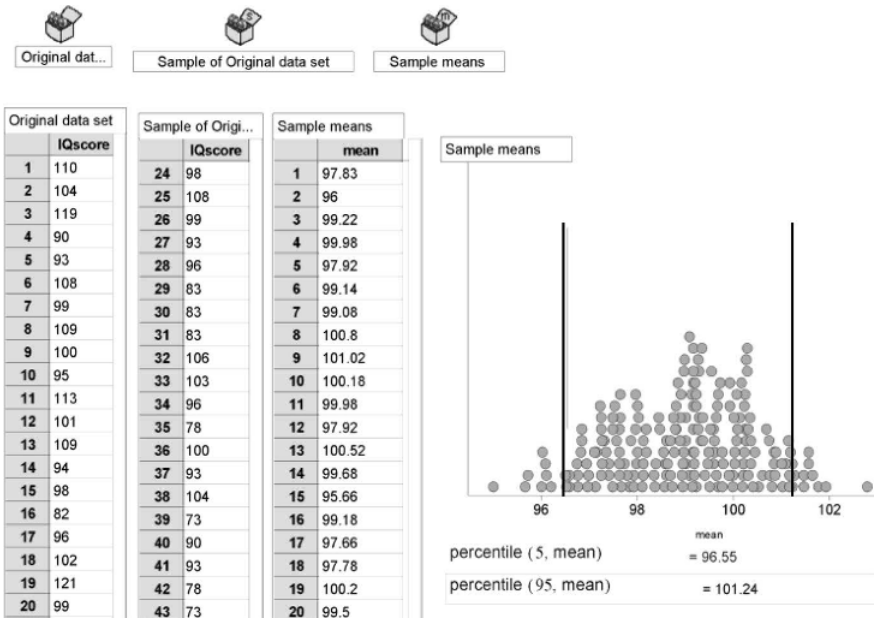


Figure 5.10. Bootstrap distribution for the sample mean with 200 resampled means

5.6.3. Updating a Prior Distribution for a Proportion

In the tea-tasting problem, instead of fixing a constant value p for the population proportion (Peter's capacity), we could analyse the problem from a Bayesian approach to inference, which is based on the subjectivist conception of probability. In this view we consider p as a random variable with a probability distribution (prior distribution) that models our *credibility* of the possible values of p before data are collected.

Learning from data with Bayes' rule
Suppose we gather some data D in an experiment and we want to compare several propositions A_i. We know the prior probability $P(A_i)$ for the propositions (our *degree of belief* in the proposition before collecting data) as well as the probabilities $P(D \mid A_i)$ or the *likelihood* to obtain the given data when proposition A_i is true (we know this likelihood for each proposition). In this situation, we use Bayes' rule (see Section 3.2) to compute the *posterior probability* $P(A_i \mid D)$ of the propositions using the prior probability and the observed data D as follows:

$$P(A_i \mid D) = \frac{P(A_i) \cdot P(D \mid A_i)}{k} \text{ with } k = \sum_{j=1}^{n} P(A_j) \cdot P(D \mid A_j).$$

We interpret the new probability on the left side of the formula as our *revised degree of belief* in the propositions in view of the particular data D (or evidence). The denominator k is known as the *total probability* and is the sum of all the products $P(A_j) \cdot P(D \mid A_j)$ of prior probabilities and likelihoods. In the following task, we fill in the details of the tea experiment.

Task 5.11. Suppose we reduce the possible values of p in the tea experiment to $p_i = 0.00, 0.10, 0.20, \ldots, 1.00$ and that we have no preference for any of these values so that all probabilities are equal, i.e., $P(p = p_i) = P(p_i) = 1/11$ (we model our credibility with equal probabilities). Peter had a success rate of 0.70 in 20 cups (he correctly classified 14 of the 20 cups). What is our updated knowledge about Peter's capacity? Determine the conditional probabilities for the values of p_i given $\hat{p} = 0.70$. Can we state that Peter has a better capacity than 0.5? What about a capacity of larger than 0.8? Can we ascribe Peter such an extraordinary capability to find out the order of tea and milk?

The initial assumption – prior probabilities – used to calculate the posterior
We assume that the propositions we want to check are the different possible values of the proportion $A_i = \{p = p_i\}$. If we insert the data $D = \hat{p}$ from the tea experiment, we just have to evaluate the probabilities of D under the various values of the parameter in order to calculate the posterior probabilities of the values p_i, i.e., the formula from above now reads as posterior distribution on p:

$$P(p_i \mid \hat{p}) = \frac{P(p_i) \cdot P(\hat{p} \mid p_i)}{k}.$$

Calculations in a spreadsheet

We illustrate the calculations with the data from the task and calculate the probability to observe 14 successes in $n = 20$ trials $(\hat{p} = 0.70)$ for the specific value of $p_i = 0.40$. In this case, the success probability equals 0.40 and the trials are independent; therefore, we have a binomial distribution for the number of successes and we get (this is a simple recall of the binomial probability function):

$$P\,(\hat{p} = 0.70 \mid 0.40) \;=\; \binom{20}{14} 0.40^{14}\; 0.60^{6} \;=\; 0.0049.$$

The further computations can be organised as shown in Table 5.3. Columns 1 and 2 include the potential values for the population proportion and their prior probabilities. The data likelihood for each value of *p* is included in column 3 and is computed with the binomial distribution as we did previously. Column 4 calculates the numerator in Bayes' formula and is the product of columns (2) and (3), i.e., the product of prior probabilities and likelihoods. Column (5) is the posterior probability that is obtained by dividing each cell in column (4) by the column total.

Table 5.3. Transforming the prior into a posterior distribution for $\hat{p} = 0.70$ *when Peter had 14 successes with 20 cups*

1	2	3	4	5
Potential values for *p*	Prior probability	Likeli-hood	Product 2 × 3	Posterior probability
0.00	1/11	0.0000	0.0000	0.0000
0.10	1/11	0.0000	0.0000	0.0000
0.20	1/11	0.0000	0.0000	0.0000
0.30	1/11	0.0002	0.0000	0.0005
0.40	1/11	0.0049	0.0004	0.0102
0.50	1/11	0.0370	0.0034	0.0776
0.60	1/11	0.1244	0.0113	0.2613
0.70	1/11	0.1916	0.0174	0.4026
0.80	1/11	0.1091	0.0099	0.2292
0.90	1/11	0.0089	0.0008	0.0186
1.00	1/11	0.0000	0.0000	0.0000
Total	1		0.0433	1.0000

The students can implement these computations in Excel and display columns 2 and 5 in bar graphs in order to perceive how data transforms the prior to a posterior distribution (Figure 5.11).

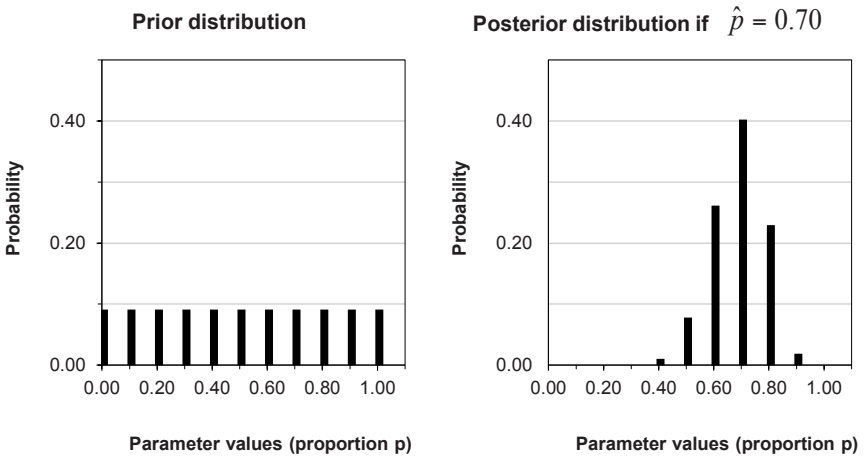


Figure 5.11. Updating a prior distribution for the population proportion if the sample proportion equals 0.70

Interpretation of the posterior probabilities

The mode of the posterior distribution on p lies at 0.70: this is the most probable value for Peter's capacity after we have seen the data (his success rate was 0.70 for twenty cups). The probability that Peter's capacity is greater than 0.50 is equal to 0.91 (the sum over the posterior probabilities for 0.60 to 1.00). We have quite a high credibility that Peter has a special ability. However, as the probability for values greater than 0.80 is only 0.02, we can ascribe to Peter a moderate or good ability to recognise the order of tea and milk but not an extraordinary expertise in doing so according to our posterior distribution.

It is important, to help students to discriminate between the *prior* and the *posterior* distribution for the parameter and to understand how Bayes' rule is used to update a prior distribution. The prior distribution is the list of possible values for a parameter and their associated probabilities fixed *before* collecting experimental data. The aim of Bayesian inference is updating this prior distribution with the new data.

The posterior distribution combines the prior distribution *and* the experimental data. It expresses our current status of knowledge of the parameter after the data were collected and may serve for different purposes: e.g., to compute the probability that the population proportion p takes specific values or lies in a specific interval. The posterior distribution can be used as a new prior distribution in further experiments where additional data will be used to update our belief about the population proportion in a sequential way.

REFERENCES

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14*(4)*,* 381–405.

Arteaga, P., Batanero, C., Contreras, J. M., & Cañadas, G. (2012). Understanding statistical graphs: A research survey. *BEIO, Boletín de Estadística e Investigación Operativa, 28*(3), 261–277.

Arteaga, P., Batanero, C., Contreras, J. M., & Cañadas, G. (2016). Evaluación de errores en la construcción de gráficos estadísticos elementales por futuros profesores (Assesing pre-service primary school teachers' errors in building statistical graphs). *Revista Latinoamericana de Matemática Educativa*, *19*(1), 15–40.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2013). *The Australian curriculum: Mathematics.* Sidney, NSW: Author.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83. Retrieved from www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Bakker.pdf

Barnett, V. (1982). *Teaching statistics in schools throughout the world*. Voorburg, The Netherlands: International Statistical Institute. Retrieved from www.iase-web.org/documents/book3/1Barnett.pdf

Barr, G. V. (1980). Some student ideas on the median and the mode. *Teaching Statistics, 2*(2), 38–41.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*(1–2), 75–97.

Batanero, C. (2001). *Didáctica de la estadística* (Statistics education). Granada, Spain: Grupo de Investigación en Educación Estadística, Universidad de Granada. Retrieved from www.ugr.es/~batanero/pages/ARTICULOS/didacticaestadistica.pdf

Batanero, C. (2016). Understanding randomness: Challenges for research and teaching. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the ninth congress of the european society for research in mathematics education* (pp. 34–49). Prague: European Society for Research in Mathematics Education. Retrieved from hal.archives-ouvertes.fr/hal-01280506/document

Batanero, C., Arteaga, P., & Ruiz, B. (2010). Statistical graphs produced by prospective teachers in comparing two distributions. In V. Durand-Guerrier, S. Soury-Lavergne, & F. Arzarello (Eds.), *Proceedings of the sixth congress of the european society for research in mathematics education* (pp. 368–377). Lyon: ERME. Retrieved from www.ife.ens-lyon.fr/publications/edition-electronique/cerme6/wg3-03-batanero.pdf

Batanero, C., Arteaga, P., Serrano, L., & Ruiz, B. (2014). Prospective primary school teachers' perception of randomness. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (pp. 345–366). New York, NY: Springer.

Batanero, C., Cañadas, G. R., Arteaga, P., & Gea, M. M. (2013). Psychology students' strategies and semiotic conflicts when assessing independence. In B. Ubuz, Ç. Haser, & M. A. Mariotti (Eds.), *Proceedings of the eighth congress of the european society for research in mathematics education* (pp. 756–765). Antalya, Turkey: ERME. Retrieved from www.mathematik.uni-dortmund.de/~erme/doc/CERME8/CERME8_2013_Proceedings.pdf

Batanero, C., Cañadas, G. R., Díaz, C., & Gea, M. M. (2015). Judgment of association between potential factors and associated risk in 2x2 tables: A study with psychology students. *The Mathematics Enthusiast*, *12*(1–3), 347–363. Retrieved from www.math.umt.edu/tmme/vol12no1thru3/25_Batanero_et_al.pdf

Batanero, C., Chernoff, E., Engel, J., Lee, H., & Sánchez, E. (2016). *Research on teaching and learning probability* (ICME-13, Topical Survey series). New York, NY: Springer.

Batanero, C., Cobo, B., & Díaz, C. (2003). Assessing secondary school students' understanding of averages. In M. A. Mariotti (Ed.), *Proceedings of the third conference of the european society for research in mathematics education* (pp. 1–9). Bellaria: ERME. Retrieved from www.mathematik.uni-dortmund.de/~erme/CERME3/Groups/TG5/TG5_batanero_cerme3.pdf

Batanero, C., & Díaz, C. (Eds.). (2011). *Estadística con proyectos* (Statistics with projects). Granada, Spain: Departamento de Didáctica de la Matemática. Retrieved from www.ugr.es/~batanero/pages/ARTICULOS/Libroproyectos.pdf

Batanero, C., Estepa, A., & Godino, J. D. (1997). Evolution of students' understanding of statistical association in a computer-based teaching environment. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 191–205). Voorburg: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/rt1996/15.Batanero.pdf

Batanero, C., Estepa, A., Godino, J. D., & Green D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27*(2), 151–169.

Batanero, C., Gea, M. M., Díaz, C., & Cañadas, G. R. (2014). Building high school pre-service teachers' knowledge to teach correlation and regression. In K. Makar (Ed.), *Proceedings of the ninth international conference on teaching statistics* (pp. 1–6). Flagstaff, AZ: International Statistical Institute. Retrieved from www.iase-web.org/icots/9/proceedings/pdfs/ICOTS9_1A3_BATANERO.pdf

Batanero, C., Godino, J. D., & Estepa, A. (1998). Building the meaning of statistical association through data analysis activities. In A. Olivier & K. Newstead (Eds.), *Proceedings of the 22nd conference of the international group for the psychology of mathematics education* (Vol. 1, pp. 221–236). Stellenbosch: University of Stellenbosch.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology*, *25*(4), 527–547.

Batanero, C., Henry, M., & Parzysz, B. (2005). The nature of chance and probability. In G. A. Jones (Ed.), *Exploring probability in school: challenges for teaching and learning* (pp. 15–37). New York, NY: Springer.

Batanero, C., Serrano, L., & Garfield, J. B. (1996). Heuristics and biases in secondary students' reasoning about probability. In L. Puig & A. Gutiérrez (Eds.), *Proceedings of the 20th conference of the international group for the psychology of mathematics education* (Vol. 2, pp. 43–50). Valencia, Spain: PME group.

Bayes, T. (1970). An essay towards solving a problem in the doctrine of chances. In E. S. Pearson & M. G. Kendall (Eds.), *Studies in the history of statistics and probability* (Vol. 1, pp. 131–153). London: Griffin (original work published in 1763).

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods 2*(2), 131–160.

Bellhouse, D. R. (2000). De Vetula: A medieval manuscript containing probability calculations. *International Statistical Review, 68*(2), 123–136.

Bennett, D. J. (1999). *Randomness.* Cambridge, MA: Harvard University Press.

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking.* Dordrecht, The Netherlands: Kluwer.

Bernoulli, J. (1987). *Ars conjectandi- 4ème* partie (N. Meunier, Trans.). Rouen: Institut de Recherche sur l'Enseignement Mathematique (original work published in 1713).

Bertin, J. (1967). *Sémiologie graphique* (Graphic semiology). Paris: Gauthier-Villars.

Biehler, R. (1986). Exploratory data analysis and the secondary stochastics curriculum. In R. Davidson & J. Swift (Eds.), *Proceedings of the second international conference on teaching statistics* (pp. 79–85). Victoria, Canada: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/icots2/Biehler.pdf

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes – Do we need a probabilistic revolution after we have taught data analysis? In J. B. Garfield (Ed.), *Research papers from ICOTS 4* (pp. 20–37). Minnesota, MN: University of Minnesota.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review, 65*(2), 167–189.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 643–689). New York, NY: Springer.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonom*y. New York, NY: Academic Press.

Blum, W. (1985). Anwendungsorientierter Mathematikunterricht in der didaktischen Diskussion (Applications orientated mathematics teaching in the didactic discussion). *Mathematische Semesterberichte, 32*(2), 195–232.

Blum, W., Galbraith, P. L., Henn, H. W., & Niss, M. (2007). *Modelling and applications in mathematics education.* New York, NY: Springer.

Borovcnik, M. (1986a). Klassische und Bayes'sche statistische Methoden. Ein kritischer Vergleich (Classical and Bayesian statistical methods – a critical comparison). *Österreichische Zeitschrift für Statistik und Informatik, 16*, 3–28.

Borovcnik, M. (1986b). Zur Rolle der beschreibenden Statistik. Teil I (On the role of descriptive statistics. Part I). *Mathematica Didactica, 9*, 177–191.

Borovcnik, M. (1987). Zur Rolle der beschreibenden Statistik. Teil II (On the role of descriptive statistics. Part II). *Mathematica Didactica*, *10*, 101–117.

Borovcnik, M. (1992). *Stochastik im Wechselspiel von Intuitionen und Mathematik* (Stochastics in the interplay between intuitions and mathematics). *Lehrbücher und Monographien zur Didaktik der Mathematik* (Vol. 10). Mannheim: Bibliographisches Institut.

Borovcnik, M. (1995, July). *Exploratory data analysis – A new approach to modelling.* Paper presented at the Seventh International Conference on Teaching Mathematical Modelling and Applications (ICTMA 7), Jordanstown, Ireland: ITM group. Retrieved from wwwg.uni-klu.ac.at/stochastik.schule/Boro/EXCEL/P_95_Modelling_EDA.pdf

Borovcnik, M. (2006a). Probabilistic and statistical thinking. In M. Bosch (Ed.), *Proceedings of the fourth congress of the european society for research in mathematics education* (pp. 484–506). Barcelona: ERME. Retrieved from www.mathematik.uni-dortmund.de/~erme/CERME4/CERME4_WG5.pdf

Borovcnik, M. (2006b). New paths in the teaching of statistics – with the help of spreadsheets. In A. Rossman & B. Chance (Eds.), *Proceedings of the seventh international conference on teaching statistics*. Salvador de Bahia, Brazil: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/icots7/C440.pdf

Borovcnik, M. (2007a, February). *On outliers, statistical risks, and a resampling approach towards statistical inference.* Paper presented at the Fifth Congress of the European Research in Mathematics Education (pp. 1–14), Larnaka, Cyprus. Retrieved from wwwg.uni-klu.ac.at/stochastik.schule/Boro/EXCEL/P_06_CERME_Borovcnik_outliers_stat_risks.pdf

Borovcnik, M. (2007b). The influence of software support on stochastics courses. In P. L. do Nascimento (Ed.), *Proceedings of the 56th session of the international statistical institute.* Lisbon: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/isi56/CPM80_Borovcnik.pdf

Borovcnik, M. (2011). Strengthening the role of probability within statistics curricula. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education: A Joint ICMI/IASE Study* (pp. 71–83). New York, NY: Springer.

Borovcnik, M. (2012a). Multiple perspectives on the concept of conditional probability. *Avances de Investigación en Educación Matemática, 2,* 5–27. Retrieved from www.aiem.es/index.php/aiem/article/view/32

Borovcnik, M. (2012b). Correlation and regression – errors and pitfalls. *Selçuk Journal of Applied Mathematics. Special issue*, 51–68.

Borovcnik, M. (2013a). E-learning or blended learning – Enriching statistics for business students. In S. Forbes & B. Phillips (Eds.), *Proceedings IASE/IAOS satellite conference on statistics education for*

progress (pp. 1–7). Macao: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/sat2013/IASE_IAOS_2013_Paper_1.3.3_Borovcnik.pdf

Borovcnik, M. (2013b). A comparative educational study of statistical inference. In X. He (Ed.), *Proceedings of the 59th world statistics congress of the international statistical institute* (pp. 1114–1119). Hong Kong: International Statistical Institute. Retrieved from www.2013.isiproceedings.org/Files/IPS112-P4-S.pdf

Borovcnik, M. (2014). Modelling and experiments – An interactive approach towards stochastics. In T. Wassong, D. Frischemeier, P. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen. Using tools for learning mathematics and statistics* (pp. 267–281). Berlin: Springer.

Borovcnik, M. (2015a). Central theorems of probability theory and their impact on probabilistic intuitions. *Didáctica de la Estadística, Probabilidad y Combinatoria, 2,* 15–35.

Borovcnik, M. (2015b). Risk and decision making: The "logic" of probability. *The Mathematics Enthusiast, 12*(1–3), 113–139.

Borovcnik, M., & Bentz, H. J. (1991). Empirical research in understanding probability. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: Probability in education* (pp. 73–105). Dordrecht, The Netherlands: Kluwer.

Borovcnik, M., & Kapadia, R. (2011). Modelling in probability and statistics – Key ideas and innovative examples. In J. Maaß & J. O'Donoghue (Eds.), *Real-world problems for secondary school students–Case studies* (pp. 1–44). Rotterdam, The Netherlands: Sense Publishers.

Borovcnik, M., & Kapadia, R. (2012, July). *Applications of probability: The Limerick experiments.* Paper presented at the 12th International Congress on Mathematics Education, Seoul. Retrieved from www.icme12.org/upload/UpFile2/TSG/0435.pdf

Borovcnik, M., & Kapadia, R. (2014a). A historical and philosophical perspective on probability. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: presenting plural perspectives* (pp. 7–34). New York, NY: Springer.

Borovcnik, M., & Kapadia, R. (2014b). From puzzles and paradoxes to concepts in probability. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (pp. 35–73). New York, NY: Springer.

Borovcnik, M, & Kapadia, R. (2015). A comparative study of statistical inference from an educational point of view. In P. L. do Nascimento (Ed.), *Proceedings of the 60th world statistics congress of the international statistical institute* (pp. 4401–4406). Rio de Janeiro, Brazil: International Statistical Institute.

Borovcnik, M., & Ossimitz, G. (1987). *Materialien zur beschreibenden Statistik und explorativen Datenanalyse* (Materials on descriptive statistics and exploratory data analysis). Wien: Hölder-Pichler-Tempsky.

Borovcnik, M., & Peard, R. (1996). Probability. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 239–288). Dordrecht, The Netherlands: Kluwer.

Bravais, A. (1846). Analyse mathématique sur les probabilités des erreurs de situation d'un point (Mathematical analysis of the probabilities for selecting a point). *Memoires presentés par divers savants à l'Académie Royale de l'Institut de France, 9*, 255–332.

Bruno, A., & Espinel, M. C. (2009). Construction and evaluation of histograms in teacher training. *International Journal of Mathematical Education in Science and Technology, 40*(4), 473–493.

Burgess, T. (2002). Investigating the 'data sense' of preservice teachers. In B. Phillips (Ed.), *Proceedings of the sixth international conference on teaching statistics* (pp. 1–6). Capetown, South Africa: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/icots6/6e4_burg.pdf

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 57–69). New York, NY: Springer.

Campbell, S. K. (1974). *Flaws and fallacies in statistical thinking.* Englewood Cliffs, NJ: Prentice-Hall.

Cañadas, G. R. (2012). *Comprensión intuitiva y aprendizaje formal de las tablas de contingencia en alumnos de psicología* (Intuitive understanding and formal learning of contingency tables by psychology students) (Unpublished Ph.D). Universidad de Granada, Spain.

Cañadas, G., Díaz, C., Batanero, C., & Estepa, A. (2013). Precisión de los estudiantes de psicología en la estimación de la asociación (Psychology students' estimation of association). *Bolema: Boletim de Educação Matemática, 27*(47), 759–778.

Carranza, P., & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Proceedings of the joint ICMI/IASE study: Teaching statistics in school mathematics. Challenges for teaching and teacher education*. Monterrey: International Commission on Mathematical Instruction and International Association for Statistical Education. Retrieved from www.iase-web.org/documents/papers/rt2008/T1P2_Carranza.pdf

Castro Sotos, A. E., Vanhoof, S., Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98–113.

CensusAtSchool (n.d.). *CensusAtSchool International*. Plymouth, UK: International Centre for Statistical Education (ICSE). Retrieved from www.censusatschool.org.uk/international-projects

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Amsterdam, The Netherlands: Kluwer.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*(3), 271–280.

Chernoff, E. J., & Sriraman B. (2014). (Eds.). *Probabilistic thinking: presenting plural perspectives*. New York, NY: Springer.

Çınlar, E. (2011). *Probability and stochastics*. New York, NY: Springer.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education, 1*(1), 1–15.

Common Core State Standards Initiative (CCSSI). (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers. Retrieved from www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Curcio, F. R. (1989). *Developing graph comprehension: Elementary and middle school activities*. Reston, VA: NCTM.

de Fermat, P., & Pascal, B. (1962). Correspondence. In F. N. David, *Games, gods and gambling* (pp. 229–253). London: Griffin (original work published in 1654).

de Finetti, B. (1974). *Theory of probability* (A. Machi & A. Smith, Trans.). London: Wiley.

de Finetti, B. (1992). Foresight: Its logical laws, its subjective sources (H. E. Kyburg Jr., Trans.). In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics. Volume I. Foundations and basic theory* (pp. 134–174). New York, NY & Berlin: Springer (original work published in 1937).

delMas, R., & Liu, Y. (2005). Exploring student's conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55–82. Retrieved from www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_delMas_Liu.pdf

Dubben, H. H., & Beck-Bornholdt, H. P. (2010). *Mit an Wahrscheinlichkeit grenzender Sicherheit. Logisches Denken und Zufall* (With a certainty close to probability. Logical thinking and chance). Reinbek: Rowohlt.

Engel, J., & Sedlmeier, P. (2011). Correlation and regression in the training of teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education: A Joint ICMI/IASE Study* (pp. 247–258). New York, NY: Springer.

Estepa, A., & Batanero, C. (1996). Judgments of correlation in scatter plots: Students' intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education, 4,* 25–41.

Estepa, A., Sánchez-Cobo, F. T., & Batanero, C. (1999). Students' understanding of regression lines. In O. Zaslavsky (Ed.), *Proceedings of the 23rd conference of the international group for the psychology of mathematics education* (Vol. 2, pp. 313–320). Haifa, Israel: Technion & Israel Institute of Technology.

Estrada, A., & Díaz, C. (2006). Computing probabilities from two way tables. An exploratory study with future teachers. In A. Rossman & B. Chance (Eds.), *Proceedings of the seventh international conference on teaching statistics* (pp. 1–4). Salvador de Bahia, Brazil: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/icots7/C413.pdf

Falk, R. (1986) Misconceptions of statistical significance. *Journal of Structural Learning, 9*(1), 83–96.

Falk, R., & Greenbaum, C. W. (1995) Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*(1), 75–98.

Falk, R., & Konold, C. (1992). The psychology of learning probability. In F. Sheldon & G. Sheldon (Eds.), *Statistics for the twenty-first century* (pp. 151–164). Washington, DC: The Mathematical Association of America.

Feller, W. (1957). An introduction to probability theory and its applications. New York, NY: Wiley.

Fidler, F., & Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. In *Proceedings of the 55th session of the international statistical institute* (pp. 1–5). Voorburg, The Netherlands: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/isi55/Fidler-Cumming.pdf

Finzer, W. (2007). *Fathom dynamic data software*. Emeryville, CA: Key Curriculum Press.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht, The Netherlands: Reidel.

Fischbein, E. (1987). *Intuition in science and mathematics. An educational approach.* Dordrecht, The Netherlands: Reidel.

Fisher, R. A. (1954). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd (original work published in 1925).

Fisher, R. A. (1956). Mathematics of a lady tasting tea. In J. R. Newman (Ed.), *The world of mathematics* (Vol. 3, pp. 1514–1521). New York, NY: Simon & Schuster.

Fisher, R. A. (1971). *The design of experiments*. Edinburgh: Oliver & Boyd (original work published in 1935).

Franklin, C., Kader. G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K-12 curriculum framework*. Alexandria, VA: American Statistical Association. Retrieved from www.amstat.org/Education/gaise/

Freedman, D., Pisani, R., & Purves, S. (1978). *Statistics*. London: W. W. Norton.

Freudenthal, H. (1972). "The empirical law of large numbers" or "The stability of frequencies". *Educational Studies in Mathematics, 4*(4), 484–490.

Friel, S., Curcio, F., & Bright, G. (2001). Making sense of graphs: critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124–158.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities (with discussion). *International Statistical Review, 70*(1), 1–51.

Gal, I. (2005). Towards "probability literacy" for all citizens: Building blocks and instructional dilemmas. In G. A. Jones (Ed.), *Exploring probability in school. Challenges for teaching and learning* (pp. 39–63). Dordrecht, The Netherlands: Kluwer.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Galton, F. (1892). *Finger prints*. London: Macmillan.

García, J. A., & Garret, A. J. (2006). On average and open-ended questions. In A. Rossman & B. Chance (Eds.), *Proceedings of the seventh international conference on teaching statistics*. Salvador de Bahia, Brazil: International Statistical Institute. Retrieved from www.stat.auckland.ac.nz/~iase/publications/17/C330.pdf

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice.* New York, NY: Springer.

Gattuso, L., & Ottaviani, M. G. (2011). Complementing mathematical thinking and statistical thinking in school mathematics. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 121–132). New York, NY: Springer.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Hamburg: Perthes und Besser.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice-versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). Chichester: Wiley.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York, NY: Simon & Schuster.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge: Cambridge University Press.

Goodchild, S. (1988). School pupils' understanding of average. *Teaching Statistics, 10*(3), 77–81.

Green, D. R. (1983). A survey of probabilistic concepts in 3000 students aged 11–16 years. In D. R. Grey, P. Holmes, V. Barnett, & G. M. Constable (Eds.), *Proceedings of the first international conference on teaching statistics* (Vol. 2, pp. 766–783). Sheffield: International Statistical Institute.

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning, 8*(1), 37–63.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability induction and statistical inference.* Cambridge: Cambridge University Press.

Hacking, I. (1990). *The taming of chance.* Cambridge: Cambridge University Press.

Hald, A. (2008). *A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935*. New York, NY: Springer.

Hall, J. (2011). Engaging teachers and students with real data: Benefits and challenges. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A joint ICMI/IASE Study* (pp. 335–346). New York, NY: Springer.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 235–246). New York, NY: Springer.

Hawkins, A., Jolliffe, F., & Glickman, L. (1992). *Teaching statistical concepts.* London: Longman.

Heitele, D. (1975). An epistemological view on fundamental stochastic ideas. *Educational Studies in Mathematics, 6*(2)*,* 187–205.

Hilbert, D. (1900). Mathematical problems. Lecture delivered before the *International Congress of Mathematicians at Paris in 1900*. Retrieved from www.aleph0.clarku.edu/~djoyce/hilbert/problems.html#

Holmes, P. (2001). Correlation: From picture to formula. *Teaching Statistics, 23*(3), 67–71.

Inhelder, B., & Piaget, J. (1955). *De la logique de l'enfant à la logique de l'adolescent: essai sur la construction des structures opératoires formelles* (From child's logic to adolescent's logic: Essay on the building of formal operative structures). Paris: Presses Universitaires de France.

Jacobbe, T., & Carvalho, C. (2011). Teachers' understanding of averages. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 199–209). New York, NY: Springer.

Jenkins, H. M., & Ward, W. C. (1965). Judgement of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79*(1), 1–17.

Jones, G. A. (Ed.). (2005). *Exploring probability in school: Challenges for teaching and learning*. New York, NY: Springer.

Jones, G. A., Langrall, C. W., & Mooney, E. S. (2007). Research in probability. Responding to classroom realities. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 909–956). Charlotte, NC: Information Age Publishing.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kolmogorov, A. N (1956). *Foundations of the theory of probability*. London: Chelsea (original work published 1933).

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59–98.

Konold, C., & Miller, C. (2011). *TinkerPlots 2.0: Dynamic data exploration*. Emeryville, CA: Key Curriculum Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259–289.

Krämer, W. (2007). *So lügt man mit Statistik* (This is the way to lie with statistics). München: Piper.

Landwehr, J. M., & Watkins, A. E. (1995). *Exploring data*. Palo Alto, CA: Dale Seymour.

Landwehr, J. M., Watkins, A. E., & Swift, J. (1987). *Exploring surveys and information from samples*. Palo Alto, CA: Dale Seymour.

Laplace, P. S. (1810). Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. *Mémoires de l' Academie de Scienes*, *1809*, 353–415.

Laplace, P. S. (1951). *Essai philosophique sur les probabilités* (A philosophical essay on probabilities). New York, NY: Dover (original work published in 1812).

Laplace, P. S. (1995). *Théorie analytique des probabilités* (Analytical theory of probabilities). Paris: Jacques Gabay (original work published in 1814).

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557–568.

Lee, C., & Meletiou-Mavrotheris, M. (2003). Some difficulties of learning histograms in introductory statistics. *Proceedings of the joint statistical meetings – Section on statistical education* (pp. 2326–2333). Alexandria, VA: American Statistical Association. Retrieved from www.statlit.org/PDF/2003LeeASA.pdf

Li, K. Y., & Shen, S .M. (1992). Students' weaknesses in statistical projects. *Teaching Statistics, 14*(1), 2–8.

Liu, Y., & Thompson, P. W. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies, 4*(2), 126–138.

Loosen, F., Lioen, M., & Lacante, M. (1985). The standard deviation: Some drawbacks of an intuitive approach. *Teaching Statistics, 7*(1), 2–5.

Lysø, K. O. (2008). Strengths and limitations of informal conceptions in introductory probability courses for future lower secondary teachers. In M. Borovcnik, D. Pratt, Y. Wu, & C. Batanero (Eds.), *Research and development in the teaching and learning of probability* (pp. 1–14). Monterrey: International Association for Statistical Education. Retrieved from www.stat.auckland.ac.nz/~iase/publications/icme11/ICME11_TSG13_11P_lyso.pdf

MacKenzie, D. A. (1981). *Statistics in Britain – 1865-1930 – The social construction of scientific knowledge*. Edinburgh: Edinburgh University Press.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1–2), 152–173.

Makar, K., Sousa, B. de, & Gould, R. (Eds.). (2014). *Proceedings of the ninth international conference on teaching statistics*. Flagstaff, AZ: International Statistical Institute.

Mayén, S., & Díaz, C. (2010). Is median an easy concept? Semiotic analysis of an open-ended task. In C. Reading (Ed.), *Proceedings of the eight international conference on teaching statistics*. Ljubljana,

Slovenia: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/icots8/ICOTS8_C265_MAYEN.pdf

McGillivray, H., & Pereira-Mendoza, L. (2011). Teaching statistical thinking through investigative projects. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 109–120). New York, NY: Springer.

Meiser, T., & Hewstone, M. (2006). Illusory and spurious correlations: Distinct phenomena or joint outcomes of exemplar-based category learning? *European Journal of Social Psychology, 36*(3), 315–336.

Mevarech, Z. R. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*(4), 415–429.

Ministerio de Educación y Ciencia, MEC. (2007). *Real decreto 1467/27, de 2 de noviembre, por el que se establece la estructura del bachillerato y se fijan sus enseñanzas mínimas* (Royal decree establishing the structure of the high-school curriculum). Madrid. Author.

Ministerio de Educación, Cultura y Deporte MECD. (2015). *Real decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la educación secundaria obligatoria y del bachillerato* (Royal decree establishing the structure of compulsory secondary-school and high-school curriculum). Madrid: Author.

Ministry of Education. (2007). *The New Zealand curriculum*. Wellington. NZ: Author. Retrieved from www.nzcurriculum.tki.org.nz/The-New-Zealand-Curriculum

Mises, R. V. (1919). Grundlagen der Wahrscheinlichkeitsrechnung (Foundations of probability theory). *Mathematische Zeitschrift, 5*, 52–99.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20–39.

Moore, D. S. (2010). *The basic practice of statistics* (5th ed.). New York, NY: Freeman.

Moore, D. S., & Cobb, G. W. (2000). Statistics and mathematics: Tension and cooperation. *The American Mathematical Monthly*, *107*(7), 615–630.

Moses, L. E. (1992). The reasoning of statistical inference. In D. C. Hoaglin & D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107–122). Washington, DC: Mathematical Association of America.

National Council of Teachers of Mathematics, NCTM. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Navas, J., Batanero, C., & Godino, J. (1997). Evaluación de concepciones sobre la noción de promedio en maestros de primaria en formación (Evaluation of primary school teachers' conceptions on averages). In H. Salmerón (Ed.), *Actas de las VII Jornadas LOGSE: Evaluación Educativa* (pp. 301–306). Granada, Spain: Universidad de Granada.

Neyman, J. (1950). *First course in probability and statistics*. New York, NY: Henry Holt.

Noll, J., & Shaughnessy, J. M. (2012). Aspects of students' reasoning about variation in empirical sampling distributions. *Journal for Research in Mathematics Education, 43*(5), 509–556.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London (A)*, *187*, 253–318.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine 5*(50), 157–175.

Pereira-Mendoza, L. (Ed.) (1993). *Introducing data analysis into schools: Who should teach it and how? Proceedings of the international statistical institute round table conference.* Lennoxville, Canada: International Statistical Institute.

Pereira-Mendoza, L., & Mellor, J. (1991). Student's concepts of bar graphs: some preliminary findings. In B. P. Dawkins, S. Carlisle, & D. Vere-Jones (Eds.), *Proceedings of the third international conference on teaching statistics* (pp. 150–157). Dunedin, New Zealand: International Statistical Institute. Retrieved www.iase-web.org/documents/papers/icots3/BOOK1/A2-5.pdf

Pérez-Echeverría, M. P. (1990). *Psicología del razonamiento probabilístico* (Psychology of probabilistic reasoning). Madrid: Ediciones de la Universidad Autónoma de Madrid.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17–46). Dordrecht, The Netherlands: Kluwer.

Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hasard chez l'enfant* (The origin of the idea of chance in children). Paris: Presses Universitaires de France.

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*(2), 191–204.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Pratt, D., Davies, N., & Connor, D. (2011). The role of technology in teaching and learning statistics. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education. A Joint ICMI/IASE Study* (pp. 97–107). New York, NY: Springer.

Raoult, J. P. (2013). La statistique dans l'enseignement secondaire en France. *Statistique et Enseignement, 4*(1), 55–69. Retrieved from www.publications-sfds.fr/index.php/StatEns/article/view/138

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review*. doi:10.1111/insr.12110

Ridgway, J., Nicholson, J., & McCusker, S. (2006). Reasoning with evidence – New opportunities in assessment. In A. Rossman & B. Chance (Eds.), *Proceedings of the seventh international conference on teaching of statistics*. Salvador de Bahia, Brazil: International Statistical Institute. Retrieved from iase-web.org/documents/papers/icots7/6D2_RIDG.pdf

Rosling, H. (2009). *Gapminder*. Stockholm: Gapminder Foundation. Retrieved from www.gapminder.org/

Ross, S. M. (2010a). *Introductory statistics* (3rd ed.). New York, NY: Academic Press.

Ross, S. M. (2010b). *Introduction to probability models* (10th ed.). Amsterdam, The Netherlands: Academic Press.

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, *7*(2), 5–19. Retrieved from www.iase-web.org/documents/SERJ/SERJ7(2)_Rossman.pdf

Russell, S. J., & Mokros, J. (1991). What's typical? Children's and teachers' ideas about average. In B. P. Dawkins, S. Carlisle, & D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 307–313). Dunedin, New Zealand: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/icots3/BOOK1/A9-3.pdf

Saldanha. L., & Thompson, P. (2007). Exploring connections between sampling distributions and statistical inference: An analysis of students' engagement and thinking in the context of instruction involving repeated sampling. *International Electronic Journal of Mathematics Education, 2*(3), 270–297.

Sánchez-Cobo, F. T., Estepa, A., & Batanero, C. (2000). Un estudio experimental de la estimación de la correlación a partir de diferentes representaciones (An experimental investigation on the estimation of correlation from different representations). *Enseñanza de las Ciencias, 18*(2), 297–310.

Scheaffer, R. (2006). Statistics and mathematics: On making a happy marriage. In G. Burrill (Ed.), *Thinking and reasoning with data and chance, 68th NCTM yearbook* (pp. 309–321). Reston, VA: National Council of Teachers of Mathematics.

Scholz, R. (1991). Psychological research on the probability concept and its acquisition. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: Probability in education* (pp. 213–254). Dordrecht, The Netherlands: Kluwer.

Scholz, R. W. (Ed.). (1983). *Decision making under uncertainty: Cognitive decision research, social interaction, development and epistemology*. Amsterdam, The Netherlands: North-Holland.

Schools Council Statistical Education Project. (1980). *Statistics in your world.* Slough, UK: Foulsham Educational.

Schuyten, G. (1991). Statistical thinking in psychology and education. In B. P. Dawkins, S. Carlisle, & D. Vere-Jones (Ed.), *Proceedings of the third international conference on teaching statistics* (Vol. 2., pp. 486–489). Dunedin, New Zealand: International Statistical Institute. Retrieved from www.iase-web.org/documents/papers/icots3/BOOK2/B9-5.pdf

Senior Secondary Board of South Australia (SSBSA). (2002). *Mathematical studies curriculum statement*. Adelaide, Australia: SSBSA.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465–494). New York, NY: Macmillan.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Charlotte, NC: Information Age Publishing.

Shaughnessy, J. M., Chance, B., & Kranendonk, H. (2009). *Focus in high school mathematics: Reasoning and sense making in statistics and probability.* Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004). Types of student reasoning on sampling tasks. In M. Johnsen Høines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Annual Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 177–184). Bergen: PME group. Retrieved from www.emis.de/proceedings/PME28/RR/RR045_Shaughnessy.pdf

Shaughnessy, J. M., Garfield, J. B., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (Vol. 1, pp. 205–237). Dordrecht: Kluwer.

Sies, H. (1988). A new parameter for sex education. *Nature, 332*, 495.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (B), 13*, 238–241.

Spiegelhalter, D. (2014). Comments on probabilistic thinking. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (Backcover). New York, NY: Springer.

Spiegelhalter, D., & Gage, J. (2014). What can education learn from real-world communication of risk and uncertainty? In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the ninth international conference on teaching statistics* (pp. 1–7). Flagstaff, AZ: International Statistical Institute. Retrieved from www.icots.info/9/proceedings/pdfs/ICOTS9_PL2_SPIEGELHALTER.pdf

Spiegelhalter, D., & Gage, J. (2015). What can education learn from real-world communication of risk and uncertainty? *The Mathematics Enthusiast, 12*(1–3), 4–10.

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education, 9*(3). Retrieved from www.amstat.org/publications/jse/v9n3/stanton.html

Stigler, S. (2002). The missing early history of contingency tables. *Annales de la Faculté des Sciences de Toulouse, 11*(4), 563–573.

Stirzaker, D. (2003). *Elementary probability*. Cambridge: Cambridge University Press.

Strauss, S., & Bichler, E. (1988). The development of children's concepts of the arithmetic average. *Journal for Research in Mathematics Education, 19*(1), 64–80.

Tanur, J. M. (Ed.). (1989). *Statistics: A guide to the unknown* (3rd ed.). San Francisco, CA: Holden Day.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.

Usiskin, Z. (2014). On the relationships between statistics and other subjects in the K-12 curriculum. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics* (pp. 1–12). Flagstaff, AZ: International Statistical Institute. Retrieved from www.iase-web.org/icots/9/proceedings/pdfs/ICOTS9_PL1_USISKIN.pdf

Vancsó, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *International Electronic Journal in Mathematics Education, 4*(3), 291–322. Retrieved from www.mathedujournal.com/dosyalar/IJEM_v4n3_7.pdf

Vere-Jones, D. (1995). The coming of age of statistical education. *International Statistical Review*, *63*(1), 3–23.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, *88*(421), 1–8.

Watson, J. M. (1997). Assessing statistical thinking using the media. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). Amsterdam: IOS Press.

Watson, J. M. (2006). *Statistical literacy at school. Growth and goals.* Mahwah, NJ: Lawrence Erlbaum.

Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology, 34*(1), 1–29.

Watson, J. M., & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2*(1–2), 11–50.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review, 67*(3), 223-265.

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference (with discussion). *Journal of the Royal Statistical Society (A)*, *174*(2), 247–295.

Wright, J. C., & Murphy, G. L. (1984). The utility of theories in intuitive statistics: The robustness of theory-based judgments. *Journal of Experimental Psychology: General, 113*(2), 301–322.

Wu, Y. (2004). Singapore secondary school students' understanding of statistical graphs. In J. Wisenbaker (Ed.), *Papers on Statistical Education presented at ICME-10* (pp 1–7). Copenhagen. Retrieved from iase-web.org/documents/papers/icme10/Yingkang.pdf

Zieffler, A., Garfield, J. B., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58. Retrieved from www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Zieffler.pdf

AUTHOR INDEX

SUBJECT INDEX