This is an annotated version of the text book with notes relevant to MATH1280 at University of the People.

Note to Bob: include my Chapter 7 notes and the Chapter 5-6 quick notes in any printed distribution.
... also add the explanations for the ch 5-6 practice test.

Introduction to Statistical Thinking
(With R, Without Calculus)

Benjamin Yakir, The Hebrew University

June, 2011

In memory of my father, Moshe Yakir, and the family he lost.

Preface

The target audience for this book is college students who are required to learn statistics, students with little background in mathematics and often no motivation to learn more. It is assumed that the students do have basic skills in using computers and have access to one. Moreover, it is assumed that the students are willing to actively follow the discussion in the text, to practice, and more importantly, to think.

Teaching statistics is a challenge. Teaching it to students who are required to learn the subject as part of their curriculum, is an art mastered by few. In the past I have tried to master this art and failed. In desperation, I wrote this book.

This book uses the basic structure of generic introduction to statistics course. However, in some ways I have chosen to diverge from the traditional approach. One divergence is the introduction of R as part of the learning process. Many have used statistical packages or spreadsheets as tools for teaching statistics. Others have used R in advanced courses. I am not aware of attempts to use R in introductory level courses. Indeed, mastering R requires much investment of time and energy that may be distracting and counterproductive for learning more fundamental issues. Yet, I believe that if one restricts the application of R to a limited number of commands, the benefits that R provides outweigh the difficulties that R engenders.

Another departure from the standard approach is the treatment of probability as part of the course. In this book I do not attempt to teach probability as a subject matter, but only specific elements of it which I feel are essential for understanding statistics. Hence, Kolmogorov's Axioms are out as well as attempts to prove basic theorems and a Balls and Urns type of discussion. On the other hand, emphasis is given to the notion of a *random variable* and, in that context, the *sample space*.

The first part of the book deals with descriptive statistics and provides probability concepts that are required for the interpretation of statistical inference. Statistical inference is the subject of the second part of the book.

The first chapter is a short introduction to statistics and probability. Students are required to have access to R right from the start. Instructions regarding the installation of R on a PC are provided.

The second chapter deals with data structures and variation. Chapter 3 provides numerical and graphical tools for presenting and summarizing the distribution of data.

The fundamentals of probability are treated in Chapters 4 to 7. The concept of a random variable is presented in Chapter 4 and examples of special types of random variables are discussed in Chapter 5. Chapter 6 deals with the Normal

random variable. Chapter 7 introduces sampling distribution and presents the Central Limit Theorem and the Law of Large Numbers. Chapter 8 summarizes the material of the first seven chapters and discusses it in the statistical context.

Chapter 9 starts the second part of the book and the discussion of statistical inference. It provides an overview of the topics that are presented in the subsequent chapter. The material of the first half is revisited.

Chapters 10 to 12 introduce the basic tools of statistical inference, namely point estimation, estimation with a confidence interval, and the testing of statistical hypothesis. All these concepts are demonstrated in the context of a single measurements.

Chapters 13 to 15 discuss inference that involve the comparison of two measurements. The context where these comparisons are carried out is that of regression that relates the distribution of a response to an explanatory variable. In Chapter 13 the response is numeric and the explanatory variable is a factor with two levels. In Chapter 14 both the response and the explanatory variable are numeric and in Chapter 15 the response in a factor with two levels.

Chapter 16 ends the book with the analysis of two case studies. These analyses require the application of the tools that are presented throughout the book.

This book was originally written for a pair of courses in the University of the People. As such, each part was restricted to 8 chapters. Due to lack of space, some important material, especially the concepts of correlation and statistical independence were omitted. In future versions of the book I hope to fill this gap.

Large portions of this book, mainly in the first chapters and some of the quizzes, are based on material from the online book "Collaborative Statistics" by Barbara Illowsky and Susan Dean (Connexions, March 2, 2010. `http://cnx.org/content/col10522/1.37/`). Most of the material was edited by this author, who is the only person responsible for any errors that where introduced in the process of editing.

Case studies that are presented in the second part of the book are taken from Rice Virtual Lab in Statistics can be found in their Case Studies section. The responsibility for mistakes in the analysis of the data, if such mistakes are found, are my own.

I would like to thank my mother Ruth who, apart from giving birth, feeding and educating me, has also helped to improve the pedagogical structure of this text. I would like to thank also Gary Engstrom for correcting many of the mistakes in English that I made.

This book is an open source and may be used by anyone who wishes to do so. (Under the conditions of the Creative Commons Attribution License (CC-BY 3.0).))

Jerusalem, June 2011   *Benjamin Yakir*

Contents

Part I

Introduction to Statistics

Chapter 1

Introduction

1.1 Student Learning Objectives

This chapter introduces the basic concepts of statistics. Special attention is given to concepts that are used in the first part of this book, the part that deals with graphical and numeric statistical ways to describe data (descriptive statistics) as well as mathematical theory of probability that enables statisticians to draw conclusions from data.

The course applies the widely used freeware programming environment for statistical analysis, known as R. In this chapter we will discuss the installation of the program and present very basic features of that system.

By the end of this chapter, the student should be able to:

- Recognize key terms in statistics and probability.

- Install the R program on an accessible computer.

- Learn and apply a few basic operations of the computational system R.

1.2 Why Learn Statistics?

You are probably asking yourself the question, "When and where will I use statistics?". If you read any newspaper or watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a news program on television, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact". Statistical methods can help you make the "best educated guess".

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques to analyze the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.
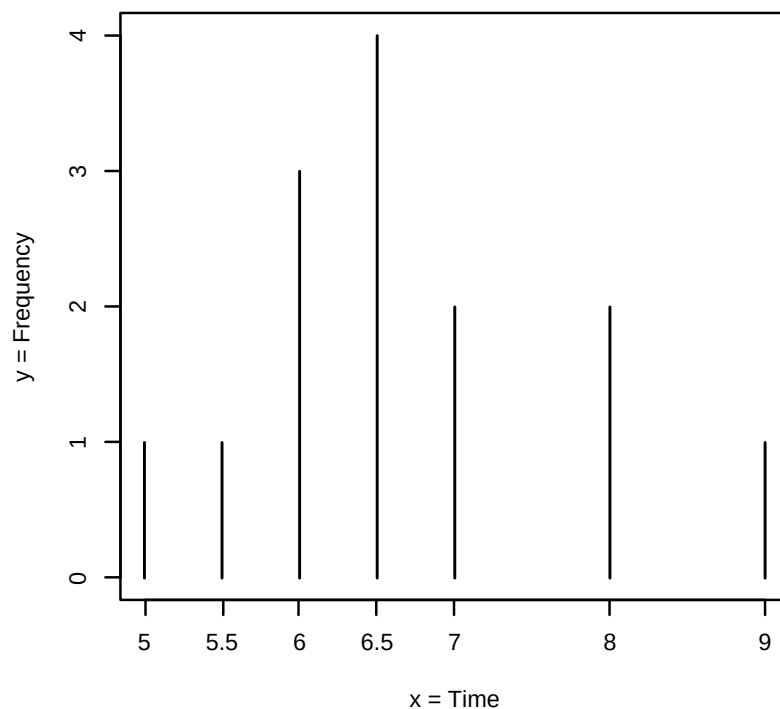
Figure 1.1: Frequency of Average Time (in Hours) Spent Sleeping per Night

Included in this chapter are the basic ideas and words of probability and statistics. In the process of learning the first part of the book, and more so in the second part of the book, you will understand that statistics and probability work together.

1.3   Statistics

The science of statistics deals with the collection, analysis, interpretation, and presentation of data. We see and use data in our everyday lives. To be able to use data correctly is essential to many professions and is in your own best self-interest.

For example, assume the average time (in hours, to the nearest half-hour) a group of people sleep per night has been recorded. Consider the following data:

$$5,\ 5.5,\ 6,\ 6,\ 6,\ 6.5,\ 6.5,\ 6.5,\ 6.5,\ 7,\ 7,\ 8,\ 8,\ 9\ .$$

In Figure 1.1 this data is presented in a graphical form (called a bar plot). A bar plot consists of a number axis (the x-axis) and bars (vertical lines) positioned

above the number axis. The length of each bar corresponds to the number of data points that obtain the given numerical value. In the given plot the frequency of average time (in hours) spent sleeping per night is presented with hours of sleep on the horizontal x-axis and frequency on vertical y-axis.

Think of the following questions:

- Would the bar plot constructed from data collected from a different group of people look the same as or different from the example? Why?

- If one would have carried the same example in a different group with the same size and age as the one used for the example, do you think the results would be the same? Why or why not?

- Where does the data appear to cluster? How could you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called descriptive statistics. Two ways to summarize data are by graphing and by numbers (for example, finding an average). In the second part of the book you will also learn how to use formal methods for drawing conclusions from "good" data. The formal methods are called inferential statistics. Statistical inference uses probabilistic concepts to determine if conclusions drawn are reliable or not.

Effective interpretation of data is based on good procedures for producing data and thoughtful examination of the data. In the process of learning how to interpret data you will probably encounter what may seem to be too many mathematical formulae that describe these procedures. However, you should always remember that the goal of statistics is not to perform numerous calculations using the formulae, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

1.4 Probability

Probability is the mathematical theory used to study uncertainty. It provides tools for the formalization and quantification of the notion of uncertainty. In particular, it deals with the chance of an event occurring. For example, if the different potential outcomes of an experiment are equally likely to occur then the probability of each outcome is taken to be the reciprocal of the number of potential outcomes. As an illustration, consider tossing a fair coin. There are two possible outcomes – a head or a tail – and the probability of each outcome is 1/2.

If you toss a fair coin 4 times, the outcomes may not necessarily be 2 heads and 2 tails. However, if you toss the same coin 4,000 times, the outcomes will be close to 2,000 heads and 2,000 tails. It is very unlikely to obtain more than 2,060 tails and it is similarly unlikely to obtain less than 1,940 tails. This is consistent with the expected theoretical probability of heads in any one toss. Even though the outcomes of a few repetitions are uncertain, there is a regular

pattern of outcomes when the number of repetitions is large. Statistics exploits this pattern regularity in order to make extrapolations from the observed sample to the entire population.

The theory of probability began with the study of games of chance such as poker. Today, probability is used to predict the likelihood of an earthquake, of rain, or whether you will get an "A" in this course. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not.

Although probability is instrumental for the development of the theory of statistics, in this introductory course we will not develop the mathematical theory of probability. Instead, we will concentrate on the philosophical aspects of the theory and use computerized simulations in order to demonstrate probabilistic computations that are applied in statistical inference.

1.5   Key Terms

In statistics, we generally want to study a population. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger population, we select a sample. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

A good sample is representative of the population. A good way to get a good sample is to select *randomly* from the *entire population.* In practice, we attempt to do this but often fall short of getting an ideal sample.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000 to 2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if the manufactured 16 ounce containers does indeed contain 16 ounces of the drink.

be sure to know the difference between a statistic and a parameter.

From the sample data, we can calculate a *statistic*. A statistic is a number that is a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic can be used as an estimate of a population *parameter*. A parameter is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a representative sample.

We will typically refer to the average as the "mean."

Two words that come up often in statistics are *average* and *proportion*. If you were to take three exams in your math classes and obtained scores of 86, 75, and 92, you calculate your average score by adding the three exam scores and dividing by three (your average score would be 84.3 to one decimal place). If, in

your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is 22/40 and the proportion of women students is 18/40. Average and proportion are discussed in more detail in later chapters.

1.6 The R Programming Environment

The R Programming Environment is a widely used open source system for statistical analysis and statistical programming. It includes thousands of functions for the implementation of both standard and exotic statistical methods and it is probably the most popular system in the academic world for the development of new statistical tools. We will use R in order to apply the statistical methods that will be discussed in the book to some example data sets and in order to demonstrate, via simulations, concepts associated with probability and its application in statistics.

The demonstrations in the book involve very basic R programming skills and the applications are implemented using, in most cases, simple and natural code. A detailed explanation will accompany the code that is used.

Learning R, like the learning of any other programming language, can be achieved only through practice. Hence, we strongly recommend that you not only read the code presented in the book but also run it yourself, in parallel to the reading of the provided explanations. Moreover, you are encouraged to play with the code: introduce changes in the code and in the data and see how the output changes as a result. One should not be afraid to experiment. At worst, the computer may crash or freeze. In both cases, restarting the computer will solve the problem . . .

You may download R from the R project home page http://www.r-project.org and install it on the computer that you are using[1].

1.6.1 Some Basic R Commands

R is an object-oriented programming system. During the session you may create and manipulate objects by the use of functions that are part of the basic installation. You may also use the R programming language. Most of the functions that are part of the system are themselves written in the R language and one may easily write new functions or modify existing functions to suit specific needs.

Let us start by opening the R Console window by double-clicking on the R icon. Type in the R Console window, immediately after the ">" prompt, the expression "1+2" and then hit the Return key. (Do not include the double quotation in the expression that you type!):

you do not type the ">" character. R shows this to indicate where your cursor is.

```
> 1+2
[1] 3
>
```

Tip: after you type a command into R, press ENTER to see the result. Then if you want to modify your previous command, press the UP arrow and edit the old command. After you enter many commands, you can press the UP arrow many times to recall old commands.

The prompt ">" indicates that the system is ready to receive commands. Writing an expression, such as "1+2", and hitting the Return key sends the expression

---

[1] Detailed explanation of how to install the system on an XP Windows Operating System may be found here: http://pluto.huji.ac.il/~msby/StatThink/install_R_WinXP.html

to be executed. The execution of the expression may produce an object, in this case an object that is composed of a single number, the number "3".

Whenever required, the R system takes an action. If no other specifications are given regarding the required action then the system will apply the pre-programmed action. This action is called the *default* action. In the case of hitting the Return key after the expression that we wrote the default is to display the produced object on the screen.

Next, let us demonstrate R in a more meaningful way by using it in order to produce the bar-plot of Figure 1.1. First we have to input the data. We will produce a sequence of numbers that form the data[2]. For that we will use the function "c" that combines its arguments and produces a sequence with the arguments as the components of the sequence. Write the expression:

```
> c(5,5.5,6,6,6,6.5,6.5,6.5,6.5,7,7,8,8,9)
```

at the prompt and hit return. The result should look like this:

```
> c(5,5.5,6,6,6,6.5,6.5,6.5,6.5,7,7,8,8,9)
 [1] 5.0 5.5 6.0 6.0 6.0 6.5 6.5 6.5 6.5 7.0 7.0 8.0 8.0 9.0
>
```

The function "c" is an example of an R function. A function has a name, "c" in this case, that is followed by brackets that include the input to the function. We call the components of the input the *arguments* of the function. Arguments are separated by commas. A function produces an output, which is typically an R object. In the current example an object of the form of a sequence was created and, according to the default application of the system, was sent to the screen and not saved.

If we want to create an object for further manipulation then we should save it and give it a name. For example, it we want to save the vector of data under the name "X" we may write the following expression at the prompt (and then hit return):

```
> X <- c(5,5.5,6,6,6,6.5,6.5,6.5,6.5,7,7,8,8,9)
>
```

The arrow that appears after the "X" is produced by typing the less than key "<" followed by the minus key "-". This arrow is the assignment operator.

Observe that you may save typing by calling and editing lines of code that were processes in an earlier part of the session. One may browse through the lines using the up and down arrows on the right-hand side of the keyboard and use the right and left arrows to move along the line presented at the prompt. For example, the last expression may be produced by finding first the line that used the function "c" with the up and down arrow and then moving to the beginning of the line with the left arrow. At the beginning of the line all one has to do is type "X <- " and hit the Return key.

Notice that no output was sent to the screen. Instead, the output from the "c" function was assigned to an object that has the name "X". A new object by the given name was formed and it is now available for further analysis. In order to verify this you may write "X" at the prompt and hit return:

---

[2]In R, a sequence of numbers is called a *vector*. However, we will use the term *sequence* to refer to vectors.
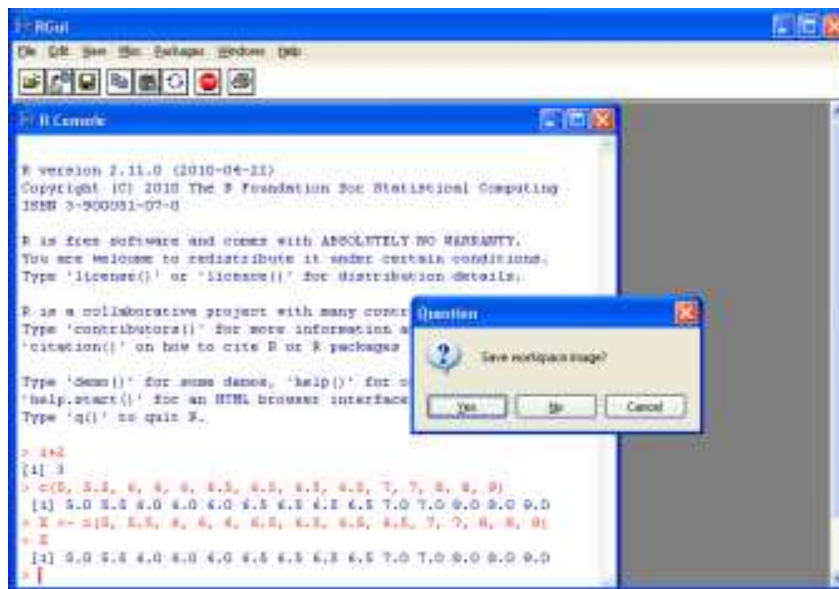
Figure 1.2: Save Workspace Dialog

```
> X
 [1] 5.0 5.5 6.0 6.0 6.0 6.5 6.5 6.5 6.5 7.0 7.0 8.0 8.0 9.0
```

The content of the object "X" is sent to the screen, which is the default output. Notice that we have not changed the given object, which is still in the memory.

The object "X" is in the memory, but it is not saved on the hard disk. With the end of the session the objects created in the session are erased unless specifically saved. The saving of all the objects that were created during the session can be done when the session is finished. Hence, when you close the R Console window a dialog box will open (See the screenshot in Figure 1.2). Via this dialog box you can choose to save the objects that were created in the session by selecting "Yes", not to save by selecting the option "No", or you may decide to abort the process of shutting down the session by selecting "Cancel". If you save the objects then they will be uploaded to the memory the next time that the R Console is opened.

We used a capital letter to name the object. We could have used a small letter just as well or practically any combination of letters. However, you should note that R distinguishes between capital and small letter. Hence, typing "x" in the console window and hitting return will produce an error message:

```
> x
Error: object "x" not found
```

If you are not accustomed to computer programming, be sure to read what R tells you. If you see an Error message, it probably means that you made a mistake. Try editing your command so that it is exactly the same as an example in the book or the notes.

An object named "x" does not exist in the R system and we have not created such object. The object "X", on the other hand, does exist.

Names of functions that are part of the system are fixed but you are free to choose a name to objects that you create. For example, if one wants to create

an object by the name "`my.vector`" that contains the numbers 3, 7, 3, 3, and -5 then one may write the expression "`my.vector <- c(3,7,3,3,-5)`" at the prompt and hit the Return key.

If we want to produce a table that contains a count of the frequency of the different values in our data we can apply the function "`table`" to the object "`X`" (which is the object that contains our data):

```
> table(X)
X
  5 5.5   6 6.5   7   8   9
  1   1   3   4   2   2   1
```

Notice that the output of the function "`table`" is a table of the different levels of the input vector and the frequency of each level. This output is yet another type of an object.

The bar-plot of Figure 1.1 can be produced by the application of the function "`plot`" to the object that is produced as an output of the function "`table`":

```
> plot(table(X))
```

Observe that a graphical window was opened with the target plot. The plot that appears in the graphical window should coincide with the plot in Figure 1.3. This plot is practically identical to the plot in Figure 1.1. The only difference is in the names given to the access. These names were changed in Figure 1.1 for clarity.

Clearly, if one wants to produce a bar-plot to other numerical data all one has to do is replace in the expression "`plot(table(X))`" the object "`X`" by an object that contains the other data. For example, to plot the data in "`my.vector`" you may use "`plot(table(my.vector))`".

1.7   Solved Exercises

Question 1.1. A potential candidate for a political position in some state is interested to know what are her chances to win the primaries of her party and be selected as parties candidate for the position. In order to examine the opinions of her party voters she hires the services of a polling agency. The polling is conducted among 500 registered voters of the party. One of the questions that the pollsters refers to the willingness of the voters to vote for a female candidate for the job. Forty two percent of the people asked said that they prefer to have a women running for the job. Thirty eight percent said that the candidate's gender is irrelevant. The rest prefers a male candidate. Which of the following is (i) a population (ii) a sample (iii) a parameter and (iv) a statistic:

1. The 500 registered voters.

2. The percentage, among all registered voters of the given party, of those that prefer a male candidate.

3. The number 42% that corresponds to the percentage of those that prefer a female candidate.

4. The voters in the state that are registered to the given party.
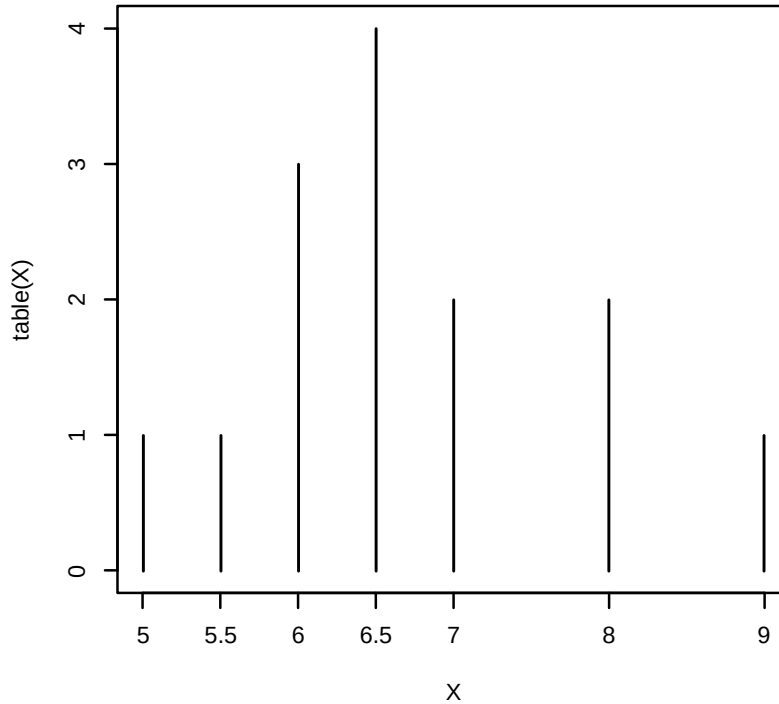
Figure 1.3: The Plot Produced by the Expression "`plot(table(X))`"

Solution (to Question 1.1.1): According to the information in the question the polling was conducted among 500 registered voters. The 500 registered voters corresponds to the sample.

Solution (to Question 1.1.2): The percentage, among all registered voters of the given party, of those that prefer a male candidate is a parameter. This quantity is a characteristic of the population.

Solution (to Question 1.1.3): It is given that 42% of the sample prefer a female candidate. This quantity is a numerical characteristic of the data, of the sample. Hence, it is a statistic.

Solution (to Question 1.1.4): The voters in the state that are registered to the given party is the target population.

Question 1.2. The number of customers that wait in front of a coffee shop at the opening was reported during 25 days. The results were:

$$4, 2, 1, 1, 0, 2, 1, 2, 4, 2, 5, 3, 1, 5, 1, 5, 1, 2, 1, 1, 3, 4, 2, 4, 3 .$$
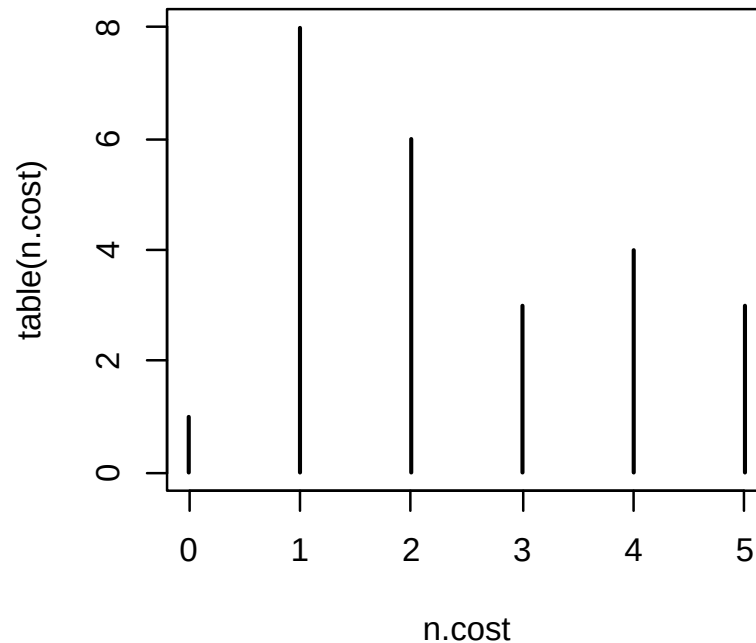
Figure 1.4: The Plot Produced by the Expression "`plot(table(n.cost))`"

1. Identify the number of days in which 5 costumers where waiting.

2. The number of waiting costumers that occurred the largest number of times.

3. The number of waiting costumers that occurred the least number of times.

Solution (to Question 1.2): One may read the data into R and create a table using the code:

be sure that you understand what the R output means. If you are not sure, check the answers on the next page or ask your instructor.

```
> n.cost <- c(4,2,1,1,0,2,1,2,4,2,5,3,1,5,1,5,1,2,1,1,3,4,2,4,3)
> table(n.cost)
n.cost
0 1 2 3 4 5
1 8 6 3 4 3
```

For convenience, one may also create the bar plot of the data using the code:

```
> plot(table(n.cost))
```

The bar plot is presented in Figure 1.4.

Solution (to Question 1.2.1): The number of days in which 5 costumers where waiting is 3, since the frequency of the value "5" in the data is 3. That can be seen from the table by noticing the number below value "5" is 3. It can also be seen from the bar plot by observing that the hight of the bar above the value "5" is equal to 3.

Solution (to Question 1.2.2): The number of waiting costumers that occurred the largest number of times is 1. The value "1" occurred 8 times, more than any other value. Notice that the bar above this value is the highest.

Solution (to Question 1.2.3): The value "0", which occurred only once, occurred the least number of times.

1.8 Summary

Glossary

Data: A set of observations taken on a sample from a population.

Statistic: A numerical characteristic of the data. A statistic estimates the corresponding population parameter. For example, the average number of contribution to the course's forum for this term is an estimate for the average number of contributions in all future terms (parameter).

Statistics The science that deals with processing, presentation and inference from data.

Probability: A mathematical field that models and investigates the notion of randomness.

Discuss in the forum

A sample is a subgroup of the population that is supposed to represent the entire population. In your opinion, is it appropriate to attempt to represent the entire population only by a sample?

When you formulate your answer to this question it may be useful to come up with an example of a question from you own field of interest one may want to investigate. In the context of this example you may identify a target population which you think is suited for the investigation of the given question. The appropriateness of using a sample can be discussed in the context of the example question and the population you have identified.

Chapter 2

Sampling and Data Structures

2.1 Student Learning Objectives

In this chapter we deal with issues associated with the data that is obtained from a sample. The variability associated with this data is emphasized and critical thinking about validity of the data encouraged. A method for the introduction of data from an external source into `R` is proposed and the data types used by `R` for storage are described. By the end of this chapter, the student should be able to:

- Recognize potential difficulties with sampled data.
- Read an external data file into `R`.
- Create and interpret frequency tables.

2.2 The Sampled Data

The aim in statistics is to learn the characteristics of a population on the basis of a sample selected from the population. An essential part of this analysis involves consideration of variation in the data.

2.2.1 <u>Variation</u> in Data

Variation is given a central role in statistics. To some extent the assessment of variation and the quantification of its contribution to uncertainties in making inference is the statistician's main concern.

Variation is present in any set of data. <u>For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid.</u> In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

$$15.8, \ 16.1, \ 15.2, \ 14.8, \ 15.8, \ 15.9, \ 16.0, \ 15.5 \ .$$

Measurements of the amount of beverage in a 16-ounce may vary because the conditions of measurement varied or because the exact amount, 16 ounces of

liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that if an investigator collects data, the data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two investigators or more, are taking data from the same source and get very different results, it is time for them to reevaluate their data-collection methods and data recording accuracy.

2.2.2   Variation in Samples

Two or more samples from the same population, all having the same characteristics as the population, may nonetheless be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students sleep each night and use all students at their college as the population. Doreen may decide to sample randomly a given number of students from the entire body of collage students. Jung, on the other hand, may decide to sample randomly a given number of classes and survey all students in the selected classes. Doreen's method is called *random sampling* whereas Jung's method is called *cluster sampling*. Doreen's sample will be different from Jung's sample even though both samples have the characteristics of the population. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Ramdom sampling is usually the easiest way to get a good sampling. A proper cluster sample requires good knowledge of the population and some advanced statistics.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (say, the average amount of time a student sleeps) would be closer to the actual population average. But still, their samples would be, most probably, different from each other.

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1200 to 1500 observations are considered large enough and good enough if the survey is random and is well done. The theory of statistical inference, that is the subject matter of the second part of this book, provides justification for these claims.

2.2.3   Frequency

The primary way of summarizing the variability of data is via the frequency distribution. Consider an example. Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below:

$$5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3 .$$

Let us create an R object by the name "work.hours" that contains these data:

This is "R-speak" that we will use in class... "create an R object."

```
> work.hours <- c(5,6,3,3,2,4,7,5,2,3,5,6,5,4,4,3,5,2,5,3)
```

Next, let us create a table that summarizes the different values of working hours and the frequency in which these values appear in the data:

```
> table(work.hours)
work.hours
2 3 4 5 6 7
3 5 3 6 2 1
```

this means: table()

Recall that the function "`table`" takes as input a sequence of data and produces as output the frequencies of the different values.

We may have a clearer understanding of the meaning of the output of the function "`table`" if we presented outcome as a frequency listing the different data values in ascending order and their frequencies. For that end we may apply the function "`data.frame`" to the output of the "`table`" function and obtain:

```
> data.frame(table(work.hours))
  work.hours Freq
2          2    3
3          3    5
4          4    3
5          5    6
6          6    2
7          7    1
```

This is one way to show the data that comes from the table() command. It shows column titles that explain what the numbers mean. The "Freq" column means "frequency" or the count of how many times the value in the "work.hours" column occurred in the data.

A frequency is the number of times a given datum occurs in a data set. According to the table above, there are three students who work 2 hours, five students who work 3 hours, etc. The total of the frequency column, 20, represents the total number of students included in the sample.

The function "`data.frame`" transforms its input into a data frame, which is the standard way of storing statistical data. We will introduce data frames in more detail in Section 2.3 below.

A relative frequency is the fraction of times a value occurs. To find the relative frequencies, divide each frequency by the total number of students in the sample – 20 in this case. Relative frequencies can be written as fractions, percents, or decimals.

As an illustration let us compute the relative frequencies in our data:

```
> freq <- table(work.hours)
> freq
work.hours
2 3 4 5 6 7
3 5 3 6 2 1
> sum(freq)
[1] 20
> freq/sum(freq)
work.hours
   2    3    4    5    6    7
0.15 0.25 0.15 0.30 0.10 0.05
```

A relative frequency shows the proportion of how many times a value appears in the data. If you multiply the relative frequency by 100, it is a percent.

The relative frequency of the number 3 is .25, which means that 25% of the values in the data are the number 3.

We stored the frequencies in an object called "`freq`". The content of the object are the frequencies 3, 5, 3, 6, 2 and 1. The function "`sum`" sums the components of its input. The sum of the frequencies is the sample size , the total number of students that responded to the survey, which is 20. Hence, when we apply the function "`sum`" to the object "`freq`" we get 20 as an output.

The outcome of dividing an object by a number is a division of each element in the object by the given number. Therefore, when we divide "`freq`" by "`sum(freq)`" (the number 20) we get a sequence of relative frequencies. The first entry to this sequence is $3/20 = 0.15$, the second entry is $5/20 = 0.25$, and the last entry is $1/20 = 0.05$. The sum of the relative frequencies should always be equal to 1:

```
> sum(freq/sum(freq))
[1] 1
```

The <u>cumulative relative frequency</u> is the accumulation of previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency of the current value. Alternatively, we may apply the function "cumsum" to the sequence of relative frequencies:

```
> cumsum(freq/sum(freq))
   2    3    4    5    6    7
0.15 0.40 0.55 0.85 0.95 1.00
```

[handwritten note in blue: A **cumulative** relative frequency shows a "running total" of all the prior relative frequencies. It should always end with 1 to indicate that 100% of the numbers have been shown.]

[handwritten note in blue (left margin): You should know how to look at the cumulative relative frequency and calculate the relative frequencies shown on the prior page. Try it now!]

Observe that the cumulative relative frequency of the smallest value 2 is the frequency of that value (0.15). The cumulative relative frequency of the second value 3 is the sum of the relative frequency of the smaller value (0.15) and the relative frequency of the current value (0.25), which produces a total of $0.15 + 0.25 = 0.40$. Likewise, for the third value 4 we get a cumulative relative frequency of $0.15 + 0.25 + 0.15 = 0.55$. The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

The computation of the cumulative relative frequency was carried out with the aid of the function "cumsum". This function takes as an input argument a numerical sequence and produces as output a numerical sequence of the same length with the cumulative sums of the components of the input sequence.

2.2.4 Critical Evaluation

Inappropriate methods of sampling and data collection may produce samples that do not represent the target population. A naïve application of statistical analysis to such data may produce misleading conclusions.

Consequently, it is important to evaluate critically the statistical analyses we encounter before accepting the conclusions that are obtained as a result of these analyses. Common problems that occurs in data that one should be aware of include:

[handwritten note in blue (left margin): Biased Samples]

Problems with Samples: <u>A sample should be representative of the popula-</u>tion. A sample that is <u>not</u> representative of the population is <u>biased</u>. Biased samples may produce results that are inaccurate and not valid.

[handwritten note in blue (left margin): these are things that can cause your sample to be biased.]

Data Quality: Avoidable errors may be introduced to the data via inaccurate handling of forms, mistakes in the input of data, etc. Data should be cleaned from such errors as much as possible.

Self-Selected Samples: Responses only by people who choose to respond, such as call-in surveys, that are often biased.

Sample Size Issues: Samples that are too small may be unreliable. Larger samples, when possible, are better. In some situations, small samples are unavoidable and can still be used to draw conclusions. Examples: Crash testing cars, medical testing for rare conditions.

[handwritten note in blue (left margin): Determining a good sample size requires some advanced calculations that we do not cover in MATH1280.]

Undue Influence: Collecting data or asking questions in a way that influences the response.

Causality: A relationship between two variables does not mean that one causes the other to occur. They may both be related (correlated) because of their relationship to a third variable.

Self-Funded or Self-Interest Studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.

Misleading Use of Data: Improperly displayed graphs and incomplete data.

Confounding: Confounding in this context means confusing. When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

2.3   Reading Data into R

In the examples so far the size of the data set was very small and we were able to input the data directly into R with the use of the function "c". In more practical settings the data sets to be analyzed are much larger and it is very inefficient to enter them manually. In this section we learn how to upload data from a file in the Comma Separated Values (CSV) format.

The file "ex1.csv" contains data on the sex and height of 100 individuals. This file is given in the CSV format. The file can be found on the internet at http://pluto.huji.ac.il/~msby/StatThink/Datasets/ex1.csv. We will discuss the process of reading data from a file into R and use this file as an illustration.

2.3.1   Saving the File and Setting the Working Directory

Before the file is read into R you may find it convenient to obtain a copy of the file and store it in some directory on the computer and read the file from that directory. We recommend that you create a special directory in which you keep all the material associated with this course. In the explanations provided below we assume that the directory to which the file is stored in called "IntroStat". (See Figure 2.1)

Files in the CSV format are ordinary text files. They can be created manually or as a result of converting data stored in a different format into this particular format. A convenient way to produce, browse and edit CSV files is by the use of a standard electronic spreadsheet programs such as Excel or Calc. The Excel spreadsheet is part of the Microsoft's Office suite. The Calc spreadsheet is part of OpenOffice suite that is freely distributed by the OpenOffice Organization.

Opening a CSV file by a spreadsheet program displays a spreadsheet with the content of the file. Values in the cells of the spreadsheet may be modified directly. (However, when saving, one should pay attention to save the file in the CVS format.) Similarly, new CSV files may be created by the entering of the data in an empty spreadsheet. The first row should include the name of the variable, preferably as a single character string with no empty spaces. The

---

Margin notes:

If you are new to computer programming, it might be difficult for you to read data into R for the first time, but it is important because we will use the data files for graded assignments.

Check the MATH1280 home page for a link to "Data Files." You should find some tips for importing the files that we will need for class.

See the MATH1280 home page and announcements for doing this.

DO NOT OPEN THE CSV FILES IN EXCEL! STUDENTS WHO DO THIS OFTEN RUIN THE FILES BECAUSE EXCEL WILL TEMPT YOU TO SAVE THE FILE IN THE WRONG FORMAT.

IF EXCEL OPENS WHEN YOU TRY TO DOWNLOAD A CSV FILE, CANCEL THE DOWNLOAD AND TRY RIGHT-CLICKING (or control-click on a Mac) TO DOWNLOAD THE FILE DIRECTLY.

The main goal when downloading a CSV file is to put it in your "working directory." To find your working directory, run the R command: getwd()

Read this

After you put the CSV files into the working directory, run the dir() command. If you do not see your CSV files listed in the output of the dir() command in R, then your files are not in the right place and you need to move them using your regular file manger (e.g., Windows explorer or Finder). If you get import errors when running the read.csv() command, you probably have a typo in file name or the file is not in your working directory.
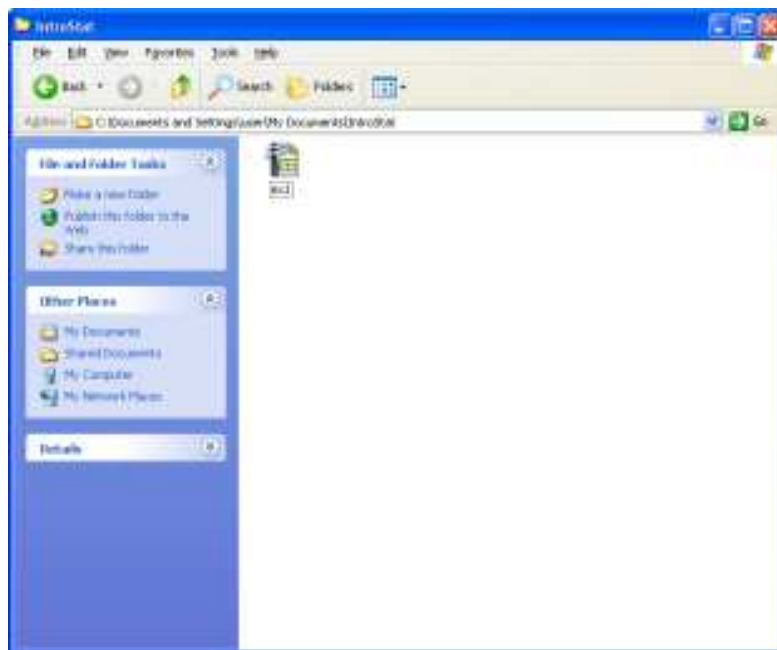
Figure 2.1: The File "read.csv"

following rows may contain the data values associated with this variable. When saving, the spreadsheet should be saved in the CSV format by the use of the "Save by name" dialog and choosing there the option of CSV in the "Save by Type" selection.

After saving a file with the data in a directory, R should be notified where the file is located in order to be able to read it. A simple way of doing so is by setting the directory with the file as R's *working directory*. The working directory is the first place R is searching for files. Files produced by R are saved in that directory. In Windows, during an active R session, one may set the working directory to be some target directory with the "File/Change Dir..." dialog. This dialog is opened by selecting the option "File" on the left hand side of the ruler on the top of the R Console window. Selecting the option of "Change Dir..." in the ruler that opens will start the dialog. (See Figure 2.2.) Browsing via this dialog window to the directory of choice, selecting it, and approving the selection by clicking the "OK" bottom in the dialog window will set the directory of choice as the working directory of R.

Rather than changing the working directory every time that R is opened one may set a selected directory to be R's working directory on opening. Again, we demonstrate how to do this on the XP Windows operating system.

The R icon was added to the Desktop when the R system was installed. The R Console is opened by double-clicking on this icon. One may change the properties of the icon so that it sets a directory of choice as R's working directory.

In order to do so click on the icon with the mouse's **right** bottom. A menu

As long as you know where your working directory is, it is not essential to change your working directory.
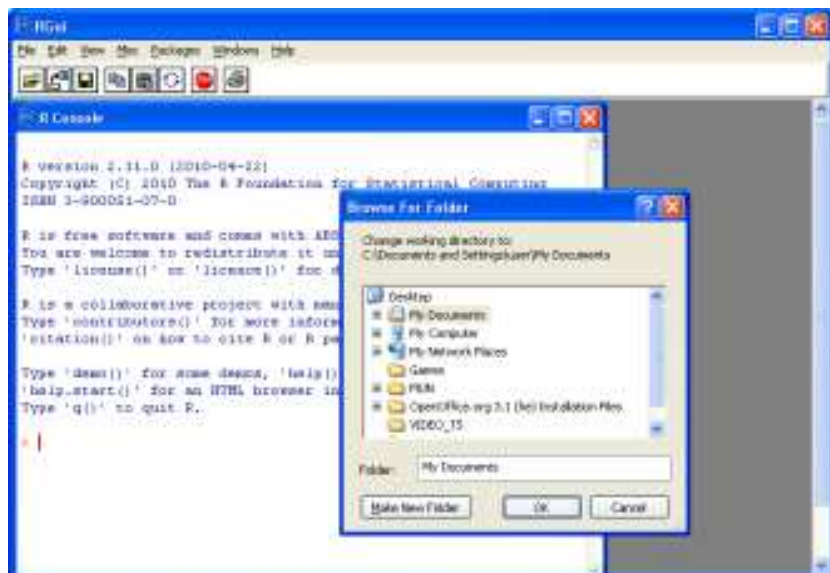
Figure 2.2: Changing The Working Directory

opens in which you should select the option "`Properties`". As a result, a dialog window opens. (See Figure 2.3.) Look at the line that starts with the words "`Start in`" and continues with a name of a directory that is the current working directory. The name of this directory is enclosed in double quotes and is given with it's full path, i.e. its address on the computer. This name and path should be changed to the name and path of the directory that you want to fix as the new working directory.

Consider again Figure 2.1. Imagine that one wants to fix the directory that contains the file "`ex1.csv`" as the permanent working directory. Notice that the full address of the directory appears at the "`Address`" bar on the top of the window. One may copy the address and paste it instead of the name of the current working directory that is specified in the "`Properties`" dialog of the R icon. One should make sure that the address to the new directory is, again, placed between double-quotes. (See in Figure 2.4 the dialog window after the changing the address of the working directory. Compare this to Figure 2.3 of the window before the change.) After approving the change by clicking the "`OK`" bottom the new working directory is set. Henceforth, each time that the `R Console` is opened by double-clicking the icon it will have the designated directory as its working directory.

In the rest of this book we assume that a designated directory is set as R's working directory and that all external files that need to be read into R, such as "`ex1.csv`" for example, are saved in that working directory. Once a working directory has been set then the history of subsequent R sessions is stored in that directory. Hence, if you choose to save the image of the session when you end the session then objects created in the session will be uploaded the next time
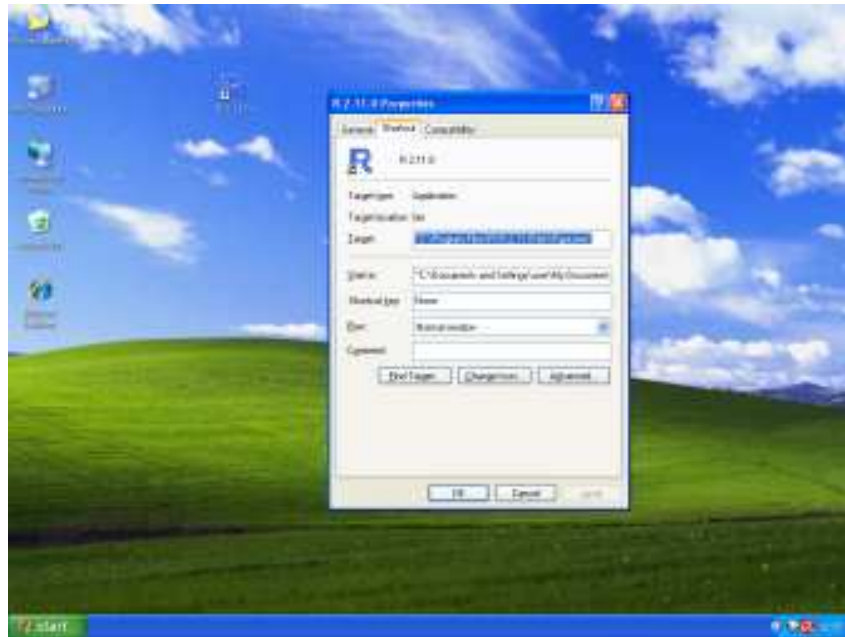
Figure 2.3: Setting the Working Directory (Before the Change)
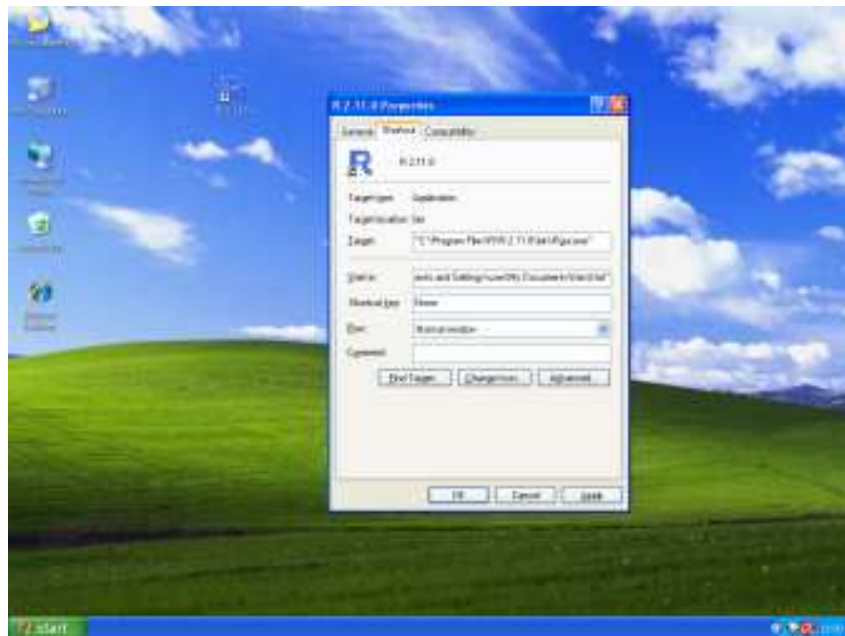
Figure 2.4: Setting the Working Directory (After the Change)

the R Console is opened.

2.3.2 Reading a CSV File into R

Now that a copy of the file "ex1.csv" is placed in the working directory we would like to read its content into R. Reading of files in the CSV format can be carried out with the R function "read.csv". To read the file of the example we run the following line of code in the R Console window:

```
> ex.1 <- read.csv("ex1.csv")
```

The function "read.csv" takes as an input argument the address of a CSV file and produces as output a *data frame* object with the content of the file. Notice that the address is placed between double-quotes. If the file is located in the working directory then giving the name of the file as an address is sufficient[1].

Consider the content of that R object "ex.1" that was created by the previous expression:

```
> ex.1
      id    sex height
1  5696379 FEMALE    182
2  3019088   MALE    168
3  2038883   MALE    172
4  1920587 FEMALE    154
5  6006813   MALE    174
6  4055945 FEMALE    176
.       .      .      .
.       .      .      .
.       .      .      .
98 9383288   MALE    195
99 1582961 FEMALE    129
100 9805356  MALE    172
>
```

You have to read what R is telling you when you run the read.csv command. If it says there is an error, it is probably because you forgot the quotes, you used upper case vs lower case, or your files are not in your working directory. You have to spell the filename exactly as the file is spelled. Some operating systems hide the file extentions by default, which can make life difficult. For the files that we use in class, they all should end with lower case ".csv" unless you renamed them while downloading.

(Noticed that we have erased the middle rows. In the R Console window you should obtain the full table. However, in order to see the upper part of the output you may need to scroll up the window.)

The object "ex.1", the output of the function "read.csv" is a *data frame.* Data frames are the standard tabular format of storing statistical data. The columns of the table are called *variables* and correspond to measurements. In this example the three variables are:

id: A 7 digits number that serves as a unique identifier of the subject.

sex: The sex of each subject. The values are either "MALE" or "FEMALE".

height: The height (in centimeter) of each subject. A numerical value.

---

[1]If the file is located in a different directory then the complete address, including the path to the file, should be provided. The file need not reside on the computer. One may provide, for example, a URL (an internet address) as the address. Thus, instead of saving the file of the example on the computer one may read its content into an R object by using the line of code "ex.1 <- read.csv("http://pluto.huji.ac.il/~msby/StatThink/Datasets/ex1.csv")" instead of the code that we provide and the working method that we recommend to follow.

When the values of the variable are numerical we say that it is a *quantitative variable* or a *numeric variable*. On the other hand, if the variable has qualitative or level values we say that it is a *factor*. In the given example, `sex` is a factor and `height` is a numeric variable.

The rows of the table are called *observations* and correspond to the subjects. In this data set there are 100 subjects, with subject number 1, for example, being a female of height 182 cm and identifying number 5696379. Subject number 98, on the other hand, is a male of height 195 cm and identifying number 9383288.

2.3.3   Data Types

The columns of `R` data frames represent variables, i.e. measurements recorded for each of the subjects in the sample. `R` associates with each variable a type that characterizes the content of the variable. The two major types are

- Factors, or Qualitative Data. The type is "`factor`".

- Quantitative Data. The type is "`numeric`".

Factors are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Qualitative data are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers and are usually the data of choice because there are many methods available for analyzing such data. Quantitative data are the result of counting or measuring attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and the number of students who take statistics are examples of quantitative data.

Quantitative data may be either discrete or continuous. All data that are the result of counting are called quantitative discrete data. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you may get results such as 0, 1, 2, 3, etc. On the other hand, data that are the result of measuring on a continuous scale are quantitative continuous data, assuming that we can measure accurately. Measuring angles in radians may result in the numbers $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Example 2.1 (Data Sample of Quantitative Discrete Data). *The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.*

Example 2.2 (Data Sample of Quantitative Continuous Data). *The data are the weights of the backpacks with the books in it. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3.*

Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Example 2.3 (Data Sample of Qualitative Data)**.** *The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.*

The distinction between continuous and discrete numeric data is not reflected usually in the statistical method that are used in order to analyze the data. Indeed, R does not distinguish between these two types of numeric data and store them both as "numeric". Consequently, we will also not worry about the specific categorization of numeric data and treat them as one. On the other hand, emphasis will be given to the difference between numeric and factors data.

One may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F. On the other hand, one may code categories of qualitative data with numerical values and report the values. The resulting data should nonetheless be treated as a factor. ← there is room for debate here.

As default, R saves variables that contain non-numeric values as factors. Otherwise, the variables are saved as numeric. The variable type is important because different statistical methods are applied to different data types. Hence, one should make sure that the variables that are analyzed have the appropriate type. Especially that factors using numbers to denote the levels are labeled as factors. Otherwise R will treat them as quantitative data.

2.4 Solved Exercises

Question 2.1. Consider the following relative frequency table on hurricanes that have made direct hits on the U.S. between 1851 and 2004 (`http://www.nhc.noaa.gov/gifs/table5.gif`). Hurricanes are given a strength category rating based on the minimum wind speed generated by the storm. Some of the entries to the table are missing.

Category	# Direct Hits	Relative Freq.	Cum. Relative Freq.
1	109		
2	72	0.2637	0.6630
3		0.2601	
4	18		0.9890
5	3	0.0110	1.0000

Table 2.1: Frequency of Hurricane Direct Hits

You definitely need to understand how to complete the numbers in this table without looking at any outside sources of information.

1. What is the relative frequency of direct hits of category 1?

2. What is the relative frequency of direct hits of category 4 or more?

Solution (to Question 2.1.1): The relative frequency of direct hits of category 1 is 0.3993. Notice that the cumulative relative frequency of category

1 and 2 hits, the sum of the relative frequency of both categories, is 0.6630. The relative frequency of category 2 hits is 0.2637. Consequently, the relative frequency of direct hits of category 1 is 0.6630 - 0.2637 = 0.3993.

Solution (to Question 2.1.2): The relative frequency of direct hits of category 4 or more is 0.0769. Observe that the cumulative relative of the value "3" is 0.6630 + 0.2601 = 0.9231. This follows from the fact that the cumulative relative frequency of the value "2" is 0.6630 and the relative frequency of the value "3" is 0.2601. The total cumulative relative frequency is 1.0000. The relative frequency of direct hits of category 4 or more is the difference between the total cumulative relative frequency and cumulative relative frequency of 3 hits: 1.0000 - 0.9231 = 0.0769.

Question 2.2. The number of calves that were born to some cows during their productive years was recorded. The data was entered into an R object by the name "`calves`". Refer to the following R code:

```
> freq <- table(calves)
> cumsum(freq)
 1  2  3  4  5  6  7
 4  7 18 28 32 38 45
```

1. How many cows were involved in this study?

2. How many cows gave birth to a total of 4 calves?

3. What is the relative frequency of cows that gave birth to at least 4 calves?

Solution (to Question 2.2.1): The total number of cows that were involved in this study is 45. The object "`freq`" contain the table of frequency of the cows, divided according to the number of calves that they had. The cumulative frequency of all the cows that had 7 calves or less, which includes all cows in the study, is reported under the number "7" in the output of the expression "`cumsum(freq)`". This number is 45.

Solution (to Question 2.2.2): The number of cows that gave birth to a total of 4 calves is 10. Indeed, the cumulative frequency of cows that gave birth to 4 calves or less is 28. The cumulative frequency of cows that gave birth to 3 calves or less is 18. The frequency of cows that gave birth to exactly 4 calves is the difference between these two numbers: 28 - 18 = 10.

Solution (to Question 2.2.3): The relative frequency of cows that gave birth to at least 4 calves is $27/45 = 0.6$. Notice that the cumulative frequency of cows that gave at most 3 calves is 18. The total number of cows is 45. Hence, the number of cows with 4 or more calves is the difference between these two numbers: 45 - 18 = 27. The relative frequency of such cows is the ratio between this number and the total number of cows: $27/45 = 0.6$.

2.5 Summary

Glossary

Population: The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Sample: A portion of the population understudy. A sample is representative if it characterizes the population being studied.

Frequency: The number of times a value occurs in the data.

Relative Frequency: The ratio between the frequency and the size of data.

Cumulative Relative Frequency: The term applies to an ordered set of data values from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Data Frame: A tabular format for storing statistical data. Columns correspond to variables and rows correspond to observations.

Variable: A measurement that may be carried out over a collection of subjects. The outcome of the measurement may be numerical, which produces a quantitative variable; or it may be non-numeric, in which case a factor is produced.

Observation: The evaluation of a variable (or variables) for a given subject.

CSV Files: A digital format for storing data frames.

Factor: Qualitative data that is associated with categorization or the description of an attribute.

Quantitative: Data generated by numerical measurements.

If you conduct a survey of 50 people and ask each of them 3 questions, you would usually save your data in a CSV file that has 50 rows and 3 columns. Each of the 50 rows represents one OBSERVATION.

Discuss in the forum

Factors are qualitative data that are associated with categorization or the description of an attribute. On the other hand, numeric data are generated by numerical measurements. A common practice is to code the levels of factors using numerical values. What do you think of this practice?

In the formulation of your answer to the question you may think of an example of factor variable from your own field of interest. You may describe a benefit or a disadvantage that results from the use of a numerical values to code the level of this factor.

Chapter 3

Descriptive Statistics

3.1   Student Learning Objectives

This chapter deals with numerical and graphical ways to describe and display data. This area of statistics is called *descriptive statistics*. You will learn to calculate and interpret these measures and graphs. By the end of this chapter, you should be able to:

- Use histograms and box plots in order to display data graphically.

- Calculate measures of central location: mean and median.

- Calculate measures of the spread: variance, standard deviation, and inter-quartile range.

- Identify outliers, which are values that do not fit the rest of the distribution.

3.2   Displaying Data

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you may ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample is often overwhelming. A better way may be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

A statistical graph is a tool that helps you learn about the shape of the distribution of a sample. The graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly.

Statisticians often start the analysis by graphing the data in order to get an overall picture of it. Afterwards, more formal tools may be applied.

In the previous chapters we used the bar plot, where bars that indicate the frequencies in the data of values are placed over these values. In this chapter

Histogram of ex.1$height
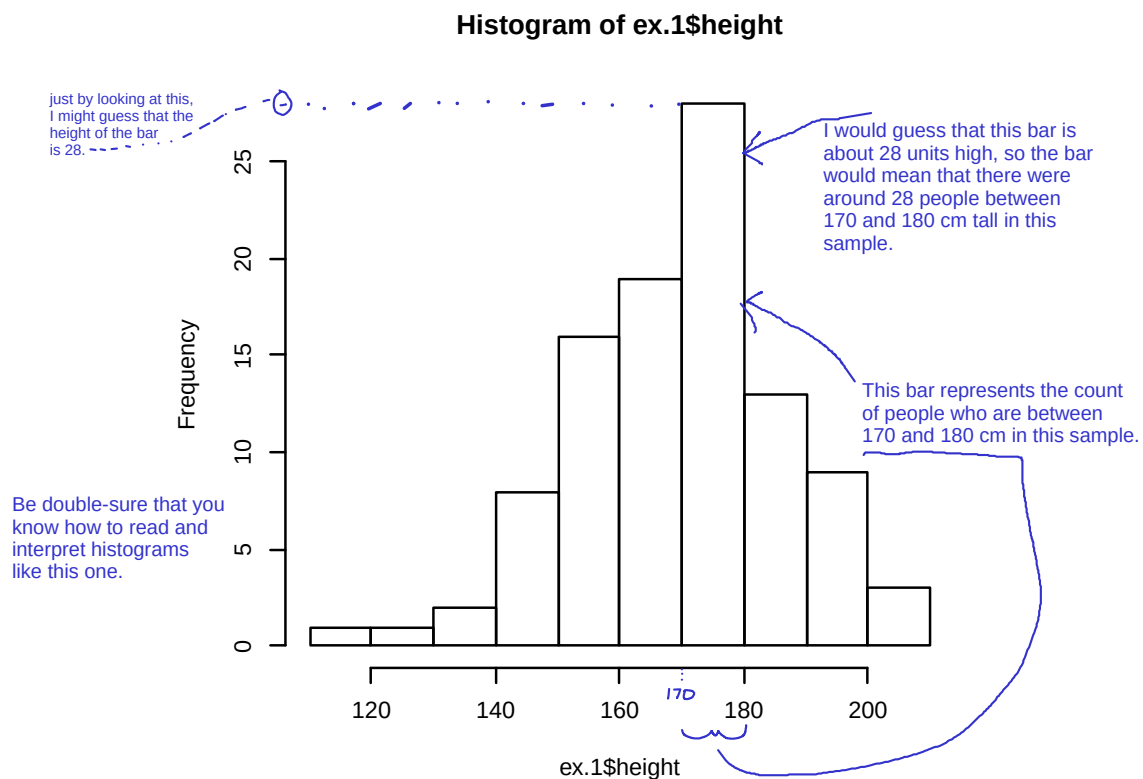
Figure 3.1: Histogram of Height

our emphasis will be on histograms and box plots, which are other types of plots. Some of the other types of graphs that are frequently used, but will not be discussed in this book, are the stem-and-leaf plot, the frequency polygon (a type of broken line graph) and the pie charts. The types of plots that will be discussed and the types that will not are all tightly linked to the notion of *frequency* of the data that was introduced in Chapter 2 and intend to give a graphical representation of this notion.

3.2.1   Histograms

The *histogram* is a frequently used method for displaying the distribution of continuous numerical data. An advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

One may produce a histogram in R by the application of the function "`hist`" to a sequence of numerical data. Let us read into R the data frame "`ex.1`" that contains data on the sex and height and create a histogram of the heights:

```
> ex.1 <- read.csv("ex1.csv")
```

```
> hist(ex.1$height)
```

The outcome of the function is a plot that apears in the graphical window and is presented in Figure 3.1.

The data set, which is the content of the CSV file "ex1.csv", was used in Chapter 2 in order to demonstrate the reading of data that is stored in a external file into R. The first line of the above script reads in the data from "ex1.csv" into a data frame object named "ex.1" that maintains the data internally in R. The second line of the script produces the histogram. We will discuss below the code associated with this second line.

A histogram consists of contiguous boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (the height, in this example). The vertical axis presents frequencies and is labeled "Frequency". By the examination of the histogram one can appreciate the shape of the data, the center, and the spread of the data.

The histogram is constructed by dividing the range of the data (the x-axis) into equal intervals, which are the bases for the boxes. The height of each box represents the count of the number of observations that fall within the interval. For example, consider the box with the base between 160 and 170. There is a total of 19 subjects with height larger that 160 but no more than 170 (that is, $160 < \texttt{height} \leq 170$). Consequently, the height of that box[1] is 19.

The input to the function "hist" should be a sequence of numerical values. In principle, one may use the function "c" to produce a sequence of data and apply the histogram plotting function to the output of the sequence producing function. However, in the current case we have already the data stored in the data frame "ex.1", all we need to learn is how to extract that data so it can be used as input to the function "hist" that plots the histogram.

Notice the structure of the input that we have used in order to construct the histogram of the variable "height" in the "ex.1" data frame. One may address the variable "variable.name" in the data frame "dataframe.name" using the format: "dataframe.name$variable.name". Indeed, when we type the expression "ex.1$height" we get as an output the values of the variable "height" from the given data frame:

```
> ex.1$height
 [1] 182 168 172 154 174 176 193 156 157 186 143 182 194 187 171
[16] 178 157 156 172 157 171 164 142 140 202 176 165 176 175 170
[31] 169 153 169 158 208 185 157 147 160 173 164 182 175 165 194
[46] 178 178 186 165 180 174 169 173 199 163 160 172 177 165 205
[61] 193 158 180 167 165 183 171 191 191 152 148 176 155 156 177
[76] 180 186 167 174 171 148 153 136 199 161 150 181 166 147 168
[91] 188 170 189 117 174 187 141 195 129 172
```

This is a numeric sequence and can serve as the input to a function that expects a numeric sequence as input, a function such as "hist". (But also other functions, for example, "sum" and "cumsum".)

---

[1] In some books an histogram is introduced as a form of a density. In densities the *area* of the box represents the frequency or the relative frequency. In the current example the height would have been $19/10 = 1.9$ if the area of the box would have represented the frequency and it would have been $(19/100)/10 = 0.019$ if the area of the box would have represented the relative frequency. However, in this book we follow the default of R in which the height represents the frequency.

Handwritten annotations:

"contiguous" ... This means that the rectangles that show in a histogram are touching each other (usually.. it is possible to have an outlier where one rectangle is far off to the side)

How to refer to columns of data in a data frame or CSV file

The dollar sign says to look inside the ex.1 R object and get the values for "height."

These are index numbers explained on the next page

After you load the ex.1 file, try these commands:

sum(ex.1$height)

cumsum(table(ex.1$height))

this one shows cumulative frequencies for the height data.

[1] Explained

There are 100 observations in the variable "`ex.1$height`". So many observations cannot be displayed on the screen on one line. Consequently, the sequence of the data is wrapped and displayed over several lines. Notice that the square brackets on the left hand side of each line indicate the position in the sequence of the first value on that line. Hence, the number on the first line is "`[1]`". The number on the second line is "`[16]`", since the second line starts with the 16th observation in the display given in the book. Notice, that numbers in the square brackets on your `R Console` window may be different, depending on the setting of the display on your computer.

3.2.2   Box Plots

The *box plot*, or box-whisker plot, gives a good graphical overall impression of the concentration of the data. It also shows how far from most of the data the extreme values are. In principle, the box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then once more in the next section.

The *median*, a number, is a way of measuring the "center" of the data. You can think of the median as the "middle value," although it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same size or smaller than the median and half the values are the same size or larger than it. For example, consider the following data that contains 14 values:

$$1, 11.5, 6, 7.2, 4, 8, 9, 10, 6.8, 8.3, 2, 2, 10, 1 \ .$$

Median: first sort the numbers in order, then find the middle value. If there is an even number of observations, there is no "middle" value, so you can find the middle of the two values closest to the center.

Ordered, from smallest to largest, we get:

$$1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5 \ .$$

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2:

$$\frac{6.8 + 7.2}{2} = 7$$

There are different techniques for handling unusual situations when calculating the median, but we will stick to these two techniques and use R for analyzing complicated data.

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile is the middle value of the lower half of the data and the third quartile is the middle value of the upper half of the data. For illustration consider the same data set from above:

the median is the 2nd quartile.
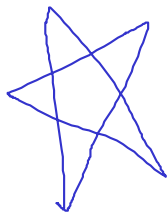
$$1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5 \ .$$

The median or second quartile is 7. The lower half of the data is:

You can also find the quartiles with the summary() command in R or the quantile() command. The summary() command is easiest to find the quartiles.

$$1, 1, 2, 2, 4, 6, 6.8 \ .$$

The middle value of the lower half is 2. The number 2, which is part of the data in this case, is the first quartile which is denoted Q1. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.
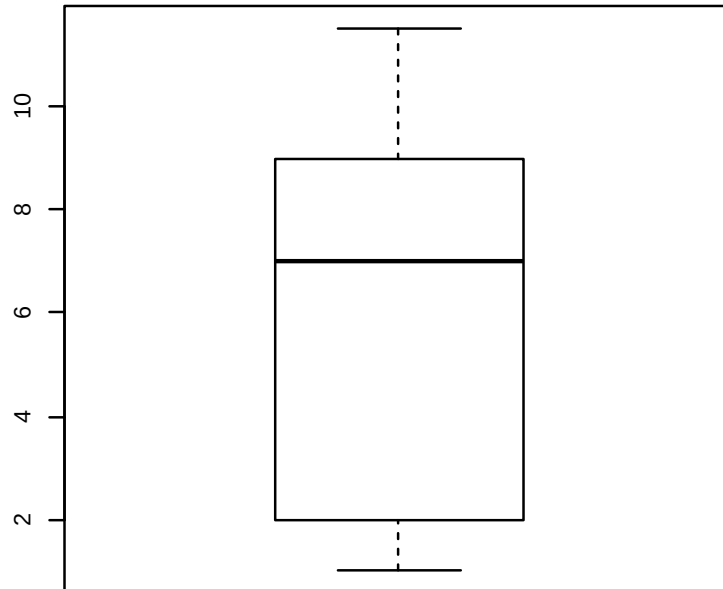
Quartiles

Figure 3.2: Box Plot of the Example

The upper half of the data is:

$$7.2, \; 8, \; 8.3, \; 9, \; 10, \; 10, \; 11.5$$

The middle value of the upper half is 9. The number 9 is the third quartile which is denoted Q3. Three-fourths of the values are less than 9 and one-fourth of the values[2] are more than 9.

Outliers are values that do not fit with the rest of the data and lie outside of the normal range. Data points with values that are much too large or much too small in comparison to the vast majority of the observations will be identified as outliers. In the context of the construction of a box plot we identify potential outliers with the help of the *inter-quartile range* (IQR). The inter-quartile range is the distance between the third quartile (Q3) and the first quartile (Q1), i.e., IQR = Q3 − Q1. A data point that is larger than the third quartile plus 1.5 times the inter-quartile range will be marked as a potential outlier. Likewise, a data point smaller than the first quartile minus 1.5 times the inter-quartile

Outliers

---

[2]The actual computation in R of the first quartile and the third quartile may vary slightly from the description given here, depending on the exact structure of the data.

Note that the IQR technique for finding outliers is a matter of personal choice and custom. There is no objective law of the universe that says that 1.5 * the IQR beyond a particular quartile is an outlier. Researchers in various fields use different criteria for outliers, but the custom explained here is a rule of thumb that might help you to find problem data.

range will also be so marked. Outliers may have a substantial effect on the outcome of statistical analysis, therefore it is important that one is alerted to the presence of outliers.

In the running example we obtained an inter-quartile range of size 9-2=7. The upper threshold for defining an outlier is $9 + 1.5 \times 7 = 19.5$ and the lower threshold is $2 - 1.5 \times 7 = -8.5$. All data points are within the two thresholds, hence there are no outliers in this data.

be sure that you understand this example.

In the construction of a box plot one uses a vertical rectangular box and two vertical "whiskers" that extend from the ends of the box to the smallest and largest data values that are not outliers. Outlier values, if any exist, are marked as points above or blow the endpoints of the whiskers. The smallest and largest non-outlier data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. The central 50% of the data fall within the box.

Interpretation of a boxplot

One may produce a box plot with the aid of the function "`boxplot`". The input to the function is a sequence of numerical values and the output is a plot. As an example, let us produce the box plot of the 14 data points that were used as an illustration:

```
> boxplot(c(1,11.5,6,7.2,4,8,9,10,6.8,8.3,2,2,10,1))
```

The resulting box plot is presented in Figure 3.2. Observe that the end points of the whiskers are 1, for the minimal value, and 11.5 for the largest value. The end values of the box are 9 for the third quartile and 2 for the first quartile. The median 7 is marked inside the box.

Next, let us examine the box plot for the height data:

```
> boxplot(ex.1$height)
```

The resulting box plot is presented in Figure 3.3. In order to assess the plot let us compute quartiles of the variable:

```
> summary(ex.1$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  117.0   158.0   171.0   170.1   180.2   208.0
```

The function "`summary`", when applied to a numerical sequence, produce the minimal and maximal entries, as well the first, second and third quartiles (the second is the Median). It also computes the average of the numbers (the Mean), which will be discussed in the next section.

Let us compare the results with the plot in Figure 3.3. Observe that the median 171 coincides with the thick horizontal line inside the box and that the lower end of the box coincides with first quartile 158.0 and the upper end with 180.2, which is the third quartile. The inter-quartile range is $180.2 - 158.0 = 22.2$. The upper threshold is $180.2 + 1.5 \times 22.2 = 213.5$. This threshold is larger than the largest observation (208.0). Hence, the largest observation is not an outlier and it marks the end of the upper whisker. The lower threshold is $158.0 - 1.5 \times 22.2 = 124.7$. The minimal observation (117.0) is less than this threshold. Hence it is an outlier and it is marked as a point below the end of the lower whisker. The second smallest observation is 129. It lies above the lower threshold and it marks the end point of the lower whisker.
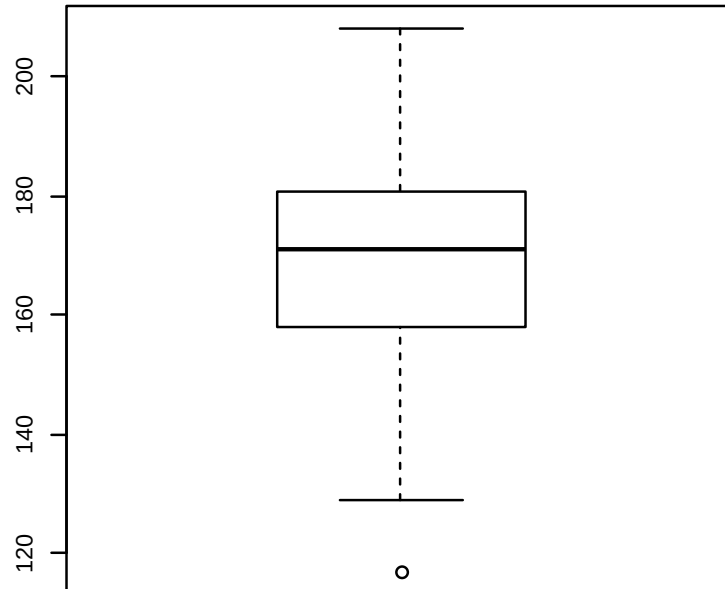
Figure 3.3: Box Plot of Height

3.3 Measures of the Center of Data

The two most widely used measures of the central location of the data are the *mean* (average) and the *median*. To calculate the average weight of 50 people one should add together the 50 weights and divide the result by 50. To find the median weight of the same 50 people, one may order the data and locate a number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. Nonetheless, the mean is the most commonly used measure of the center.

We shall use small Latin letters such as x to mark the sequence of data. In such a case we may mark the sample mean by placing a bar over the x: $\bar{x}$ (pronounced "x bar").

The mean can be calculated by averaging the data points or it also can be calculated with the relative frequencies of the values that are present in the data. In the latter case one multiplies each distinct value by its relative frequency and then sum the products across all values. To see that both ways of calculating

x-bar is a symbol used to represent the mean of a sample. Later, we will use the Greek letter mu to represent the mean of a population.

x-bar IS the mean of a sample...
nothing more and nothing less.

Histogram of x

symmetric

Histogram of x

"Skewed to the left" because it is
"pulled to the left"

Histogram of x

"Skewed to the right" because it is
"pulled to the right"
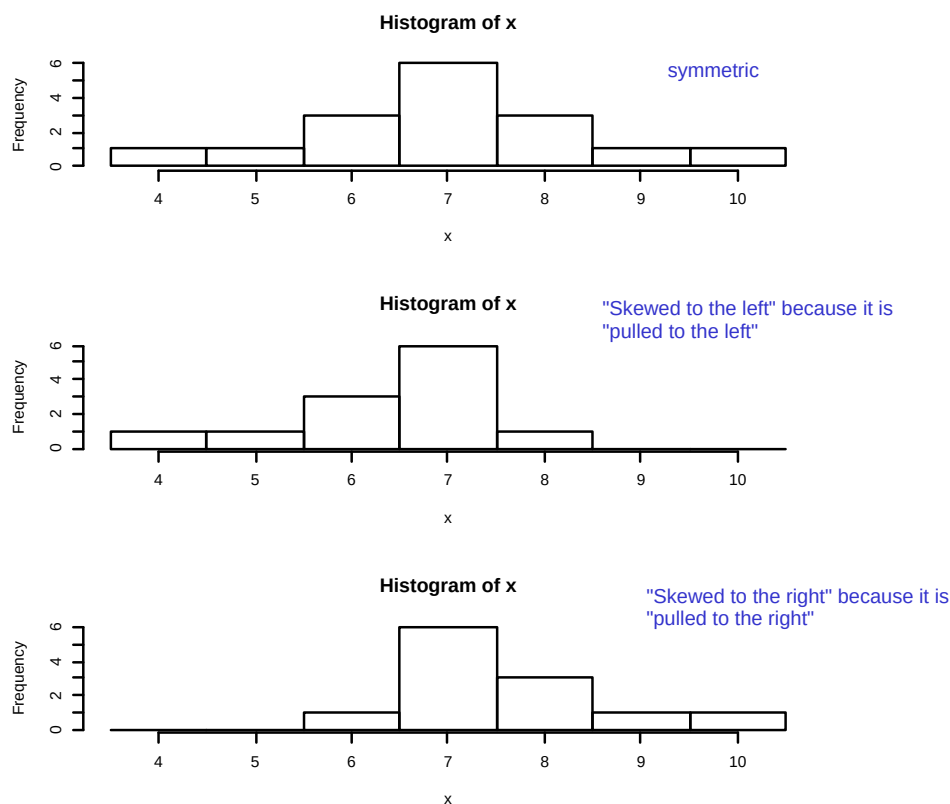
Figure 3.4: Three Histograms

the mean are the same, consider the data:

$$1, \ 1, \ 1, \ 2, \ 2, \ 3, \ 4, \ 4, \ 4, \ 4, \ 4 \ .$$

In the first way of calculating the mean we get:

Two ways of calculating the
mean. The second technique
will be most important in
Chapters 4 and beyond.

$$\bar{x} = \frac{1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4}{11} = 2.7 \ .$$

Alternatively, we may note that the distinct values in the sample are 1, 2, 3, and 4 with relative frequencies of 3/11, 2/11, 1/11 and 5/11, respectively. The alternative method of computation produces:

$$\bar{x} = 1 \times \frac{3}{11} + 2 \times \frac{2}{11} + 3 \times \frac{1}{11} + 4 \times \frac{5}{11} = 2.7 \ .$$

3.3.1   Skewness, the Mean and the Median

Consider the following data set:

$$4, \ 5, \ 6, \ 6, \ 6, \ 7, \ 7, \ 7, \ 7, \ 7, \ 7, \ 8, \ 8, \ 8, \ 9, \ 10$$

This data produces the upper most histogram in Figure 3.4. Each interval has width one and each value is located at the middle of an interval. The histogram displays a symmetrical distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and to the right of the vertical line are mirror images of each other.

Let us compute the mean and the median of this data:

R functions for
mean and median.

```
> x <- c(4,5,6,6,6,7,7,7,7,7,7,8,8,8,9,10)
> mean(x)
[1] 7
> median(x)
[1] 7
```

The mean and the median are each 7 for these data. In a perfectly symmetrical distribution, the mean and the median are the same[3].

The functions "`mean`" and "`median`" were used in order to compute the mean and median. Both functions expect a numeric sequence as an input and produce the appropriate measure of centrality of the sequence as an output.

The histogram for the data:

$$4, \ 5, \ 6, \ 6, \ 6, \ 7, \ 7, \ 7, \ 7, \ 7, \ 7, \ 8$$

is not symmetrical and is displayed in the middle of Figure 3.4. The right-hand side seems "chopped off" compared to the left side. The shape of the distribution is called skewed to the left because it is pulled out towards the left.

Let us compute the mean and the median for this data:

```
> x <- c(4,5,6,6,6,7,7,7,7,7,7,8)
> mean(x)
[1] 6.416667
> median(x)
[1] 7
```

(Notice that the original data is replaced by the new data when object `x` is reassigned.) The median is still 7, but the mean is less than 7. The relation between the mean and the median reflects the skewing.

Consider yet another set of data:

$$6, \ 7, \ 7, \ 7, \ 7, \ 7, \ 7, \ 8, \ 8, \ 8, \ 9, \ 10$$

The histogram for the data is also not symmetrical and is displayed at the bottom of Figure 3.4. Notice that it is skewed to the right. Compute the mean and the median:

```
> x <- c(6,7,7,7,7,7,7,8,8,8,9,10)
> mean(x)
[1] 7.583333
> median(x)
[1] 7
```

---

[3]In the case of a symmetric distribution the vertical line of symmetry is located at the mean, which is also equal to the median.

The median is yet again equal to 7, but this time the mean is greater than 7. Again, the mean reflects the skewing.

In summary, if the distribution of data is skewed to the left then the mean is less than the median. If the distribution of data is skewed to the right then the median is less than the mean.

This is not always the case, but it is usually the case.

Examine the data on the height in "`ex.1`":

```
> mean(ex.1$height)
[1] 170.11
> median(ex.1$height)
[1] 171
```

Observe that the histogram of the height (Figure 3.1) is skewed to the left. This is consistent with the fact that the mean is less than the median.

skewed left

3.4   Measures of the Spread of Data

One measure of the spread of the data is the inter-quartile range that was introduced in the context of the box plot. However, the most important measure of spread is the standard deviation.

"subject-i"
If you collect data from 50 people, then you have 50 observations. If you want the total score for the 7th person, then "i" would be 7. If the sum of the values for th 7th person is 32, then:

$$\sum x_7 = 32$$

where the $\sum$ symbol is the Greek letter sigma, and it means "sum"

Before dealing with the standard deviation let us discuss the calculation of the variance. If x_i is a data value for subject i and $\bar{x}$ is the sample mean, then $x_i - \bar{x}$ is called the deviation of subject i from the mean, or simply the deviation. In a data set, there are as many deviations as there are data values. The variance is in principle the average of the squares of the deviations.

Consider the following example: In a fifth grade class, the teacher was interested in the average age and the standard deviation of the ages of her students. Here are the ages of her students to the nearest half a year:

9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5 .

In order to explain the computation of the variance of these data let us create an object x that contains the data:

do not type these chars
> +
R puts them there.

```
> x <- c(9,9.5,9.5,10,10,10,10,10.5,10.5,10.5,10.5,11,11,11,11,11,
+ 11,11.5,11.5,11.5)
> length(x)
[1] 20
```

If you have real data in R, you can use the length() command to tell you how many observations you have.

Pay attention to the fact that we did not write the "`+`" at the beginning of the second line. That symbol was produced by R when moving to the next line to indicate that the expression is not complete yet and will not be executed. Only after inputting the right bracket and the hitting of the Return key does R carry out the command and creates the object "x". When you execute this example yourself on your own computer make sure not to copy the "`+`" sign. Instead, if you hit the return key after the last comma on the first line, the plus sign will be produced by R as a new prompt and you can go on typing in the rest of the numbers.

The function "`length`" returns the length of the input sequence. Notice that we have a total of 20 data points.

The next step involves the computation of the deviations:

```
> x.bar <- mean(x)
> x.bar
[1] 10.525
> x - x.bar
 [1] -1.525 -1.025 -1.025 -0.525 -0.525 -0.525 -0.525 -0.025
 [9] -0.025 -0.025 -0.025  0.475  0.475  0.475  0.475  0.475
[17]  0.475  0.975  0.975  0.975
```

(handwritten: The mean of x is 10.525)

(blue margin note: The results show you the difference between each value of x and the mean. Check it yourself so you understand what R is doing. The first value in "x" on the prior page is 9, and if you subtract the mean of 10.525, you get -1.525.)

The average of the observations is equal to 10.525 and when we delete this number from each of the components of the sequence x we obtain the deviations. For example, the first deviation is obtained as 9 - 10.525 = -1.525, the second deviation is 9.5 - 10.525 = -1.025, and so forth. The 20th deviation is 11.5 - 10.525 = 0.975, and this is the last number that is presented in the output.

From a more technical point of view observe that the expression that computed the deviations, "x - x.bar", involved the deletion of a single value (x.bar) from a sequence with 20 values (x). The expression resulted in the deletion of the value from each component of the sequence. This is an example of the general way by which R operates on sequences. The typical behavior of R is to apply an operation to each component of the sequence.

(handwritten: ignore the "delete" comment)

As yet another illustration of this property consider the computation of the squares of the deviations:

```
> (x - x.bar)^2
```
(blue note: We will eventually use this in our calculation of variance and standard deviation.)
```
 [1] 2.325625 1.050625 1.050625 0.275625 0.275625 0.275625
 [7] 0.275625 0.000625 0.000625 0.000625 0.000625 0.225625
[13] 0.225625 0.225625 0.225625 0.225625 0.225625 0.950625
[19] 0.950625 0.950625
```

Recall that "x - x.bar" is a sequence of length 20. We apply the square function to this sequence. This function is applied to each of the components of the sequence. Indeed, for the first component we have that $(-1.525)^2 = 2.325625$, for the second component $(-1.025)^2 = 1.050625$, and for the last component $(0.975)^2 = 0.950625$.

For the variance we sum the square of the deviations and divide by the total number of data values minus one $(n-1)$. The standard deviation is obtained by taking the square root of the variance:

(handwritten: Variance)
```
> sum((x - x.bar)^2)/(length(x)-1)
[1] 0.5125
```
(handwritten: Std. deviation)
```
> sqrt(sum((x - x.bar)^2)/(length(x)-1))
[1] 0.715891
```

(blue note: This says "the sum of "x minus x-bar" squared, divided by "n - 1." Note that the length() command gives you the number of observations, and in other equations, that is represented by the letter "n." For "sample variance" we put "n - 1" in the denominator, but for population variance, we put just "n" in the denominator (Chapter 4).)

If the variance is produced as a result of dividing the sum of squares by the number of observations minus one then the variance is called the *sample variance*.

The function "var" computes the sample variance and the function "sd" computes the standard deviations. The input to both functions is the sequence of data values and the outputs are the sample variance and the standard deviation, respectively:

```
> var(x)
[1] 0.5125
```

(blue note: WARNING: use the var() and sd() R commands only if you have all the data from a sample. In the next chapter, you might have a relative frequency table that describes a sample, and you can not just slam those numbers into the var() command because the relative frequency numbers have a different interpretation that what the var() or sd() commands will accept.)

```
> sd(x)
[1] 0.715891
```

In the computation of the variance we divide the sum of squared deviations by the number of deviations minus one and not by the number of deviations. The reason for that stems from the theory of statistical inference that will be discussed in Part II of this book. Unless the size of the data is small, dividing by n or by $n-1$ does not introduce much of a difference.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

The sample standard deviation, s, is either zero or is larger than zero. When $s = 0$, there is no spread and the data values are equal to each other. When s is a lot larger than zero, the data values are very spread out about the mean. Outliers can make s very large.

The standard deviation is a number that measures how far data values are from their mean. For example, if the data contains the value 7 and if the mean of the data is 5 and the standard deviation is 2, then the value 7 is one standard deviation from its mean because $5 + 1 \times 2 = 7$. We say, then, that 7 is one standard deviation larger than the mean 5 (or also say "to the right of 5"). If the value 1 was also part of the data set, then 1 is two standard deviations smaller than the mean (or two standard deviations to the left of 5) because $5 - 2 \times 2 = 1$.

The standard deviation, when first presented, may not be too simple to interpret. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation is less so. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value.

3.5   Solved Exercises

Question 3.1. Three sequences of data were saved in 3 R objects named "x1", "x2" and "x3", respectively. The application of the function "summary" to each of these objects is presented below:

be sure that you understand what the summary() command is telling you.

```
> summary(x1)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   2.498   3.218   3.081   3.840   4.871
> summary(x2)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
0.0001083 0.5772000 1.5070000 1.8420000 2.9050000 4.9880000
> summary(x3)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.200   3.391   4.020   4.077   4.690   6.414
```

In Figure 3.5 one may find the histograms of these three data sequences, given in a random order. In Figure 3.6 one may find the box plots of the same data, given in yet a different order.
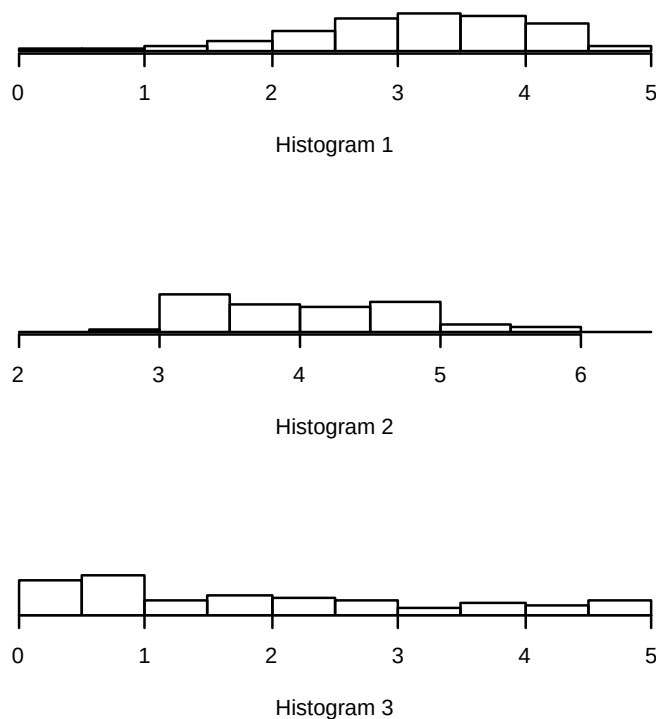
Figure 3.5: Three Histograms

1. Match the summary result with the appropriate histogram and the appropriate box plot.

2. Is the value 0.000 in the sequence "x1" an outlier?

3. Is the value 6.414 in the sequence "x3" an outlier?

Solution (to Question 3.1.1): Consider the data "x1". From the summary we see that it is distributed in the range between 0 and slightly below 5. The central 50% of the distribution are located between 2.5 and 3.8. The mean and median are approximately equal to each other, which suggests an approximately symmetric distribution. Consider the histograms in Figure 3.5. Histograms 1 and 3 correspond to a distributions in the appropriate range. However, the distribution in Histogram 3 is concentrated in lower values than suggested by the given first and third quartiles. Consequently, we match the summary of "x1" with Histograms 1.

Consider the data "x2". Again, the distribution is in the range between 0 and slightly below 5. The central 50% of the distribution are located between 0.6 and 1.8. The mean is larger than the median, which suggests a distribution skewed
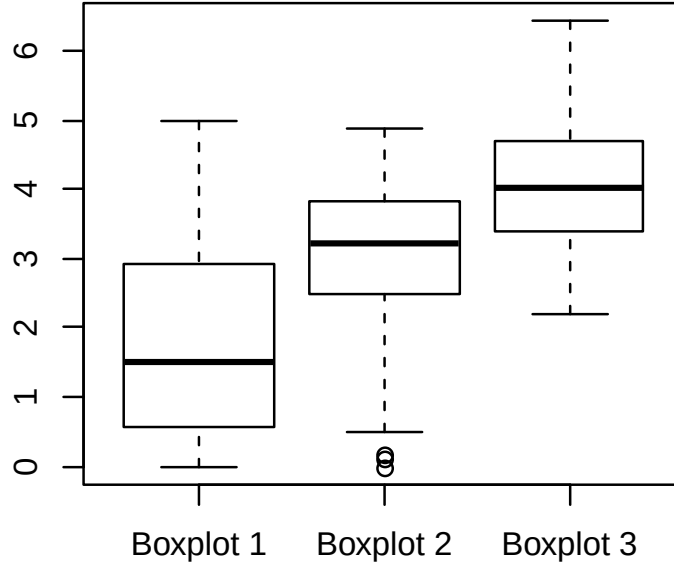
Figure 3.6: Three Box Plots

to the right. Therefore, we match the summary of "x2" with Histograms 3.

For the data in "x3" we may note that the distribution is in the range between 2 and 6. The histogram that fits this description is Histograms 2.

The box plot is essentially a graphical representation of the information presented by the function "summary". Following the rational of matching the summary with the histograms we may obtain that Histogram 1 should be matched with Box-plot 2 in Figure 3.6, Histogram 2 matches Box-plot 3, and Histogram 3 matches Box-plot 1. Indeed, it is easier to match the box plots with the summaries. However, it is a good idea to practice the direct matching of histograms with box plots.

Solution (to Question 3.1.2): The data in "x1" fits Box-plot 2 in Figure 3.6. The value 0.000 is the smallest value in the data and it corresponds to the smallest point in the box plot. Since this point is below the bottom whisker it follows that it is an outlier. More directly, we may note that the inter-quartile range is equal to $IQR = 3.840 - 2.498 = 1.342$. The lower threshold is equal to $2.498 - 1.5 \times 1.342 = 0.485$, which is larger that the given value. Consequently, the given value 0.000 is an outlier.

Solution (to Question 3.1.3): Observe that the data in "x3" fits Box-plot 3 in Figure 3.6. The vale 6.414 is the largest value in the data and it corresponds to the endpoint of the upper whisker in the box plot and is not an outlier. Alternatively, we may note that the inter-quartile range is equal to $IQR = 4.690 - 3.391 = 1.299$. The upper threshold is equal to $4.690 + 1.5.299 = 6.6385$, which is larger that the given value. Consequently, the given value 6.414 is not an outlier.

Question 3.2. The number of toilet facilities in 30 buildings were counted. The results are recorded in an R object by the name "x". The frequency table of the data "x" is:

```
> table(x)
x
 2  4  6  8 10
10  6 10  2  2
```

1. What is the mean ($\bar{x}$) of the data?

2. What is the sample standard deviation of the data?

3. What is the median of the data?

4. What is the inter-quartile range (IQR) of the data?

5. How many standard deviations away from the mean is the value 10?

Solution (to Question 3.2.1): In order to compute the mean of the data we may write the following simple R code:

```
> x.val <- c(2,4,6,8,10)
> freq <- c(10,6,10,2,2)
> rel.freq <- freq/sum(freq)
> x.bar <- sum(x.val*rel.freq)
> x.bar
[1] 4.666667
```

This shows how to calculate the mean (x.bar) using a list of values and frequencies.

We created an object "x.val" that contains the unique values of the data and an object "freq" that contains the frequencies of the values. The object "rel.freq" contains the relative frequencies, the ratios between the frequencies and the number of observations. The average is computed as the sum of the products of the values with their relative frequencies. It is stored in the objects "x.bar" and obtains the value 4.666667.

An alternative approach is to reconstruct the original data from the frequency table. A simple trick that will do the job is to use the function "rep". The first argument to this function is a sequence of values. If the second argument is a sequence of the same length that contains integers then the output will be composed of a sequence that contains the values of the first sequence, each repeated a number of times indicated by the second argument. Specifically, if we enter to this function the unique value "x.val" and the frequency of the values "freq" then the output will be the sequence of values of the original sequence "x":

```
> x <- rep(x.val,freq)
> x
 [1]  2  2  2  2  2  2  2  2  2  2  4  4  4  4  4  4  6  6  6
[20]  6  6  6  6  6  6  6  8  8 10 10
> mean(x)
[1] 4.666667
```

Observe that when we apply the function "`mean`" to "`x`" we get again the value 4.666667.

Solution (to Question 3.2.2): In order to compute the sample standard deviation we may compute first the sample variance and then take the square root of the result:

```
> var.x <- sum((x.val-x.bar)^2*freq)/(sum(freq)-1)
> sqrt(var.x)
[1] 2.425914
```

You can use this to find the variance and standard deviation when you are given a list of values and *frequencies* (NOT relative frequencies).

Notice that the expression "`sum((x.val-x.bar)^2*freq)`" compute the sum of square deviations. The expression "`(sum(freq)-1)`" produces the number of observations minus 1 $(n-1)$. The ratio of the two gives the sample variance.

Alternatively, had we produced the object "`x`" that contains the data, we may apply the function "`sd`" to get the sample standard deviation:

```
> sd(x)
[1] 2.425914
```

Observe that in both forms of computation we obtain the same result: 2.425914.

Solution (to Question 3.2.3): In order to compute the median one may produce the table of cumulative relative frequencies of "`x`":

```
> data.frame(x.val,cumsum(rel.freq))
  x.val cumsum.rel.freq.
1     2        0.3333333
2     4        0.5333333
3     6        0.8666667
4     8        0.9333333
5    10        1.0000000
```

For cumulative relative frequency, the last number should be 1 if you have the full frequency table.

Recall that the object "`x.val`" contains the unique values of the data. The expression "`cumsum(rel.freq)`" produces the cumulative relative frequencies. The function "`data.frame`" puts these two variables into a single data frame and provides a clearer representation of the results.

Notice that more that 50% of the observations have value 4 or less. However, strictly less than 50% of the observations have value 2 or less. Consequently, the median is 4. (If the value of the cumulative relative frequency at 4 would have been exactly 50% then the median would have been the average between 4 and the value larger than 4.)

In the case that we produce the values of the data "`x`" then we may apply the function "`summary`" to it and obtain the median this way

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   4.000   4.667   6.000  10.000
```

Solution (to Question 3.2.4): As for the inter-quartile range (IQR) notice that the first quartile is 2 and the third quartile is 6. Hence, the inter-quartile range is equal to 6 - 2 = 4. The quartiles can be read directly from the output of the function "summary" or can be obtained from the data frame of the cumulative relative frequencies. For the later observe that more than 25% of the data are less or equal to 2 and more 75% of the data are less or equal to 6 (with strictly less than 75% less or equal to 4).

Solution (to Question 3.2.5): In order to answer the last question we conduct the computation: $(10 - 4.666667)/2.425914 = 2.198484$. We conclude that the value 10 is approximately 2.1985 standard deviations above the mean.

3.6 Summary

Glossary

Median: A number that separates ordered data into halves: half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Quartiles: The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Outlier: An observation that does not fit the rest of the data.

Interquartile Range (IQR) : The distance between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 - Q1.

Mean: A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\bar{x}$) is

$$\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}} .$$

(Sample) Variance: Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1:

$$s^2 = \frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample} - 1} .$$

There is a small difference between "sample variance" and "population variance." In Chapter 4 where the variance of a random variable is described, that is population variance. The only difference is that with population variance, the denominator contains just "n" instead of "n - 1."

(Sample) Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. $s = \sqrt{s^2}$.

Discuss in the forum

An important practice is to check the validity of any data set that you are supposed to analyze in order to detect errors in the data and outlier observations. Recall that outliers are observations with values outside the normal range of values of the rest of the observations.

It is said by some that outliers can help us understand our data better. What is your opinion?

When forming your answer to this question you may give an example of how outliers may provide insight or, else, how they may abstract our understanding. For example, consider the price of a stock that tend to go up or go down at most 2% within each trading day. A sudden 5% drop in the price of the stock may be an indication to reconsidering our position with respect to this stock.

I think the Sigma character here is the Greek upper case sigma, which looks entirely different from the Greek lower case sigman.

Commonly Used Symbols

Please review these symbols. They really aren't too difficult.

↙ *Sigma*

- The symbol $\sum$ means to add or to find the sum.

Sigma just means "add the values." "Sigma" is the letter that represents "Sum."

- n = the number of data values in a sample.

- $\bar{x}$ = the sample mean. *X−bar*

- s = the sample standard deviation.

- f = frequency.

- f/n = relative frequency.

- x = numerical value.

If 25% of the values in your sample are the #3, then $x \times (f_x/n)$ = 3 × .25 = $\frac{3}{4}$

Commonly Used Expressions

- $x \times (f_x/n)$ = A value multiplied by its respective relative frequency.

- $\sum_{i=1}^{n} x_i$ = The sum of the data values. *this just says "add all the values in the sample."*

The second part of this expression is the same as what I wrote above. The sigma in front says to calculate the value above for each observation and sum them (add them).

- $\sum_{x}(x \times f_x/n)$ = The sum of values multiplied by their respective relative frequencies.

- $x - \bar{x}$ = Deviations from the mean (how far a value is from the mean).

- $(x - \bar{x})^2$ = Deviations squared.

Formulas:

If you have a list of five numbers, take each one, one at a time, and subtract the mean from it, and put your answer in a new list. You should eventually create a list of five new numbers that represent "deviations from the mean" (plural for "deviations").

- Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{x} \left(x \times (f_x/n) \right)$

- Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n}{n-1} \sum_{x} \left((x - \bar{x})^2 \times (f_x/n) \right)$ *"sample variance"*

- Standard Deviation: $s = \sqrt{s^2}$

If you look at the pieces of this equation, each piece is not too difficult. It is easier to look at the examples in the "solved problems" section above, and after you solve some problems a few times the expression here will start to make more sense.

Chapter 4

Probability

4.1 Student Learning Objective

This section extends the notion of variability that was introduced in the context of data to other situations. The variability of the entire *population* and the concept of a *random variable* is discussed. These concepts are central for the development and interpretation of statistical inference. By the end of the chapter the student should:

- Consider the distribution of a variable in a population and compute parameters of this distribution, such as the mean and the standard deviation.

- Become familiar with the concept of a random variable.

- Understand the relation between the distribution of the population and the distribution of a random variable produced by sampling a random subject from the population.

- Identify the distribution of the random variable in simple settings and compute its expectation and variance.

4.2 Different Forms of Variability

In the previous chapters we examined the variability in data. In the statistical context, data is obtained by selecting a sample from the target population and measuring the quantities of interest for the subjects that belong to the sample. Different subjects in the sample may obtain different values for the measurement, leading to variability in the data.

This variability may be summarized with the aid of a *frequency table,* a table of *relative frequency,* or via the *cumulative relative frequency*. A graphical display of the variability in the data may be obtained with the aid of the *bar plot*, the *histogram*, or the *box plot*.

Numerical summaries may be computed in order to characterize the main features of the variability. We used the *mean* and the *median* in order to identify the location of the distribution. The *sample variance*, or better yet the *sample standard deviation*, as well as the *inter-quartile range* were all described as tools to quantify the overall spread of the data.

.

The aim of all these graphical representations and numerical summaries is to investigate the variability of the data.

The subject of this chapter is to introduce two other forms of variability, variability that is not associated, at least not directly, with the data that we observe. The first type of variability is the *population variability*. The other type of variability is the variability of a *random variable*.

The notions of variability that will be presented are abstract, they are not given in terms of the data that we observe, and they have a mathematical-theoretical flavor to them. At first, these abstract notions may look to you as a waste of your time and may seem to be unrelated to the subject matter of the course. The opposite is true. The very core of statistical thinking is relating observed data to theoretical and abstract models of a phenomena. Via this comparison, and using the tools of statistical inference that are presented in the second half of the book, statisticians can extrapolate insights or make statements regarding the phenomena on the basis of the observed data. Thereby, the abstract notions of variability that are introduced in this chapter, and are extended in the subsequent chapters up to the end of this part of the book, are the essential foundations for the practice of statistics.

variability of a population is easy: we know that people are different heights: the variability of "height data" reflects the variation in the heights of real people.

The first notion of variability is the variability that is associated with the population. It is similar in its nature to the variability of the data. The difference between these two types of variability is that the former corresponds to the variability of the quantity of interest across all members of the population and not only for those that were selected to the sample.

Note: is some of our exercises, we will refer to a big data file as a "population." It is often up to the researcher to determine the "population of interest." In other words, if I am interested on only those people who live in my city, then that is my population. If I sample from people from my city, but I want to make assumptions about the global population, then my population would be everyone on earth--the researcher chooses what is the "population."

In Chapters 2 and 3 we examined the data set "`ex.1`" which contained data on the sex and height of a sample of 100 observations. In this chapter we will consider the sex and height of *all* the members of the population from which the sample was selected. The size of the relevant population is 100,000, including the 100 subjects that composed the sample. When we examine the values of the height across the entire population we can see that different people may have different heights. This variability of the heights is the population variability.

Here is one way to think of a random variable. First start with the variability of a population: the variability of the height of humans. We might count all the people and say that there are 1 billion people taller than 190 cm.

The other abstract type of variability, the variability of a random variable, is a mathematical concept. The aim of this concept is to model the notion of randomness in measurements or the uncertainty regarding the outcome of a measurement. In particular we will initially consider the variability of a random variable in the context of selecting one subject at random from the population.

If we think of a random variable that is similar, we might say that "10% of the elements in the population are greater than 190." By making abstract statements about the probability of randomly selecting a particular value from the random variable, we can describe the random variable in a way similar to how you might describe the variability in the heights of real people.

Imagine we have a population of size 100,000 and we are about to select at random one subject from this population. We intend to measure the height of the subject that will be selected. Prior to the selection and measurement we are not certain what value of height will be obtained. One may associate the notion of variability with uncertainty — different subjects to be selected may obtain different evaluations of the measurement and we do not know before hand which subject will be selected. The resulting variability is the variability of a random variable.

Random variables can be defined for more abstract settings. Their aim is to provide models for randomness and uncertainty in measurements. Simple examples of such abstract random variables will be provided in this chapter. More examples will be introduced in the subsequent chapters. The more abstract examples of random variables need not be associated with a specific population. Still, the same definitions that are used for the example of a random variable that emerges as a result of sampling a single subject from a population will

apply to the more abstract constructions.

All types of variability, the variability of the data we dealt with before as well as the other two types of variability, can be displayed using graphical tools and characterized with numerical summaries. Essentially the same type of plots and numerical summaries, possibly with some modifications, may and will be applied.

A point to remember is that the variability of the data relates to a concrete list of data values that is presented to us. In contrary to the case of the variability of the data, the other types of variability are not associated with quantities we actually get to observe. The data for the sample we get to see but not the data for the rest of the population. Yet, we can still discuss the variability of a population that is out there, even though we do not observe the list of measurements for the entire population. (The example that we give in this chapter of a population was artificially constructed and serves for illustration only. In the actual statistical context one does not obtain measurements from the entire population, only from the subjects that went into the sample.) The discussion of the variability in this context is theoretical in its nature. Still, this theoretical discussion is instrumental for understanding statistics.

4.3   A Population

In this section we introduce the variability of a population and present some numerical summaries that characterizes this variability. Before doing so, let us review with the aid of an example some of the numerical summaries that were used for the characterization of the variability of data.

Recall the file "`ex1.csv`" that contains data on the height and sex of 100 subjects. (The data file can be obtained from `http://pluto.huji.ac.il/ ~msby/StatThink/Datasets/ex1.csv`.) We read the content of the file into a data frame by the name "`ex.1`" and apply the function "`summary`" to the data frame:

```
> ex.1 <- read.csv("ex1.csv")
> summary(ex.1)
      id              sex           height
 Min.   :1538611   FEMALE:54   Min.   :117.0
 1st Qu.:3339583   MALE  :46   1st Qu.:158.0
 Median :5105620               Median :171.0
 Mean   :5412367               Mean   :170.1
 3rd Qu.:7622236               3rd Qu.:180.2
 Max.   :9878130               Max.   :208.0
```

A population can be any set of things that you want to study: the height of all people on the planet, the height of all people in your country, the height of all people in your school, the height of people in your statistics class -- you (as the researcher) decide what you want to study and call it the population. Usually a population is "big," but there is no exact requirement for a population to be a certain size.

We saw in the previous chapter that, when applied to a numeric sequence, the function "`summary`" produces the smallest and largest values in the sequence, the three quartiles (including the median) and the mean. If the input of the same function is a factor then the outcome is the frequency in the data of each of the levels of the factor. Here "`sex`" is a factor with two levels. From the summary we can see that 54 of the subjects in the sample are female and 46 are male.

Notice that when the input to the function "`summary`" is a data frame, as is the case in this example, then the output is a summary of each of the variables

Here is one way to make sense of this plot. I can imagine that the entire shaded area represents 100% of the things in this sample. I can see that a relatively small percentage of values are less than 137 and a small percentage are more than 200. That means that if I randomly selected a value from this data, there is only a small chance that I would find a value less than 137 or more than 200. The tallest bars are around 170, so if I select a value randomly from this data, I would be more likely to find values near 170 than I would to find values near 200.

The tallest bar is about 3,600 units high, meaning that about 3,600 people in the sample were very close to 170 cm tall (probably rounded to the neares cm).
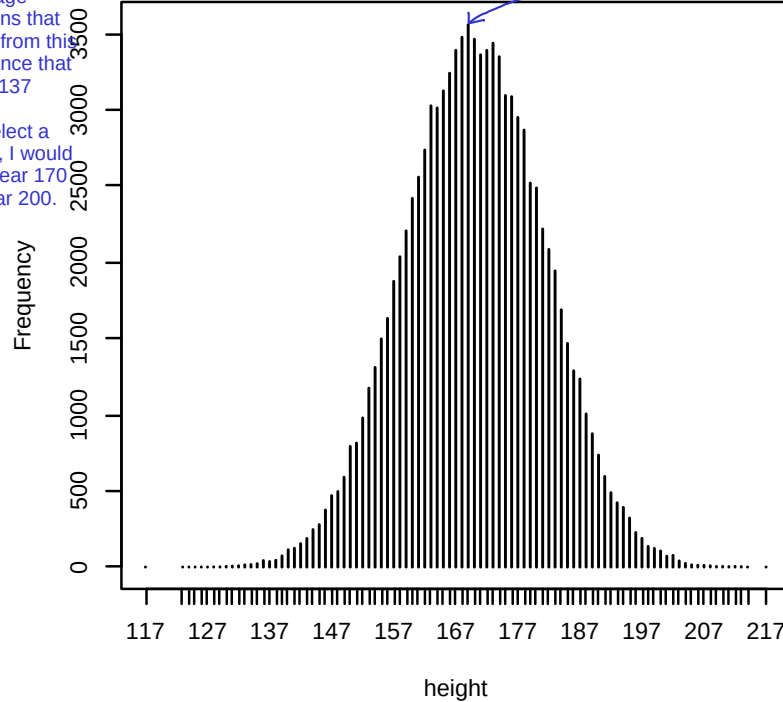
Figure 4.1: Bar Plot of Height

of the data frame. In this example two of the variables are numeric ("id" and "height") and one variable is a factor ("sex").

Recall that the mean is the arithmetic average of the data which is computed by summing all the values of the variable and dividing the result by the number of observations. Hence, if n is the number of observations ($n = 100$ in this example) and x_i is the value of the variable for subject i, then one may write the mean in a formula form as

$$\bar{x} = \frac{\text{Sum of all values in the data}}{\text{Number of values in the data}} = \frac{\sum_{i=1}^{n} x_i}{n} \, ,$$

where $\bar{x}$ corresponds to the mean of the data and the symbol "$\sum_{i=1}^{n} x_i$" corresponds to the sum of all values in the data.

The median is computed by ordering the data values and selecting a value that splits the ordered data into two equal parts. The first and third quartile are obtained by further splitting each of the halves into two quarters.

Let us discuss the variability associated with an entire target population. The file "pop1.csv" that contains the population data can be found on the internet (http://pluto.huji.ac.il/~msby/StatThink/Datasets/pop1.csv). It

is a CSV file that contains the information on sex and height of an entire adult population of some imaginary city. (The data in "ex.1" corresponds to a sample from this city.) Read the population data into R and examine it:

```
> pop.1 <- read.csv(file="pop1.csv")
> summary(pop.1)
       id                sex          height
 Min.   : 1000082   FEMALE:48888   Min.   :117.0
 1st Qu.: 3254220   MALE  :51112   1st Qu.:162.0
 Median : 5502618                  Median :170.0
 Mean   : 5502428                  Mean   :170.0
 3rd Qu.: 7757518                  3rd Qu.:178.0
 Max.   : 9999937                  Max.   :217.0
```

third quartile → 3rd Qu.:178.0
for height data

You sould practice reading CSV data files. Read the output that R gives you when you enter the command. If you see error messages, that means that you might have a typo or that you do not have the data file in your working directory. You must type the command EXACTLY as shown and have a file that has the exact same name in your working directory.

The object "pop.1" is a data frame of the same structure as the data frame "ex.1". It contains three variables: a unique identifier of each subject (id), the sex of the subject (sex), and its height (height). Applying the function "summary" to the data frame produces the summary of the variables that it contains. In particular, for the variable "sex", which is a factor, it produces the frequency of its two categories – 48,888 female and 51,112 – a total of 100,000 subjects. For the variable "height", which is a numeric variable, it produces the extreme values, the quartiles, and the mean.

Let us concentrate on the variable "height". A bar plot of the distribution of the heights in the entire population is given in Figure 4.1[1]. Recall that a vertical bar is placed above each value of height that appears in the population, with the height of the bar representing the frequency of the value in the population. One may read out of the graph or obtain from the numerical summaries that the variable takes integer values in the range between 117 and 217 (heights are rounded to the nearest centimeter). The distribution is centered at 170 centimeter, with the central 50% of the values spreading between 162 and 178 centimeters.

The mean of the height in the entire population is equal to 170 centimeter. This mean, just like the mean for the distribution of data, is obtained by the summation of all the heights in the population divided by the population size. Let us denote the size of the entire population by N. In this example $N = 100,000$. (The size of the sample for the data was called n and was equal to $n = 100$ in the parallel example that deals with the data of a sample.) The mean of an entire population is denoted by the Greek letter μ and is read "*mew*". (The average for the data was denoted $\bar{x}$). The formula of the population mean is:

The character that looks like a "u" is the lower case Greek letter "mu." It represents the population mean.

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}} = \frac{\sum_{i=1}^{N} x_i}{N} .$$

Observe the similarity between the definition of the mean for the data and the definition of the mean for the population. In both cases the arithmetic average is computed. The only difference is that in the case of the mean of the data the computation is with respect to the values that appear in the sample whereas for the population all the values in the population participate in the computation.

---

[1]Such a bar plot can be produced with the expression "plot(table(pop.1$height))".

In actual life, we will not have all the values of a variable in the entire population. Hence, we will not be able to compute the actual value of the population mean. However, it is still meaningful to talk about the population mean because this number exists, even though we do not know what its value is. As a matter of fact, one of the issues in statistics is to try to estimate this unknown quantity on the basis of the data we do have in the sample.

A characteristic of the distribution of an entire population is called a *parameter.* Hence, μ, the population average, is a parameter. Other examples of parameters are the population median and the population quartiles. These parameters are defined exactly like their data counterparts, but with respect to the values of the entire population instead of the observations in the sample alone.

parameters = characteristic of the full population, whereas a "statistic" is a characteristic of a sample.

Another example of a parameter is the population variance. Recall that the sample variance was defined with the aid of the deviations $x_i - \bar{x}$, where x_i is the value of the measurement for the ith subject and $\bar{x}$ is the mean for the data. In order to compute the sample variance these deviations were squared to produce the squared deviations. The squares were summed up and then divided by the sample size minus one $(n - 1)$. The sample variance, computed from the data, was denoted s^2.

The population variance is defined in a similar way. First, the deviations from the population mean $x_i - \mu$ are considered for each of the members of the population. These deviations are squared and the average of the squares is computed. We denote this parameter by σ^2 (read "*sigma square*"). A minor difference between the sample variance and the population variance is that for the latter we should divide the sum of squared deviations by the population size (N) and not by the population size minus one $(N - 1)$:

population variance

This symbol is important. The lower-case sigma squared represents the population variance. Remember that the upper case sigma means "sum."

$$\sigma^2 = \text{The average square deviation in the population}$$
$$= \frac{\text{Sum of the squares of the deviations in the population}}{\text{Number of values in the population}}$$
$$= \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}\ .$$

little sigma without the "squared" is the population standard deviation.

The standard deviation of the population, yet another parameter, is denoted by σ and is equal to the square root of the variance. The standard deviation summarizes the overall variability of the measurement across the population. Again, the typical situation is that we do not know what the actual value of the standard deviation of the population is. Yet, we may refer to it as a quantity and we may try to estimate its value based on the data we do have from the sample.

For the height of the subjects in our imaginary city we get that the variance is equal to $\sigma^2 = 126.1576$. The standard deviation is equal to $\sigma = \sqrt{126.1576} = 11.23199$. These quantities can be computed in this example from the data frame "`pop.1`" with the aid of the functions "`var`" and "`sd`", respectively[2].

---

[2] Observe that the function "`var`" computes the sample variance. Consequently, the sum of squares is divided by $N - 1$. We can correct that when computing the population variance by multiplying the result by $N - 1$ and dividing by N. Notice that the difference between the two quantities is negligible for a large population. Henceforth we will use the functions "`var`" and "`sd`" to compute the variance and standard deviations of populations without the application of the correction.

4.4 Random Variables

In the previous section we dealt with the variability of the population. Next we consider the variability of a random variable. As an example, consider taking a sample of size $n = 1$ from the population (a single person) and measuring his/her height.

The object `pop.1$height` is a sequence with 100,000 entries. Think of it as a population. We will apply the function "`sample`" to this sequence:

```
> sample(pop.1$height,1)
[1] 162
```

This command extracts one value randomly from the height data. This time, R gave you the value 162 from the height data. In the paragraph below, R randomly selected the value 192.

The first entry to the function is the given sequence of heights. When we set the second argument to 1 then the function selects one of the entries of the sequence at random, with each entry having the same likelihood of being selected. Specifically, in this example an entry that contains the value 162 was selected. Let us run the function again:

```
> sample(pop.1$height,1)
[1] 192
```

In this instance an entry with a different value was selected. Try to run the command several times yourself and see what you get. Would you necessarily obtain a different value in each run?

Now let us enter the same command without pressing the return key:

```
> sample(pop.1$height,1)
```

Can you tell, before pressing the key, what value will you get?

The answer to this question is of course *"No"*. There are 100,000 entries with a total of 94 distinct values. In principle, any of the values may be selected and there is no way of telling in advance which of the values will turn out as an outcome.

A random variable is the future outcome of a measurement, **before** the measurement is taken. It does not have a specific value, but rather a collection of potential values with a distribution over these values. After the measurement is taken and the specific value is revealed then the random variable ceases to be a random variable! Instead, it becomes data.

Although one is not able to say what the outcome of a random variable will turn out to be. Still, one may identify patterns in this potential outcome. For example, knowing that the distribution of heights in the population ranges between 117 and 217 centimeter one may say in advance that the outcome of the measurement must also be in that interval. Moreover, since there is a total of 3,476 subjects with height equal to 168 centimeter and since the likelihood of each subject to be selected is equal then the likelihood of selecting a subject of this height is $3,476/100,000 = 0.03476$. In the context of random variables we call this likelihood *probability*. In the same vain, the frequency of subjects with hight 192 centimeter is 488, and therefore the probability of measuring such a height is 0.00488. The frequency of subjects with height 200 centimeter or above is 393, hence the probability of obtaining a measurement in the range between 200 and 217 centimeter is 0.00393.

A random variable is a set of measurements (values) with corresponding probabilities. You could define a random variable that contains only the integers from 1 to 6, each of which having a probability of 1/6. This could be used to model the results of throwing one die (six-sided cube with dots that represent 1 through 6). The random variable is NOT a data sample. You could also define a random variable like the ones in Chapter 5, one of which is a uniform variable that can have any number within a given range, and each number has an equal chance of being randomly selected.

4.4.1   Sample Space and Distribution

Let us turn to the formal definition of a random variable: A random variable re-
fer to numerical values, typically the outcome of an observation, a measurement,
or a function thereof.

A random variable is characterized via the collection of potential values it
may obtain, known as the *sample space* and the likelihood of obtaining each of
the values in the sample space (namely, the probability of the value). In the
given example, the sample space contains the 94 integer values that are marked
in Figure 4.1. The probability of each value is the height of the bar above the
value, divided by the total frequency of 100,000 (namely, the relative frequency
in the population).

We will denote random variables with capital Latin letters such as X, Y,
and Z. Values they may obtain will be marked by small Latin letters such as
x, y, z. For the probability of values we will use the letter "P". Hence, if we
denote by X the measurement of height of a random individual that is sampled
from the given population then:

$$P(X = 168) = 0.03476$$

this says that the probability of randomly
selecting the value 168 from the population
is .03476 (which is about 3.5%).

and

$$P(X \geq 200) = 0.00393 .$$

Consider, as yet another example, the probability that the height of a random
person sampled from the population differs from 170 centimeter by no more
than 10 centimeters. (In other words, that the height is between 160 and 180
centimeters.) Denote by X the height of that random person. We are interested
in the probability $P(|X - 170| \leq 10)$.[3]

this one refers to the probabilty that a randomly selcted value will
deviate from 170
by no more than
10 units. In other
words, the prob-
ability of being
between 160 and
180.

The random person can be any of the subjects of the population with equal
probability. Thus, the sequence of the heights of the 100,000 subjects represents
the distribution of the random variable X:

```
> pop.1 <- read.csv(file="pop1.csv")
> X <- pop.1$height
```

Notice that the object "X" is a sequence of lenght 100,000 that stores all the
heights of the population. The probability we seek is the relative frequency in
this sequency of values between 160 and 180. First we compute the probability
and then explain the method of computation:

```
> mean(abs(X-170) <= 10)
[1] 0.64541
```

The mean of the absolute value of the "difference between X and 170"
is .64541. This is the mean "absolute deviation from the mean".

We get that the height of a person randomly sampled from the population is
between 160 and 180 centimeters with probability 0.64541.

Let us produce a small example that will help us explain the computation
of the probability. We start by forming a sequence with 10 numbers:

```
> Y <- c(6.3, 6.9, 6.6, 3.4, 5.5, 4.3, 6.5, 4.7, 6.1, 5.3)
```

---

[3]The expression $\{|X - 170| \leq 10\}$ reads as "the absolute value of the difference between X
and 170 is no more that 10". In other words, $\{-10 \leq X - 170 \leq 10\}$, which is equivalent to
the statement that $\{160 \leq X \leq 180\}$. It follows that $P(|X - 170| \leq 10) = P(160 \leq X \leq 180)$.

The goal is to compute the proportion of numbers that are in the range $[4, 6]$ (or, equivalently, $\{|Y - 5| \leq 1\}$).

The function "abs" computes the absolute number of its input argument. When the function is applied to the sequence "Y-5" it produces a sequence of the same length with the distances between the components of "Y" and the number 5:

```
> abs(Y-5)
 [1] 1.3 1.9 1.6 1.6 0.5 0.7 1.5 0.3 1.1 0.3
```

Look at the Y values at the bottom of the prior page, then subtract 5 from each one and you get the values in the list here:

Compare the resulting output to the original sequence. The first value in the input sequence is 6.3. Its distance from 5 is indeed 1.3. The fourth value in the input sequence is 3.4. The difference 3.4 - 5 is equal to -1.6, and when the absolute value is taken we get a distance of 1.6.

The function "<=" expects an argument to the right and an argument to the left. It compares each component to the left with the parallel component to the right and returns a logical value, "TRUE" or "FALSE", depending on whether the relation that is tested holds or not:

```
> abs(Y - 5) <= 1
 [1] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
```

This part might seem geeky to business majors. It is not essential to know this, but you should know how to apply it, as in here...

Observe the in this example the function "<=" produced 10 logical values, one for each of the elements of the sequence to the left of it. The first input in the sequence "Y" is 6.3, which is more than one unit away from 5. Hence, the first output of the logical expression is "FALSE". On the other hand, the last input in the sequence "Y" is 5.3, which is within the range. Therefore, the last output of the logical expression is "TRUE".

Next, we compute the proportion of "TRUE" values in the sequence:

```
> mean(abs(Y - 5) <= 1)
 [1] 0.4
```

Note that the things inside the outer ( ) are a "logical comparison," which means that you are asking R to check if the "absolute value of Y - 5" is less than or equal to 1, and if it is, count all the TRUE values as 1 and find the mean of the whole thing. So the .4 number means that 40% of values are between 4 and 6. You can try variations of this with ">=" or two equals signs: "==" where the two equals signs ask R to answer "TRUE" of the two things are equal.

When a sequence with logical values is entered into the function "mean" then the function replaces the TRUE's by 1 and the FALSE's by 0. The average produces then the relative frequency of TRUE's in the sequence as required. Specifically, in this example there are 4 TRUE's and 6 FALSE's. Consequently, the output of the final expression is $4/10 = 0.4$.

The computation of the probability that the sampled height falls within 10 centimeter of 170 is based on the same code. The only differences are that the input sequence "Y" is replaced by the sequence of population heights "X" as input. the number "5" is replaced by the number "170" and the number "1" is replaced by the number "10". In both cases the result of the computation is the relative proportion of the times that the values of the input sequence fall within a given range of the indicated number.

The probability function of a random variable is defined for any value that the random variable may obtain and produces the *distribution* of the random variable. The probability function may emerge as a relative frequency as in the given example or it may be a result of theoretical modeling. Examples of theoretical random variables are presented mainly in the next two chapters.

A distribution

Consider an example of a random variable. The sample space and the probability function specify the distribution of the random variable. For example,

assume it is known that a random variable X may obtain the values 0, 1, 2, or
3. Moreover, imagine that it is known that $P(X = 1) = 0.25$, $P(X = 2) = 0.15$,
and $P(X = 3) = 0.10$. What is $P(X = 0)$, the probability that X is equal to 0?

The sample space, the collection of possible values that the random variable
may obtain is the collection $\{0, 1, 2, 3\}$. Observe that the sum over the positive
values is:

$$P(X > 0) = P(X = 1) + P(X = 2) + P(X = 3) = 0.25 + 0.15 + 0.10 = 0.50 .$$

It follows, since the sum of probabilities over the entire sample space is equal
to 1, that $P(X = 0) = 1 - 0.5 = 0.5$.

Value	Probability	Cum. Prob.
0	0.50	0.50
1	0.25	0.75
2	0.15	0.90
3	0.10	1.00

Table 4.1: The Distribution of X

Table 4.1 summarizes the distribution of the random variable X. Observe
the similarity between the probability function and the notion of relative fre-
quency that was discussed in Chapter 2. Both quantities describe distribution.
Both are non-negative and sum to 1. Likewise, notice that one may define the
cumulative probability the same way cumulative relative frequency is defined:
Ordering the values of the random variable from smallest to largest, the cumu-
lative probability at a given value is the sum of probabilities for values less or
equal to the given value.

Knowledge of the probabilities of a random variable (or the cumulative prob-
abilities) enables the computation of other probabilities that are associated with
the random variable. For example, considering the random variable X of Ta-
ble 4.1, we may calculate the probability of X falling in the interval $[0.5, 2.3]$.
Observe that the given range contains two values from the sample space, 1 and
2, therefore:

$$P(0.5 \leq X \leq 2.3) = P(X = 1) + P(X = 2) = 0.25 + 0.15 = 0.40 .$$

Likewise, we may produce the probability of X obtaining an odd value:

$$P(X = \text{odd}) = P(X = 1) + P(X = 3) = 0.25 + 0.10 = 0.35 .$$

Observe that both $\{0.5 \leq X \leq 2.3\}$ and $\{X = \text{odd}\}$ refer to subsets of values
of the sample space. Such subsets are denoted *events*. In both examples the
probability of the event was computed by the summation of the probabilities
associated with values that belong to the event.

4.4.2   Expectation and Standard Deviation

We may characterize the center of the distribution of a random variable and the
spread of the distribution in ways similar to those used for the characterization
of the distribution of data and the distribution of a population.

The *expectation* marks the center of the distribution of a random variable. It is equivalent to the data average $\bar{x}$ and the population average μ, which was used in order to mark the location of the distribution of the data and the population, respectively.

Recall from Chapter 3 that the average of the data can be computed as the weighted average of the values that are present in the data, with weights given by the relative frequency. Specifically, we saw for the data

$$1, \ 1, \ 1, \ 2, \ 2, \ 3, \ 4, \ 4, \ 4, \ 4, \ 4$$

that

$$\frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 1 \times \frac{3}{11} + 2 \times \frac{2}{11} + 3 \times \frac{1}{11} + 4 \times \frac{5}{11} \ ,$$

producing the value of $\bar{x} = 2.727$ in both representations. Using a formula, the equality between the two ways of computing the mean is given in terms of the equation:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{x} \left(x \times (f_x/n) \right) \ .$$

This equation describes what was done above. Take each x value and multiply it by its relative frequency, then add them all together.

In the first representation of the arithmetic mean, the average is computed by the summation of all data points and dividing the sum by the sample size. In the second representation, that uses a weighted sum, the sum extends over all the unique values that appear in the data. For each unique value the value is multiplied by the relative frequency of the value in the data. These multiplications are summed up to produce the mean.

The expectation of a random variable is computed in the spirit of the second formulation. The expectation of a random variable is marked with the letter "E" and is define via the equation:

Expectation

$$\mathrm{E}(X) = \sum_{x} \left(x \times \mathrm{P}(x) \right) \ .$$

This says almost the same as the equation above, but P(x) refers to the probability that the particular x-value will occur, and that probability might be hypothetical (it might not refer to real data--just a mathematical possibility).

In this definition all the unique values of the sample space are considered. For each value a product of the value and the probability of the value is taken. The expectation is obtained by the summation of all these products. In this definition the probability $\mathrm{P}(x)$ replaces the relative frequency f_x/n but otherwise, the definition of the expectation and the second formulation of the mean are identical to each other.

Consider the random variable X with distribution that is described in Table 4.1. In order to obtain its expectation we multiply each value in the sample space by the probability of the value. Summation of the products produces the expectation (see Table 4.2):

this is an example of the previous equation for expectation.

$$\mathrm{E}(X) = 0 \times 0.5 + 1 \times 0.25 + 2 \times 0.15 + 3 \times 0.10 = 0.85 \ .$$

In the example of height we get that the expectation is equal to 170.035 centimeter. Notice that this expectation is equal to μ, the mean of the population[4]. This is no accident. The expectation of a potential measurement of a randomly selected subject from a population is equal to the average of the measurement across all subjects.

---

[4]The mean of the population can be computed with the expression "`mean(pop.1$height)`"

X   P(X)

Value	Probability	$x \times \mathrm{P}(X = x)$
0	0.50	0.00
1	0.25	0.25
2	0.15	0.30
3	0.10	0.30
		$\mathrm{E}(X) = 0.85$

Table 4.2: The Expectation of X

The sample variance (s^2) is obtained as the sum of the squared deviations from the average, divided by the sample size (n) minus 1:

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

A second formulation for the computation of the same quantity is via the use of relative frequencies. The formula for the sample variance takes the form

$$s^2 = \frac{n}{n - 1} \sum_x \left((x - \bar{x})^2 \times (f_x/n) \right) \ .$$

In this formulation one considers each of the unique value that are present in the data. For each value the deviation between the value and the average is computed. These deviations are then squared and multiplied by the relative frequency. The products are summed up. Finally, the sum is multiplied by the ratio between the sample size n and $n - 1$ in order to correct for the fact that in the sample variance the sum of squared deviations is divided by the sample size minus 1 and not by the sample size.

In a similar way, the variance of a random variable may be defined via the probability of the values that make the sample space. For each such value one computes the deviation from the expectation. This deviation is then squared and multiplied by the probability of the value. The multiplications are summed up in order to produce the variance:

x - E(X) is similar to x - x-bar.

Population Variance $$\mathrm{Var}(X) = \sum_x \left((x - \mathrm{E}(X))^2 \times \mathrm{P}(x) \right) \ .$$

Notice that the formula for the computation of the variance of a random variable is very similar to the second formulation for the computation of the sample variance. Essentially, the mean of the data is replaced by the expectation of the random variable and the relative frequency of a value is replaced by the probability of the value. Another difference is that the correction factor is not used for the variance of a random variable.

As an example consider the variance of the random variable X. The computation of the variance of this random variable is carried out in Table 4.3). The sample space, the values that the random variable may obtain, are given in the first column and the probabilities of the values are given in the second column. In the third column the deviation of the value from the expectation $\mathrm{E}(X) = 0.85$ is computed for each value. The 4th column contains the square of these deviations and the 5th and last column involves the product of the square deviations and the probabilities. The variance is obtained by summing up the

Value	Prob.	$x - \mathrm{E}(X)$	$(x - \mathrm{E}(X))^2$	$(x - \mathrm{E}(X))^2 \times \mathrm{P}(X = x)$
0	0.50	-0.85	0.7225	0.361250
1	0.25	0.15	0.0225	0.005625
2	0.15	1.15	1.3225	0.198375
3	0.10	2.15	4.6225	0.462250
				$\mathrm{Var}(X) = 1.027500$

Table 4.3: The Variance of X

products in the last column. In the given example:

$$\begin{aligned}
\mathrm{Var}(X) =& (0 - 0.85)^2 \times 0.5 + (1 - 0.85)^2 \times 0.25 \\
& + (2 - 0.85)^2 \times 0.15 + (3 - 0.85)^2 \times 0.10 = 1.0275 \ .
\end{aligned}$$

The standard deviation of a random variable is the square root of the variance. The standard deviation of X is $\sqrt{\mathrm{Var}(X)} = \sqrt{1.0275} = 1.013657$.

In the example that involves the height of a subject selected from the population at random we obtain that the variance is 126.1576, equal to the population variance, and the standard deviation is 11.23199, the square root of the variance.

Other characterization of the distribution that were computed for data, such as the median, the quartiles, etc., may also be defined for random variables.

4.5 Probability and Statistics

Modern science may be characterized by a systematic collection of empirical measurements and the attempt to model laws of nature using mathematical language. The drive to deliver better measurements led to the development of more accurate and more sensitive measurement tools. Nonetheless, at some point it became apparent that measurements may not be perfectly reproducible and any repeated measurement of presumably the exact same phenomena will typically produce variability in the outcomes. On the other hand, scientists also found that there are general laws that govern this variability in repetitions. For example, it was discovered that the average of several independent repeats of the measurement is less variable and more reproducible than each of the single measurements themselves.

Probability was first introduced as a branch of mathematics in the investigation of uncertainty associated with gambling and games of chance. During the early 19th century probability began to be used in order to model variability in measurements. This application of probability turned out to be very successful. Indeed, one of the major achievements of probability was the development of the mathematical theory that explains the phenomena of reduced variability that is observed when averages are used instead of single measurements. In Chapter **??** we discuss the conclusions of this theory.

Statistics study method for inference based on data. Probability serves as the mathematical foundation for the development of statistical theory. In this chapter we introduced the probabilistic concept of a random variable. This concept is key for understanding statistics. In the rest of Part I of this book we discuss the probability theory that is used for statistical inference. Statistical inference itself is discussed in Part II of the book.
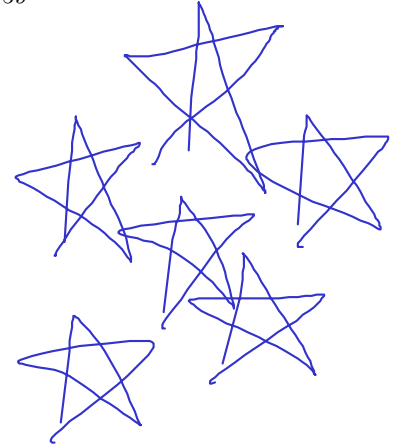
Value	Probability
0	p
1	$2p$
2	$3p$
3	$4p$
4	$5p$
5	$6p$

Table 4.4: The Distribution of Y

4.6   Solved Exercises

Question 4.1. Table 4.6 presents the probabilities of the random variable Y. These probabilities are a function of the number p, the probability of the value "0". Answer the following questions:

1. What is the value of p?

2. $\mathrm{P}(Y < 3) = ?$

3. $\mathrm{P}(Y = \text{odd}) = ?$

4. $\mathrm{P}(1 \leq Y < 4) = ?$

5. $\mathrm{P}(|Y - 3| < 1.5) = ?$

6. $\mathrm{E}(Y) = ?$

7. $\mathrm{Var}(Y) = ?$

8. What is the standard deviation of Y.

Solution (to Question 4.1.1): Consult Table 4.6. The probabilities of the different values of Y are $\{p, 2p, \dots, 6p\}$. These probabilities sum to 1, consequently

$$p + 2p + 3p + 4p + 5p + 6p = (1 + 2 + 3 + 4 + 5 + 6)p = 21p = 1 \implies p = 1/21 .$$

Solution (to Question 4.1.2): The event $\{Y < 3\}$ contains the values 0, 1 and 2. Therefore,

$$\mathrm{P}(Y < 3) = \mathrm{P}(Y = 0) + \mathrm{P}(Y = 1) + \mathrm{P}(Y = 2) = \frac{1}{21} + \frac{2}{21} + \frac{3}{21} = \frac{6}{21} = 0.2857 .$$

Solution (to Question 4.1.3): The event $\{Y = \text{odd}\}$ contains the values 1, 3 and 5. Therefore,

$$\mathrm{P}(Y = \text{odd}) = \mathrm{P}(Y = 1) + \mathrm{P}(Y = 3) + \mathrm{P}(Y = 5) = \frac{2}{21} + \frac{4}{21} + \frac{6}{21} = \frac{12}{21} = 0.5714 .$$

Solution (to Question 4.1.4): The event $\{1 \leq Y < 4\}$ contains the values 1, 2 and 3. Therefore,

$$P(1 \leq Y < 4) = P(Y = 1) + P(Y = 2) + P(Y = 3) = \frac{2}{21} + \frac{3}{21} + \frac{4}{21} = \frac{9}{21} = 0.4286 \ .$$

Solution (to Question 4.1.5): The event $\{|Y - 3| < 1.5\}$ contains the values 2, 3 and 4. Therefore,

$$P(|Y-3| < 1.5) = P(Y = 2) + P(Y = 3) + P(Y = 4) = \frac{3}{21} + \frac{4}{21} + \frac{5}{21} = \frac{12}{21} = 0.5714 \ .$$

Solution (to Question 4.1.6): The values that the random variable Y obtains are the numbers 0, 1, 2, ..., 5, with probabilities $\{1/21, 2/21, \ldots, 6/21\}$, respectively. The expectation is obtained by the multiplication of the values by their respective probabilities and the summation of the products. Let us carry out the computation in R:

```
> Y.val <- c(0,1,2,3,4,5)
> P.val <- c(1,2,3,4,5,6)/21
> E <- sum(Y.val*P.val)
> E
[1] 3.333333
```

note that when you divide these numbers by their sum, the set of values for P.val will be decimals that sum to 1. That represents the probability that the corresponding value from Y.val will occur. Probabilities for a random variable should always add to 1, so run an R command like sum(P.val) and confirm that it is 1.

We obtain an expectation $E(Y) = 3.3333$.

Solution (to Question 4.1.7): The values that the random variable Y obtains are the numbers 0, 1, 2, ..., 5, with probabilities $\{1/21, 2/21, \ldots, 6/21\}$, respectively. The expectation is equal to $E(Y) = 3.333333$. The variance is obtained by the multiplication of the squared deviation from the expectation of the values by their respective probabilities and the summation of the products. Let us carry out the computation in R:

```
> Var <- sum((Y.val-E)^2*P.val)
> Var
[1] 2.222222
```

The R objects called "E" and "P.val"... come from above.

*** You can use this technique to calculate the variance of a random variable when you have a set of values for the sample space (Y.val) and corresponding relative frequencies (P.val).

We obtain a variance $Var(Y) = 2.2222$.

Solution (to Question 4.1.8): The standard deviation is the square root of the variance: $\sqrt{Var(Y)} = \sqrt{2.2222} = 1.4907$.

Question 4.2. One invests \$2 to participate in a game of chance. In this game a coin is tossed three times. If all tosses end up "Head" then the player wins \$10. Otherwise, the player losses the investment.

1. What is the probability of winning the game?

2. What is the probability of loosing the game?

3. What is the expected gain for the player that plays this game? (Notice that the expectation can obtain a negative value.)

Solution (to Question 4.2.1): An outcome of the game of chance may be represented by a sequence of length three composed of the letters "H" and "T". For example, the sequence "THH" corresponds to the case where the first toss produced a "Tail", the second a "Head" and the third a "Head".

With this notation we obtain that the possible outcomes of the game are $\{HHH, THH, HTH, TTH, HHT, THT, HTT, TTT\}$. All outcomes are equally likely. There are 8 possible outcomes and only one of which corresponds to winning. Consequently, the probability of winning is $1/8$.

Solution (to Question 4.2.2): Consider the previous solution. One looses if any other of the outcomes occurs. Hence, the probability of loosing is $7/8$.

Solution (to Question 4.2.3): Denote the gain of the player by X. The random variable X may obtain two values: 10-2 $= 8$ if the player wins and -2 if the player looses. The probabilities of these values are $\{1/8, 7/8\}$, respectively. Therefore, the expected gain, the expectation of X is:

$$\mathrm{E}(X) = 8 \times \frac{1}{8} + (-2) \times \frac{7}{8} = -0.75 \ .$$

4.7   Summary

Glossary

Random Variable: The probabilistic model for the value of a measurement, before the measurement is taken.

Sample Space: The set of all values a random variable may obtain.

Probability: A number between 0 and 1 which is assigned to a subset of the sample space. This number indicates the likelihood of the random variable obtaining a value in that subset.

Expectation: The central value for a random variable. The expectation of the random variable X is marked by $\mathrm{E}(X)$.

Variance: The (squared) spread of a random variable. The variance of the random variable X is marked by $\mathrm{V}ar(X)$. The standard deviation is the square root of the variance.

Discussion in the Forum

Random variables are used to model situations in which the outcome, before the fact, is uncertain. One component in the model is the sample space. The sample space is the list of all possible outcomes. It includes the outcome that took place, but also all other outcomes that could have taken place but never did materialize. The rationale behind the consideration of the sample space is

the intention to put the outcome that took place in context. What do you think of this rationale?

When forming your answer to this question you may give an example of a situation from you own field of interest for which a random variable can serve as a model. Identify the sample space for that random variable and discuss the importance (or lack thereof) of the correct identification of the sample space.

For example, consider a factory that produces car parts that are sold to car makers. The role of the QA personnel in the factory is to validate the quality of each batch of parts before the shipment to the client.

To achieve that, a sample of parts may be subject to a battery of quality test. Say that 20 parts are selected to the sample. The number of those among them that will not pass the quality testing may be modeled as a random variable. The sample space for this random variable may be any of the numbers 0, 1, 2, ..., 20.

The number 0 corresponds to the situation where all parts in the sample passed the quality testing. The number 1 corresponds to the case where 1 part did not pass and the other 19 did. The number 2 describes the case where 2 of the 20 did not pass and 18 did pass, etc.

Summary of Formulas

Population Size: $N =$ the number of people, things, etc. in the population.

Population Average: $\mu = (1/N) \sum_{i=1}^{N} x_i$ We usually say "population mean" instead of "population average."

Expectation of a Random Variable: $E(X) = \sum_x \left(x \times P(x) \right)$

Population Variance: $\sigma^2 = (1/N) \sum_{i=1}^{N} (x_i - \mu)^2$

Variance of a Random Variable: $\mathrm{Var}(X) = \sum_x \left((x - E(X))^2 \times P(x) \right)$

Chapter 5

Random Variables

5.1   Student Learning Objective

This section introduces some important examples of random variables. The distributions of these random variables emerge as mathematical models of real-life settings. In two of the examples the sample space is composed of integers. In the other two examples the sample space is made of continuum of values. For random variables of the latter type one may use the density, which is a type of a histogram, in order to describe the distribution.

By the end of the chapter the student should:

- Identify the Binomial, Poisson, Uniform, and Exponential random variables, relate them to real life situations, and memorize their expectations and variances.

- Relate the plot of the density/probability function and the cumulative probability function to the distribution of a random variable.

- Become familiar with the R functions that produce the density/probability of these random variables and their cumulative probabilities.

- Plot the density and the cumulative probability function of a random variable and compute probabilities associated with random variables.

5.2   Discrete Random Variables

In the previous chapter we introduced the notion of a random variable. A random variable corresponds to the outcome of an observation or a measurement prior to the actual making of the measurement. In this context one can talk of all the values that the measurement may potentially obtain. This collection of values is called the *sample space*. To each value in the sample space one may associate the *probability* of obtaining this particular value. Probabilities are like relative frequencies. All probabilities are positive and the sum of the probabilities that are associated with all the values in the sample space is equal to one.

A random variable is defined by the identification of its sample space and the probabilities that are associated with the values in the sample space. For

remember that "sample space" refers to the set of numbers that are allowed to appear in the random variable (or distribution).

"Discrete" means that the values in the distribution are from a finite set of things and you can not automatically assume that all decimal values between the minimum and maximum are valid. In this class, all of our discrete variables will be comprised of integers -- that means that decimal values in these distributions are illegal.

65

each type of random variable we will identify first the sample space — the values it may obtain — and then describe the probabilities of the values. Examples of situations in which each type of random variable may serve as a model of a measurement will be provided. The R system provides functions for the computation of probabilities associated with specific types of random variables. We will use these functions in this and in proceeding chapters in order to carry out computations associated with the random variables and in order to plot their distributions.

The distribution of a random variable, just like the distribution of data, can be characterized using numerical summaries. For the latter we used summaries such as the mean and the sample variance and standard deviation. The mean is used to describe the central location of the distribution and the variance and standard deviation are used to characterize the total spread. Parallel summaries are used for random variable. In the case of a random variable the name *expectation* is used for the central location of the distribution and the *variance* and the *standard deviation* (the square root of the variation) are used to summarize the spread. In all the examples of random variables we will identify the expectation and the variance (and, thereby, also the standard deviation).

Random variables are used as probabilistic models of measurements. Theoretical considerations are used in many cases in order to define random variables and their distribution. A random variable for which the values in the sample space are separated from each other, say the values are integers, is called a *discrete random variable*. In this section we introduce two important integer-valued random variables: The *Binomial* and the *Poisson* random variables. These random variables may emerge as models in contexts where the measurement involves counting the number of occurrences of some phenomena.

Many other models, apart from the Binomial and Poisson, exist for discrete random variables. An example of such model, the Negative-Binomial model, will be considered in Section 5.4. Depending on the specific context that involves measurements with discrete values, one may select the Binomial, the Poisson, or one of these other models to serve as a theoretical approximation of the distribution of the measurement.

5.2.1   The Binomial Random Variable   (this is a discrete distribution... integers only)

We usually associate the binomial distribution with a coin-toss experiment, but remember that the probability of success can be anything between 0 and 1.

The Binomial random variable is used in settings in which a trial that has two possible outcomes is repeated several times. Let us designate one of the outcomes as "Success" and the other as "Failure". Assume that the probability of success in each trial is given by some number p that is larger than 0 and smaller than 1. Given a number n of repeats of the trial and given the probability of success, the actual number of trials that will produce "Success" as their outcome is a random variable. We call such random variable *Binomial*. The fact that a random variable X has such a distribution is marked by the expression: "$X \sim \text{Binomial}(n, p)$".

As an example consider tossing 10 coins. Designate "Head" as success and "Tail" as failure. For fair coins the probability of "Head" is 1/2. Consequently, if X is the total number of "Heads" then $X \sim \text{Binomial}(10, 0.5)$, where $n = 10$ is the number of trials and $p = 0.5$ is the probability of success in each trial.

It may happen that all 10 coins turn up "Tail". In this case X is equal to 0. It may also be the case that one of the coins turns up "Head" and the others

based on an eyeball measurement, it looks like the tallest bar is at around .24 (a more precise value is on the next page)

The tallest bar indicates that the value 5 is the most frequent value, and that it occurs about 24% of the time in this random variable.

This plot shows that there are 11 values in the distribution: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Because we know that there are 10 experiments in this example, You can not possibly have 11 or more successes.

Figure 5.1: The Binomial(10,0.5) Distribution

turn up "Tail". The random variable X will obtain the value 1 in such a case. Likewise, for any integer between 0 and 10 it may be the case that the number of "Heads" that turn up is equal to that integer with the other coins turning up "Tail". Hence, the sample space of X is the set of integers $\{0, 1, 2, \ldots, 10\}$. The probability of each outcome may be computed by an appropriate mathematical formula that will not be discussed here[1].

The probabilities of the various possible values of a Binomial random variable may be computed with the aid of the R function "dbinom" (that uses the mathematical formula for the computation). The input to this function is a sequence of values, the value of n, and the value of p. The output is the sequence of probabilities associated with each of the values in the first input.

For example, let us use the function in order to compute the probability that the given Binomial obtains an odd value. A sequence that contains the odd values in the Binomial sample space can be created with the expression "c(1,3,5,7,9)". This sequence can serve as the input in the first argument of the function "dbinom". The other arguments are "10" and "0.5", respectively:

---

[1] If $X \sim$ Binomial(n, p) then $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$, for $x = 0, 1, \ldots, n$.

This R code does many things all at once. The set of numbers 1, 3, 5, 7, 9, all combined in the c() command, are asking R to calculate the dbinom results for 1, 3, 5, 7, and 9, just as would happen if you ran each value separately.

The last two arguments are saying that there will be 10 experiments, each of which have a probability of success of .5 (meaning 50%).

the tallest bar on the prior page is actually this tall.

```
> dbinom(c(1,3,5,7,9),10,0.5)
[1] 0.009765625 0.117187500 0.246093750 0.117187500 0.009765625
```

The value .009765725 is the binomial probability of getting exactly one success in 10 trials of an experiment where each trial has a .50 probablity of success.

The value .246093750 is the third value in the answer, so it corresponds to the third number in the input, which is 5. So the probability of getting exactly 5 successes in this experiment is around .246.

Observe that the output of the function is a sequence of the same length as the first argument. This output contains the Binomial probabilities of the values in the first argument. In order to obtain the probability of the event {X is odd} we should sum up these probabilities, which we can do by applying the function "sum" to the output of the function that computes the Binomial probabilities:

```
> sum(dbinom(c(1,3,5,7,9),10,0.5))
[1] 0.5
```

When you put the dbinom() function inside the sum() function, you are telling R to first calculate all the values shown above, then add them all. The answer means that 50% of all the results will have either 1, 3, 5, 7, or 9 successes.

Observe that the probability of obtaining an odd value in this specific case is equal to one half.

Another example is to compute all the probabilities of all the potential values of a Binomial(10, 0.5) random variable:

```
> x <- 0:10
> dbinom(x,10,0.5)
 [1] 0.0009765625 0.0097656250 0.0439453125 0.1171875000
 [5] 0.2050781250 0.2460937500 0.2050781250 0.1171875000
 [9] 0.0439453125 0.0097656250 0.0009765625
```

The expression "start.value:end.value" produces a sequence of numbers that initiate with the number "start.value" and proceeds in jumps of size one until reaching the number "end.value". In this example, "0:10" produces the sequence of integers between 0 and 10, which is the sample space of the current Binomial example. Entering this sequence as the first argument to the function "dbinom" produces the probabilities of all the values in the sample space.

One may display the distribution of a discrete random variable with a bar plot similar to the one used to describe the distribution of data. In this plot a vertical bar representing the probability is placed above each value of the sample space. The hight of the bar is equal to the probability. A bar plot of the Binomial(10, 0.5) distribution is provided in Figure 5.1.

Another useful function is "pbinom", which produces the cumulative probability of the Binomial:

```
> pbinom(x,10,0.5)
 [1] 0.0009765625 0.0107421875 0.0546875000 0.1718750000
 [5] 0.3769531250 0.6230468750 0.8281250000 0.9453125000
 [9] 0.9892578125 0.9990234375 1.0000000000
> cumsum(dbinom(x,10,0.5))
 [1] 0.0009765625 0.0107421875 0.0546875000 0.1718750000
 [5] 0.3769531250 0.6230468750 0.8281250000 0.9453125000
 [9] 0.9892578125 0.9990234375 1.0000000000
```

The output of the function "pbinom" is the cumulative probability $P(X \leq x)$ that the random variable is less than or equal to the input value. Observe that this cumulative probability is obtained by summing all the probabilities associated with values that are less than or equal to the input value. Specifically, the cumulative probability at $x = 3$ is obtained by the summation of the

probabilities at $x = 0$, $x = 1$, $x = 2$, and $x = 3$:

$$P(X \leq 3) = 0.0009765625 + 0.009765625 + 0.0439453125 + 0.1171875 = 0.171875$$

The numbers in the sum are the first 4 values from the output of the function "dbinom(x,10,0.5)", which computes the probabilities of the values of the sample space.

In principle, the expectation of the Binomial random variable, like the expectation of any other (discrete) random variable is obtained from the application of the general formulae:

These are the generic equations for expectation and variance from the prior chapter.

$$E(X) = \sum_x \left(x \times P(X = x) \right), \quad \mathrm{Var}(X) = \sum_x \left((x - E(X))^2 \times P(x) \right).$$

However, in the specific case of the Binomial random variable, in which the probability $P(X = x)$ obeys the specific mathematical formula of the Binomial distribution, the expectation and the variance reduce to the specific formulae:

We can use this calculation (which is more efficient then the generic calculation in the previous paragraph) when we are working with a binomial random variable (distribution)

$$E(X) = np, \quad \mathrm{Var}(X) = np(1 - p).$$

Expectation and Variance of a Binomial

Hence, the expectation is the product of the number of trials n with the probability of success in each trial p. In the variance the number of trials is multiplied by the product of a probability of success (p) with the probability of a failure $(1 - p)$.

As illustration, let us compute for the given example the expectation and the variance according to the general formulae for the computation of the expectation and variance in random variables and compare the outcome to the specific formulae for the expectation and variance in the Binomial distribution:

X.val is the sample space, and P.val is the list of probabilities for the corresponding X.val.

```
> X.val <- 0:10
> P.val <- dbinom(X.val,10,0.5)
> EX <- sum(X.val*P.val)
> EX
[1] 5 = expectation
> sum((X.val-EX)^2*P.val)
[1] 2.5 = Variance
```

0:10 creates a list of integers from 0 - 10.

This is how to calculate the expectation and variance using the old method, and you can also use:
*E = 10 * .5*
*V = 10 * .5 * (1 - .5)*
You should get the same answers as those found on the lef if there are no typos.

This agrees with the specific formulae for Binomial variables, since $10 \times 0.5 = 5$ and $10 \times 0.5 \times (1 - 0.5) = 2.5$.

Recall that the general formula for the computation of the expectation calls for the multiplication of each value in the sample space with the probability of that value, followed by the summation of all the products. The object "X.val" contains all the values of the random variable and the object "P.val" contains the probabilities of these values. Hence, the expression "X.val*P.val" produces the product of each value of the random variable times the probability of that value. Summation of these products with the function "sum" gives the expectation, which is saved in an object that is called "EX".

X.val contains the complete sample space

The general formula for the computation of the variance of a random variable involves the product of the squared deviation associated with each value with the probability of that value, followed by the summation of all products. The expression "(X.val-EX)^2" produces the sequence of squared deviations from the expectation for all the values of the random variable. Summation of the

You can calculate the variance using this trick when you are given a table of values (the sample space) and the corresponding relative frequencies (probabilities).

70   CHAPTER 5. RANDOM VARIABLES

product of these squared deviations with the probabilities of the values (the outcome of "`(X.val-EX)^2*P.val`") gives the variance.

When the value of p changes (without changing the number of trials n) then the probabilities that are assigned to each of the values of the sample space of the Binomial random variable change, but the sample space itself does not. For example, consider rolling a die 10 times and counting the number of times that the face 3 was obtained. Having the face 3 turning up is a "Success". The probability p of a success in this example is $1/6$, since the given face is one out of 6 equally likely faces. The resulting random variable that counts the total number of success in 10 trials has a Binomial$(10, 1/6)$ distribution. The sample space is yet again equal to the set of integers $\{0, 1, \ldots, 10\}$. However, the probabilities of values are different. These probabilities can again be computes with the aid of the function "`dbinom`":

The "1/6" means that the probability is around 0.16666.

```
> dbinom(x,10,1/6)
 [1] 1.615056e-01 3.230112e-01 2.907100e-01 1.550454e-01
 [5] 5.426588e-02 1.302381e-02 2.170635e-03 2.480726e-04
 [9] 1.860544e-05 8.269086e-07 1.653817e-08
```

In this case, x was previously set to:
x <- 0:10
which is the list of numbers from 0 to 10 inclusive.

In this case smaller values of the random variable are assigned higher probabilities and larger values are assigned lower probabilities..

In Figure 5.2 the probabilities for Binomial$(10, 1/6)$, the Binomial$(10, 1/2)$, and the Binomial$(10, 0.6)$ distributions are plotted side by side. In all these 3 distributions the sample space is the same, the integers between 0 and 10. However, the probabilities of the different values differ. (Note that all bars should be placed on top of the integers. For clarity of the presentation, the bars associated with the Binomial$(10, 1/6)$ are shifted slightly to the left and the bars associated with the Binomial$(10, 0.6)$ are shifted slightly to the right.)

The expectation of the Binomial$(10, 0.5)$ distribution is equal to $10 \times 0.5 = 5$. Compare this to the expectation of the Binomial$(10, 1/6)$ distribution, which is $10 \times (1/6) = 1.666667$ and to the expectation of the Binomial$(10, 0.6)$ distribution which equals $10 \times 0.6 = 6$.

The variance of the Binomial$(10, 0.5)$ distribution is $10 \times 0.5 \times 0.5 = 2.5$. The variance when $p = 1/6$ is $10 \times (1/6) \times (5/6) = 1.388889$ and the variance when $p = 0.6$ is $10 \times 0.6 \times 0.4 = 2.4$.

Example 5.1. *As an application of the Binomial distribution consider a pre-election poll. A candidate is running for office and is interested in knowing the percentage of support in the general population in its candidacy. Denote the probability of support by p. In order to estimate the percentage a sample of size 300 is selected from the population. Let X be the count of supporters in the sample. A natural model for the distribution of X is the* Binomial$(300, p)$ *distribution, since each subject in the sample may be a supporter ("Success") or may not be a supporter ("Failure"). The probability that a subject supports the candidate is p and there are n = 300 subjects in the sample.*

Example 5.2. *As another example consider the procedure for quality control that is described in Discussion Forum of Chapter 4. According to the procedure 20 items are tested and the number of faulty items is recorded. If p is the probability that an item is identified as faulty then the distribution of the total number of faulty items may be modeled by the* Binomial$(20, p)$ *distribution.*
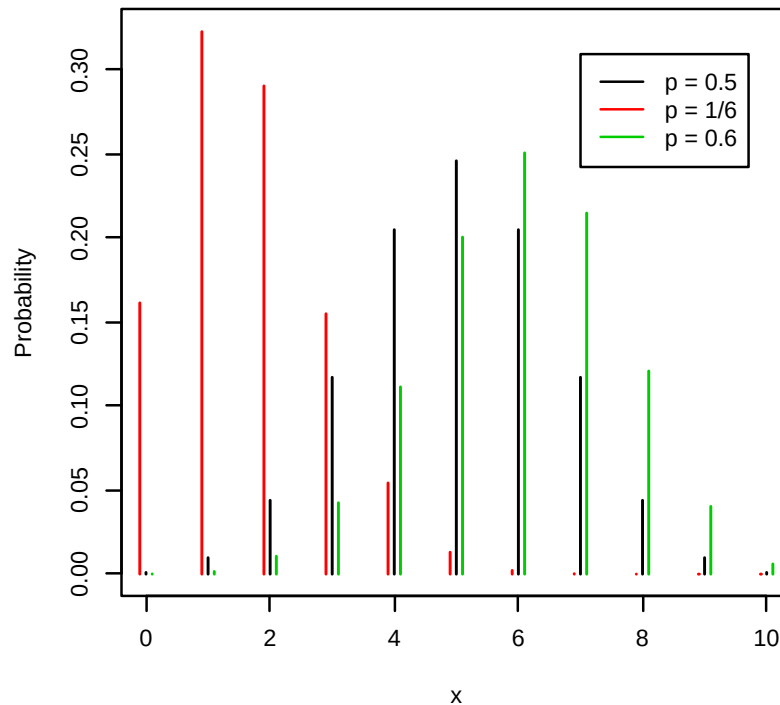
Figure 5.2: The Binomial Distribution for Various Probability of "Success" p

In both examples one may be interested in making statements on the probability p based on the sample. Statistical inference relates the actual count obtained in the sample to the theoretical Binomial distribution in order to make such statements.

5.2.2 The Poisson Random Variable This is also a discrete variable: only integers

The *Poisson* distribution is used as an approximation of the total number of occurrences of rare events. Consider, for example, the Binomial setting that involves n trials with p as the probability of success of each trial. Then, if p is small but n is large then the number of successes X has, approximately, the Poisson distribution.

The sample space of the Poisson random variable is the unbounded collection of integers: $\{0, 1, 2, \ldots\}$. Any integer value is assigned a positive probability. Hence, the Poisson random variable is a convenient model when the maximal number of occurrences of the events in a-priori unknown or is very large. For example, one may use the Poisson distribution to model the number of phone calls that enter a switchboard in a given interval of time or the number of

Note that the shape of the Poisson distribution will change with different vales for lambda.
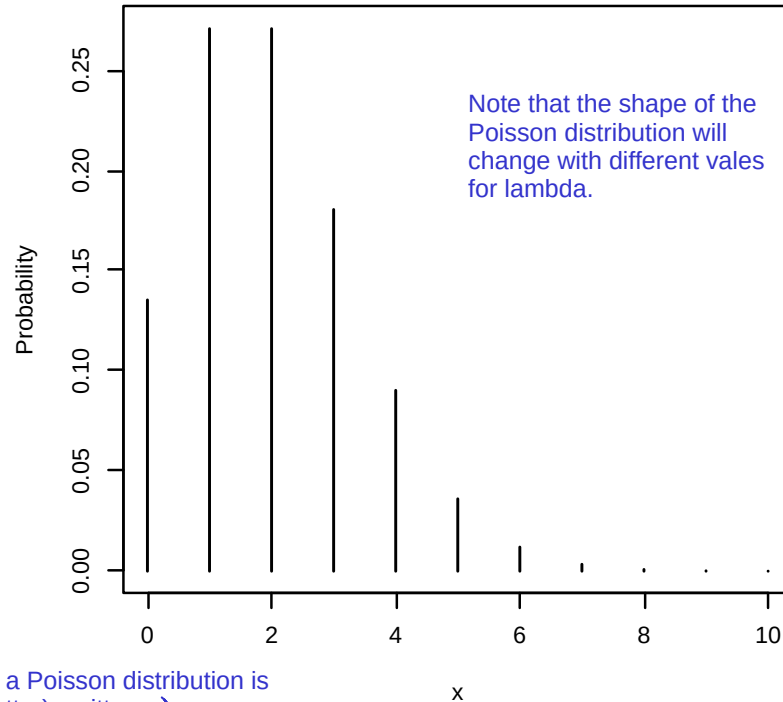
The expectation of a Poisson distribution is lambda (a Greek letter), written: λ

Figure 5.3: The Poisson(2) Distribution

malfunctioning components in a shipment of some product.

The Binomial distribution was specified by the number of trials n and probability of success in each trial p. The Poisson distribution is specified by its expectation, which we denote by λ. The expression "$X \sim \text{Poisson}(\lambda)$" states that the random variable X has a Poisson distribution[2] with expectation $E(X) = \lambda$. The function "dpois" computes the probability, according to the Poisson distribution, of values that are entered as the first argument to the function. The expectation of the distribution is entered in the second argument. The function "ppois" computes the cumulative probability. Consequently, we can compute the probabilities and the cumulative probabilities of the values between 0 and 10 for the Poisson(2) distribution via:

The 2 means that we are considering a Poisson distribution that has an expectation of 2.

```
> x <- 0:10
> dpois(x,2)
 [1] 1.353353e-01 2.706706e-01 2.706706e-01 1.804470e-01
 [5] 9.022352e-02 3.608941e-02 1.202980e-02 3.437087e-03
 [9] 8.592716e-04 1.909493e-04 3.818985e-05
```

Remember that 0:10 tells R to create the list of integers from 0 - 10.

you can ignore this equation for this class.

[2]If $X \sim \text{Poisson}(\lambda)$ then $P(X = x) = e^{-\lambda}\lambda^x/x!$, for $x = 0, 1, 2, \ldots$.

The input list was 0:10, which contains 11 numbers. The last number of the output corresponds to the probability of getting exactly 10 success when you have a Poisson distribution with expectation (lambda) of 2.

3.818985e-05 is a notation that means approximately $3.8 * 10^{-5}$ which is 0.00003818985
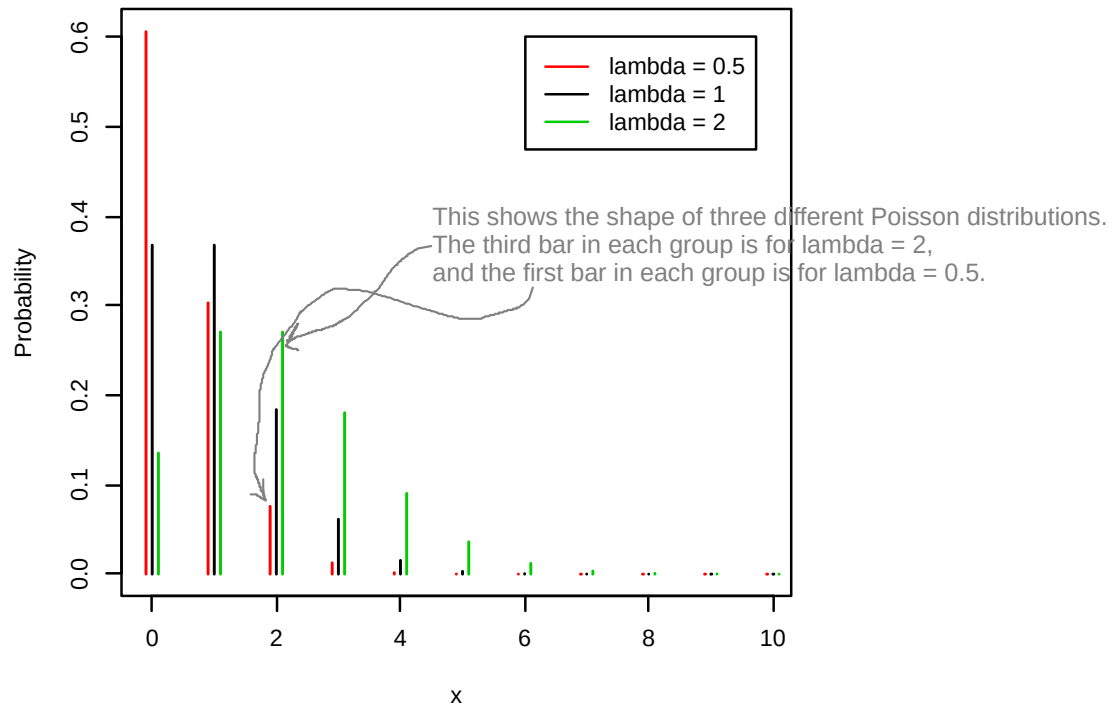
Figure 5.4: The Poisson Distribution for Various Values of λ

The ppois function tells you the cumulative probability of getting less than
or equal to the input value when the expectation is 2. In this case by the time we get
to 6, the cumulative probability is .9954 or 99.54%.

```
> ppois(x,2)
 [1] 0.1353353 0.4060058 0.6766764 0.8571235 0.9473470 0.9834364
 [7] 0.9954662 0.9989033 0.9997626 0.9999535 0.9999917
```

The probability function of the Poisson distribution with $\lambda = 2$, in the range
between 0 and 10, is plotted in Figure 5.3. Observe that in this example prob-
abilities of the values 8 and beyond are very small. As a matter of fact, the
cumulative probability at $x = 7$ (the 8th value in the output of "ppois(x,2)")
is approximately 0.999, out of the total cumulative probability of 1.000, leaving
a total probability of about 0.001 to be distributed among all the values larger
than 7.

Let us compute the expectation of the given Poisson distribution:

```
> X.val <- 0:10
> P.val <- dpois(X.val,2)
> sum(X.val*P.val)
[1] 1.999907
```

This calculation of the expectation uses the generic
equation from Chapter 4.

Observe that the outcome is almost, but not quite, equal to 2.00, which is the
actual value of the expectation. The reason for the inaccuracy is the fact that

we have based the computation in R on the first 11 values of the distribution only, instead of the infinite sequence of values. A more accurate result may be obtained by the consideration of the first 101 values:

```
> X.val <- 0:100
> P.val <- dpois(X.val,2)
> EX <- sum(X.val*P.val)
> EX
[1] 2
> sum((X.val-EX)^2*P.val)
[1] 2
```

In the last expression we have computed the variance of the Poisson distribution and obtained that it is equal to the expectation. This results can be validated mathematically. For the Poisson distribution it is always the case that the variance is equal to the expectation, namely to λ:

$$\mathrm{E}(X) = \mathrm{Var}(X) = \lambda \ .$$

For a Poisson distribution, the expectation equals the variance.

In Figure 5.4 you may find the probabilities of the Poisson distribution for $\lambda = 0.5$, $\lambda = 1$ and $\lambda = 2$. Notice once more that the sample space is the same for all the Poisson distributions. What varies when we change the value of λ are the probabilities. Observe that as λ increases then probability of larger values increases as well.

Example 5.3. *A radio active element decays by the release of subatomic particles and energy. The decay activity is measured in terms of the number of decays per second in a unit mass. A typical model for the distribution of the number of decays is the Poisson distribution. Observe that the number of decays in a second is a integer and, in principle, it may obtain any integer value larger or equal to zero. The event of a radio active decay of an atom is a relatively rare event. Therefore, the Poisson model is likely to fit this phenomena[3].*

Example 5.4. *Consider an overhead power line suspended between two utility poles. During rain, drops of water may hit the power line. The total number of drops that hit the line in a one minute period may be modeled by a Poisson random variable.*

Note: the binomial and Poisson distributions are "discrete" and only accept integers. Now we are going to cover continuous distributions that can take any decimal value within the specified domain.
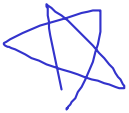
5.3 Continuous Random Variable

Many types of measurements, such as height, weight, angle, temperature, etc., may in principle have a continuum of possible values. Continuous random variables are used to model uncertainty regarding future values of such measurements.

The main difference between discrete random variables, which is the type we examined thus far, and continuous random variable, that are added now to the list, is in the sample space, i.e., the collection of possible outcomes. The former

---

[3]The number of decays may also be considered in the Binomial(n, p) setting. The number n is the total number of atoms in the unit mass and p is the probability that an atom decays within the given second. However, since n is very large and p is very small we get that the Poisson distribution is an appropriate model for the count.

type is used when the possible outcomes are separated from each other as the integers are. The latter type is used when the possible outcomes are the entire line of real numbers or when they form an interval (possibly an open ended one) of real numbers.

The difference between the two types of sample spaces implies differences in the way the distribution of the random variables is being described. For discrete random variables one may list the probability associated with each value in the sample space using a table, a formula, or a bar plot. For continuous random variables, on the other hand, probabilities are assigned to intervals of values, and not to specific values. Thence, densities are used in order to display the distribution.

Densities are similar to histograms, with areas under the plot corresponding to probabilities. We will provide a more detailed description of densities as we discuss the different examples of continuous random variables.

In continuous random variables integration replaces summation and the density replaces the probability in the computation of quantities such as the probability of an event, the expectation, and the variance.

Hence, if the expectation of a discrete random variable is given in the formula $E(X) = \sum_x \big(x \times P(x)\big)$, which involves the summation over all values of the product between the value and the probability of the value, then for continuous random variable the definition becomes:

$$E(X) = \int \big(x \times f(x)\big)dx \, ,$$

you do not need to know this for MATH1280.

where $f(x)$ is the density of X at the value x. Therefore, in the expectation of a continuous random variable one multiplies the value by the density at the value. This product is then integrated over the sample space.

Likewise, the formula $\mathrm{Var}(X) = \sum_x \big((x - E(X))^2 \times P(x)\big)$ for the variance is replaced by:

$$\mathrm{Var}(X) = \int \big((x - E(X))^2 \times f(x)\big)dx \, .$$

You don't need this for MATH1280

Nonetheless, the intuitive interpretation of the expectation as the central value of the distribution that identifies the location and the interpretation of the standard deviation (the square root of the variance) as the summary of the total spread of the distribution is still valid.

In this section we will describe two types of continuous random variables: Uniform and Exponential. In the next chapter another example – the Normal distribution – will be introduced.

5.3.1   The Uniform Random Variable

The Uniform distribution is used in order to model measurements that may have values in a given interval, with all values in this interval equally likely to occur.

For example, consider a random variable X with the Uniform distribution over the interval $[3, 7]$, denoted by "$X \sim \mathrm{Uniform}(3, 7)$". The density function at given values may be computed with the aid of the function "`dunif`". For instance let us compute the density of the $\mathrm{Uniform}(3, 7)$ distribution over the integers $\{0, 1, \ldots, 10\}$:
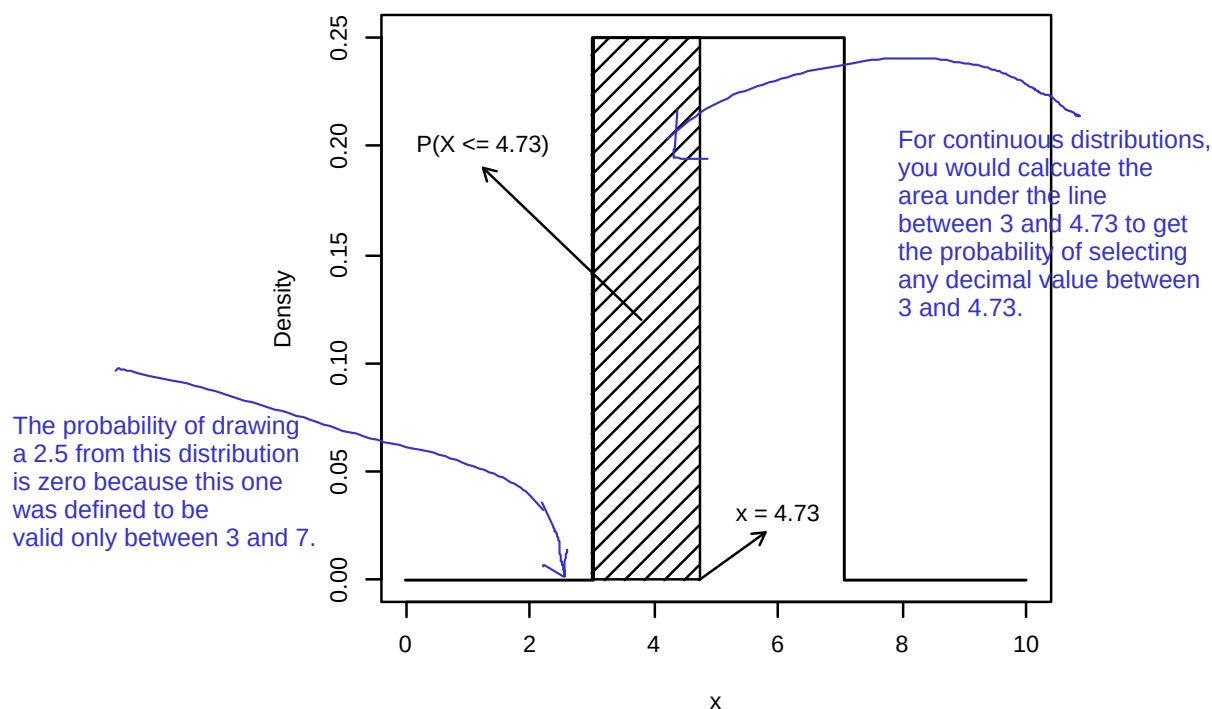
Figure 5.5: The Uniform(3,7) Distribution

We won't use dunif() much, but if you want to use it, you have to
use two points and use the output as the height of the
rectangle (see Figure 5.5 above).

```
> dunif(0:10,3,7)
 [1] 0.00 0.00 0.00 0.25 0.25 0.25 0.25 0.25 0.00 0.00 0.00
```

Notice that for the values 0, 1, and 2, and the values 8, 9 and 10 that are outside
of the interval the density is equal to zero, indicating that such values cannot
occur in the given distribution. The values of the density at integers inside the
interval are positive and equal to each other. The density is not restricted to
integer values. For example, at the point 4.73 we get that the density is positive
and of the same height:

```
> dunif(4.73,3,7)
[1] 0.25
```

A plot of the Uniform$(3, 7)$ density is given in Figure 5.5 in the form of a
solid line. Observe that the density is positive over the interval $[3, 7]$ where its
height is $1/4$. Area under the curve in the density corresponds to probability.
Indeed, the fact that the total probability is one is reflected in the total area
under the curve being equal to 1. Over the interval $[3, 7]$ the density forms a
rectangle. The base of the rectangle is the length of the interval $7 - 3 = 4$. The

height of the rectangle is thus equal to 1/4 in order to produce a total area of $4 \times (1/4) = 1$.

The function "`punif`" computes the cumulative probability of the uniform distribution. The probability $P(X \leq 4.73)$, for $X \sim \text{Uniform}(3, 7)$, is given by:

```
> punif(4.73,3,7)
[1] 0.4325
```

This will tell you the probability of randomly selecting a value that is less than or equal to 4.73 from a uniform distribution that goes from 3. to 7.

This probability corresponds to the marked area to the left of the point $x = 4.73$ in Figure 5.5. This area of the marked rectangle is equal to the length of the base 4.73 - 3 = 1.73, times the height of the rectangle 1/(7-3) = 1/4. Indeed:

```
> (4.73-3)/(7-3)
[1] 0.4325
```

is the area of the marked rectangle and is equal to the probability.

Let us use R in order to plot the density and the cumulative probability functions of the Uniform distribution. We produce first a large number of points in the region we want to plot. The points are produced with aid of the function "`seq`". The output of this function is a sequence with equally spaced values. The starting value of the sequence is the first argument in the input of the function and the last value is the second argument in the input. The argument "`length=1000`" sets the length of the sequence, 1,000 values in this case:

```
> x <- seq(0,10,length=1000)
> den <- dunif(x,3,7)
> plot(x,den)
```

The seq() command says to generate a list of decimals between 0 and 10 so that there are 1,000 items in the list. I prefer to use an odd number, like 1,001.

The object "`den`" is a sequence of length 1,000 that contains the density of the Uniform$(3, 7)$ evaluated over the values of "`x`". When we apply the function "`plot`" to the two sequences we get a scatter plot of the 1,000 points that is presented in the upper panel of Figure 5.6.

A scatter plot is a plot of points. Each point in the scatter plot is identify by its horizontal location on the plot (its "x" value) and by its vertical location on the plot (its y value). The horizontal value of each point in the plot is determined by the first argument to the function "`plot`" and the vertical value is determined by the second argument. For example, the first value in the sequence "`x`" is 0. The value of the Uniform density at this point is 0. Hence, the first value of the sequence "`den`" is also 0. A point that corresponds to these values is produced in the plot. The horizontal value of the point is 0 and the vertical value is 0. In a similar way the other 999 points are plotted. The last point to be plotted has a horizontal value of 10 and a vertical value of 0.

The number of points that are plotted is large and they overlap each other in the graph and thus produce an impression of a continuum. In order to obtain nicer looking plots we may choose to connect the points to each other with segments and use smaller points. This may be achieved by the addition of the argument "`type="l"`", with the letter `"l"` for line, to the plotting function:

```
> plot(x,den,type="l")
```

The output of the function is presented in the second panel of Figure 5.6. In the last panel the cumulative probability of the Uniform$(3, 7)$ is presented. This function is produced by the code:
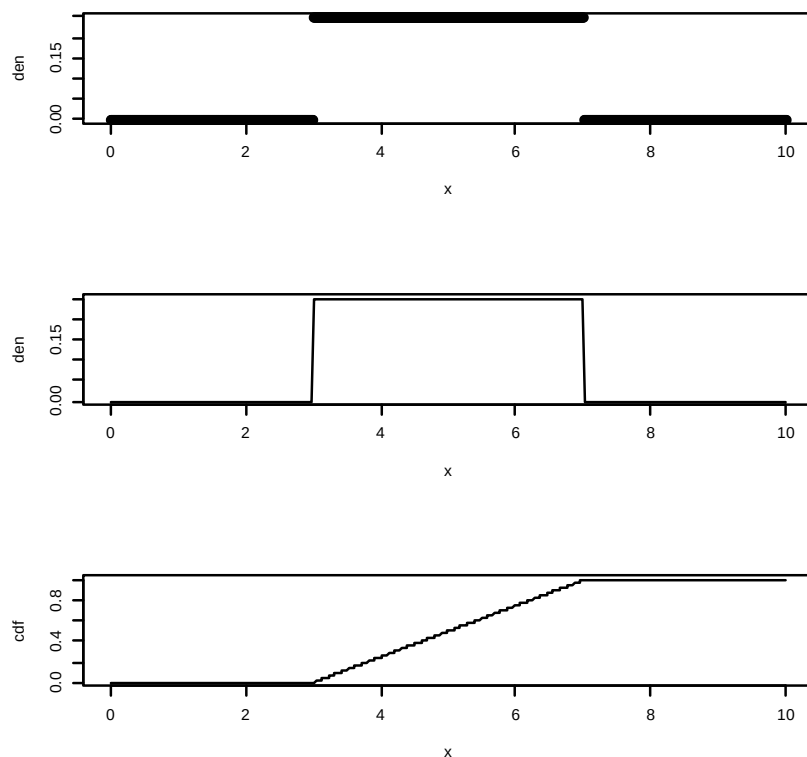
Figure 5.6: The Density and Cumulative Probability of Uniform(3,7)

```
> cdf <- punif(x,3,7)
> plot(x,cdf,type="l")
```

How to find the expectation of a uniform distr.

One can think of the density of the Uniform as an histogram[4]. The expectation of a Uniform random variable is the middle point of it's histogram. Hence, if $X \sim \text{Uniform}(a, b)$ then:

$$\text{E}(X) = \frac{a+b}{2} .$$

To get the expectation of a uniform distribution, a = the minimum, b = the max, follow the instructions.

For the $X \sim \text{Uniform}(3, 7)$ distribution the expectation is $\text{E}(X) = (3+7)/2 = 5$. Observe that 5 is the center of the Uniform density in Plot 5.5.

It can be shown that the variance of the Uniform(a, b) is equal to

$$\text{Var}(X) = \frac{(b-a)^2}{12} ,$$

with the standard deviation being the square root of this value. Specifically, for $X \sim \text{Uniform}(3, 7)$ we get that $\text{Var}(X) = (7-3)^2/12 = 1.333333$. The standard deviation is equal to $\sqrt{1.333333} = 1.154701$.

How to find the variance of a uniform dist.

[4]If $X \sim \text{Uniform}(a, b)$ then the density is $f(x) = 1/(b - a)$, for $a \leq x \leq b$, and it is equal to 0 for other values of x.

Example 5.5. *In Example 5.4 we considered rain drops that hit an overhead power line suspended between two utility poles. The* **number** *of drops that hit the line can be modeled using the Poisson distribution. The* **position** *between the two poles where a rain drop hits the line can be modeled by the Uniform distribution. The rain drop can hit any position between the two utility poles. Hitting one position along the line is as likely as hitting any other position.*

Example 5.6. *Meiosis is the process in which a diploid cell that contains two copies of the genetic material produces an haploid cell with only one copy (sperms or eggs, depending on the sex). The resulting molecule of genetic material is linear molecule (chromosome) that is composed of consecutive segments: a segment that originated from one of the two copies followed by a segment from the other copy and vice versa. The border points between segments are called points of crossover. The Haldane model for crossovers states that the position of a crossover between two given loci on the chromosome corresponds to the Uniform distribution and the total number of crossovers between these two loci corresponds to the Poisson distribution.*

5.3.2 The Exponential Random Variable

The Exponential distribution is frequently used to model times between events. For example, times between incoming phone calls, the time until a component becomes malfunction, etc. We denote the Exponential distribution via "$X \sim$ Exponential(λ)", where λ is a parameter that characterizes the distribution and is called the *rate* of the distribution. The overlap between the parameter used to characterize the Exponential distribution and the one used for the Poisson distribution is deliberate. The two distributions are tightly interconnected. As a matter of fact, it can be shown that if the distribution between occurrences of a phenomena has the Exponential distribution with rate λ then the total number of the occurrences of the phenomena within a unit interval of time has a Poisson(λ) distribution.

lambda is the "rate"

The sample space of an Exponential random variable contains all non-negative numbers. Consider, for example, $X \sim$ Exponential(0.5). The density of the distribution in the range between 0 and 10 is presented in Figure 5.7. Observe that in the Exponential distribution smaller values are more likely to occur in comparison to larger values. This is indicated by the density being larger at the vicinity of 0. The density of the exponential distribution given in the plot is positive, but hardly so, for values larger than 10.

The density of the Exponential distribution can be computed with the aid of the function "`dexp`"[5]. The cumulative probability can be computed with the function "`pexp`". For illustration, assume $X \sim$ Exponential(0.5). Say one is interested in the computation of the probability $P(2 < X \leq 6)$ that the random variable obtains a value that belongs to the interval $(2,6]$. The required probability is indicated as the marked area in Figure 5.7. This area can be computed as the difference between the probability $P(X \leq 6)$, the area to the left of 6, and the probability $P(X \leq 2)$, the area to the left of 2:

Do you see all those stars?

```
> pexp(6,0.5) - pexp(2,0.5)
[1] 0.3180924
```

The first part of this says to find the probability of getting less than or equal to 6, then the second part removes something that we do not want to see: the probability of getting less than or equal to 2. The result is the probability of getting between 2 and 6 in an exponential distribution that has a rate of 0.5.

---

[5]If $X \sim$ Exponential(λ) then the density is $f(x) = \lambda e^{-\lambda x}$, for $0 \leq x$, and it is equal to 0 for $x < 0$.

you do not need to know this equation for MATH1280

when you are trying to do this at home, and your answer is a negative number, that mean you made a mistake because there is no such thing as negative probability.
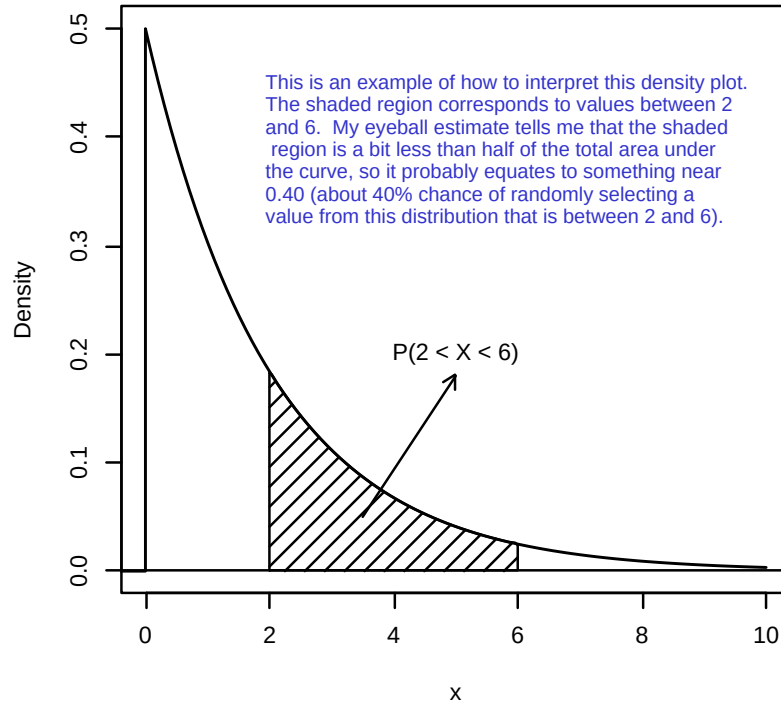
Figure 5.7: The Exponential(0.5) Distribution

How to find the expectation of an exponential distribution.

The difference is the probability of belonging to the interval, namely the area marked in the plot.

The expectation of X, when $X \sim$ Exponential(λ), is given by the equation:

$$E(X) = 1/\lambda , \qquad also \quad \lambda = \frac{1}{E(X)}$$

and the variance is given by:

$$Var(X) = 1/\lambda^2 . \qquad \lambda = \sqrt{\frac{1}{Var(X)}} = \frac{1}{\sigma_{\!x-}} \cdots standard\ deviation$$

The standard deviation is the square root of the variance, namely $1/\lambda$. Observe that the larger is the rate the smaller are the expectation and the standard deviation.

How to find the variance of an exponential distribution.

In Figure 5.8 the densities of the Exponential distribution are plotted for $\lambda = 0.5$, $\lambda = 1$, and $\lambda = 2$. Notice that with the increase in the value of the parameter then the values of the random variable tends to become smaller. This inverse relation makes sense in connection to the Poisson distribution. Recall that the Poisson distribution corresponds to the total number of occurrences in a unit interval of time when the time between occurrences has an Exponential
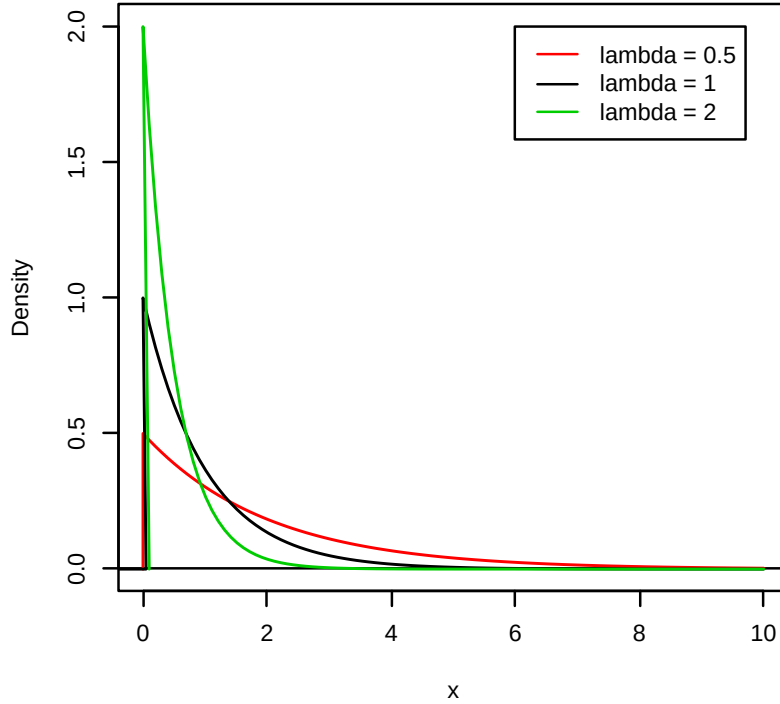
Figure 5.8: The Exponential Distribution for Various Values of λ

distribution. A larger expectation λ of the Poisson corresponds to a larger number of occurrences that are likely to take place during the unit interval of time. The larger is the number of occurrences the smaller are the time intervals between occurrences.

Example 5.7. *Consider Examples 5.4 and 5.5 that deal with rain dropping on a power line. The times between consecutive hits of the line may be modeled by the Exponential distribution. Hence, the time to the first hit has an Exponential distribution. The time between the first and the second hit is also Exponentially distributed, and so on.*

Example 5.8. *Return to Example 5.3 that deals with the radio activity of some element. The total count of decays per second is model by the Poisson distribution. The times between radio active decays is modeled according to the Exponential distribution. The rate λ of that Exponential distribution is equal to the expectation of the total count of decays in one second, i.e. the expectation of the Poisson distribution.*

5.4   Solved Exercises

Question 5.1. A particular measles vaccine produces a reaction (a fever higher that 102 Fahrenheit) in each vaccinee with probability of 0.09. A clinic vaccinates 500 people each day.

1. What is the expected number of people that will develop a reaction each day?

2. What is the standard deviation of the number of people that will develop a reaction each day?

3. In a given day, what is the probability that more than 40 people will develop a reaction?

4. In a given day, what is the probability that the number of people that will develop a reaction is between 50 and 45 (inclusive)?

Solution (to Question 5.1.1): The Binomial distribution is a reasonable model for the number of people that develop high fever as result of the vaccination. Let X be the number of people that do so in a give day. Hence, $X \sim \text{Binomial}(500, 0.09)$. According to the formula for the expectation in the Binomial distribution, since $n = 500$ and $p = 0.09$, we get that:

$$\text{E}(X) = np = 500 \times 0.09 = 45 \ .$$

The E(X) = np equation is on page 69 … for binomial distributions.

Solution (to Question 5.1.2): Let $X \sim \text{Binomial}(500, 0.09)$. Using the formula for the variance for the Binomial distribution we get that:

The Var(X) eq… page 69. $$\text{Var}(X) = np(1-p) = 500 \times 0.09 \times 0.91 = 40.95 \ .$$

Hence, since $\sqrt{Var(X)} = \sqrt{40.95} = 6.3992$, the standard deviation is 6.3992.

Solution (to Question 5.1.3): Let $X \sim \text{Binomial}(500, 0.09)$. The probability that more than 40 people will develop a reaction may be computed as the difference between 1 and the probability that 40 people or less will develop a reaction:

$$\text{P}(X > 40) = 1 - \text{P}(X \le 40) \ .$$

The probability can be computes with the aid of the function "`pbinom`" that produces the cumulative probability of the Binomial distribution:

```
> 1 - pbinom(40,500,0.09)
[1] 0.7556474
```
We use "1 - pbinom(…)" for "more than"

Solution (to Question 5.1.4): The probability that the number of people that will develop a reaction is between 50 and 45 (inclusive) is the difference between $\text{P}(X \le 50)$ and $\text{P}(X < 45) = \text{P}(X \le 44)$. Apply the function "`pbinom`" to get:

```
> pbinom(50,500,0.09) - pbinom(44,500,0.09)
[1] 0.3292321
```

Question 5.2. The Negative-Binomial distribution is yet another example of a
discrete, integer valued, random variable. The sample space of the distribution
are all non-negative integers $\{0, 1, 2, \ldots\}$. The fact that a random variable X
has this distribution is marked by "$X \sim$ Negative-Binomial(r, p)", where r and
p are parameters that specify the distribution.

this one is an application of what you have learned. Do this one last after you have mastered everything else.

Consider 3 random variables from the Negative-Binomial distribution:

- $X_1 \sim$ Negative-Binomial$(2, 0.5)$

- $X_2 \sim$ Negative-Binomial$(4, 0.5)$

- $X_3 \sim$ Negative-Binomial$(8, 0.8)$

The bar plots of these random variables are presented in Figure 5.9, re-organizer
in a random order.

1. Produce bar plots of the distributions of the random variables X_1, X_2, X_3
   in the range of integers between 0 and 15 and thereby identify the pair of
   parameters that produced each one of the plots in Figure 5.9. Notice that
   the bar plots can be produced with the aid of the function "`plot`" and the
   function "`dnbinom(x,r,p)`", where "`x`" is a sequence of integers and "`r`"
   and "`p`" are the parameters of the distribution. Pay attention to the fact
   that you should use the argument "`type = "h"`" in the function "`plot`"
   in order to produce the horizontal bars.

2. Below is a list of pairs that includes an expectation and a variance. Each
   of the pairs is associated with one of the random variables X_1, X_2, and
   X_3:

   (a) $E(X) = 4$, $Var(X) = 8$.
   (b) $E(X) = 2$, $Var(X) = 4$.
   (c) $E(X) = 2$, $Var(X) = 2.5$.

   Use Figure 5.9 in order to match the random variable with its associated
   pair. Do not use numerical computations or formulae for the expectation
   and the variance in the Negative-Binomial distribution in order to carry
   out the matching[6]. Use, instead, the structure of the bar-plots.

Solution (to Question 5.2.1): The plots can be produced with the following
code, which should be run one line at a time:

```
> x <- 0:15
> plot(x,dnbinom(x,2,0.5),type="h")
> plot(x,dnbinom(x,4,0.5),type="h")
> plot(x,dnbinom(x,8,0.8),type="h")
```

The first plot, that corresponds to $X_1 \sim$ Negative-Binomial$(2, 0.5)$, fits
Barplot 3. Notice that the distribution tends to obtain smaller values and that
the probability of the value "0" is equal to the probability of the value "1".

The second plot, the one that corresponds to $X_2 \sim$ Negative-Binomial$(4, 0.5)$,
is associated with Barplot 1. Notice that the distribution tends to obtain larger

---

[6]It can be shown, or else found on the web, that if $X \sim$ Negative-Binomial(r, p) then
$E(X) = r(1 - p)/p$ and $Var(X) = r(1 - p)/p^2$.

values. For example, the probability of the value "10" is substantially larger than zero, where for the other two plots this is not the case.

The third plot, the one that corresponds to $X_3 \sim$ Negative-Binomial$(8, 0.8)$, matches Barplot 2. Observe that this distribution tends to produce smaller probabilities for the small values as well as for the larger values. Overall, it is more concentrated than the other two.

Solution (to Question 5.2.2): Barplot 1 corresponds to a distribution that tends to obtain larger values than the other two distributions. Consequently, the expectation of this distribution should be larger. The conclusion is that the pair $E(X) = 4$, $Var(X) = 8$ should be associated with this distribution.

Barplot 2 describes a distribution that produce smaller probabilities for the small values as well as for the larger values and is more concentrated than the other two. The expectations of the two remaining distributions are equal to each other and the variance of the pair $E(X) = 2$, $Var(X) = 2.5$ is smaller. Consequently, this is the pair that should be matched with this box plot.

This leaves only Barplot 3, that should be matched with the pair $E(X) = 2$, $Var(X) = 4$.

5.5   Summary

Glossary

Binomial Random Variable: The number of successes among n repeats of independent trials with a probability p of success in each trial. The distribution is marked as Binomial(n, p).

Poisson Random Variable: An approximation to the number of occurrences of a rare event, when the expected number of events is λ. The distribution is marked as Poisson(λ).

Density: Histogram that describes the distribution of a continuous random variable. The area under the curve corresponds to probability.

Uniform Random Variable: A model for a measurement with equally likely outcomes over an interval $[a, b]$. The distribution is marked as Uniform(a, b).

Exponential Random Variable: A model for times between events. The distribution is marked as Exponential(λ).

Discuss in the Forum

This unit deals with two types of discrete random variables, the Binomial and the Poisson, and two types of continuous random variables, the Uniform and the Exponential. Depending on the context, these types of random variables may serve as theoretical models of the uncertainty associated with the outcome of a measurement.

In your opinion, is it or is it not useful to have a theoretical model for a situation that occurs in real life?

When forming your answer to this question you may give an example of a situation from you own field of interest for which a random variable, possibly from one of the types that are presented in this unit, can serve as a model. Discuss the importance (or lack thereof) of having a theoretical model for the situation.

For example, the Exponential distribution may serve as a model for the time until an atom of a radio active element decays by the release of subatomic particles and energy. The decay activity is measured in terms of the number of decays per second. This number is modeled as having a Poisson distribution. Its expectation is the rate of the Exponential distribution. For the radioactive element Carbon-14 (^{14}C) the decay rate is 3.8394×10^{-12} particles per second. Computations that are based on the Exponential model may be used in order to date ancient specimens.

Summary of Formulas

Discrete Random Variable:

$$E(X) = \sum_x \left(x \times P(x) \right)$$

$$Var(X) = \sum_x \left((x - E(X))^2 \times P(x) \right)$$

Continuous Random Variable:

$$E(X) = \int (x \times f(x)) dx$$

$$Var(X) = \int ((x - E(X))^2 \times f(x)) dx$$

Binomial: $E(X) = np$, $Var(X) = np(1 - p)$

Poisson: $E(X) = \lambda$, $Var(X) = \lambda$

Uniform: $E(X) = (a + b)/2$, $Var(X) = (b - a)^2/12$

Exponential: $E(X) = 1/\lambda$, $Var(X) = 1/\lambda^2$
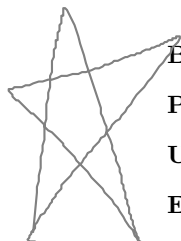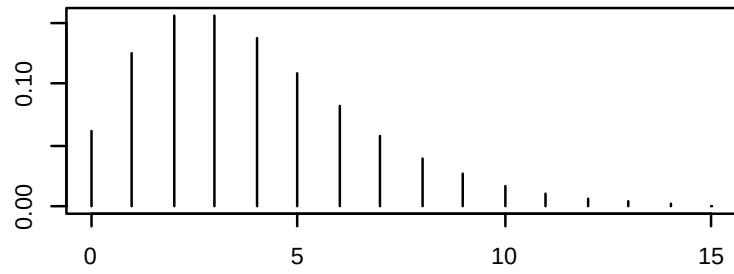
Barplot 1
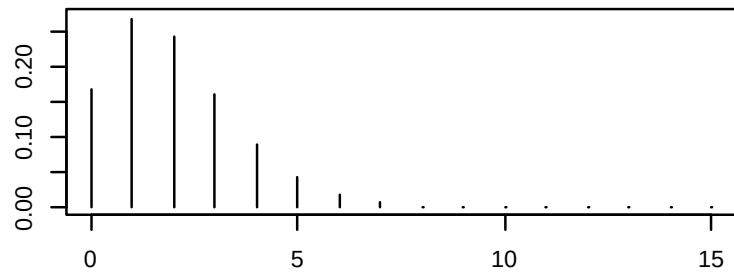
Barplot 2
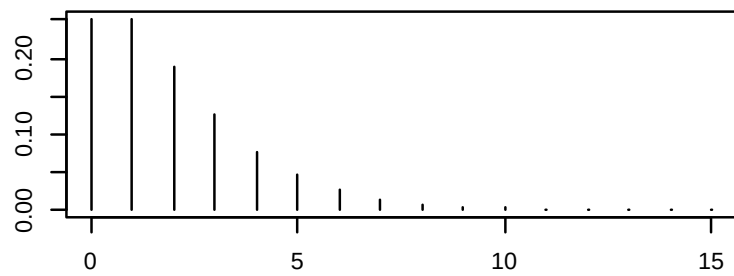
Barplot 3

Figure 5.9: Bar Plots of the Negative-Binomial Distribution

Chapter 6

The Normal Random Variable

6.1 Student Learning Objective

This chapter introduces a very important bell-shaped distribution known as the Normal distribution. Computations associated with this distribution are discussed, including the percentiles of the distribution and the identification of intervals of subscribed probability. The Normal distribution may serve as an approximation to other distributions. We demonstrate this property by showing that under appropriate conditions the Binomial distribution can be approximated by the Normal distribution. This property of the Normal distribution will be picked up in the next chapter where the mathematical theory that establishes the Normal approximation is demonstrated. By the end of this chapter, the student should be able to:

- Recognize the Normal density and apply `R` functions for computing Normal probabilities and percentiles.

- Associate the distribution of a Normal random variable with that of its standardized counterpart, which is obtained by centering and re-scaling.

- Use the Normal distribution to approximate the Binomial distribution.

6.2 The Normal Random Variable

The Normal distribution is the most important of all distributions that are used in statistics. In many cases it serves as a generic model for the distribution of a measurement. Moreover, even in cases where the measurement is modeled by other distributions (i.e. Binomial, Poisson, Uniform, Exponential, etc.) the Normal distribution emerges as an approximation of the distribution of numerical characteristics of the data produced by such measurements.

P(0 < X < 5)

The shaded area under the curve represents the probability of randomly selecting a value between 0 and 5 from this distribution.
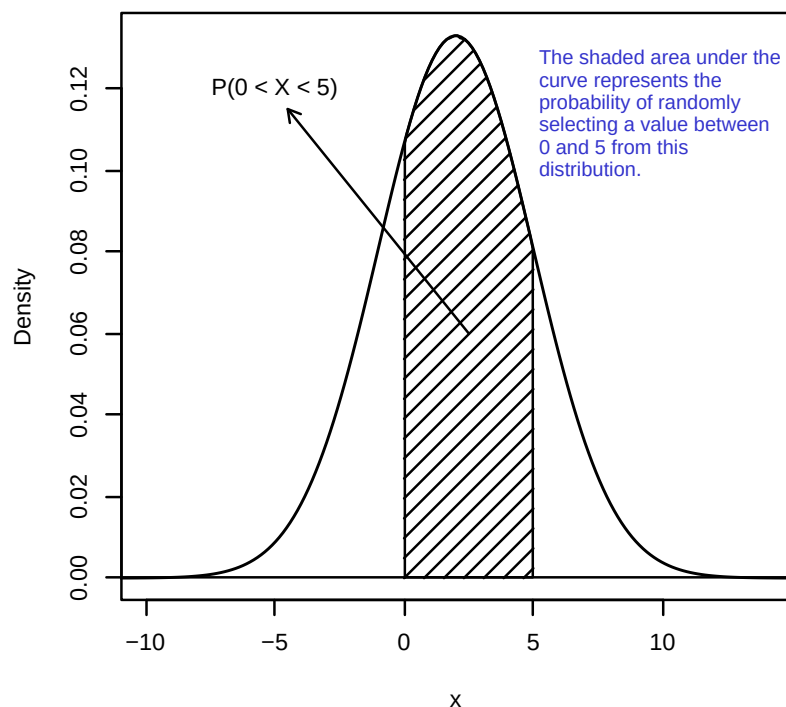
Figure 6.1: The Normal(2,9) Distribution

6.2.1   The Normal Distribution

A Normal random variable has a continuous distribution over the sample space of all numbers, negative or positive. We denote the Normal distribution via "$X \sim \text{Normal}(\mu, \sigma^2)$", where $\mu = \text{E}(X)$ is the expectation of the random variable and $\sigma^2 = \text{Va}r(X)$ is it's variance[1].

Consider, for example, $X \sim \text{Normal}(2, 9)$. The density of the distribution is presented in Figure 6.1. Observe that the distribution is symmetric about the expectation 2. The random variable is more likely to obtain its value in the vicinity of the expectation. Values much larger or much smaller than the expectation are substantially less likely.

The density of the Normal distribution can be computed with the aid of the function "dnorm". The cumulative probability can be computed with the function "pnorm". For illustrating the use of the latter function, assume that $X \sim \text{Normal}(2, 9)$. Say one is interested in the computation of the probability $\text{P}(0 < X \leq 5)$ that the random variable obtains a value that belongs to the

---

[1]If $X \sim \text{Normal}(\mu, \sigma^2)$ then the density of X is given by the formula $f(x) = \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}/\sqrt{2\pi\sigma^2}$, for all x.

$$\rightarrow N\left(\mu, \sigma^2\right)$$

that means a normal distribution with population mean μ and population variance σ^2.

This says that there is a normal distribution that has a mean of 2 and a variance of 9.
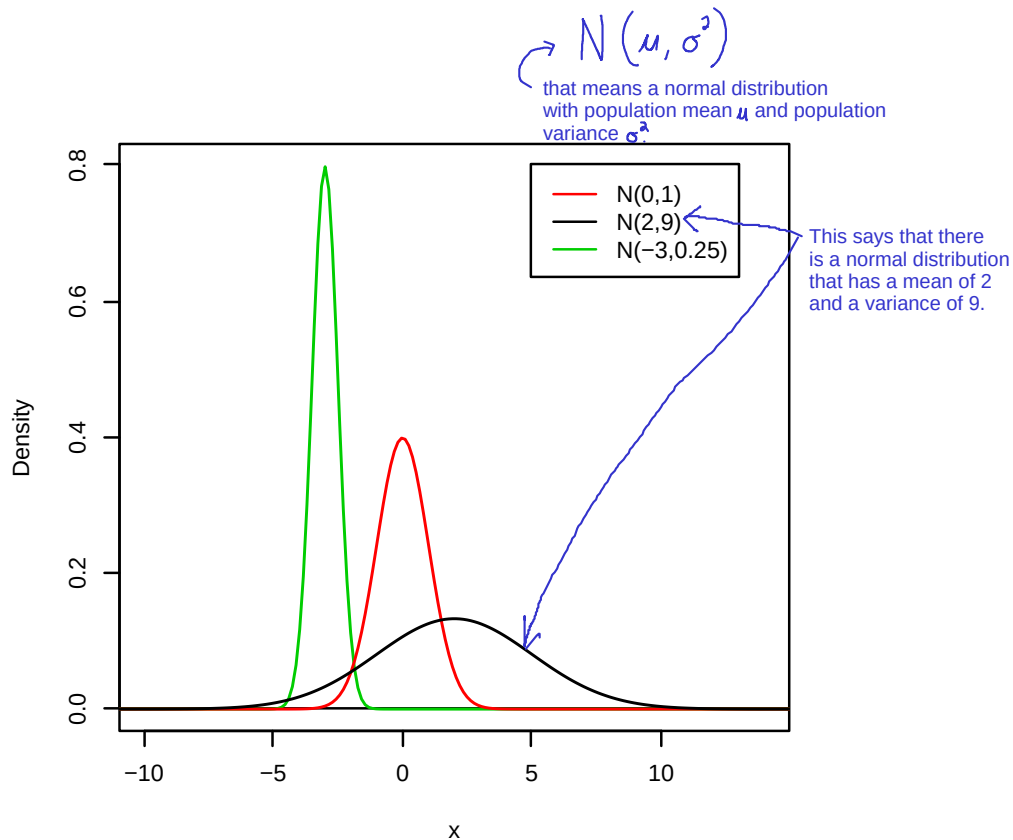
Figure 6.2: The Normal Distribution for Various Values of μ and σ^2

interval $(0, 5]$. The required probability is indicated by the marked area in Figure 6.1. This area can be computed as the difference between the probability $P(X \leq 5)$, the area to the left of 5, and the probability $P(X \leq 0)$, the area to the left of 0:

```
> pnorm(5,2,3) - pnorm(0,2,3)
[1] 0.5888522
```

This tells you the probability of randomly selecting between 0 and 5 from a normal distribution that has a mean of 2 and a std deviation of 3.

The difference is the indicated area that corresponds to the probability of being inside the interval, which turns out to be approximately equal to 0.589. Notice that the expectation μ of the Normal distribution is entered as the second argument to the function. The third argument to the function is the standard deviation, i.e. the square root of the variance. In this example, the standard deviation is $\sqrt{9} = 3$.

Figure 6.2 displays the densities of the Normal distribution for the combinations $\mu = 0$, $\sigma^2 = 1$ (the *red* line); $\mu = 2$, $\sigma^2 = 9$ (the *black* line); and $\mu = -3$, $\sigma^2 = 1/4$ (the *green* line). Observe that the smaller the variance the more concentrated is the distribution of the random variable about the expectation.

Example 6.1. *IQ tests are a popular (and controversial) mean for measuring intelligence. They are produced as (weighted) average of a response to a long list of questions, designed to test different abilities. The score of the test across the entire population is set to be equal to 100 and the standard deviation is set to 15. The distribution of the score is Normal. Hence, if X is the IQ score of a random subject then $X \sim \text{Normal}(100, 15^2)$.*

Example 6.2. *Any measurement that is produced as a result of the combination of many independent influencing factors is likely to poses the Normal distribution. For example, the hight of a person is influenced both by genetics and by the environment in which that person grew up. Both the genetic and the environmental influences are a combination of many factors. Thereby, it should not come as a surprise that the heights of people in a population tend to follow the Normal distribution.*

6.2.2   The Standard Normal Distribution

The standard normal distribution is a normal distribution of standardized values, which are called z-scores. A z-score is the original measurement measured in units of the standard deviation from the expectation. For example, if the expectation of a Normal distribution is 2 and the standard deviation is $3 = \sqrt{9}$, then the value of 0 is $2/3$ standard deviations smaller than (or to the left of) the expectation. Hence, the z-score of the value 0 is -2/3. The calculation of the z-score emerges from the equation:

$$(0 =) \; x = \mu + z \cdot \sigma \; (= 2 + z \cdot 3)$$

The z-score is obtained by solving the equation

$$0 = 2 + z \cdot 3 \quad \Longrightarrow \quad z = (0 - 2)/3 = -2/3 \;.$$

In a similar way, the z-score of the value $x = 5$ is equal to 1, following the solution of the equation $5 = 2 + z \cdot 3$, which leads to $z = (5 - 2)/3 = 1$.

The standard Normal distribution is the distribution of a standardized Normal measurement. The expectation for the standard Normal distribution is 0 and the variance is 1. When $X \sim N(\mu, \sigma^2)$ has a Normal distribution with expectation μ and variance σ^2 then the transformed random variable $Z = (X - \mu)/\sigma$ produces the standard Normal distribution $Z \sim N(0, 1)$. The transformation corresponds to the reexpression of the original measurement in terms of a new "zero" and a new unit of measurement. The new "zero" is the expectation of the original measurement and the new unit is the standard deviation of the original measurement.

Computation of probabilities associated with a Normal random variable X can be carried out with the aid of the standard Normal distribution. For example, consider the computation of the probability $P(0 < X \leq 5)$ for $X \sim N(2, 9)$, that has expectation $\mu = 2$ and standard deviation $\sigma = 3$. Consider X's standardized values: $Z = (X - 2)/3$. The boundaries of the interval $[0, 5]$, namely 0 and 5, have standardized z-scores of $(0 - 2)/3 = -2/3$ and $(5 - 2)/3 = 1$, respectively. Clearly, the original measurement X falls between the original boundaries ($0 < X \leq 5$) if, and only if, the standardized measurement Z falls

---

Margin notes:

z- score:
We can describe a number in a sample by saying that it is a certain number of standard deviations from the mean... it is "z" standard deviations from the mean.

Example. If the mean is 7 and the std. deviation is 2 and my sample point is 11, then I can say that my sample point is at z=2 (it is 2 std dev above mu). If z=2, then I know that the probability of being at least 2 standard deviations from the mean is: pnorm(2) = .977. I know that few observations will be above z=2.

"transformed random variable" means that I can take all the numbers in my sample, subract the mean, divide by the std deviation and havea list of z-scores.
R:
mu = 7
sigma = 2
x <- c(11, 1,8,7,5)
z = (x - mu) / sigma
z

If you do not know the population mean or sd, you can estimate it:

x <- c(2,7,3,2,6)
z = (x - mean(x)) /sd(x)
z
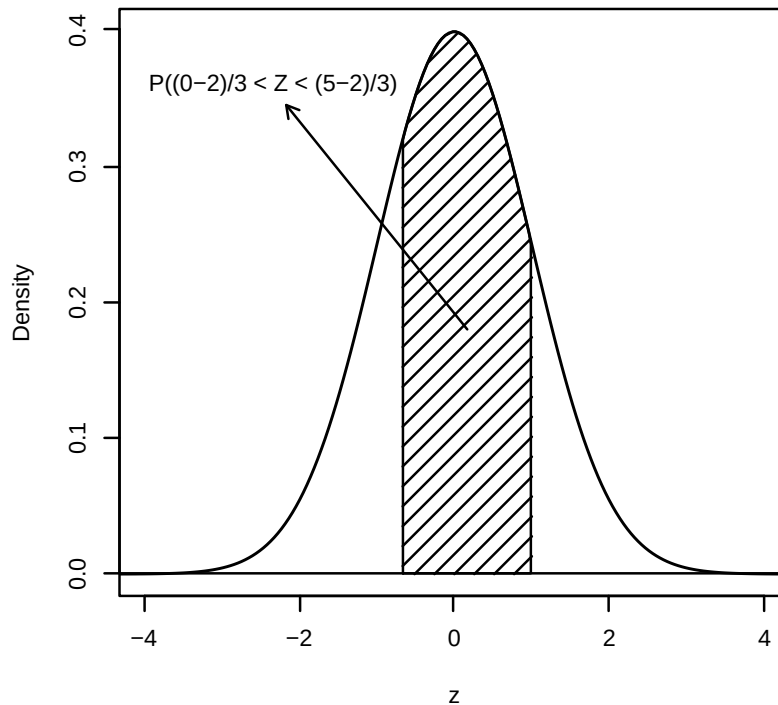
Figure 6.3: The Standard Normal Distribution

between the standardized boundaries $(-2/3 < Z \le 1)$. Therefore, the probability that X obtains a value in the range $[0, 5]$ is equal to the probability that Z obtains a value in the range $[-2/3, 1]$.

The function "pnorm" was used in the previous subsection in order to compute that probability that X obtains values between 0 and 5. The computation produced the probability 0.5888522. We can repeat the computation by the application of the same function to the standardized values:

```
> pnorm((5-2)/3) - pnorm((0-2)/3)
[1] 0.5888522
```
Test yourself: what does this R command mean?

The value that is being computed, the area under the graph for the standard Normal distribution, is presented in Figure 6.3. Recall that 3 arguments where specified in the previous application of the function "pnorm": the x value, the expectation, and the standard deviation. In the given application we did not specify the last two arguments, only the first one. (Notice that the output of the expression "(5-2)/3" is a single number and, likewise, the output of the expression "(0-2)/3" is also a single number.)

Most R function have many arguments that enables flexible application in a

wide range of settings. For convenience, however, default values are set to most of these arguments. These default values are used unless an alternative value for the argument is set when the function is called. The default value of the second argument of the function "`pnorm`" that specifies the expectation is "`mean=0`", and the default value of the third argument that specifies the standard deviation is "`sd=1`". Therefore, if no other value is set for these arguments the function computes the cumulative distribution function of the standard Normal distribution.

6.2.3   Computing Percentiles

You will need to know how to compute percentiles for quizzes. You might be asked to find the "central 95%" or "middle 95%" (same thing), or maybe find the top or bottom X%.

Consider the issue of determining the range that contains 95% of the probability for a Normal random variable. We start with the standard Normal distribution. Consult Figure 6.4. The figure displays the standard Normal distribution with the central region shaded. The area of the shaded region is 0.95.

We may find the z-values of the boundaries of the region, denoted in the figure as z_0 and z_1 by the investigation of the cumulative distribution function. Indeed, in order to have 95% of the distribution in the central region one should leave out 2.5% of the distribution in each of the two tails. That is, 0.025 should be the area of the unshaded region to the right of z_1 and, likewise, 0.025 should be the area of the unshaded region to the left of z_0. In other words, the cumulative probability up to z_0 should be 0.025 and the cumulative distribution up to z_1 should be 0.975.

You are usually asked to find an x-value that acts as the "cutoff point" or "criterion value," which means you want to find a value from the distribution that is bigger than or smaller than X% of the values in the distribution.

In general, given a random variable X and given a percent p, the x value with the property that the cumulative distribution up to x is equal to the probability p is called the *p-percentile* of the distribution. Here we seek the 2.5%-percentile and the 97.5%-percentile of the standard Normal distribution.

The percentiles of the Normal distribution are computed by the function "`qnorm`". The first argument to the function is a probability (or a sequence of probabilities), the second and third arguments are the expectation and the standard deviations of the normal distribution. The default values to these arguments are set to 0 and 1, respectively. Hence if these arguments are not provided the function computes the percentiles of the standard Normal distribution. Let us apply the function in order to compute z_1 and z_0:

1.95 here is a z-score. It comes from qnorm() with default settings, which means the mean = 0 and sd =1, which is the standard normal distribution.

z-score example

```
> qnorm(0.975)
[1] 1.959964
> qnorm(0.025)
[1] -1.959964
```

This says that if you are 1.959964 standard deviations from the mean, you are at the 97.5 th percentile, meaning that you have a big value for this distribution. See the plot on the next page.

Observe that z_1 is practically equal to 1.96 and $z_0 = -1.96 = -z_1$. The fact that z_0 is the negative of z_1 results from the symmetry of the standard Normal distribution about 0. As a conclusion we get that for the standard Normal distribution 95% of the probability is concentrated in the range $[-1.96, 1.96]$.

The problem of determining the central range that contains 95% of the distribution can be addresses in the context of the original measurement X (See Figure 6.5). We seek in this case an interval centered at the expectation 2, which is the center of the distribution of X, unlike 0 which was the center of the standardized values Z. One way of solving the problem is via the application of the function "`qnorm`" with the appropriate values for the expectation and the standard deviation:

z-score demonstration

*** Read section 6.2.2 first!! then look at this plot ***

The z_o refers to the lower z cutoff point.

$P(z_0 < Z < z_1) = 0.95$

this applies to the qnorm() example on the prior page.

The shaded region shows that most values are between plus or minus 2 standard deviations from the mean (this is true regardless of the mean or variance of the particular distribution.

$z_0 = -1.96$

$z_1 = 1.96$

"cutoff points" (criterion values) such as finding the top or bottom 5% of values... use qnorm().

Density

z

The values along the x-axis show z-scores.

Figure 6.4: Central 95% of the Standard Normal Distribution

You can also write this as qnorm(0.975, mean=2, sd=3)

The value 7.879892 is at the 97.5 percentile of a normal distribution that has a mean of 2 and a standard deviation of 3. The answer is a VALUE not a probability!!!

```
> qnorm(0.975,2,3)
[1] 7.879892
> qnorm(0.025,2,3)
[1] -3.879892
```

There are "q" functions for all the other distributions, so to find the value associated with the 97.5 percentail for an exponential distribution, you would use qexp(). Also see qpois() for Poisson and qbinom() for binomial.

Hence, we get that $x_0 = -3.88$ has the property that the total probability to its left is 0.025 and $x_1 = 7.88$ has the property that the total probability to its right is 0.025. The total probability in the range $[-3.88, 7.88]$ is 0.95.

An alternative approach for obtaining the given interval exploits the interval that was obtained for the standardized values. An interval $[-1.96, 1.96]$ of standardized z-values corresponds to an interval $[2 - 1.96 \cdot 3, 2 + 1.96 \cdot 3]$ of the original x-values:

"qnorm(.975)" with no extra arguments gives us "the z-score that is higher than 97.5% of observations in any normal distribution."

```
> 2 + qnorm(0.975)*3
[1] 7.879892
> 2 + qnorm(0.025)*3
[1] -3.879892
```

Note: this is a continuation of the example at the top of section 6.2.2 where the mean is 2 and the standard deviation is 3.

Hence, we again produce the interval $[-3.88, 7.88]$, the interval that was obtained before as the central interval that contains 95% of the distribution of the

P(x_0< X < x_1) = 0.95

This is asking you to find the x-values that enclose the middle 95%

x_0 = ?

x_1 = ?

This tells us that we have a normal distribution with a population mean of 2 and a population variance of 9.

The mean is about here
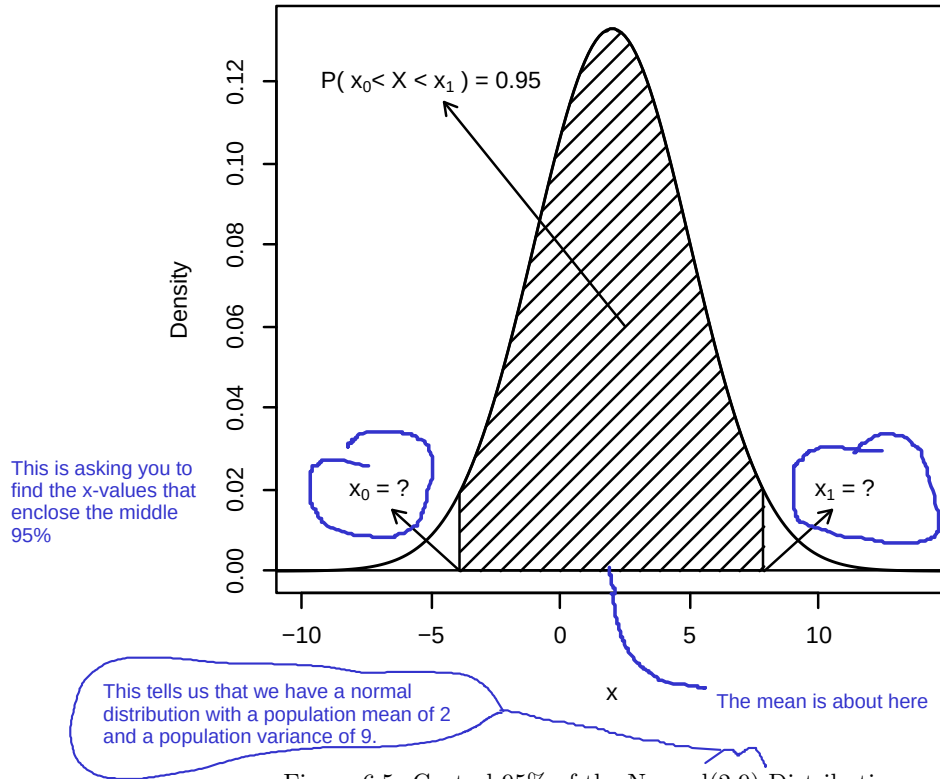
Figure 6.5: Central 95% of the Normal(2,9) Distribution

The paragraph below says that if you add 1.96 std. dev and subtract 1.96 std. dev from the mean, those two points will enclose the central 95% of the normal distribution. This works only if the numbers are normally distributed.
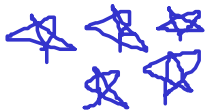
Normal$(2, 9)$ random variable.

In general, if $X \sim N(\mu, \sigma^2)$ is a Normal random variable then the interval $[\mu - 1.96 \cdot \sigma, \mu + 1.96 \cdot \sigma]$ contains 95% of the distribution of the random variable. Frequently one uses the notation $\mu \pm 1.96 \cdot \sigma$ to describe such an interval.

6.2.4   Outliers and the Normal Distribution

Consider, next, the computation of the interquartile range in the Normal distribution. Recall that the interquartile range is the length of the central interval that contains 50% of the distribution. This interval starts at the first quartile (Q1), the value that splits the distribution so that 25% of the distribution is to the left of the value and 75% is to the right of it. The interval ends at the third quartile (Q3) where 75% of the distribution is to the left and 25% is to the right.

For the standard Normal the third and first quartiles can be computed with the aid of the function "qnorm":

```
> qnorm(0.75)
```

```
[1] 0.6744898
> qnorm(0.25)
[1] -0.6744898
```

Because the mean and standard deviation are not entered into the qnorm() function, it means that we are using the standard normal distribution, which has mean 0 and standard deviation 1. The qnorm(0.25) command is asking R to tell you the number of standard deviations from the mean that is greater than 25% of values in the distribution.

Observe that for the standard Normal distribution one has that 75% of the distribution is to the left of the value 0.6744898, which is the third quartile of this distribution. Likewise, 25% of the standard Normal distribution are to the left of the value -0.6744898, which is the first quartile. the interquartile range is the length of the interval between the third and the first quartiles. In the case of the standard Normal distribution this length is equal to $0.6744898 - (-0.6744898) = 1.348980$.

In Chapter 3 we considered box plots as a mean for the graphical display of numerical data. The box plot includes a vertical rectangle that initiates at the first quartile and ends at the third quartile, with the median marked within the box. The rectangle contains 50% of the data. Whiskers extends from the ends of this rectangle to the smallest and to the largest data values that are not outliers. Outliers are values that lie outside of the normal range of the data. Outliers are identified as values that are more then 1.5 times the interquartile range away from the ends of the central rectangle. Hence, a value is an outlier if it is larger than the third quartile plus 1.5 times the interquartile range or if it is less than the first quartile minus 1.5 times the interquartile range.

How likely is it to obtain an outlier value when the measurement has the standard Normal distribution? We obtained that the third quartile of the standard Normal distribution is equal to 0.6744898 and the first quartile is minus this value. The interquartile range is the difference between the third and first quartiles. The upper and lower thresholds for the defining outliers are:

```
> qnorm(0.75) + 1.5*(qnorm(0.75)-qnorm(0.25))
[1] 2.697959
> qnorm(0.25) - 1.5*(qnorm(0.75)-qnorm(0.25))
[1] -2.697959
```

This is the long way to identify outliers. I don't know of anyone who would do this in the real world. The cutoff points for outliers are ultimately a matter of choice, but we use some techniques as a matter of convention.

Hence, a value larger than 2.697959 or smaller than -2.697959 would be identified as an outlier.

The probability of being less than the upper threshold 2.697959 in the standard Normal distribution is computed with the expression "`pnorm(2.697959)`". The probability of being above the threshold is 1 minus that probability, which is the outcome of the expression "`1-pnorm(2.697959)`".

By the symmetry of the standard Normal distribution we get that the probability of being below the lower threshold -2.697959 is equal to the probability of being above the upper threshold. Consequently, the probability of obtaining an outlier is equal to twice the probability of being above the upper threshold:

```
> 2*(1-pnorm(2.697959))
[1] 0.006976603
```

We get that for the standard Normal distribution the probability of an outlier is approximately 0.7%.

6.3   Approximation of the Binomial Distribution

The Normal distribution emerges frequently as an approximation of the distribution of data characteristics. The probability theory that mathematically establishes such approximation is called the Central Limit Theorem and is the subject of the next chapter. In this section we demonstrate the Normal approximation in the context of the Binomial distribution.

6.3.1   Approximate Binomial Probabilities and Percentiles

Consider, for example, the probability of obtaining between 1940 and 2060 heads when tossing 4,000 fair coins. Let X be the total number of heads. The tossing of a coin is a trial with two possible outcomes: "Head" and "Tail." The probability of a "Head" is 0.5 and there are 4,000 trials. Let us call obtaining a "Head" in a trial a "Success". Observe that the random variable X counts the total number of successes. Hence, $X \sim \text{Binomial}(4000, 0.5)$.

The probability $P(1940 \leq X \leq 2060)$ can be computed as the difference between the probability $P(X \leq 2060)$ of being less or equal to 2060 and the probability $P(X < 1940)$ of being strictly less than 1940. However, 1939 is the largest integer that is still strictly less than the integer 1940. As a result we get that $P(X < 1940) = P(X \leq 1939)$. Consequently, $P(1940 \leq X \leq 2060) = P(X \leq 2060) - P(X \leq 1939)$.

Applying the function "`pbinom`" for the computation of the Binomial cumulative probability, namely the probability of being less or equal to a given value, we get that the probability in the range between 1940 and 2060 is equal to

```
> pbinom(2060,4000,0.5) - pbinom(1939,4000,0.5)
[1] 0.9442883
```

This is an exact computation. The Normal approximation produces an approximate evaluation, not an exact computation. The Normal approximation replaces Binomial computations by computations carried out for the Normal distribution. The computation of a probability for a Binomial random variable is replaced by computation of probability for a Normal random variable that has the same expectation and standard deviation as the Binomial random variable.

Notice that if $X \sim \text{Binomial}(4000, 0.5)$ then the expectation is $E(X) = 4,000 \times 0.5 = 2,000$ and the variance is $\text{Var}(X) = 4,000 \times 0.5 \times 0.5 = 1,000$, with the standard deviation being the square root of the variance. Repeating the same computation that we conducted for the Binomial random variable, but this time with the function "`pnorm`" that is used for the computation of the Normal cumulative probability, we get:

```
> mu <- 4000*0.5
> sig <- sqrt(4000*0.5*0.5)
> pnorm(2060,mu,sig) - pnorm(1939,mu,sig)
[1] 0.9442441
```

Observe that in this example the Normal approximation of the probability (0.9442441) agrees with the Binomial computation of the probability (0.9442883) up to 3 significant digits.

Normal computations may also be applied in order to find approximate percentiles of the Binomial distribution. For example, let us identify the central

region that contains for a Binomial$(4000, 0.5)$ random variable (approximately) 95% of the distribution. Towards that end we can identify the boundaries of the region for the Normal distribution with the same expectation and standard deviation as that of the target Binomial distribution:

```
> qnorm(0.975,mu,sig)
[1] 2061.980
> qnorm(0.025,mu,sig)
[1] 1938.020
```

The value 2061 is at the 97.5 th percentile of the NORMAL distribution that was specified here, and this will be an ESTIMATION of the binomial distribution in the next block of code.

The value 1938.020 is at the 2.5 percentail of the distribution.

After rounding to the nearest integer we get the interval $[1938, 2062]$ as a proposed central region.

In order to validate the proposed region we may repeat the computation under the actual Binomial distribution:

```
> qbinom(0.975,4000,0.5)
[1] 2062
> qbinom(0.025,4000,0.5)
[1] 1938
```

qbinom example. The value 2062 is the correct answer for the 97.5 th percentil of the binomial distribution that has 4,000 experiments that each have a .5 probability of success.

Again, we get the interval $[1938, 2062]$ as the central region, in agreement with the one proposed by the Normal approximation. Notice that the function "`qbinom`" produces the percentiles of the Binomial distribution. It may not come as a surprise to learn that "`qpois`", "`qunif`", "`qexp`" compute the percentiles of the Poisson, Uniform and Exponential distributions, respectively.

The ability to approximate one distribution by the other, when computation tools for both distributions are handy, seems to be of questionable importance. Indeed, the significance of the Normal approximation is not so much in its ability to approximate the Binomial distribution as such. Rather, the important point is that the Normal distribution may serve as an approximation to a wide class of distributions, with the Binomial distribution being only one example. Computations that are based on the Normal approximation will be valid for all members in the class of distributions, including cases where we don't have the computational tools at our disposal or even in cases where we do not know what the exact distribution of the member is! As promised, a more detailed discussion of the Normal approximation in a wider context will be presented in the next chapter.

On the other hand, one need not assume that any distribution is well approximated by the Normal distribution. For example, the distribution of wealth in the population tends to be skewed, with more than 50% of the people possessing less than 50% of the wealth and small percentage of the people possessing the majority of the wealth. The Normal distribution is not a good model for such distribution. The Exponential distribution, or distributions similar to it, may be more appropriate.

6.3.2   Continuity Corrections

In order to complete this section let us look more carefully at the Normal approximations of the Binomial distribution.

In principle, the Normal approximation is valid when n, the number of independent trials in the Binomial distribution, is large. When n is relatively small

This is supposed to show that the binomial distribution is discrete and only has integer values, but the normal distribution is continuous and can obtain decimal values.

If we use the normal distribution, we treat the binomial integers as extending for a width of 1 unit, so the number 6 extends from 5.5 to 6.5.

You can envision the binomial distribution as having bars of width "1-unit," so if you try to use the normal distribution to simulate the binomial distribution, you would force everything to be on left or right side of 1-unit bars that are centred in integers. That means that the 1-unit bars end on the half digits, like 1.5, 2.5, 3.5 ...

Figure 6.6: Normal Approximation of the Binomial Distribution

the approximation may not be so good. Indeed, take $X \sim \text{Binomial}(30, 0.3)$ and consider the probability $\text{P}(X \leq 6)$. Compare the actual probability to the Normal approximation:

```
> pbinom(6,30,0.3)
[1] 0.1595230
> pnorm(6,30*0.3,sqrt(30*0.3*0.7))
[1] 0.1159989
```

The Normal approximation, which is equal to 0.1159989, is not too close to the actual probability, which is equal to 0.1595230.

A naïve application of the Normal approximation for the Binomial(n, p) distribution may not be so good when the number of trials n is small. Yet, a small modification of the approximation may produce much better results. In order to explain the modification consult Figure 6.6 where you will find the bar plot of the Binomial distribution with the density of the approximating Normal distribution superimposed on top of it. The target probability is the sum of heights of the bars that are painted in *red*. In the naïve application of the Normal approximation we used the area under the normal density which is to

the left of the bar associated with the value $x = 6$.

Alternatively, you may associate with each bar located at x the area under the normal density over the interval $[x - 0.5, x + 0.5]$. The resulting correction to the approximation will use the Normal probability of the event $\{X \leq 6.5\}$, which is the area shaded in *red*. The application of this approximation, which is called *continuity correction* produces:

```
> pnorm(6.5,30*0.3,sqrt(30*0.3*0.7))
[1] 0.1596193
```

The 30 * .3 is the mean of a binomial distribution with n= 30 and p = .3.
The sqrt() thing is the std deviation of a binomial from the middle of page 69 (and take the square root of the variance to get std deviation).

Observe that the corrected approximation is much closer to the target probability, which is 0.1595230, and is substantially better that the uncorrected approximation which was 0.1159989. Generally, it is recommended to apply the continuity correction to the Normal approximation of a discrete distribution.

Consider the Binomial(n, p) distribution. Another situation where the Normal approximation may fail is when p, the probability of "Success" in the Binomial distribution, is too close to 0 (or too close to 1). Recall, that for large n the Poisson distribution emerged as an approximation of the Binomial distribution in such a setting. One may expect that when n is large and p is small then the Poisson distribution may produce a better approximation of a Binomial probability. When the Poisson distribution is used for the approximation we call it a *Poisson Approximation*.

Let us consider an example. Let us analyze 3 Binomial distributions. The expectation in all the distributions is equal to 2 but the number of trials, n, vary. In the first case $n = 20$ (and hence $p = 0.1$), in the second $n = 200$ (and $p = 0.01$), and in the third $n = 2,000$ (and $p = 0.001$. In all three cases we will be interested in the probability of obtaining a value less or equal to 3.

The Poisson approximation replaces computations conducted under the Binomial distribution with Poisson computations, with a Poisson distribution that has the same expectation as the Binomial. Since in all three cases the expectation is equal to 2 we get that the same Poisson approximation is used to the three probabilities:

```
> ppois(3,2)
[1] 0.8571235
```

The ppois(3,2) command tells you the probability of randomly selecting a value that is 3 or less from the Poisson distribution that has an expectation of 2.

The actual Binomial probability in the first case ($n = 20$, $p = 0.1$) and a Normal approximation thereof are:

```
> pbinom(3,20,0.1)
```

The pbinom(3, 20, 0.1) command will tell you the probability of getting 3 successes from 20 experiments if each experiment has a probability of success of 0.1.

```
[1] 0.8670467
> pnorm(3.5,2,sqrt(20*0.1*0.9))
[1] 0.8682238
```

Observe that the Normal approximation (with a continuity correction) is better than the Poisson approximation in this case.

In the second case ($n = 200$, $p = 0.01$) the actual Binomial probability and the Normal approximation of the probability are:

```
> pbinom(3,200,0.01)
[1] 0.858034
> pnorm(3.5,2,sqrt(200*0.01*0.99))
[1] 0.856789
```

Observe that the Poisson approximation that produces 0.8571235 is slightly closer to the target than the Normal approximation. The greater accuracy of the Poisson approximation for the case where n is large and p is small is more pronounced in the final case ($n = 2000$, $p = 0.001$) where the target probability and the Normal approximation are:

```
> pbinom(3,2000,0.001)
[1] 0.8572138
> pnorm(3.5,2,sqrt(2000*0.001*0.999))
[1] 0.8556984
```

Compare the actual Binomial probability, which is equal to 0.8572138, to the Poisson approximation that produced 0.8571235. The Normal approximation, 0.8556984, is slightly off, but is still acceptable.

6.4   Solved Exercises

Question 6.1. Consider the problem of establishing regulations concerning the maximum number of people who can occupy a lift. In particular, we would like to assess the probability of exceeding maximal weight when 8 people are allowed to use the lift simultaneously and compare that to the probability of allowing 9 people into the lift.

   Assume that the total weight of 8 people chosen at random follows a normal distribution with a mean of 560kg and a standard deviation of 57kg. Assume that the total weight of 9 people chosen at random follows a normal distribution with a mean of 630kg and a standard deviation of 61kg.

1. What is the probability that the total weight of 8 people exceeds 650kg?

2. What is the probability that the total weight of 9 people exceeds 650kg?

3. What is the central region that contains 80% of distribution of the total weight of 8 people?

4. What is the central region that contains 80% of distribution of the total weight of 9 people?

Solution (to Question 6.1.1): Let X be the total weight of 8 people. By the assumption, $X \sim \text{Normal}(560, 57^2)$. We are interested in the probability $P(X > 650)$. This probability is equal to the difference between 1 and the probability $P(X \leq 650)$. We use the function "**pnorm**" in order to carry out the computation:

This finds the probability of being ABOVE 650.

```
> 1 - pnorm(650,560,57)
[1] 0.05717406
```

We get that the probability that the total weight of 8 people exceeds 650kg is equal to 0.05717406.

Solution (to Question 6.1.2): Let Y be the total weight of 9 people. By the assumption, $Y \sim \text{Normal}(630, 61^2)$. We are interested in the probability $P(Y > 650)$. This probability is equal to the difference between 1 and the probability $P(Y \leq 650)$. We use again the function "**pnorm**" in order to carry out the computation:

```
> 1 - pnorm(650,630,61)
[1] 0.3715054
```

We get that the probability that the total weight of 9 people exceeds 650kg is much higher and is equal to 0.3715054.

Solution (to Question 6.1.3): Again, $X \sim \text{Normal}(560, 57^2)$, where X is the total weight of 8 people. In order to find the central region that contains 80% of the distribution we need to identify the 10%-percentile and the 90%-percentile of X. We use the function "`qnorm`" in the code:

```
> qnorm(0.1,560,57)        486.9516 is at the 10th percentile (the cutoff point for the lowest 10%)
[1] 486.9516
> qnorm(0.9,560,57)
[1] 633.0484               633.0484 is at the 90th percentile (the cutoff for the top 10%).
```

The requested region is the interval [486.9516, 633.0484].

Solution (to Question 6.1.4): As before, $Y \sim \text{Normal}(630, 61^2)$, where Y is the total weight of 9 people. In order to find the central region that contains 80% of the distribution we need to identify the 10%-percentile and the 90%-percentile of Y. The computation this time produces:

```
> qnorm(0.1,630,61)
[1] 551.8254
> qnorm(0.9,630,61)
[1] 708.1746
```

and the region is [551.8254, 708.1746].

Question 6.2. Assume $X \sim \text{Binomial}(27, 0.32)$. We are interested in the probability $P(X > 11)$.

1. Compute the (exact) value of this probability.

2. Compute a Normal approximation to this probability, without a continuity correction.

3. Compute a Normal approximation to this probability, with a continuity correction.

4. Compute a Poisson approximation to this probability.

Solution (to Question 6.2.1): The probability $P(X > 11)$ can be computed as the difference between 1 and the probability $P(X \leq 11)$. The latter probability can be computed with the function "`pbinom`":

```
> 1 - pbinom(11,27,0.32)
[1] 0.1203926
```

Therefore, $P(X > 11) = 0.1203926$.

Solution (to Question 6.2.2): Refer again to the probability $P(X > 11)$. A formal application of the Normal approximation replaces in the computation

the Binomial distribution by the Normal distribution with the same mean and variance. Since $E(X) = n \cdot p = 27 \cdot 0.32 = 8.64$ and $Var(X) = n \cdot p \cdot (1 - p) = 27 \cdot 0.32 \cdot 0.68 = 5.8752$. If we take $X \sim \text{Normal}(8.64, 5.8752)$ and use the function "pnorm" we get:

```
> 1 - pnorm(11,27*0.32,sqrt(27*0.32*0.68))
[1] 0.1651164
```

Therefore, the current Normal approximation proposes $P(X > 11) \approx 0.1651164$.

Solution (to Question 6.2.3): The continuity correction, that consider interval of range 0.5 about each value, replace $P(X > 11)$, that involves the values $\{12, 13, \ldots, 27\}$, by the event $P(X > 11.5)$. The Normal approximation uses the Normal distribution with the same mean and variance. Since $E(X) = 8.64$ and $Var(X) = 5.8752$. If we take $X \sim \text{Normal}(8.64, 5.8752)$ and use the function "pnorm" we get:

```
> 1 - pnorm(11.5,27*0.32,sqrt(27*0.32*0.68))
[1] 0.1190149
```

The Normal approximation with continuity correction proposes $P(X > 11) \approx 0.1190149$.

Solution (to Question 6.2.4): The Poisson approximation replaces the Binomial distribution by the Poisson distribution with the same expectation. The expectation is $E(X) = n \cdot p = 27 \cdot 0.32 = 8.64$. If we take $X \sim \text{Poisson}(8.64)$ and use the function "ppois" we get:

```
> 1 - ppois(11,27*0.32)
[1] 0.1635232
```

Therefore, the Poisson approximation proposes $P(X > 11) \approx 0.1651164$.

6.5   Summary

Glossary

Normal Random Variable: A bell-shaped distribution that is frequently used to model a measurement. The distribution is marked with $\text{Normal}(\mu, \sigma^2)$.

Standard Normal Distribution: The $\text{Normal}(0, 1)$. The distribution of standardized Normal measurement.

Percentile: Given a percent $p \cdot 100\%$ (or a probability p), the value x is the percentile of a random variable X if it satisfies the equation $P(X \leq x) = p$.

Normal Approximation of the Binomial: Approximate computations associated with the Binomial distribution with parallel computations that use the Normal distribution with the same expectation and standard deviation as the Binomial.

Poisson Approximation of the Binomial: Approximate computations associated with the Binomial distribution with parallel computations that use the Poisson distribution with the same expectation as the Binomial.

Discuss in the Forum

Mathematical models are used as tools to describe reality. These models are supposed to characterize the important features of the analyzed phenomena and provide insight. Random variables are mathematical models of measurements. Some people claim that there should be a perfect match between the mathematical characteristics of a random variable and the properties of the measurement it models. Other claim that a partial match is sufficient. What is your opinion?

When forming your answer to this question you may give an example of a situation from you own field of interest for which a random variable can serve as a model. Identify discrepancies between the theoretical model and actual properties of the measurement. Discuss the appropriateness of using the model in light of these discrepancies.

Consider, for example, testing IQ. The score of many IQ tests are modeled as having a Normal distribution with an expectation of 100 and a standard deviation of 15. The sample space of the Normal distribution is the entire line of real numbers, including the negative numbers. In reality, IQ tests produce only positive values.

Chapter 7

The Sampling Distribution

7.1 Student Learning Objective

In this section we integrate the concept of *data* that is extracted from a sample with the concept of a *random variable*. The new element that connects between these two concepts is the notion of *sampling distribution*. The data we observe results from the specific sample that was selected. The sampling distribution, in a similar way to random variables, corresponds to all samples that could have been selected. (Or, stated in a different tense, to the sample that will be selected prior to the selection itself.) Summaries of the distribution of the data, such as the sample mean and the sample standard deviation, become random variables when considered in the context of the sampling distribution. In this section we investigate the sampling distribution of such data summaries. In particular, it is demonstrated that (for large samples) the sampling distribution of the sample average may be approximated by the Normal distribution. The mathematical theorem that proves this approximation is called the *Central Limit Theory*. By the end of this chapter, the student should be able to:

- Comprehend the notion of sampling distribution and simulate the sampling distribution of the sample average.

- Relate the expectation and standard deviation of a measurement to the expectation and standard deviation of the sample average.

- Apply the Central Limit Theorem to the sample averages.

7.2 The Sampling Distribution

In Chapter 5 the concept of a random variable was introduced. As part of the introduction we used an example that involved the selection of a random person from the population and the measuring of his/her height. Prior to the action of selection, the height of that person is a *random variable*. It has the potential of obtaining any of the heights that are present in the population, which is the *sample space* of this example, with a distribution that reflects the relative frequencies of each of the heights in the population: the *probabilities* of the values. After the selection of the person and the measuring of the height

we get a particular value. This is the *observed value* and is no longer a random variable. In this section we extend the concept of a random variable and define the concept of *a random sample*.

7.2.1   A Random Sample

The relation between the random sample and the data is similar to the relation between a random variable and the observed value. The data is the observed values of a sample taken from a population. The content of the data is known. The random sample, similarly to a random variable, is the data that *will* be selected when taking a sample, prior to the selection itself. The content of the random sample is unknown, since the sample has not yet been taken. Still, just like for the case of the random variable, one is able to say what the possible evaluations of the sample may be and, depending on the mechanism of selecting the sample, what are the probabilities of the different potential evaluations. The collection of all possible evaluations of the sample is the *sample space of the random sample* and the probabilities of the different evaluations produce the *distribution* of the random sample.

(Alternatively, if one prefers to speak in past tense, one can define the sample space of a random sample to be the evaluations of the sample that could have taken place, with the distribution of the random sample being the probabilities of these evaluations.)

A *statistic* is a function of the data. Example of statistics are the average of the data, the sample variance and standard deviation, the median of the data, etc. In each case a given formula is applied to the data. In each type of statistic a different formula is applied.

The same formula that is applied to the observed data may, in principle, be applied to random samples. Hence, for example, one may talk of the sample average, which is the average of the elements in the data. The average, considered in the context of the observed data, is a number and its value is known. However, if we think of the average in the context of a random sample then it becomes a random variable. Prior to the selection of the actual sample we do not know what values it will include. Hence, we cannot tell what the outcome of the average of the values will be. However, due to the identification of all possible evaluations that the sample can possess we may say in advance what is the collection of values the sample average can have. This is the sample space of the sample average. Moreover, from the sampling distribution of the random sample one may identify the probability of each value of the sample average, thus obtaining the *sampling distribution* of the sample average.

The same line of argumentation applies to any statistic. Computed in the context of the observed data, the statistic is a known number that may, for example, be used to characterize the variation in the data. When thinking of a statistic in the context of a random sample it becomes a random variable. The distribution of the statistic is called the sampling distribution of the statistic. Consequently, we may talk of the sampling distribution of the median, the sample distribution of the sample variance, etc.

Random variables are also applied as models for uncertainty in future measurements in more abstract settings that need not involve a specific population. Specifically, we introduced the Binomial and Poisson random variables for settings that involve counting and the Uniform, Exponential, and Normal random

variables for settings where the measurement is continuous.

The notion of a sampling distribution may be extended to a situation where one is taking several measurements, each measurement taken independently of the others. As a result one obtains a *sequence* of measurements. We use the term "sample" to denote this sequence. The distribution of this sequence is also called the sampling distribution. If all the measurements in the sequence are Binomial then we call it a *Binomial sample*. If all the measurements are Exponential we call it an *Exponential sample* and so forth.

Again, one may apply a formula (such as the average) to the content of the random sequence and produce a random variable. The term *sampling distribution* describes again the distribution that the random variable produced by the formula inherits from the sample.

In the next subsection we examine an example of a sample taken from a population. Subsequently, we discuss examples that involves a sequence of measurements from a theoretical model.

7.2.2 Sampling From a Population

The pop1 file is a big file. We arbitrarily treat it as a population.

Consider taking a sample from a population. Let us use again for the illustration the file "pop1.csv" like we did in Chapter 4. The data frame produced from the file contains the sex and hight of the 100,000 members of some imaginary population. Recall that in Chapter 4 we applied the function "sample" to randomly sample the height of a single subject from the population. Let us apply the same function again, but this time in order to sample the heights of 100 subjects:

```
> pop.1 <- read.csv("pop1.csv")
> X.samp <- sample(pop.1$height,100)
> X.samp
  [1] 168 177 172 174 154 179 145 160 188 172 175 174 176 144 164
 [16] 171 167 158 181 165 166 173 184 174 169 176 168 154 167 175
 [31] 178 179 175 187 160 171 175 172 178 167 181 193 163 181 168
 [46] 153 200 168 169 194 177 182 167 183 177 155 167 172 176 168
 [61] 164 162 188 163 166 156 163 185 149 163 157 155 161 177 176
 [76] 153 162 180 177 156 162 197 183 166 185 178 188 198 175 167
 [91] 185 160 148 160 174 162 161 178 159 168
```

The sample() command will extract 100 items from the height data to simulate a new experiment in which we collect 100 observations.

In the first line of code we produce a data frame that contains the information on the entire population. In the second line we select a sample of size 100 from the population, and in the third line we present the content of the sample.

The first argument to the function "sample" that selects the sample is the sequence of length 100,000 with the list of heights of all the members of the population. The second argument indicates the sample size, 100 in this case. The outcome of the random selection is stored in the object "X.samp", which is a sequence that contains 100 heights.

Typically, a researcher does not get to examine the entire population. Instead, measurements on a sample from the population are made. In relation to the imaginary setting we simulate in the example, the typical situation is that the research does not have the complete list of potential measurement evaluations, i.e. the complete list of 100,000 heights in "pop.1$height", but only a sample of measurements, namely the list of 100 numbers that are stored in

"`X.samp`" and are presented above. The role of statistics is to make inference on the parameters of the unobserved population based on the information that is obtained from the sample.

For example, we may be interested in estimating the mean value of the heights in the population. A reasonable proposal is to use the sample average to serve as an estimate:

```
> mean(X.samp)
[1] 170.73
```
the mean of our sample was very close to the real mean, but not exact.

In our artificial example we can actually compute the true population mean:

```
> mean(pop.1$height)
[1] 170.035
```

Hence, we may see that although the match between the estimated value and the actual value is not perfect still they are close enough.

The actual estimate that we have obtained resulted from the specific sample that was collected. Had we collected a different subset of 100 individuals we would have obtained different numerical value for the estimate. Consequently, one may wonder: Was it pure luck that we got such good estimates? How likely is it to get estimates that are close to the target parameter?

Notice that in realistic settings we do not know the actual value of the target population parameters. Nonetheless, we would still want to have at least a probabilistic assessment of the distance between our estimates and the parameters they try to estimate. The sampling distribution is the vehicle that may enable us to address these questions.

In order to illustrate the concept of the sampling distribution let us select another sample and compute its average:

```
> X.samp <- sample(pop.1$height,100)
> X.bar <- mean(X.samp)
> X.bar
[1] 171.87
```
The sample() command tells R to select 100 values at random from the height data. This simulates a real experiment where you collect 100 observations of data.

and do it once more:

```
> X.samp <- sample(pop.1$height,100)
> X.bar <- mean(X.samp)
> X.bar
[1] 171.02
```
The selection process is affected by random variability, so the next sample is a bit different. Your results will vary.

In each case we got a different value for the sample average. In the first of the last two iterations the result was more than 1 centimeter away from the population average, which is equal to 170.035, and in the second it was within the range of 1 centimeter. Can we say, prior to taking the sample, what is the probability of falling within 1 centimeter of the population mean?

Chapter 4 discussed the random variable that emerges by randomly sampling a single number from the population presented by the sequence "`pop.1$height`". The distribution of the random variable resulted from the assignment of the probability 1/100,000 to each one of the 100,000 possible outcomes. The same principle applies when we randomly sample 100 individuals. Each possible outcome is a collection of 100 numbers and each collection is assigned equal probability. The resulting distribution is called *the sampling distribution*.

The distribution of the average of the sample emerges from this distribution: With each sample one may associate the average of that sample. The probability assigned to that average outcome is the probability of the sample. Hence, one may assess the probability of falling within 1 centimeter of the population mean using the sampling distribution. Each sample produces an average that either falls within the given range or not. The probability of the sample average falling within the given range is the proportion of samples for which this event happens among the entire collection of samples.

However, we face a technical difficulty when we attempt to assess the sampling distribution of the average and the probability of falling within 1 centimeter of the population mean. Examination of the distribution of a sample of a single individual is easy enough. The total number of outcomes, which is 100,000 in the given example, can be handled with no effort by the computer. However, when we consider samples of size 100 we get that the total number of ways to select 100 number out of 100,000 numbers is in the order of 10^{342} (1 followed by 342 zeros) and cannot be handled by any computer. Thus, the probability cannot be computed.

As a compromise we will approximate the distribution by selecting a large number of samples, say 100,000, to represent the entire collection, and use the resulting distribution as an approximation of the sampling distribution. Indeed, the larger the number of samples that we create the more accurate the approximation of the distribution is. Still, taking 100,000 repeats should produce approximations which are good enough for our purposes.

Consider the sampling distribution of the sample average. We simulated above a few examples of the average. Now we would like to simulate 100,000 such examples. We do this by creating first a sequence of the length of the number of evaluations we seek (100,000) and then write a small program that produces each time a new random sample of size 100 and assigns the value of the average of that sample to the appropriate position in the sequence. Do first and explain later[1]:

```
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- sample(pop.1$height,100)
+     X.bar[i] <- mean(X.samp)
+ }
> hist(X.bar)
```

See the MATH1280Notes.pdf file for a detailed explanation of this.

In the first line we produce a sequence of length 100,000 that contains zeros. The function "`rep`" creates a sequence that contains repeats of its first argument a number of times that is specified by its second argument. In this example, the numerical value 0 is repeated 100,000 times to produce a sequence of zeros of the length we seek.

---

[1]Running this simulation, and similar simulations of the same nature that will be considered in the sequel, demands more of the computer's resources than the examples that were considered up until now. Beware that running times may be long and, depending on the strength of your computer and your patience, too long. You may save time by running less iterations, replacing, say, "`10^5`" by "`10^4`". The results of the simulation will be less accurate, but will still be meaningful.
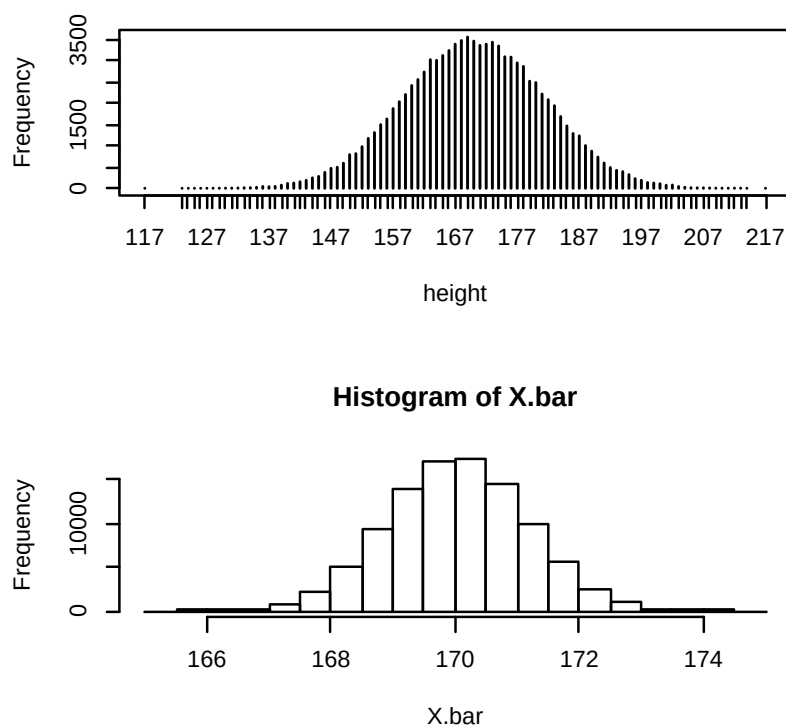
Figure 7.1:   Distribution of Height and the Sampling Distribution of Averages

The main part of the program is a "`for`" loop. The argument of the function "`for`" takes the special form: "*index.name* `in` *index.values*", where *index.name* is the name of the running index and *index.values* is the collection of values over which the running index is evaluated. In each iteration of the loop the running index is assigned a value from the collection and the expression that follows the brackets of the "`for`" function is evaluated with the given value of the running index.

In the given example the collection of values is produced by the expression "`1:n`". Recall that the expression "`1:n`" produces the collection of integers between `1` and `n`. Here, `n` = 100,000. Hence, in the given application the collection of values is a sequence that contains the integers between 1 and 100,000. The running index is called "`i`". the expression is evaluated 100,000 times, each time with a different integer value for the running index "`i`".

The R system treats a collection of expressions enclosed within curly brackets as one entity. Therefore, in each iteration of the "`for`" loop, the lines that are within the curly brackets are evaluated. In the first line a random sample of size 100 is produced and in the second line the average of the sample is computed and stored in the *i*-th position of the sequence "`X.bar`". Observe that the specific

position in the sequence is referred to by using square brackets.

The program changes the original components of the sequence, from 0 to the average of a random sample, one by one. When the loop ends all values are changed and the sequence "`X.bar`" contains 100,000 evaluations of the sample average. The last line, which is outside the curly brackets and is evaluated after the "`for`" loop ends, produces an histogram of the averages that were simulated. The histogram is presented in the lower panel of Figure 7.1.

Compare the distribution of the sample average to the distribution of the heights in the population that was presented first in Figure 4.1 and is currently presented in the upper panel of Figure 7.1. Observe that both distributions are centered at about 170 centimeters. Notice, however, that the range of values of the sample average lies essentially between 166 and 174 centimeters, whereas the range of the distribution of heights themselves is between 127 and 217 centimeter. Broadly speaking, the sample average and the original measurement are centered around the same location but the sample average is less spread.

Specifically, let us compare the expectation and standard deviation of the sample average to the expectation and standard deviation of the original measurement:

```
> mean(pop.1$height)
[1] 170.035
> sd(pop.1$height)
[1] 11.23205
> mean(X.bar)
[1] 170.037
> sd(X.bar)
[1] 1.122116
```

Observe that the expectation of the population and the expectation of the sample average, are practically the same, the standard deviation of the sample average is about 10 times smaller than the standard deviation of the population. This result is not accidental and actually reflects a general phenomena that will be seen below in other examples.

We may use the simulated sampling distribution in order to compute an approximation of the probability of the sample average falling within 1 centimeter of the population mean. Let us first compute the relevant probability and then explain the details of the computation:

```
> mean(abs(X.bar - mean(pop.1$height)) <= 1)
[1] 0.62589
```

Hence we get that the probability of the given event is about 62.6%.

The object "`X.bar`" is a sequence of length 100,000 that contains the simulated sample averages. This sequence represents the distribution of the sample average. The expression "`abs(X.bar - mean(pop.1$height)) <= 1`" produces a sequence of logical "`TRUE`" or "`FALSE`" values, depending on the value of the sample average being less or more than one unit away from the population mean. The application of the function "`mean`" to the output of the last expression results in the computation of the relative frequency of `TRUE`s, which corresponds to the probability of the event of interest.

Example 7.1. *A poll for the determination of the support in the population for a candidate was describe in Example 5.1. The proportion in the population of supporters was denoted by p. A sample of size n = 300 was considered in order to estimate the size of p. We identified that the distribution of X, the number of supporters in the sample, is* Binomial(300, p). *This distribution is the sampling distribution[2] of X. One may use the proportion in the sample of supporters, the number of supporters in the sample divided by 300, as an estimate to the parameter p. The sampling distribution of this quantity, X/300, may be considered in order to assess the discrepancy between the estimate and the actual value of the parameter.*

7.2.3   Theoretical Models

Sampling distribution can also be considered in the context of theoretical distribution models. For example, take a measurement $X \sim$ Binomial(10, 0.5) from the Binomial distribution. Assume 64 independent measurements are produced with this distribution: $X_1, X_2, \ldots, X_{64}$. The sample average in this case corresponds to the distribution of the random variable produced by averaging these 64 random variables:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{64}}{64} = \frac{1}{64} \sum_{i=1}^{64} X_i \;.$$

Again, one may wonder what is the distribution of the sample average $\bar{X}$ in this case?

We can approximate the distribution of the sample average by simulation. The function "`rbinom`" produces a random sample from the Binomial distribution. The first argument to the function is the sample size, which we take in this example to be equal to 64. The second and third arguments are the parameters of the Binomial distribution, 10 and 0.5 in this case. We can use this function in the simulation:

```
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- rbinom(64,10,0.5)
+     X.bar[i] <- mean(X.samp)
+ }
```

Observe that in this code we created a sequence of length 100,000 with evaluations of the sample average of 64 Binomial random variables. We start with a sequence of zeros and in each iteration of the "`for`" loop a zero is replaced by the average of a random sample of 64 Binomial random variables.

Examine the sampling distribution of the Binomial average:

```
> hist(X.bar)
```

---

[2]Mathematically speaking, the Binomial distribution is only an approximation to the sampling distribution of X. Actually, the Binomial is an exact description to the distribution only in the case where each subject has the chance be represented in the sample more than once. However, only when the size of the sample is comparable to the size of the population would the Binomial distribution fail to be an adequate approximation to the sampling distribution.
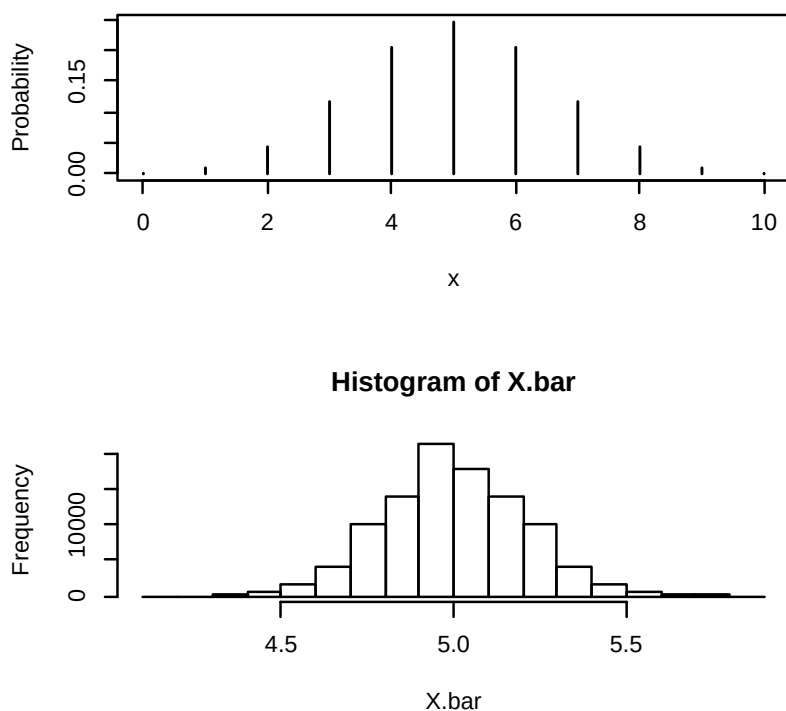
Figure 7.2: Distributions of an Average and a Single Binomial(10,0.5)

```
> mean(X.bar)
[1] 4.999074
> sd(X.bar)
[1] 0.1982219
```

The histogram of the sample average is presented in the lower panel of Figure 7.2. Compare it to the distribution of a single Binomial random variable that appears in the upper panel. Notice, once more, that the center of the two distributions coincide but the spread of the sample average is smaller. The sample space of a single Binomial random variable is composed of integers. The sample space of the average of 64 Binomial random variables, on the other hand, contains many more values and is closer to the sample space of a random variable with a continuous distribution.

Recall that the expectation of a Binomial$(10, 0.5)$ random variable is $E(X) = 10 \cdot 0.5 = 5$ and the variance is $\text{Var}(X) = 10 \cdot 0.5 \cdot 0.5 = 2.5$ (thus, the standard deviation is $\sqrt{2.5} = 1.581139$). Observe that the expectation of the sample average that we got from the simulation is essentially equal to 5 and the standard deviation is 0.1982219.

One may prove mathematically that the expectation of the sample mean is equal to the theoretical expectation of its components:

$$\mathrm{E}(\bar{X}) = \mathrm{E}(X) .$$

The results of the simulation for the expectation of the sample average are consistent with the mathematical statement. The mathematical theory of probability may also be used in order to prove that the variance of the sample average is equal to the variance of each of the components, divided by the sample size:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n ,$$

The variance of the sampling distribution of the mean equals the variance of the underlying distribution divided by the number of observations.

here n is the number of observations in the sample. Specifically, in the Binomial example we get that $\mathrm{Var}(\bar{X}) = 2.5/64$, since the variance of a Binomial component is 2.5 and there are 64 observations. Consequently, the standard deviation is $\sqrt{2.5/64} = 0.1976424$, in agreement, more or less, with the results of the simulation (that produced 0.1982219 as the standard deviation).

Note that the variance for the sampling distribution of the mean changes when you change the sample size.

Consider the problem of identifying the central interval that contains 95% of the distribution. In the Normal distribution we were able to use the function "`qnorm`" in order to compute the percentiles of the theoretical distribution. A function that can be used for the same purpose for simulated distribution is the function "`quantile`". The first argument to this function is the sequence of simulated values of the statistic, "`X.bar`" in the current case. The second argument is a number between 0 and 1, or a sequence of such numbers:

```
> quantile(X.bar,c(0.025,0.975))
    2.5%    97.5%
4.609375 5.390625
```

You have asked R to show you the 2.5 percentile and the 97.5 percentile using a data object that holds sample means from many simulated experiments.

We used the sequence "`c(0.025,0.975)`" as the input to the second argument. As a result we obtained the output 4.609375, which is the 2.5%-percentile of the sampling distribution of the average, and 5.390625, which is the 97.5%-percentile of the sampling distribution of the average.

Of interest is to compare these percentiles to the parallel percentiles of the Normal distribution with the same expectation and the same standard deviation as the average of the Binomials:

```
> qnorm(c(0.025,0.975),mean(X.bar),sd(X.bar))
[1] 4.611456 5.389266
```

This version gets the 2.5 and 97.5 percentiles from a normal distributoin following the equations for mean and standard deviation at the top of this page. Note that our simulation produce almost the same results.

Observe the similarity between the percentiles of the distribution of the average and the percentiles of the Normal distribution. This similarity is a reflection of the Normal approximation of the sampling distribution of the average, which is formulated in the next section under the title: *The Central Limit Theorem.*

Example 7.2. *The distribution of the number of events of radio active decay in a second was modeled in Example 5.3 according to the Poisson distribution. A quantity of interest is* λ*, the expectation of that Poisson distribution. This quantity may be estimated by measuring the total number of decays over a period of time and dividing the outcome by the number of seconds in that period of time. Let n be this number of second. The procedure just described corresponds to taking the sample average of* $\mathrm{Poisson}(\lambda)$ *observations for a sample of size n.*
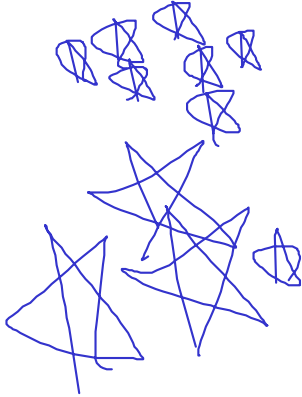
The expectation of the sample average is λ and the variance is λ/n, leading to a standard deviation of size $\sqrt{\lambda/n}$. The Central Limit Theorem states that the sampling distribution of this average corresponds, approximately, to the Normal distribution with this expectation and standard deviation.

7.3 Law of Large Numbers and Central Limit Theorem

The Law of Large Numbers and the Central Limit Theorem are mathematical theorems that describe the sampling distribution of the average for large samples.

7.3.1 The Law of Large Numbers

The Law of Large Numbers states that, as the sample size becomes larger, the sampling distribution of the sample average becomes more and more concentrated about the expectation.

the law of large numbers is a statement about the mean of the sampling distribution. That's all !!

Let us demonstrate the Law of Large Numbers in the context of the Uniform distribution. Let the distribution of the measurement X be Uniform$(3, 7)$. Consider three different sample sizes n: $n = 10$, $n = 100$, and $n = 1000$. Let us carry out a simulation similar to the simulations of the previous section. However, this time we run the simulation for the three sample sizes in parallel:

```
> unif.10 <- rep(0,10^5)
> unif.100 <- rep(0,10^5)
> unif.1000 <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp.10 <- runif(10,3,7)
+     unif.10[i] <- mean(X.samp.10)
+     X.samp.100 <- runif(100,3,7)
+     unif.100[i] <- mean(X.samp.100)
+     X.samp.1000 <- runif(1000,3,7)
+     unif.1000[i] <- mean(X.samp.1000)
+ }
```

Observe that we have produced 3 sequences of length 100,000 each: "`unif.10`", "`unif.100`", and "`unif.1000`". The first sequence is an approximation of the sampling distribution of an average of 10 independent Uniform measurements, the second approximates the sampling distribution of an average of 100 measurements and the third the distribution of an average of 1000 measurements. The distribution of single measurement in each of the examples is Uniform$(3, 7)$.

Consider the expectation of sample average for the three sample sizes:

```
> mean(unif.10)
[1] 4.999512
> mean(unif.100)
[1] 4.999892
> mean(unif.1000)
[1] 4.99996
```

For all sample size the expectation of the sample average is equal to 5, which is the expectation of the Uniform$(3, 7)$ distribution.

Recall that the variance of the Uniform(a, b) distribution is $(b - a)^2/12$. Hence, the variance of the given Uniform distribution is $\mathrm{Var}(X) = (7-3)^2/12 = 16/12 \approx 1.3333$. The variances of the sample averages are:

```
> var(unif.10)
[1] 0.1331749
> var(unif.100)
[1] 0.01333089
> var(unif.1000)
[1] 0.001331985
```

Notice that the variances decrease with the increase of the sample sizes. The decrease is according to the formula $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$.

The variance is a measure of the spread of the distribution about the expectation. The smaller the variance the more concentrated is the distribution around the expectation. Consequently, in agreement with the Law of Large Numbers, the larger the sample size the more concentrated is the sampling distribution of the sample average about the expectation.

7.3.2   The Central Limit Theorem (CLT)

The Law of Large Numbers states that the distribution of the sample average tends to be more concentrated as the sample size increases. The Central Limit Theorem (CLT in short) provides an approximation of this distribution.

The deviation between the sample average and the expectation of the measurement tend to decreases with the increase in sample size. In order to obtain a refined assessment of this deviation one needs to magnify it. The appropriate way to obtain the magnification is to consider the standardized sample average, in which the deviation of the sample average from its expectation is divided by the standard deviation of the sample average:

This equation and the next one are not essential as long as you know how to answer the questions in this chapter.

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\sqrt{\mathrm{Var}(\bar{X})}} \ .$$

Recall that the expectation of the sample average is equal to the expectation of a single random variable ($\mathrm{E}(\bar{X}) = \mathrm{E}(X)$) and that the variance of the sample average is equal to the variance of a single observation, divided by the sample size ($\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$). Consequently, one may rewrite the standardized sample average in the form:

$$Z = \frac{\bar{X} - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)/n}} = \frac{\sqrt{n}(\bar{X} - \mathrm{E}(X))}{\sqrt{\mathrm{Var}(X)}} \ .$$

The second equality follows from placing in the numerator the square root of n which *divides* the term in the denominator. Observe that with the increase of the sample size the decreasing difference between the average and the expectation is magnified by the square root of n.

The Central Limit Theorem states that, with the increase in sample size, the sample average converges (after standardization) to the standard Normal distribution.

The central limit theorem is a statment about the shape of the distribution of a sampling distribution.
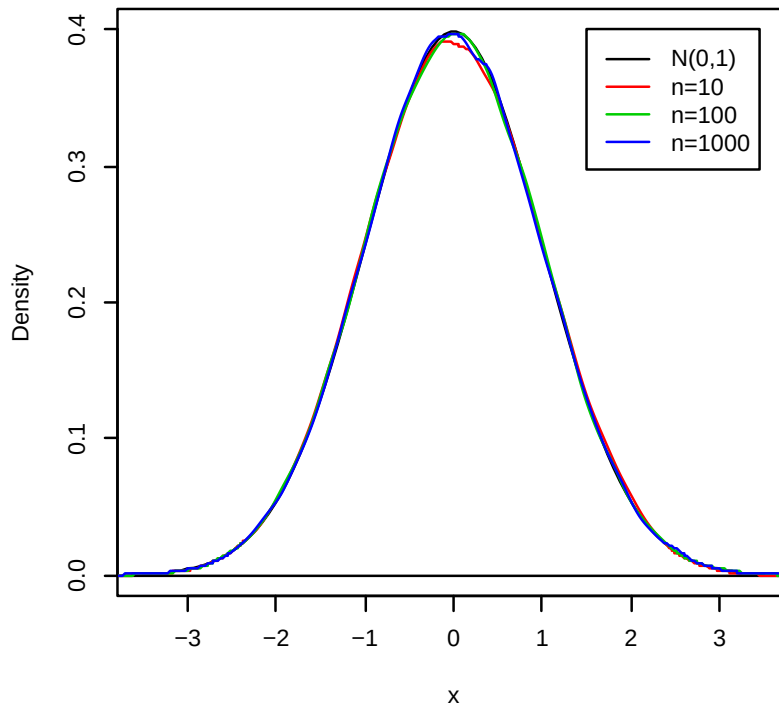
Figure 7.3: The CLT for the Uniform(3,7) Distribution

Let us examine the Central Normal Theorem in the context of the example of the Uniform measurement. In Figure 7.3 you may find the (approximated) density of the standardized average for the three sample sizes based on the simulation that we carried out previously (as *red*, *green*, and *blue* lines). Along side with these densities you may also find the theoretical density of the standard Normal distribution (as a *black* line). Observe that the four curves are almost one on top of the other, proposing that the approximation of the distribution of the average by the Normal distribution is good even for a sample size as small as $n = 10$.

However, before jumping to the conclusion that the Central Limit Theorem applies to any sample size, let us consider another example. In this example we repeat the same simulation that we did with the Uniform distribution, but this time we take Exponential(0.5) measurements instead:

```
> exp.10 <- rep(0,10^5)
> exp.100 <- rep(0,10^5)
> exp.1000 <- rep(0,10^5)
> for(i in 1:10^5)
```
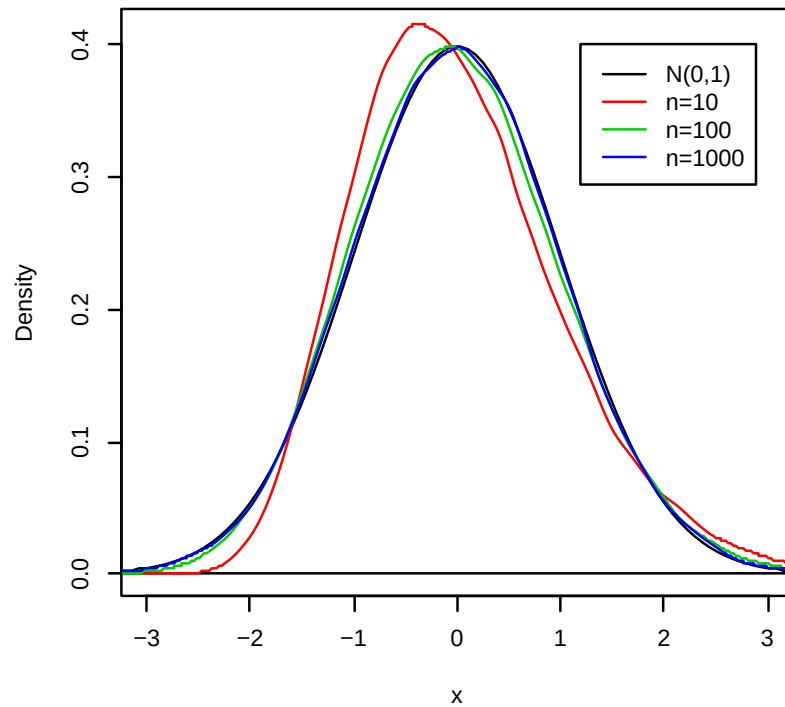
Figure 7.4: The CLT for the Exponential(0.5) Distribution

```
+ {
+     X.samp.10 <- rexp(10,0.5)
+     exp.10[i] <- mean(X.samp.10)
+     X.samp.100 <- rexp(100,0.5)
+     exp.100[i] <- mean(X.samp.100)
+     X.samp.1000 <- rexp(1000,0.5)
+     exp.1000[i] <- mean(X.samp.1000)
+ }
```

The expectation of an Exponential(0.5) random variable is $E(X) = 1/\lambda = 1/0.5 = 2$ and the variance is $Var(X) = 1/\lambda^2 = 1/(0.5)^2 = 4$. Observe below that the expectations of the sample averages are equal to the expectation of the measurement and the variances of the sample averages follow the relation $Var(\bar{X}) = Var(X)/n$:

```
> mean(exp.10)
[1] 1.999888
> mean(exp.100)
[1] 2.000195
```

```
> mean(exp.1000)
[1] 1.999968
```

So the expectations of the sample average are all equal to 2. For the variance we get:

```
> var(exp.10)
[1] 0.4034642
> var(exp.100)
[1] 0.03999479
> var(exp.1000)
[1] 0.004002908
```

Which is in agreement with the decrease proposed by the theory,

However, when one examines the densities of the sample averages in Figure 7.4 one may see a clear distinction between the sampling distribution of the average for a sample of size 10 and the normal distribution (compare the *red* curve to the *black* curve. The match between the *green* curve that corresponds to a sample of size $n = 100$ and the *black* line is better, but not perfect. When the sample size is as large as $n = 1000$ (the *blue* curve) then the agreement with the normal curve is very good.

7.3.3 Applying the Central Limit Theorem

The conclusion of the Central Limit Theorem is that the sampling distribution of the sample average can be approximated by the Normal distribution, regardless what is the distribution of the original measurement, but provided that the sample size is large enough. This statement is very important, since it allows us, in the context of the sample average, to carry out probabilistic computations using the Normal distribution even if we do not know the actual distribution of the measurement. All we need to know for the computation are the expectation of the measurement, its variance (or standard deviation) and the sample size.

The theorem can be applied whenever probability computations associated with the sampling distribution of the average are required. The computation of the approximation is carried out by using the Normal distribution with the same expectation and the same standard deviation as the sample average.

An example of such computation was conducted in Subsection 7.2.3 where the central interval that contains 95% of the sampling distribution of a Binomial average was required. The 2.5%- and the 97.5%-percentiles of the Normal distribution with the same expectation and variance as the sample average produced boundaries for the interval. These boundaries were in good agreement with the boundaries produced by the simulation. More examples will be provided in the Solved Exercises of this chapter and the next one.

With all its usefulness, one should treat the Central Limit Theorem with a grain of salt. The approximation may be valid for large samples, but may be bad for samples that are not large enough. When the sample is small a careless application of the Central Limit Theorem may produce misleading conclusions.

7.4   Solved Exercises

Question 7.1. The file "`pop2.csv`" contains information associated to the blood pressure of an imaginary population of size 100,000. The file can be found on the internet (`http://pluto.huji.ac.il/~msby/StatThink/Datasets/pop2.csv`). The variables in this file are:

id: A numerical variable. A 7 digits number that serves as a unique identifier of the subject.

sex: A factor variable. The sex of each subject. The values are either "`MALE`" or "`FEMALE`".

age: A numerical variable. The age of each subject.

bmi: A numerical variable. The body mass index of each subject.

systolic: A numerical variable. The systolic blood pressure of each subject.

diastolic: A numerical variable. The diastolic blood pressure of each subject.

group: A factor variable. The blood pressure category of each subject. The values are "`NORMAL`" both the systolic blood pressure is within its normal range (between 90 and 139) and the diastolic blood pressure is within its normal range (between 60 and 89). The value is "`HIGH`" if either measurements of blood pressure are above their normal upper limits and it is "`LOW`" if either measurements are below their normal lower limits.

Our goal in this question is to investigate the sampling distribution of the sample average of the variable "`bmi`". We assume a sample of size $n = 150$.

1. Compute the population average of the variable "`bmi`".

2. Compute the population standard deviation of the variable "`bmi`".

3. Compute the expectation of the sampling distribution for the sample average of the variable.

4. Compute the standard deviation of the sampling distribution for the sample average of the variable.

5. Identify, using simulations, the central region that contains 80% of the sampling distribution of the sample average.

6. Identify, using the Central Limit Theorem, an approximation of the central region that contains 80% of the sampling distribution of the sample average.

Solution (to Question 7.1.1): After placing the file "`pop2.csv`" in the working directory one may produce a data frame with the content of the file and compute the average of the variable "`bmi`" using the code:

```
> pop.2 <- read.csv(file="pop2.csv")
> mean(pop.2$bmi)
[1] 24.98446
```

We obtain that the population average of the variable is equal to 24.98446.

Solution (to Question 7.1.2): Applying the function "`sd`" to the sequence of population values produces the population standard deviation:

```
> sd(pop.2$bmi)
[1] 4.188511
```

In turns out that the standard deviation of the measurement is 4.188511.

Solution (to Question 7.1.3): In order to compute the expectation under the sampling distribution of the sample average we conduct a simulation. The simulation produces (an approximation) of the sampling distribution of the sample average. The sampling distribution is represented by the content of the sequence "`X.bar`":

```
> X.bar <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- sample(pop.2$bmi,150)
+     X.bar[i] <- mean(X.samp)
+ }
> mean(X.bar)
[1] 24.98681
```

Initially, we produce a vector of zeros of the given lenght (100,000). In each iteration of the "`for`" loop a random sample of size 150 is selected from the population. The sample average is computed and stored in the sequence "`X.bar`". At the end of all the iterations all the zeros are replaced by evaluations of the sample average.

The expectation of the sampling distribution of the sample average is computed by the application of the function "`mean`" to the sequence that represents the sampling distribution of the sample average. The result for the current is 24.98681, which is vary similar[3] to the population average 24.98446.

Solution (to Question 7.1.4): The standard deviation of the sample average under the sampling distribution is computed using the function "`sd`":

```
> sd(X.bar)
[1] 0.3422717
```

The resulting standard deviation is 0.3422717. Recall that the standard deviation of a single measurement is equal to 4.188511 and that the sample size is $n = 150$. The ratio between the standard deviation of the measurement and the square root of 150 is $4.188511/\sqrt{150} = 0.3419905$, which is similar in value to the standard deviation of the sample average[4].

---

[3]Theoretically, the two numbers should coincide. The small discrepancy follows from the fact that the sequence "`X.bar`" is only an approximation of the sampling distribution.

[4]It can be shown mathematically that the variance of the sample average, in the case of sampling from a population, is equal to $[(N - n)/(N - 1)] \cdot \mathrm{Var}(X)/n$, where $\mathrm{Var}(X)$ is the population variance of the measurement, n is the sample size, and N is the population size. The factor $[(N - n)/(N - 1)]$ is called *the finite population correction*. In the current setting the finite population correction is equal to 0.99851, which is practically equal to one.

Solution (to Question 7.1.5): The central region that contains 80% of the sampling distribution of the sample average can be identified with the aid of the function "`quantile`":

```
> quantile(X.bar,c(0.1,0.9))
     10%      90%
24.54972 25.42629
```

The value 24.54972 is the 10%-percentile of the sampling distribution. To the left of this value are 10% of the distribution. The value 25.42629 is the 90%-percentile of the sampling distribution. To the right of this value are 10% of the distribution. Between these two values are 80% of the sampling distribution.

Solution (to Question 7.1.6): The Normal approximation, which is the conclusion of the Central Limit Theorem substitutes the sampling distribution of the sample average by the Normal distribution with the same expectation and standard deviation. The percentiles are computed with the function "`qnorm`":

```
> qnorm(c(0.1,0.9),mean(X.bar),sd(X.bar))
[1] 24.54817 25.42545
```

Observe that we used the expectation and the standard deviation of the sample average in the function. The resulting interval is $[24.54817, 25.42545]$, which is similar to the interval $[24.54972, 25.42629]$ which was obtained via simulations.

Question 7.2. A subatomic particle hits a linear detector at random locations. The length of the detector is 10 nm and the hits are uniformly distributed. The location of 25 random hits, measured from a specified endpoint of the interval, are marked and the average of the location computed.

1. What is the expectation of the average location?

2. What is the standard deviation of the average location?

3. Use the Central Limit Theorem in order to approximate the probability the average location is in the left-most third of the linear detector.

4. The central region that contains 99% of the distribution of the average is of the form $5 \pm c$. Use the Central Limit Theorem in order to approximate the value of c.

Solution (to Question 7.2.1): Denote by X the distance from the specified endpoint of a random hit. Observe that $X \sim \text{Uniform}(0, 10)$. The 25 hits form a sample $X_1, X_2, \ldots, X_{25}$ from this distribution and the sample average $\bar{X}$ is the average of these random locations. The expectation of the average is equal to the expectation of a single measurement. Since $\text{E}(X) = (a + b)/2 = (0 + 10)/2 = 5$ we get that $\text{E}(\bar{X}) = 5$.

Solution (to Question 7.2.2): The variance of the sample average is equal to the variance of a single measurement, divided by the sample size. The variance of the Uniform distribution is $\text{Var}(X) = (a+b)^2/12 = (10-0)^2/12 = 8.333333$. The standard deviation of the sample average is equal to the standard deviation

of the sample average is equal to the standard deviation of a single measurement, divided by the square root of the sample size. The sample size is $n = 25$. Consequently, the standard deviation of the average is $\sqrt{8.333333/25} = 0.5773503$.

Solution (to Question 7.2.3): The left-most third of the detector is the interval to the left of 10/3. The distribution of the sample average, according to the Central Limit Theorem, is Normal. The probability of being less than 10/3 for the Normal distribution may be computed with the function "`pnorm`":

```
> mu <- 5
> sig <- sqrt(10^2/(12*25))
> pnorm(10/3,mu,sig)
[1] 0.001946209
```

The expectation and the standard deviation of the sample average are used in computation of the probability. The probability is 0.001946209, about 0.2%.

Solution (to Question 7.2.3): The central region in the Normal(μ, σ^2) distribution that contains 99% of the distribution is of the form $\mu \pm$ `qnorm(0.995)` $\cdot \sigma$, where "`qnorm(0.995)`" is the 99.5%-percentile of the Standard Normal distribution. Therefore, $c =$ `qnorm(0.995)` $\cdot \sigma$:

```
> qnorm(0.995)*sig
[1] 1.487156
```

We get that $c = 1.487156$.

7.5 Summary

Glossary

Random Sample: The probabilistic model for the values of a measurements in the sample, before the measurement is taken.

Sampling Distribution: The distribution of a random sample.

Sampling Distribution of a Statistic: A statistic is a function of the data; i.e. a formula applied to the data. The statistic becomes a random variable when the formula is applied to a random sample. The distribution of this random variable, which is inherited from the distribution of the sample, is its sampling distribution.

Sampling Distribution of the Sample Average: The distribution of the sample average, considered as a random variable.

The Law of Large Numbers: A mathematical result regarding the sampling distribution of the sample average. States that the distribution of the average of measurements is highly concentrated in the vicinity of the expectation of a measurement when the sample size is large.

The Central Limit Theorem: A mathematical result regarding the sampling distribution of the sample average. States that the distribution of the average is approximately Normal when the sample size is large.

Discussion in the Forum

Limit theorems in mathematics deal with the convergence of some property to a limit as some indexing parameter goes to infinity. The Law of Large Numbers and the Central Limit Theorem are examples of limit theorems. The property they consider is the sampling distribution of the sample average. The indexing parameter that goes to infinity is the sample size n.

Some people say that the Law of Large Numbers and the Central Limit Theorem are useless for practical purposes. These theorems deal with a sample size that goes to infinity. However, all sample sizes one finds in reality are necessarily finite. What is your opinion?

When forming your answer to this question you may give an example of a situation from your own field of interest in which conclusions of an abstract mathematical theory are used in order to solve a practical problem. Identify the merits and weaknesses of the application of the mathematical theory.

For example, in making statistical inference one frequently needs to make statements regarding the sampling distribution of the sample average. For instant, one may want to identify the central region that contains 95% of the distribution. The Normal distribution is used in the computation. The justification is the Central Limit Theorem.

Summary of Formulas

Expectation of the sample average: $\mathrm{E}(\bar{X}) = \mathrm{E}(X)$

Variance of the sample average: $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$

Chapter 8

Overview and Integration

8.1 Student Learning Objective

This section provides an overview of the concepts and methods that where presented in the first part of the book. We attempt to relate them to each other and put them in prospective. Some problems are provided. The solutions to these problems require combinations of many of the tools that were presented in previous chapters. By the end of this chapter, the student should be able to:

- Have a better understanding of the relation between descriptive statistics, probability, and inferential statistics.

- Distinguish between the different uses of the concept of variability.

- Integrate the tools that were given in the first part of the book in order to solve complex problems.

8.2 An Overview

The purpose of the first part of the book was to introduce the fundamentals of statistics and teach the concepts of probability which are essential for the understanding of the statistical procedures that are used to analyze data. These procedures are presented and discussed in the second part of the book.

Data is typically obtained by selecting a sample from a population and taking measurements on the sample. There are many ways to select a sample, but all methods for such selection should not violate the most important characteristic that a sample should posses, namely that it represents the population it came from. In this book we concentrate on simple random sampling. However, the reader should be aware of the fact that other sampling designs exist and may be more appropriate in specific applications. Given the sampled data, the main concern of the science of statistics is in making inference on the parameter of the population on the basis of the data collected. Such inferences are carried out with the aid of statistics, which are functions of the data.

Data is frequently stored in the format of a data frame, in which columns are the measured variable and the rows are the observations associated with the selected sample. The main types of variables are numeric, either discrete or not,

125

and factors. We learned how one can produce data frames and read data into `R` for further analysis.

Statistics is geared towards dealing with variability. Variability may emerge in different forms and for different reasons. It can be summarized, analyzed and handled with many tools. Frequently, the same tool, or tools that have much resemblance to each other, may be applied in different settings and for different forms of variability. In order not to loose track it is important to understand in each scenario the source and nature of the variability that is being examined.

An important split in term of the source of variability is between descriptive statistics and probability. Descriptive statistics examines the distribution of data. The frame of reference is the data itself. Plots, such as the bar plots, histograms and box plot; tables, such as the frequency and relative frequency as well as the cumulative relative frequency; and numerical summaries, such as the mean, median and standard deviation, can all serve in order to understand the distribution of the given data set.

In probability, on the other hand, the frame of reference is not the data at hand but, instead, it is all data sets that could have been sampled (the sample space of the sampling distribution). One may use similar plots, tables, and numerical summaries in order to analyze the distribution of functions of the sample (statistics), but the meaning of the analysis is different. As a matter of fact, the relevance of the probabilistic analysis to the data actually sampled is indirect. The given sample is only one realization within the sample space among all possible realizations. In the probabilistic context there is no special role to the observed realization in comparison to all other potential realizations.

The fact that the relation between probabilistic variability and the observed data is not direct does not make the relation unimportant. On the contrary, this indirect relation is the basis for making statistical inference. In statistical inference the characteristics of the data may be used in order to extrapolate from the sampled data to the entire population. Probabilistic description of the distribution of the sample is then used in order to assess the reliability of the extrapolation. For example, one may try to estimate the value of population parameters, such as the population average and the population standard deviation, on the basis of the parallel characteristics of the data. The variability of the sampling distribution is used in order to quantify the accuracy of this estimation. (See Example 5 below.)

Statistics, like many other empirically driven forms of science, uses theoretical modeling for assessing and interpreting observational data. In statistics this modeling component usually takes the form of a probabilistic model for the measurements as random variables. In the first part of this book we have encountered several such models. The model of simple sampling assumed that each subset of a given size from the population has equal probability to be selected as the sample. Other, more structured models, assumed a specific form to the distribution of the measurements. The examples we considered were the Binomial, the Poisson, the Uniform, the Exponential and the Normal distributions. Many more models may be found in the literature and may be applied when appropriate. Some of these other models have `R` functions that can be used in order to compute the distribution and produce simulations.

A statistic is a function of sampled data that is used for making statistical inference. When a statistic, such as the average, is computed on a random sample then the outcome, from a probabilistic point of view, is a random vari-

able. The distribution of this random variable depends on the distribution of the measurements that form the sample but is not identical to that distribution. Hence, for example, the distribution of an average of a sample from the Uniform distribution does not follow the Uniform distribution. In general, the relation between the distribution of a measurement and the distribution of a statistic computed from a sample that is generated from that distribution may be complex. Luckily, in the case of the sample average the relation is rather simple, at least for samples that are large enough.

The Central Limit Theorem provides an approximation of the distribution of the sample average that typically improves with the increase in sample size. The expectation of the sample average is equal to the expectation of a single measurement and the variance is equal to the variance of a single measurement, divided by the sample size. The Central Limit Theorem adds to this observation the statement that the distribution of the sample average may be approximated by the Normal distribution (with the same expectation and standard deviation as those of the sample average). This approximation is valid for practically any distribution of the measurement. The conclusion is, at least in the case of the sample average, that the distribution of the statistic depends on the underlying distribution of the measurements only through their expectation and variance but not through other characteristics of the distribution.

The conclusion of the theorem extends to quantities proportional to the sample average. Therefore, since the sum of the sample is obtained by multiplying the sample average by the sample size n, we get that the theorem can be used in order to approximate the distribution of sums. As a matter of fact, the theorem may be generalized much further. For example, it may be shown to hold for a smooth function of the sample average, thereby increasing the applicability of the theorem and its importance.

In the next section we will solve some practical problems. In order to solve these problems you are required to be familiar with the concepts and tools that were introduced throughout the first part of the book. Hence, we strongly recommend that you read again and review all the chapters of the book that preceded this one before moving on to the next section.

8.3   Integrated Applications

The main message of the Central Limit Theorem is that for the sample average we may compute probabilities based on the Normal distribution and obtain reasonable approximations, provided that the sample size is not too small. All we need to figure out for the computations are the expectation and variance of the underlying measurement. Otherwise, the exact distribution of that measurement is irrelevant. Let us demonstrate the applicability of the Central Limit Theorem in two examples.

8.3.1   Example 1

A study involving stress is done on a college campus among the students. The stress scores follow a (continuous) Uniform distribution with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:

1. The probability that the average stress score for the 75 students is less than 2.

2. The 90th percentile for the average stress score for the 75 students.

3. The probability that the total of the 75 stress scores is less than 200.

4. The 90th percentile for the total stress score for the 75 students.

Solution:

Denote by X the stress score of a random student. We are given that $X \sim$ Uniform$(1, 5)$. We use the formulas $\mathrm{E}(X) = (a+b)/2$ and $\mathrm{Var}(X) = (b-a)^2/12$ in order to obtain the expectation and variance of a single observation and then we use the relations $\mathrm{E}(\bar{X}) = \mathrm{E}(X)$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$ to translated these results to the expectation and variance of the sample average:

```
> a <- 1
> b <- 5
> n <- 75
> mu.bar <- (a+b)/2
> sig.bar <- sqrt((b-a)^2/(12*n))
> mu.bar
[1] 3
> sig.bar
[1] 0.1333333
```

These equations come from the uniform distribution section of Chapter 5, and are also shown above.

After obtaining the expectation and the variance of the sample average we can forget about the Uniform distribution and proceed only with the R functions that are related to the Normal distribution. By the Central Limit Theorem we get that the distribution of the sample average is approximately Normal(μ, σ^2), with $\mu = $ `mu.bar` and $\sigma = $ `sig.bar`.

In the Question 1.1 we are asked to find the value of the cumulative distribution function of the sample average at $x = 2$:

```
> pnorm(2,mu.bar,sig.bar)
[1] 3.190892e-14
```

The goal of Question 1.2 is to identify the 90%-percentile of the sample average:

```
> qnorm(0.9,mu.bar,sig.bar)
[1] 3.170874
```

In other words, to identify the 90th percentile, use the qnorm() function.

The sample average is equal to the total sum divided by the number of observations, $n = 75$ in this example. The total sum is less than 200 if, and only if the average is less than $200/n$. Therefore, for Question 1.3:

```
> pnorm(200/n,mu.bar,sig.bar)
[1] 0.006209665
```

Finally, if 90% of the distribution of the average is less that 3.170874 then 90% of the distribution of the total sum is less than $3.170874\,n$. In Question 1.4 we get:

```
> n*qnorm(0.9,mu.bar,sig.bar)
[1] 237.8155
```

8.3.2   Example 2

Consider again the same stress study that was described in Example 1 and answer the same questions. However, this time assume that the stress score may obtain only the values 1, 2, 3, 4 or 5, with the same likelihood for obtaining each of the values.

Solution:

Denote again by X the stress score of a random student. The modified distribution states that the sample space of X are the integers $\{1, 2, 3, 4, 5\}$, with equal probability for each value. Since the probabilities must sum to 1 we get that $P(X = x) = 1/5$, for all x in the sample space. In principle we may repeat the steps of the solution of previous example, substituting the expectation and standard deviation of the continuous measurement by the discrete counterpart:

```
> x <- 1:5
> p <- rep(1/5,5)
> n <- 75
> mu.X <- sum(x*p)
> sig.X <- sum((x-mu.X)^2*p)
> mu.bar <- mu.X
> sig.bar <- sqrt(sig.X/n)
> mu.bar
[1] 3
> sig.bar
[1] 0.1632993
```

Notice that the expectation of the sample average is the same as before but the standard deviation is somewhat larger due to the larger variance in the distribution of a single response.

We may apply the Central Limit Theorem again in order to conclude that distribution of the average is approximately Normal(μ, σ^2), with $\mu = $ `mu.bar` as before and for the new $\sigma = $ `sig.bar`.

For Question 2.1 we compote that the cumulative distribution function of the sample average at $x = 2$ is approximately equal:

```
> pnorm(2,mu.bar,sig.bar)
[1] 4.570649e-10
```

and the 90%-percentile is:

```
> qnorm(0.9,mu.bar,sig.bar)
[1] 3.209276
```

which produces the answer to Question 2.2.

Similarly to the solution of Question 1.3 we may conclude that the total sum is less than 200 if, and only if, the average is less than $200/n$. Therefore, for Question 2.3:

```
> pnorm(200/n,mu.bar,sig.bar)
[1] 0.02061342
```

Observe that in the current version of the question we have the score is integer-valued. Clearly, the sum of scores is also integer valued. Hence we may choose to apply the continuity correction for the Normal approximation whereby we approximate the probability that the sum is less than 200 (i.e. is less than or equal to 199) by the probability that a Normal random variable is less than or equal to 199.5. Translating this event back to the scale of the average we get the approximation[1]:

```
> pnorm(199.5/n,mu.bar,sig.bar)
[1] 0.01866821
```

Finally, if 90% of the distribution of the average is less that 3.170874 then 90% of the distribution of the total sum is less than $3.170874n$. Therefore:

```
> n*qnorm(0.9,mu.bar,sig.bar)
[1] 240.6957
```

or, after rounding to the nearest integer we get for Question 2.4 the answer 241.

8.3.3   Example 3

Suppose that a market research analyst for a cellular phone company conducts a study of their customers who exceed the time allowance included on their basic cellular phone contract. The analyst finds that for those customers who exceed the time included in their basic contract, the excess time used follows an exponential distribution with a mean of 22 minutes. Consider a random sample of 80 customers and find

1. The probability that the average excess time used by the 80 customers in the sample is longer than 20 minutes.

2. The 95th percentile for the average excess time for samples of 80 customers who exceed their basic contract time allowances.

Solution:

Let X be the excess time for customers who exceed the time included in their basic contract. We are told that $X \sim \text{Exponential}(\lambda)$. For the Exponential distribution $\text{E}(X) = 1/\lambda$. Hence, given that $\text{E}(X) = 22$ we can conclude that $\lambda = 1/22$. For the Exponential we also have that $\text{Var}(X) = 1/\lambda^2$. Therefore:

```
> lam <- 1/22
> n <- 80
> mu.bar <- 1/lam
> sig.bar <- sqrt(1/(lam^2*n))
> mu.bar
[1] 22
> sig.bar
[1] 2.459675
```

---

[1]As a matter of fact, the continuity correction could have been applied in the previous two sections as well, since the sample average has a discrete distribution.

Like before, we can forget at this stage about the Exponential distribution and refer henceforth to the Normal Distribution. In Question 2.1 we are asked to compute the probability above $x = 20$. The total probability is 1. Hence, the required probability is the difference between 1 and the probability of being less or equal to $x = 20$:

```
> 1-pnorm(20,mu.bar,sig.bar)
[1] 0.7919241
```

The goal in Question 2.2 is to find the 95%-percentile of the sample average:

```
> qnorm(0.95,mu.bar,sig.bar)
[1] 26.04580
```

8.3.4 Example 4

A beverage company produces cans that are supposed to contain 16 ounces of beverage. Under normal production conditions the expected amount of beverage in each can is 16.0 ounces, with a standard deviation of 0.10 ounces.

As a quality control measure, each hour the QA department samples 50 cans from the production during the previous hour and measures the content in each of the cans. If the average content of the 50 cans is below a control threshold then production is stopped and the can filling machine is re-calibrated.

1. Compute the probability that the amount of beverage in a random can is below 15.95.

2. Compute the probability that the amount of beverage in a sample average of 50 cans is below 15.95.

3. Find a threshold with the property that the probability of stopping the machine in a given hour is 5% when, in fact, the production conditions are normal.

4. Consider the data in the file "QC.csv"[2]. It contains measurement results of 8 hours. Assume that we apply the threshold that was obtained in Question 4.3. At the end of which of the hours the filling machine needed re-calibration?

5. Based on the data in the file "QC.csv", which of the hours contains measurements which are suspected outliers in comparison to the other measurements conducted during that hour?

Solution

The only information we have on the distribution of each measurement is its expectation (16.0 ounces under normal conditions) and its standard deviation (0.10, under the same condition). We do not know, from the information provided in the question, the actual distribution of a measurement. (The fact that the production conditions are normal does not imply that the distribution

---

[2]URL for the file: http://pluto.huji.ac.il/~msby/StatThink/Datasets/QC.csv
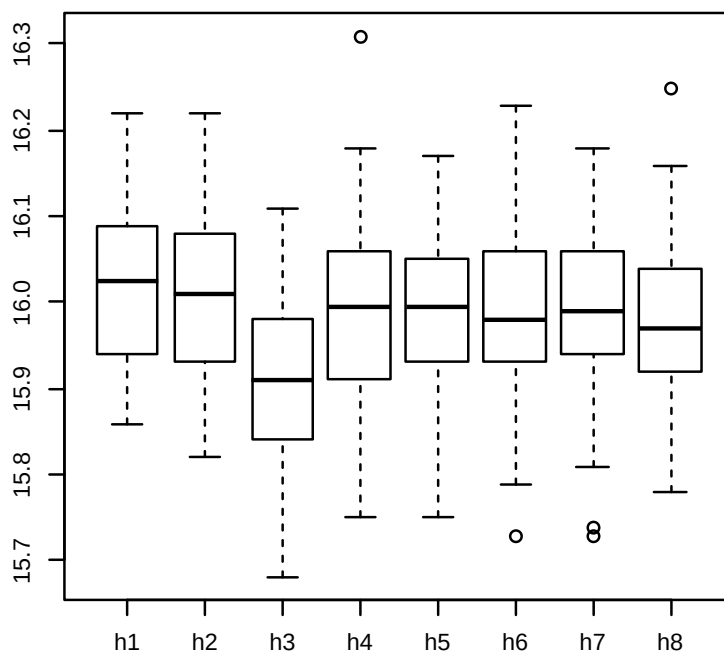
Figure 8.1: Box Plots

of the measurement in the Normal distribution!) Hence, the correct answer to Question 4.1 is that there is not enough information to calculate the probability.

When we deal with the sample average, on the other hand, we may apply the Central Limit Theorem in order to obtain at least an approximation of the probability. Observe that the expectation of the sample average is 16.0 ounces and the standard deviation is $0.1/\sqrt{50}$. The distribution of the average is approximately the Normal distribution:

```
> pnorm(15.95,16,0.1/sqrt(50))
[1] 0.000203476
```

Hence, we get that the probability of the average being less than 15.95 ounces is (approximately) 0.0002, which is a solution to Question 4.2.

In order to solve Question 4.3 we may apply the function "qnorm" in order to compute the 5%-percentile of the distribution of the average:

```
> qnorm(0.05,16,0.1/sqrt(50))
[1] 15.97674
```

Consider the data in the file "`QC.csv`". Let us read the data into a data frame by the by the name "`QC`" and apply the function "`summary`" to obtain an overview of the content of the file:

```
> QC <- read.csv("QC.csv")
> summary(QC)
      h1               h2               h3               h4
 Min.   :15.86   Min.   :15.82   Min.   :15.68   Min.   :15.75
 1st Qu.:15.94   1st Qu.:15.93   1st Qu.:15.84   1st Qu.:15.91
 Median :16.02   Median :16.01   Median :15.91   Median :15.99
 Mean   :16.02   Mean   :16.01   Mean   :15.91   Mean   :15.99
 3rd Qu.:16.09   3rd Qu.:16.08   3rd Qu.:15.98   3rd Qu.:16.06
 Max.   :16.22   Max.   :16.22   Max.   :16.11   Max.   :16.31
      h5               h6               h7               h8
 Min.   :15.75   Min.   :15.73   Min.   :15.73   Min.   :15.78
 1st Qu.:15.93   1st Qu.:15.93   1st Qu.:15.94   1st Qu.:15.92
 Median :15.99   Median :15.98   Median :15.99   Median :15.97
 Mean   :15.99   Mean   :15.98   Mean   :15.99   Mean   :15.97
 3rd Qu.:16.05   3rd Qu.:16.06   3rd Qu.:16.05   3rd Qu.:16.04
 Max.   :16.17   Max.   :16.23   Max.   :16.18   Max.   :16.25
```

Observe that the file contains 8 quantitative variables that are given the names h1, ..., h8. Each of these variables contains the 50 measurements conducted in the given hour.

Observe that the mean is computed as part of the summary. The threshold that we apply to monitor the filling machine is 15.97674. Clearly, the average of the measurements at the third hour "h3" is below the threshold. Not enough significance digits of the average of the 8th hour are presented to be able to say whether the average is below or above the threshold. A more accurate presentation of the computed mean is obtained by the application of the function "`mean`" directly to the data:

```
> mean(QC$h8)
[1] 15.9736
```

Now we can see that the average is below the threshold. Hence, the machine required re-calibration after the 3rd and the 8th hours, which is the answer to Question 4.4.

In Chapter 3 it was proposed to use box plots in order to identify points that are suspected to be outliers. We can use the expression "`boxplot(QC$h1)`" in order to obtain the box plot of the data of the first hour and go through the names of the variable one by one in order to screen all variable. Alternatively, we may apply the function "`boxplot`" directly to the data frame "`QC`" and get a plot with box plots of all the variables in the data frame plotted side by side (see Figure 8.1):

```
> boxplot(QC)
```

Examining the plots we may see that evidence for the existence of outliers can be spotted on the 4th, 6th, 7th, and 8th hours, providing an answer to Question 4.5

8.3.5   Example 5

A measurement follows the Uniform$(0, b)$, for an unknown value of b. Two statisticians propose two distinct ways to estimate the unknown quantity b with the aid of a sample of size $n = 100$. Statistician A proposes to use twice the sample average $(2\bar{X})$ as an estimate. Statistician B proposes to use the largest observation instead.

The motivation for the proposal made by Statistician A is that the expectation of the measurement is equal to $E(X) = b/2$. A reasonable way to estimate the expectation is to use the sample average $\bar{X}$. Thereby, a reasonable way to estimate b, twice the expectation, is to use $2\bar{X}$. A motivation for the proposal made by Statistician B is that although the largest observation is indeed smaller that b, still it may not be much smaller than that value.

In order to choose between the two options they agreed to prefer the statistic that tends to have values that are closer to b. (with respect to the sampling distribution). They also agreed to compute the expectation and variance of each statistic. The performance of a statistic is evaluated using the *mean square error* (MSE), which is defined as the sum of the variance and the squared difference between the expectation and b. Namely, if T is the statistic (either the one proposed by Statistician A or Statistician B) then

$$MSE = \mathrm{Var}(T) + (E(T) - b)^2 \ .$$

A smaller mean square error corresponds to a better, more accurate, statistic.

1. Assume that the actual value of b is 10 ($b = 10$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician A.

2. Assume that the actual value of b is 10 ($b = 10$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician B. (Hint: the maximal value of a sequence can be computed with the function "`max`".)

3. Assume that the actual value of b is 13.7 ($b = 13.7$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician A.

4. Assume that the actual value of b is 13.7 ($b = 13.7$). Use simulations to compute the expectation, the variance and the MSE of the statistic proposed by Statistician B. (Hint: the maximal value of a sequence can be computed with the function "`max`".)

5. Based on the results in Questions 5.1–4, which of the two statistics seems to be preferable?

Solution

In Questions 5.1 and 5.2 we take the value of b to be equal to 10. Consequently, the distribution of a measurement is Uniform$(0, 10)$. In order to generate the sampling distributions we produce two sequences, "`A`" and "`B`", both of length 100,000, with the evaluations of the statistics:

```
> A <- rep(0,10^5)
> B <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- runif(100,0,10)
+     A[i] <- 2*mean(X.samp)
+     B[i] <- max(X.samp)
+ }
```

Observe that in each iteration of the "for" loop a sample of size $n = 100$ from the Uniform$(0, 10)$ distribution is generated. The statistic proposed by Statistician A ("2*mean(X.samp)") is computed and stored in sequence "A" and the statistic proposed by Statistician B ("max(X.samp)") is computed and stored in sequence "B".

Consider the statistic proposed by Statistician A:

```
> mean(A)
[1] 9.99772
> var(A)
[1] 0.3341673
> var(A) + (mean(A)-10)^2
[1] 0.3341725
```

The expectation of the statistic is 9.99772 and the variance is 0.3341673. Consequently, we get that the mean square error is equal to

$$0.3341673 + (9.99772 - 10)^2 = 0.3341725 \ .$$

Next, deal with the statistic proposed by Statistician B:

```
> mean(B)
[1] 9.901259
> var(B)
[1] 0.00950006
> var(B) + (mean(B)-10)^2
[1] 0.01924989
```

The expectation of the statistic is 9.901259 and the variance is 0.00950006. Consequently, we get that the mean square error is equal to

$$0.00950006 + (9.901259 - 10)^2 = 0.01924989 \ .$$

Observe that the mean square error of the statistic proposed by Statistician B is smaller.

For Questions 5.3 and 5.4 we run the same type of simulations. All we change is the value of b (from 10 to 13.7):

```
> A <- rep(0,10^5)
> B <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+     X.samp <- runif(100,0,13.7)
+     A[i] <- 2*mean(X.samp)
+     B[i] <- max(X.samp)
+ }
```