Quantitative Methods in Economics

Ryan T. Godwin

Visit https://ryantgodwin.com/quantecon/ for the most up to date version of this book.

This book is completely free to use. Please contact me at ryan.godwin@umanitoba.ca if you've used this book for your course, and for any comments, corrections, or suggestions.

Winnipeg, Manitoba, Canada
ISBN: ZZZ

Contents

1. Introduction

1.1 About this Book

This book is intended for a second year undergraduate course in an Economics program. It is a short text, focusing on specific needs. Most of the material in the book should be covered in a single semester course.

1.2 Quantitative Methods

Quantitative methods encompass the collection of data, and what is done with it. It includes rearranging, summarizing, and visualizing data; calculating statistics; describing relationships between variables using scatterplots and correlation coefficients; conducting hypothesis tests and calculating confidence intervals; and estimating models using least-squares.

There is much overlap between statistics, econometrics, and quantitative methods. This book focusses less on theory than a statistics or econometrics book would, and more on explaining how to accomplish methods in practice. However, this book overlaps quite a bit with introductory statistics and econometrics.

The ultimate goal is to build the tools necessary to begin analyzing *causal* effects. We are careful throughout the book to always remember that the methods discussed can never confirm *causation*.

1.3 Objectives

Some objectives of this text are the following:

- Explore and describe data used to inform decisions in economics.
- Review and expand basic concepts of probability and random variables.
- Draw conclusions about a population or process from sample data.
- Model a response based on an explanatory variable.
- Perform quantitative analysis using R.

1.4 Format of this Book

Definitions, quotations, examples, and R code are formatted separately from the main text.

Definitions in the text. Important definitions and points will appear in these boxes.

> "How do I know when something is a quote, or an important question?" asked the student.

Example 1.1 This is an example of the examples you will see in this book. They will appear in these boxes.

```
print("R Code will be displayed in these boxes.")
```

```
[1] "R Code will be displayed in these boxes."
```

The upper box contains the input, and the lower box contains the output.

1.5 Acknowledgements

Janelle Mann for arranging funding for the book, for help with outline, content, and edits. University of Manitoba for providing financial support. Cover images and chapter heading images produced by NASA and the Space Telescope Science Institute (STScI). Statistics performed using R and RStudio. Figures produced using R and Inkscape.

2. The R Programming Language

2.1 What is R?

Although R is a programming language, it is unlike most others. It is designed to analyse data. It isn't too difficult to learn, and is extremely popular. R has the advantage that it is free and open-source, and that thousands of users have contributed "add-on" packages that are readily downloadable by anyone.

R is found in all areas of academia that encounter data, and in many private and public organizations. R is great for any job or task that uses data.

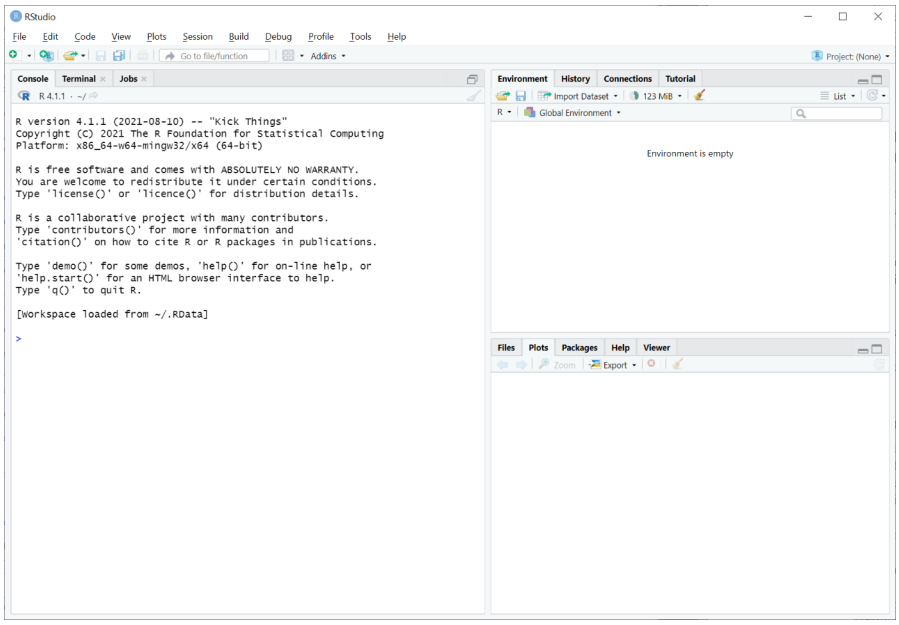
2.2 Where to get R

In this book, we will use R and RStudio. Both are free and open-source. Download and install R first: https://cran.r-project.org/bin/windows/base/ (for Windows) or https://cran.r-project.org/bin/macosx/ (for Mac). Then, download and install RStudio from https://www.rstudio.com/products/rstudio/download/.

2.3 Getting started with RStudio

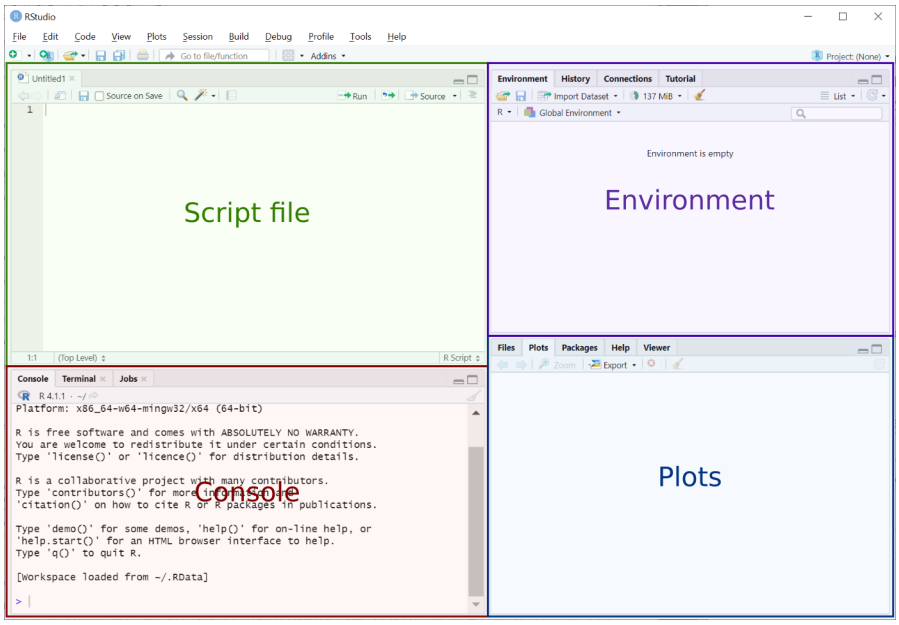
2.3.1 Open RStudio

Search your computer for RStudio.exe and open the application. It should look something like this:



2.3.2 Create a "script" file

A script file is a file where you can type and save your R computer code. To open a script file, click on "File", "New File", "R Script".



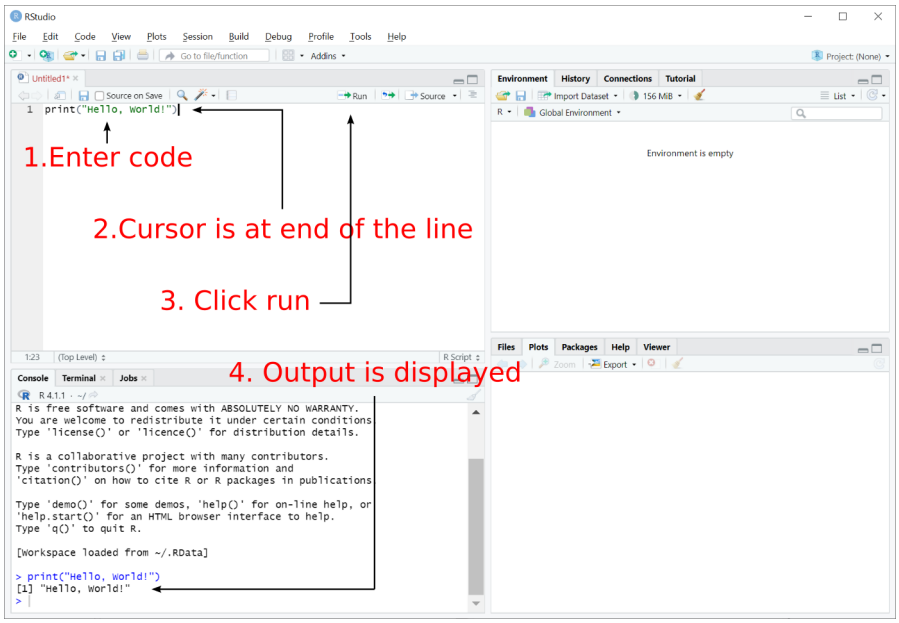
- In the top left is your Script file. R commands can be run from the R Script file, and saved at any time.
- In the bottom left is the Console window. Output is displayed here. R commands can be run from the Console, but not saved.
- In the top right is the Environment. Data and variables will be visible here.
- The bottom right will display graphics (e.g. histograms and scatterplots).

2.3.3 Running R code

Copy and paste the following R code into the script window:

```
print("Hello, World!")
```

Run the code by highlighting it, or making sure the cursor is active at the end of the line, and clicking "Run" (you can also press `Ctrl + Enter` on PC or `Cmd + Return` on Mac).



Often we will display R output in boxes. The output from your program is reproduced in the box below:

```
[1] "Hello, World!"
```

2.4 Use R as a calculator

R's arithmetic operators include:

Operator	Function
+	addition
-	subtraction
*	multiplication
/	division
^	exponentiation

Example 2.1 — Arithmetic in R. Use R to perform the following arithmetic operations:

1. $3 + 5$

```
3 + 5
```

```
[1] 8
```

2. $12 - 4$

```
12 - 4
```

```
[1] 8
```

3. 2×13

```
2 * 13
```

```
[1] 26
```

4. $16/4$

```
16 / 4
```

```
[1] 4
```

5. 2^8

```
2 ^ 8
```

```
[1] 256
```

6. $\frac{10+6}{2}$

```
(10 + 6) / 2
```

```
[1] 8
```

2.5 Create an object

You can create objects in R. Objects can be vectors, matrices, character strings, data frames, scalars etc. Create two different scalars. Give them any name you like, but object names cannot start with a number and cannot include certain characters like "!":

```
a <- 3
b <- 5
```

We have created two new objects called `a` and `b`, and have assigned them values using the assignment operator `<-` (the "less than" symbol followed by the "minus" symbol). Notice that `a` and `b` pop up in the top-right of your screen (the Environment window). We can now refer to these objects by name:

```
a * b
```

```
[1] 15
```

produces the output `15`. To create a vector in R we use the "combine" function, `c()`:

```
myvector <- c(1, 2, 4, 6, 7)
```

Notice that after creating it, the `myvector` object appears in the top-right Environment window. `myvector` is just a list of numbers:

$$\texttt{myvector} = \begin{bmatrix} 1 \\ 2 \\ 4 \\ 6 \\ 7 \end{bmatrix}$$

2.6 Simple functions in R

Function (programming). Similar to a function in mathematics, a function in computer programming takes an input and produces an output.

Table 2.1: Simple R functions.

Function
`sum()`
`mean()`
`var()`
`summary()`

A "function" in computer coding is much like a function in mathematics; it takes an input, performs an operation, and then provides an output. In R, we type the name of the function and then type the input inside of parentheses: `function.name(input)`. After we click the "Run" button, we get the output. The function could be as simple as adding up two numbers, estimating a very complicated statistical model, or producing a graph. There are thousands of functions in R, and you can even make your own! We'll try a few simple ones to begin with:

Example 2.2 To add up all of the numbers in `myvector` we would run:

```
sum(myvector)
```

```
[1] 20
```

which provides the output `20`. We have asked the computer to add up an object by calling the function `sum()`, and putting the name of the object `myvector` inside of the parentheses. Try all of the functions in Table 2.1 on `myvector`.

2.7 Logical operators

Logical operators. Logical operators can check which values of a variable satisfy a certain condition, allowing us to create "subsets" of data.

Logical operators are used to determine whether something is `TRUE` or `FALSE`. Some logical operators are:

Operator	Function
>	greater than
==	equal to
<	less than
>=	greater than or equal to
<=	less than or equal to
!=	not equal to

Logical operators are useful for creating "subsamples" or "subsets" from our data. Using logical operators, we can calculate statistics separately for ethnicities, treatment group vs. control group, developed vs. developing countries, etc. (we will see how to do this later). For now, let's try some simple logical operations. Try entering and running each of the following lines of code one by one:

```
8 > 4
```

```
[1] TRUE
```

```
b == 6
```

```
[1] FALSE
```

```
b > 2
```

```
[1] TRUE
```

To check to see which elements in `myvector` are greater than 3 we use:

```
myvector > 3
```

```
[1] FALSE FALSE TRUE TRUE TRUE
```

2.7.1 Multiple logical operators

Sometimes we would like to create subsets in our data based on multiple conditions or characteristics. For example, we might want to study a subset of our data consisting of only single or widowed women with 1 child or more. The "and" / "or" operators are useful in these situations:

Operator	Function
&	"and"
|	"or"

For example, the following line of code:

```
myvector > 3 & myvector < 7
```

```
[1] FALSE FALSE TRUE TRUE FALSE
```

checks to see whether each element in `myvector` is greater than 3 *and* less than 7.

2.8 Loading data into R

CSV format. A common and simple format for data files. These data files have the extension `.csv` and can be opened in applications like Excel, and in most econometrics and statistical software packages.

There are several ways to load data into R. We cover three of them here. In this book, we work mostly with the *comma-separated values* file format (CSV format).

2.8.1 Directly from the internet

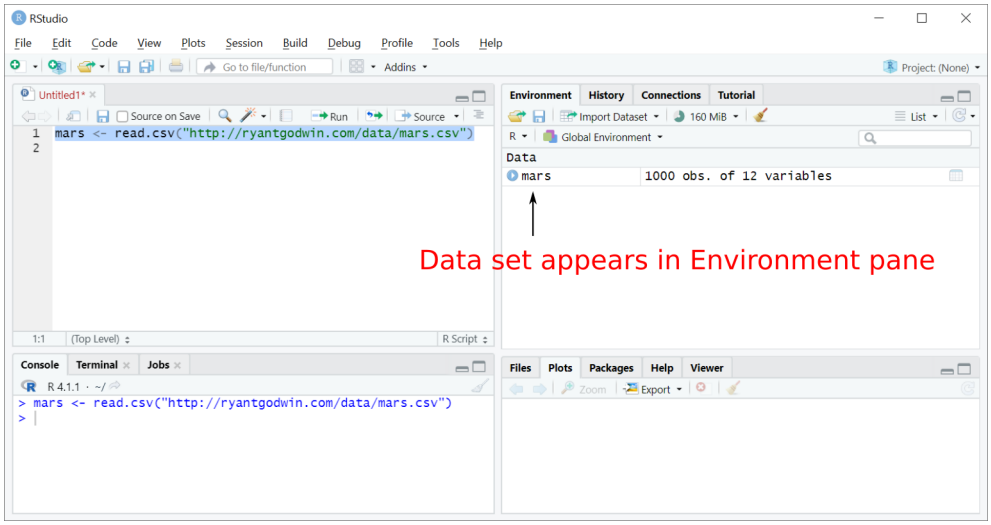
We can use the R code:

```
mydata <- read.csv("file location.csv")
```

We need to replace `file location` with the actual location of the file, either on the internet or on your computer. We can also replace the name of the data set `mydata` with any name we like. For example, to load data directly from the internet into R, try the following:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

After running the above line of code, you should see the data set appear in the top-right of RStudio (the environment pane).



2.8.2 From a location on your computer

After saving a `.csv` file to your computer, you can use the `read.csv()` command to load the file from its location on your computer. For example:

```
mars <- read.csv("c:/data/mars.csv")
```

loads a file from the location `c:/data/`.

2.8.3 `file.choose()`

Using the `file.choose()` command will prompt you to select the file using file explorer:

```
mars <- read.csv(file.choose())
```

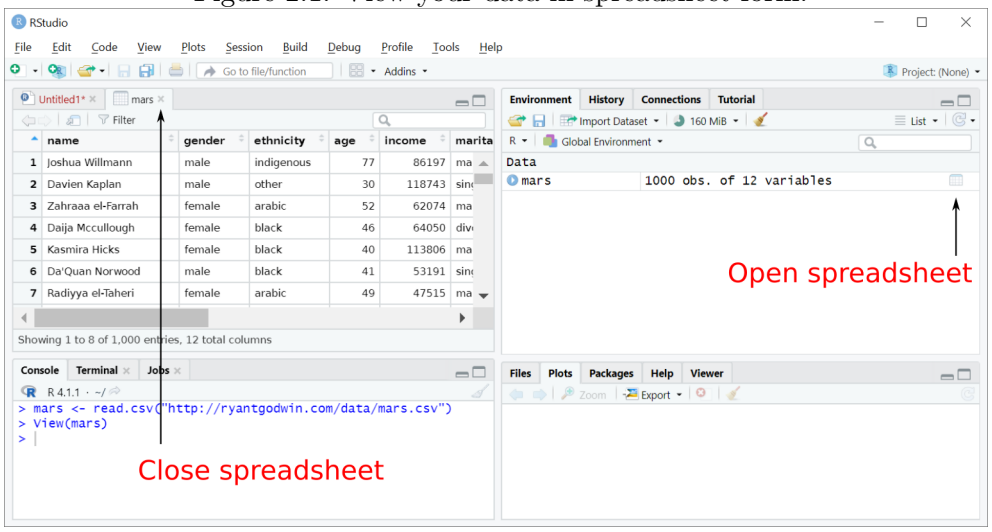
2.9 View your data in spreadsheet form

Click on the spreadsheet icon next to your `mars` data set, or run the following command:

```
View(mars)
```

Note the uppercase `V` (R is *case sensitive*). This command allows you to view your data in spreadsheet form. See Figure 2.1.

Figure 2.1: View your data in spreadsheet form.



2.10 Scientific notation in R output

R's default is to report numbers with many digits in *scientific notation*. For example, the number 1 million (1000000) is written in scientific notation as 1×10^6. We can see this notation in R using:

```
my.number <- 1000000
my.number
```

```
[1] 1e+06
```

The `e` in the output signifies an exponent to base 10. Similarly, the number 0.0000001 would be output as `1e-06` (note the negative sign on the exponent).

The scientific notation can be difficult to read at times, and you can suppress this notation using `options(scipen=999)`. Try this option, and print out `my.number` again:

```
options(scipen=999)
my.number
```

```
[1] 1000000
```

3. Collecting Data

Quantitative and statistical studies are often trying to provide answers. The key feature that differentiates quantitative analysis from other methods that attempt to provide answers, is the use of data collected through experimentation, or by simply observing what happens.

Where does data come from? In this chapter, we discuss various sources of data, and issues involved with collecting or obtaining data. Not all data are created equally. Some are better at answering specific questions than others, and some may not be useful at all!

Several aspects of the data collection process can lead to *sampling bias*. In this chapter, we will discuss situations in which the data set might not represent the population, in ways that create misleading statistical conclusions.

3.1 Data sources

The usefulness of data can depend on how it was created or collected. Some data sets are much better than others. For example, *experimental data* is almost always better than *observational data*. *Anecdotal data* is not very useful for quantitative and statistical analyses.

Data is often used to infer some property of the *population*. In most cases however, it is not feasible to collect information on every member of the population, and so a *sample* must instead be used. How the sample is determined can also affect the quality of the data.

In this section, we discuss various data sources. Then, we will define the terms *sample* and *population*, before talking about *sampling bias* and the importance of *simple random samples*. A lot of key words have just been used! We will sort out what all of them mean in the coming sections.

3.1.1 Anecdotal evidence

Anecdotal evidence. Anecdotal evidence is based on individual experiences.

Other people's explanations or accounts of their experiences, when used to form an opinion or come to a conclusion or answer some interesting question, is called *anecdotal evidence*. Anecdotal evidence is very important for many areas of research such as history, but is not very useful for economists and other wielders of quantitative methods.

Anecdotal evidence is not very useful in statistical analysis because the *sample size* is very low (typically 1 or 2). Later on, we will learn the importance of having a sample size as large as possible, and that when our sample is small we cannot be very confident about the conclusions that we draw.

An anecdote for example, is:

> "My friend Homer has an anti-tiger rock, and has never been attacked by a tiger."

Another anecdote:

> "My friend Pooh doesn't have an anti-tiger rock, and is continuously attacked by a tiger."

From these anecdotes, one might be tempted to draw the conclusion that anti-tiger rocks prevent tiger attacks. From the standpoint of the statistician, this information is not very valuable because it only contains 2 data points. The data set from these anecdotes might look something like Table 3.1.

Table 3.1: Data set from anecdotal evidence.

name	**anti-tiger rock?**	**tiger attack?**
Homer	Yes	No
Pooh	No	Yes

There is a perfect negative correlation (-1)[1] between the two variables in the data set. There is no way for a statistician to *disprove* the notion that rocks prevent tiger attacks using this data set. However, if more information was collected on tiger attacks (providing a bigger data set), we would likely see that there is no relationship between rocks and tiger attacks at all!

Not only are sample sizes typically too small, anecdotal evidence may only exist because the personal experiences are unusual or memorable in some way. In this chapter we will talk about *random sampling* from a *population*. Anecdotes might just contain the most extreme cases in a population, and so might not be very representative of the population itself.

3.1.2 Experimental data

Experimental data is often considered the *best* kind of data for estimating *causal effects*. In an experiment, the researcher can randomly assign individuals to a *treatment group* or a *control group*. Random assignment is an extremely powerful tool in statistical analyses of causation.

"Treatment" can be defined quite broadly. Traditionally it meant treatment with a drug or medical procedure, but the concept has expanded to include education, labour training programs, health insurance, or anything which my effect an *outcome* of interest.

[1] Assigning a numerical value for "Yes", and a different numerical value for "No" (1 and 0 are natural choices) would give a correlation of -1 between the variables `rock` and `tiger.attack`. We will talk about correlation in Section 5.10.

The treatment group are those individuals that receive the "treatment"; the control group does not receive treatment (in a medical study they might receive a "placebo"). The effect of the treatment can sometimes be determined by comparing the *outcomes* of the two groups. An outcome is something particular that is thought to be influenced by the treatment. See Table 3.2 for examples of treatments and outcomes.

Table 3.2: Examples of treatments and outcomes.

Treatment	**Outcome**
LIPITOR	cholesterol levels
university education	wages
carbon tax	CO_2 emissions
universal health care	life expectancy
universal basic income	unemployment rate

The outcomes in Table 3.2 could differ depending on the researcher. Rather than wages, a criminologist might be interested in the effect of education (treatment) on crime rates (outcome). Governments may wonder whether adopting universal health care (treatment) might reduce health care costs (outcome). The labelling of things as treatment or outcome is part of a framework that allows researchers to try to figure out *cause* and *effect*.

Lurking variable. A lurking variable is an unobserved variable which influences both the probability of an individual receiving treatment, and the outcome associated with the treatment.

Randomly assigning individuals to treatment or control groups prevents individuals from *choosing* whether they receive treatment or not. Random assignment is important because the *choice* to get treated might affect the outcome. Sometimes this problem is expressed in terms of a *lurking variable*. A lurking variable is unobserved and influences both the decision for an individual to seek treatment, and the outcome from the treatment itself.

With random assignment, the lurking variables no longer have power to influence the data that we observe. That is, other factors that influence the outcome (besides the treatment that we are interested in), do not matter on average, in an experiment with random assignment. Below we consider an example to try to solidify some of the terminology we have used.

Example 3.1 How could we use an experiment to determine the value (in terms of wages), of a university education? We could randomly select 10 individuals, and then randomly choose 5 to receive a free university education (this is the treatment group). The other 5 are not allowed to receive an education. 20 years later, we measure the wages of the individuals (wage is the *outcome*). The experimental data is displayed in the table below.

name	education	wage (in thousands)
Raven	university	101
Gary	university	70
Roberto	high school	59
Amanda	university	144
Justin	high school	135
Hadeel	high school	126
Mudrika	university	124
Dewarren	high school	69
Jacob	university	98
Melinda	high school	80

One way to quantify the effect of the treatment is to calculate the sample average outcome between the *treatment group* (university) and the *control group* (high school). We will talk about the sample average in depth in a later chapter, but you should be able to calculate this difference now. The sample average wage for the group with a university education is:

$$\bar{wage}_{university} = \frac{101 + 70 + 144 + 124 + 89}{5} = 105.6$$

Similarly, the sample average wage for the group without a university education is:

$$\bar{wage}_{highschool} = \frac{59 + 135 + 126 + 69 + 80}{5} = 93.8$$

Taking the difference between these two sample averages ($105.6 - 93.8 = 11.8$) might lead us to conclude that one of the effects of an education is to increase wages by \$11,800 on average. Better yet, we might express this increase as a percentage instead. That is, we estimate that wages increase by $^{11.8}/_{93.8} = 12.6\%$ due to a university education.

Can you identify any problems with this experiment? The sample size is probably too small for us to have much confidence in our result, and we would want to include many more individuals in this experiment (but we wanted to fit the table on the page). More importantly, conducting this experiment would be very expensive, and would be unethical. We would have to pay for the university education of each member in the treatment group. Each member in the control group would be denied an education, the access to which is a human right. This experiment, like many that would be useful in economics, is too expensive to perform and would not pass an ethics board!

3.1.3 Observational data

Although less useful than experimental data, observational data is much more commonly used in economic analysis. The experiments we would need to conduct in economics are often too expensive, and are unethical (see the example in the previous section).

Observational data. Observational data is data that is collected by observing and recording the universe as it unfolds, without intervening.

Observational data is recorded without being able to apply any *control* over whether the individuals in the data are in the treatment group, or in the control group. We simply observe the choices that people make, and the outcomes that occur. There is little to no influence over the behaviour or actions of the individuals in the data set. There is no random assignment in observational data[2].

The lack of random assignment and control, means that individuals have some degree of choice in whether or not they are in the treatment or control group. This can have very serious consequences when trying to make causal statements using observational data. As an example, we will reconsider the link between education and wage, but in a setting where the individuals in the data have *chosen* to obtain an education.

Example 3.2 Suppose now that there is no experiment available to determine the effect of education on wages. Instead, we merely observe an individual's wage and educational attainment. We are powerless over which individuals obtain an education. Consider that the same individuals who were enrolled in the previous experiment were instead allowed to live their lives free of interference. Some chose to get a university education, some did not.

name	education	wage (in thousands)
Raven	university	101
Gary	high school	62
Roberto	high school	59
Amanda	university	144
Justin	university	152
Hadeel	university	142
Mudrika	university	124
Dewarren	high school	69
Jacob	high school	87
Melinda	high school	80

Justin and Hadeel *decided* to obtain an education (opposite to the experiment where they were *assigned* to have *no* education). Gary and Jacob decided not to obtain an education (in the experiment they were assigned to receive an education). Why did their decisions contrast to what happened in the experiment?

Labour economics has several explanations as to why the individual decisions to obtain an education might be linked to the *anticipated* or *predicted* wage. For the purposes of this example, let's assume a simple reason. Suppose that the *true* increase in wage due to an education is 12.6% (this is what was revealed by the experiment). Having some knowledge of this, individuals with a higher earning potential will be more attracted to a university education. They have more to gain.

[2] *Natural experiments* are one exception.

Let's compare the sample averages between the two groups again. We get:

$$\bar{wage}_{university} = \frac{101 + 144 + 152 + 142 + 124}{5} = 132.6$$

and

$$\bar{wage}_{highschool} = \frac{62 + 59 + 69 + 87 + 80}{5} = 71.4$$

so that the average increase in wages is $^{132.6 - 71.4}/_{71.4} = 85.7\%$! This is much more than what was indicated using the *experimental data* (12.6%). What happened here? Those individuals who had more to gain (a higher base salary) *chose* to get an education.

In this example, education is not just increasing wages, it is *indicating* the earning potential (base salary) of individuals. This makes it impossible to attribute the increase in wages between the two groups to the difference in education. Here, the *lurking variable* is an individuals perceived benefit of obtaining an education (their self assessed earning potential).

Endogeneity. In economics, endogeneity can refer to a situation where an individual's anticipation of an outcome influences the choices that they make. It can also refer to a situation where there is some factor (possibly unobserved) driving various decisions.

Observational data, such as in the above example, often involve something economists refer to as *endogeneity*. A large part of econometrics is dedicated to being able to make causal statements (such as how much education causes an increase in wages) in the face of "threats" of endogeneity. In this textbook we will not tackle such issues, but we will be working primarily with observational data. We need to be aware of the limitations of observational data, especially when attempting to infer causality.

3.1.4 Available data

Available data is data that has already been recorded for some specific or general purpose. Most of the data that economists use is already available. When trying to answer a specific research question, it is rare to have the opportunity to collect and create a new data set. Researchers typically start by looking for observational (or sometimes experimental) data that already exists.

For example, Statistics Canada collects and distributes demographic and economic data, which is used extensively in economics research and policy analysis. Most countries have similar agencies, for example the United States Census Bureau. For labour related issues, such as determining the effect of education on wages (see the previous two examples), a popular source of available data in the U.S. is the Current Population Survey. The World Bank provides development data for countries. These are a few examples; there are thousands of data sets available free and online.

3.2 Populations and Samples

Most data is collected by *sampling* from a population. A *sample* is in contrast to a *census*. In a census, all individuals in the population are contacted. In a sample, only a portion of the population is contacted. The main reason for using a sample is that it is usually too costly (or it is impossible) to record information on an entire population.

Every 5 years, the Canadian Census of Population attempts to contact every household in Canada, costing more than half a billion dollars. While census data is important, most economics researchers have a much smaller budget, and so must rely on a *sample*. In addition, a census may require too much time to collect, and may be less accurate than a carefully collected sample.

3.2.1 Population

Population. The population contains every member of a group of interest.

A population contains all cases, units, or members that we are interested in. In economics a "member" or a "case" is usually an individual, a firm, or a country. The terminology case/unit/member just refers to a single component of the population.

If we are interested in the effect of education on wage, the population consists of every working individual, and a case refers to each individual. If we are comparing GDP between countries then the population consists of all countries in the world, and each case/unit/member is a separate country. If we are describing increasing food prices in Manitoba, then the population might be every grocery store in the province. In the following discussion, we will often refer to a "member" or a "case" as an "individual", but the discussions are valid whether we are talking about individuals, businesses, schools, institutions, countries, etc.

3.2.2 Sample

Sample. A sample collects data on a subset of members from the population.

A sample is simply a subset of the population. It usually consists of far fewer cases or members than the entire population. Information in a sample is meant to reflect the properties and characteristics of the population of interest. The sample contains those members of the population that are actually examined, and from which the data set is created. A sample is in contrast to a census, where there is an attempt to contact every member of the population.

Census. In a census, there is an attempt to contact and record data on every member of a population.

In most situations in economics, a sample, not a census, is used to conduct quantitative analysis[3].

3.3 Sampling bias

The way in which the sample data is collected is very important. A bad sample, one that does not represent the population of interest, leads to bad results. The sample is

[3] When comparing economic indicators such as GDP, usually the entire population (all countries) are used.

only useful in describing the population if it is a fair and unbiased representation of the population. Bad sample data can occur for several reasons, some of which are defined below.

Sample biases.

- *sample selection bias* - when characteristics of the members of the sample do not represent the characteristics of the members of the population.
- *non-response bias* - when individuals, who have something in common with each other, choose not to respond to a survey or poll.
- *misreporting* - when individuals report inaccurate information.

A common and highly recommended way of constructing a sample is by *randomly* selecting members from the population. Random selection prevents links and commonalities between those that are sampled. Random sampling can help to prevent *sample selection bias*.

Survey / Poll. A survey or a poll provides a sample of data by asking people questions.

Surveys, also called polls, are used for many purposes and are an important source of data. Although it is usually better to observe information about the individuals in the sample directly, rather than ask those individuals to report that information, surveys/polls are often the only option to collect data. For example, we might collect data on individual's income either by asking them how much they make (poll/survey), or by observing their pay cheques from their employers. Polls and surveys suffer from the possibility of *non-response* and *misreporting*. You might anticipate that when a person is asked "how much do you make?", they may refuse to answer, or lie.

In this section, we will further explore some ways in which samples can be collected, and how the problems of sample selection bias, non-response bias, and misreporting can arise.

3.3.1 Sample selection bias

Suppose that we want to know how Manitobans are going to vote in the next election. We go outside the classroom and ask the first 30 people how they are going to vote. Only 6 of them say they will vote conservative. Should we predict that the next government will not be conservative? Are these 30 individuals a fair representation of the voting population? Probably not. Professors in social science overwhelming vote on the left[3], and this tendency likely extends to students. While collecting this sample might be *convenient* for us, a university campus is not a subset that fairly reflects the political views of the larger population. Inferences drawn from university campus samples may not be correct and susceptible to *sample selection bias*.

Example 3.3 An infamous example of the failure of sampling is that of the *Literary Digest* poll of 1936. Some 10 million questionnaire cards were mailed out, 2.4 million of which were returned. Based on the data in the returned questionnaires the *Literary Digest* mistakenly predicted that Landon (Republican), not Roosevelt, would win the presidential election. Many academics have since held that the poll failed so miserably due to the *Digest* selecting it's sample from telephone books and

car registries [4], which contained more affluent individuals (those that could afford a telephone and a car), and who tended to vote Republican.

Voluntary response sampling and on-line surveys are also prone to sample selection bias. Who are the type of people who would answer an on-line survey? Likely it is those individuals most passionate, and holding extreme views, that are willing to take the time and effort to voluntarily provide information.

3.3.2 Non-response bias

Non-response bias can also lead to a sample failing to fairly reflect the population. If some people do not respond to the poll or survey, that is fine. But if there is an underlying reason for non-response, that is also linked to the answers that people provide, then the results inferred from the sample will be *biased*.

Example 3.4 In 2016, polls predicted that Hillary Clinton would likely win the presidential election, putting her probability of winning around 90%[2]. How did the polls get it so wrong? One theory is *non-response bias*. The sample was biased in the sense that Trump supporters simply refused to respond. This theory is backed by findings that individuals with lower education, and anti-government views, are less likely to respond to surveys.

Example 3.5 The view that the *Literary Digest* disproportionately sampled Republican voters (see Example 3.3) has been challenged[4]. *Non-response bias* is an alternate suspected culprit. ⅓ of Landon's supporters answered the survey, compared to only ⅕ of Roosevelt supporters. Most of the 7.6 million unanswered surveys were from Democrats!

3.3.3 Misreporting

With any survey, *misreporting* is a concern. Misreporting is when a survey or poll respondent does not provide accurate information. The reasons for this can be many. For example, the "Shy Trump Hypothesis" supposes that the 2016 polls failed due to Trump supporters feeling that their views were unaccepted by society. Individuals may be too embarrassed to report truthfully, may be worried about social stigma, may not understand the questions, or may not recall information accurately. If there is systematic misreporting (in the sense that there is a pattern or a commonality among the people who report), then inferences drawn from such surveys can be *biased*.

Example 3.6 The Current Population Survey (CPS) is an important survey that is used in a variety of quantitative analyses, and that has hundreds of thousands of citations in economics research.

The CPS asks respondents questions on enrolment in food stamp programs. This information is important for understanding poverty, and ways to mitigate poverty. A study investigating misreporting in CPS data has found that approximately 50% of households on food stamps do not report it on the CPS[5], and that theories such as *stigma* may explain the misreporting. When individuals feel that they may be judged, they may not answer survey questions accurately.

3.4 Simple random samples

Simple random sample. A simple random sample is collected by randomly selecting members from the population.

In order to avoid sample selection bias, *simple random samples* are often recommended. A simple random sample is when members of the population of interest are selected at random. Each member has an equal chance of being selected. Imagine a bowl containing pieces of paper with everyone in the population's name written on. Pulling out n pieces of paper from the bowl, and contacting those selected, would create a simplae random sample of size n.

Simple random sampling is in contrast to convenience sampling, voluntary sampling, and on-line polls. In a simple random sample, information and opinions will not be skewed by those individuals who are the most motivated or the most willing to participate in a study. There will be no underlying link between the members in the sample.

There are more complicated versions of random sampling. For example, a *stratified random sample* selects members from *subgroups* of a population. In this way, members with certain characteristics have a higher probability of being sampled.

Example 3.7 — Stratified sample. Suppose that we want the portion of ethnicities in our sample to perfectly reflect the portion of ethnicities in the population. Suppose that we know that the population contains only 3% of a certain ethnicity. If we take a sample of 100 from the population, what is the probability that no one in the sample is from that ethnicity? It turns out to be approximately 5%.[a] We might completely miss this group! Instead of pure random sampling, we could randomly select a certain number of individuals from each ethnicity, where the number that we select is based on their proportions in the population. That is, we could randomly select exactly 3 people (if our sample size is going to be 100) from the ethnicity that comprises 3% of the population.

[a] Assuming that the population is very large, the probability of *not* drawing the certain ethnic group is maintained at 97% for each draw, and the probability of 0 draws is $0.97^100 = 0.048$.

3.5 Data ethics

Although experiments are fairly rare in economics, it is worth noting the ethics behind designing experiments. We have already seen one example where an economics experiment would be unethical (wages and education), but who determines what is ethical or not? In most cases, this is determined by a *review board*. Most experiments are subject to ethical approval before they can proceed. In order to secure approval, most experiments will require informed consent (the participants in the experiment must understand the consequences of being experimented on and agree to be subjected to an experiment). In addition, most experiments must preserve confidentiality, so that although the results of the experiment may be made public, the public cannot obtain sensitive information about the participants.

4. Describing Data

In this chapter, we will begin to describe the variables in our data set. We start by explaining the structure of a data set. Each row in a data set corresponds to a different *observation*, and each column is a different *variable*. We then discuss some basic characteristics of the variables, such as whether they are quantitative or categorical, and whether they are continuous or discrete.

Such considerations not only help us understand our data set, but also inform the type of graph that we should use to visualize the data. We will learn about the following ways to graph a *single* variable in this chapter:

- pie charts
- bar graphs
- histograms
- time plots

Creating graphics from data is a powerful way to learn about the *distribution* of a variable. Graphics are also used to convey information, to make a point, or to try to convince the reader of some hypothesis. In this chapter we will match the appropriate type of graph to the different types of variables, and learn how to create those graphs in R.

By graphing a *quantitative* variable in a *histogram*, we can learn about the *shape*, *location*, and *spread* of its distribution. These are important considerations that help to characterize the population that we are studying. Graphs help us look for patterns and exceptions to the pattern.

Finally, we will discuss the *scatterplot*. A scatterplot graphs *two* variables at once (sometimes more!), and is a powerful way to begin to describe the *relationship* between two variables that may be related to each other. We can use a scatterplot to describe the *direction*, *form*, and *strength* of a relationship, whether the relationship is *linear* or *nonlinear*, and to see if a relationship even exists!

4.1 How data is arranged

A data set is typically arranged with each *observation* taking a different row, and each *variable* taking a different column. Each row represents a single observational unit. Each column is a different type of information on the observations. In Figure 4.1, the observations are on people (each row represents a different person), and the variables are age, gender, income, etc. That is, each column contains a different type of information about the people in the sample.

Figure 4.1: Data set on Mars colonists.

variables

observations

name	gender	ethnicity	age	income	marital.status
Ciara Stavish	female	white	24	79321	single
William Cooper	nonbinary	black	41	56589	married
Soo Ho Causey	male	asian	35	125077	married
Shaqeeq el-Chahine	male	arabic	50	114230	married
Abdur Razzaaq el-Masri	male	arabic	50	73342	divorced
Khadijah Chau	female	asian	32	42383	married
Noora el-Shahin	female	arabic	50	52655	divorced
Katheer al-Salik	male	arabic	55	58723	married
Siobhan Lapioli	female	black	39	109679	married
Tammassha Madrid	female	indigenous	64	147720	married
Hannah Sublette	female	asian	35	48872	married

The number of observations, or the number of rows in the data set, is called the *sample size* and is denoted n. It is always better to have a larger n!

Sample size. The sample size is the number of observations (rows) in the data set, and is denoted by n.

Example 4.1 — Data example: Mars has been colonized. At several points in the book we will use data on Mars colonists (see Figure 4.1 for a few rows and columns of the data set). Mars has been colonized, with 720,720 individuals thriving on Mars City. Due to the importance that Mars City represents for the survival of humanity, detailed information on the inhabitants is available. People who want to live on Mars are subjected to intense scrutiny and have agreed to allow detailed information about themselves to be available. The data is of course fake (randomly generated), but has variables that mimic many real data sets, such as the Current Population Survey.

4.2 Types of observations

Observations may also be called *cases*, *units of analysis*, or *experimental units* (if the data were obtained by an experiment). The type of observation depends on the nature of the data. In general data describes people, places, things, or situations. So, each observation could be a different person (as we saw in Figure 4.1), or a different country, province, firm, university, or even a moment in time! In the example below, we see a

data set where each observation is a different country.

Figure 4.2: Data from the 2019 World Happiness report. Each observation (row) is a different country. The variables (columns) are the average Happiness Score, and GDP per capita. The name of the first column reveals that we have observations on *countries*, rather than individuals, provinces, businesses etc.

Country.name	Happiness.score	Log.GDP.per.capita
Finland	7.858107	10.636060
Denmark	7.648786	10.755594
Switzerland	7.508587	10.975945
Netherlands	7.463097	10.809204
Norway	7.444262	11.085626
Austria	7.396002	10.741893
Sweden	7.374792	10.766932

Example 4.2 — Data example: 2019 World Happiness Report. We will use the World Happiness report for several examples throughout the book. The First World Happiness report was prepared in 2013, in support of a United Nations High-Level Meeting on "Well-Being and Happiness: Defining a New Economic Paradigm." The World Happiness Reports are funded and supported by many individuals and institutions, and based on a wide variety of data. The most important source of data, however, is the Gallup World Poll question of life evaluations. The English wording is:

> "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"

The responses can be averaged so that each country is ranked (see Figure 4.2 for the happiest countries in the world!) By including other variables in the data set for each country, researchers have an opportunity to investigate what factors lead to differences in happiness between countries (differences such as GDP per capita). In this data set, GDP per capita is in terms of Purchasing Power Parity adjusted to constant 2011 international dollars.

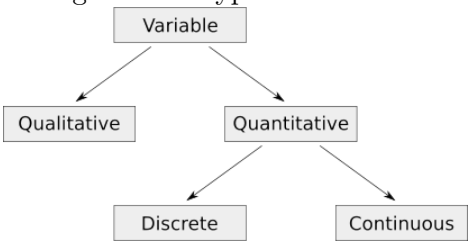
4.3 Types of variables

In order to know what type of graph or statistical technique should be used, it is helpful to categorize variables into different types. For example, we would not display the information on *marital status* in the same type of graph as we would an individual's income, or their education level. Similarly, some statistics formulas cannot be used with certain types of variables. It is important to be able to classify a variable for these reasons.

In the first few rows and columns of the data set on the Martian colonists (see Figure 4.1), we see several different *types* of variables. The first column `name` tells you that the type of observation is an individual. The name of the individual allows you to locate a specific row. A row number would do the job just as well. That is, the name of the person is not particularly useful and is not a variable; it just serves as an identifier.

The variable `gender` is what we call a categorical or qualitative variable. Similarly, `ethnicity` and `marital.status` are qualitative variables. In contrast, `age` and `income` are quantitative variables, and we might go further to say that `age` is a discrete variable whereas `income` is a continuous variable. See Figure 4.3 for an overview of how we will classify variables in this section.

Figure 4.3: Types of variables.



4.3.1 Qualitative / categorical variables

Qualitative variable. A variable that describes a quality of the observation, and does not have natural numerical meaning.

A qualitative variable is one that takes on two or more possible qualitative values (qualitative variables are also called categorical variables). When we say qualitative we mean something that is not necessarily numerical, but that has a quality or a property. For example, red is a quality, three is a quantity. The colour of someone's eyes or hair could be a quality that fits into one of several categories, whereas their height or weight could be quantified. We could say that one person is twice as tall as another, but we can't make the same kind of algebraic comparisons for eye colour.

Some typical examples of qualitative variables encountered in the social sciences are:

- gender
- treatment
- ethnicity
- province or territory of residence
- marital status
- political affiliation
- exchange rate regime

For most of the examples above, the categorical variable can take on one of several different possible values. A key feature of a categorical variable is that its categories must be *exhaustive*. That is, each observation must be able to fit in one of the categories. A simple way to ensure this is to have an "other" category that acts as a catch-all for observations that are not easily categorized.

Ethnicity is a categorical variable reported in many data sets that collect information at the individual level. "Ethnicity" as a categorical variable is problematic in terms of developing appropriate concepts, avoiding ambiguity, and avoiding offensive constructs and terminology (for example Eskimo in reference to Inuit). However, the international meeting on the Challenges of Measuring an Ethnic World (Ottawa, 1992) noted that ethnicity is a fundamental factor of human life inherent in human experience, and that data on ethnicity is in high demand by a diverse audience. Statistics

Figure 4.4: Statistics Canada ethnic categories.

Classification structure

Code	Category
1	North American Aboriginal origins
2	Other North American origins
3	European origins
4	Caribbean origins
5	Latin, Central and South American origins
6	African origins
7	Asian origins
8	Oceania origins

Canada has a standard that classifies individuals in one of eight categories: See Figure 4.4.

The number of categories that a categorical variable can take is often up to the discretion of the researcher, and can vary. For example, countries must decide how to manage their currency on the foreign exchange market. A categorical variable could be used to describe which *regime* (currency exchange system) each country follows. There are three basic types, so for example each country could have a variable called `exchange.regime` which takes on one of three values: `floating.exchange`, `fixed.exchange` and `pegged.float.exchange`. However, the IMF classifies countries in 1 of 8 exchange rate regime categories, so the `exchange.regime` variable could instead take on one of eight possible values.

Finally, why are categorical variables used? They are important for predicting, modelling, and understanding the *differences between groups*. Is a drug effective? We can compare the outcomes between the *treated* and *placebo/control* groups. The categorical variable will identify which individuals belong to which group. Do women earn less than men? To be able to investigate, and perhaps ultimately solve discrimination by gender or race, we first need a way to identify differences between groups; this task is greatly aided by categorical variables.

Dummy variables

Gender was traditionally considered a *binary* or *dummy* variable in the social sciences. A dummy variable is a special kind of categorical variable that can take on one of only *two* values (binary refers to a number system with a base of 2). Historically, a `gender` categorical variable could take on the values either "male" or "female"; each person was forced to belong to one of the two categories. With the more common understanding that gender is a spectrum rather than a binary, more contemporary statistical analyses try to recognize broader categories, such as non-binary, trans, and possibly dozens others. For example, Statistics Canada has slightly broadened its sex and gender classifications. A person's sex can be "male", "female" or "intersex", and a person can be "Cisgender", "Transgender", "Male gender", "Female gender" or "Gender diverse". With more than two categories, gender is no longer the quintessential "dummy" variable in the social sciences.

Dummy variable. A dummy variable, also called a binary variable, is a categorical variable that takes on one of two values.

Better examples of dummy variables are in "yes" or "no" situations. For example,

did the subject receive the "treatment"? The `treatment` variable could take on values `yes` or `no`. Numbers are typically assigned to these dummy variables: 1 indicates "yes" and 0 indicates "no". Don't be fooled by the numerical values! The numbers don't actually mean anything, other than to provide a key to the categories.

Other examples of dummy variables in economics include whether a firm is "domestic" or "foreign", whether an individual has participated in a social program or not, whether a person has ever received social assistance, whether an individual or country has ever defaulted on a loan, whether an individual has ever committed a crime, etc.

Ordinal variables

Ordinal variables rank observations (order them) relative to one another. For example, the position that an athlete places in a race (1st for gold, 2nd for silver, etc.) is an ordinal variable. The ranking of countries by happiness (see Figure 4.2) is an ordinal variable. Ordinal variables do not contain as much information as quantitative variables, and are not considered as useful. The *magnitudes* of ordinal variables don't have much meaning. Did the athlete who received a silver medal (`position = 2`) take twice as long to complete the race as the athlete that received gold (`position = 1`)? Ordinal variables provide a type of *qualitative* information.

Ordinal variable. An ordinal variable ranks each observation among all the observations.

Ordinal variables usually occur due to the ordering of some other *latent* or *hidden* variable. In the case of the athletes, the `time` to complete the race is the underlying variable that generates the ordinal `position` variable. It would always be better to have the underlying variable `time` instead. The ordinal variable does not contain as much information. Similarly, we would rather know the actual `Happiness.score` of each country rather than their happiness rank. Ordinal variables are used when no such *quantitative* alternative exists.

4.3.2 Quantitative variables

Usually, when we think of a variable, we think of it being able to take on different numbers, not different categories. In this sense, quantitative variables may seem more natural or comfortable than the qualitative variables discussed above.

Quantitative variable. A quantitative variable takes on numerical values and measures the magnitude of something.

A quantitative variable takes on different numbers, and the *magnitude* of the variable is important (whether the number is small or large). Depending on the nature of the variable, it may have a certain *domain*. A domain is all the possible places the variable can occur or "live". For example, income cannot be below 0, so an income variable might be confined to the set of *positive real numbers*. A variable measuring temperature on Earth might realistically be confined between -100 and 70 degrees Celsius. In some situations, the domain might be the entire real line, so that the variable might take on any value between negative and positive infinity!

In Figure 4.1 we see that `age` and `income` are quantitative variables. Yet, there is something different in the nature of these two variables. In fact, quantitative variables can be divided into two types: *discrete* and *continuous*. In the Mars colonist data

example, `age` is a *discrete* variable,[1] and income is a *continuous* variable[2].

Discrete variables

A discrete variable can take on a *countable* number of values. For example, we can count the number of values that the `age` variable can take. Some other examples of where we can count the numbers of things:

- Times a customer might visit a store.
- Students in an Econ 2040 class.
- Children in a family.
- Years of education.
- Individuals in Canada.

The key property of a discrete variable is that we can *count* all the possibilities[3]. By contrast, *continuous* variables can take on an *uncountable* number of values!

Discrete variable. A discrete variable is a type of quantitative variable. It takes on a countable number of values, which are usually non-negative integers: $(0, 1, 2, 3, \dots)$.

Continuous variables

A continuous variable is obtained by measuring, and can take any value over its range. Even if the range is not infinity, a continuous variable has an uncountable number of possibilities! For example, the possible heights of an individual are uncountable, even though the possibilities are between 0m and 3m, for example. The person could be 1.63m tall. What about 1.63001m tall? Or 1.630000001m tall? We could keep adding zeros. The possibilities are uncountably infinite. In Figure 4.2, the `Happiness.score` and `Log.GDP.per.capita` variables are continuous. They can take on any values in a range, but we can't count all the possible values.

The distinction between discrete and continuous variables leads to very important mathematical considerations in statistical modelling. For example, where a discrete variable might be added up, a continuous variable would be integrated. Similarly, we could find the derivative for a function of a continuous variable, but we can't take the derivative of a function of a discrete variable. We do not get into these topics in this book, but rather focus on the consequences that these differences have for the way in which we *graph* the variable.

4.4 Graphing categorical data

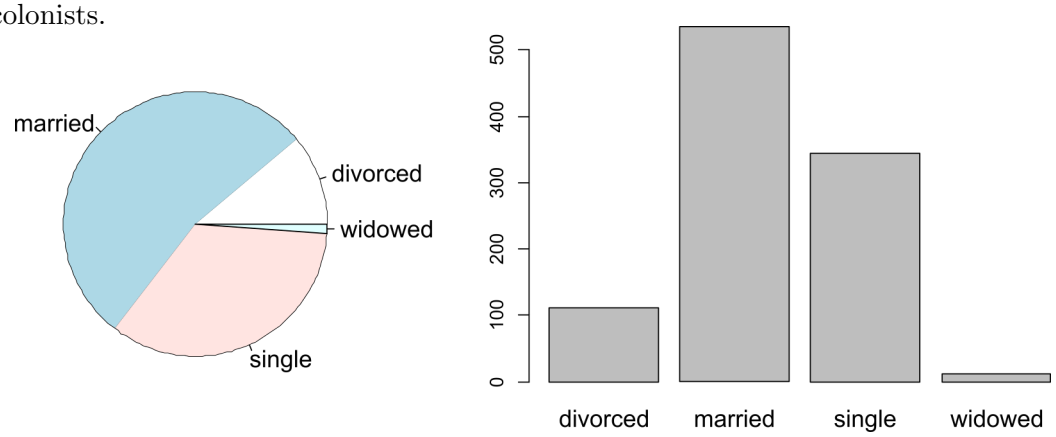
Categorical data may be graphed using a *pie chart* or *bar plot*. To construct these graphs, the number of observations in each category must be calculated. For a pie chart, these numbers are converted into percentages by dividing by the *sample size* (and multiplying by 100). The entire pie represents 100%, with the size of each slice representing the percentage of observations in each category.

[1] `age` is a discrete variable because it is measured in *years*. Measuring it in finer units (down to the second or millisecond) would make it essentially continuous.

[2] It can be argued that `income` is not truly a continuous variable, since salaries are for example paid down to the cent, and only have a maximum number of decimal places of two. Thus, there are a countable number of different incomes that each person can have. However, due to all measurements of any continuous variable being subject to a certain degree of human accuracy, the same argument could be made for many "continuous" variables.

[3] Some variables are countably *infinite*, meaning that even if they can take on an infinite number of possibilities, we could list them all.

Figure 4.5: Pie chart (left) and bar plot (right) of marital status for a sample of Mars colonists.



Similar to a pie chart is the bar plot. A bar plot simply uses the number of observation in a category for the height of a bar. The bar plot has the added benefit that it conveys the actual number of observations in each category. For example, in Figure 4.5 we can see that approximately 100 individuals in the sample are divorced.

Example 4.3 — Pie chart for marital status in Mars city. Let's recreate Figure 4.5. We'll make a pie chart and bar plot for the marital status of a sample of 1000 Mars colonists. First, load the data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

Next, look at a table of the `marital.status` variable:

```
table(mars$marital.status)
```

```
divorced married  single widowed
     111     534     344      11
```

To make the pie chart, we can use:

```
pie(table(mars$marital.status))
```

and to make the bar chart, we use:

```
barplot(table(mars$marital.status))
```

Note that you can "export" the images that you create (that's how we got them into this book!).

Typically, a pie chart *or* a bar plot is used, not both. In fact, it is questionable if these graphs are even needed for categorical data. The table below, upon which the graphs are based, takes up very little space and conveys a lot of information:

marital status	divorced	married	single	widowed
number of observations	111	534	344	11

4.5 Graphing quantitative data

Commonly, *histograms* are used to graph continuous variables, and *bar plots* or *histograms* are used for discrete variables. These graphs provide a visual representation of the *distribution* of the variable. A distribution describes the values (the *range*) that the variable can take, and conveys how often (or the probability) the variable takes on certain values.

Later we will talk about the *Normal* distribution (and others), but for now let's develop terminology that allows us to describe what we see when viewing a graphical representation of a distribution.

4.5.1 Histograms

Histogram. A common graphic for portraying the distribution of a continuous variable. A histogram "bins" the variable, and draws the height of each bin by using the number of observations in that bin.

A histogram is created by breaking up the range of a variable into several "bins", counting the number of observations that fall into each bin, and then graphing the heights of the bins. This gives us a visual representation of how often the variable takes on ranges of values. The histogram tells us if there are extreme values, if the variable is spread out or tightly packed, and which values the variable tends to take. Example 4.4 illustrates how a histogram is produced.

Example 4.4 — Histogram of IQ scores. Here is a sample of 84 different IQ scores:

112	108	126	81	133	106	76	86	101	100	66	111	92	103
108	123	110	117	88	94	106	101	85	81	97	111	105	83
89	89	72	94	100	114	125	101	128	101	121	122	101	77
99	97	131	103	108	125	106	81	91	127	97	95	109	115
100	86	114	117	93	114	68	75	126	112	93	129	94	76
103	83	107	74	108	101	103	112	90	104	117	105	96	97

To create a histogram, the computer will "bin" this data, count how many scores fall into each bin, and then use the number of values in the bin to graph its height. For bins of size 10, we could have:

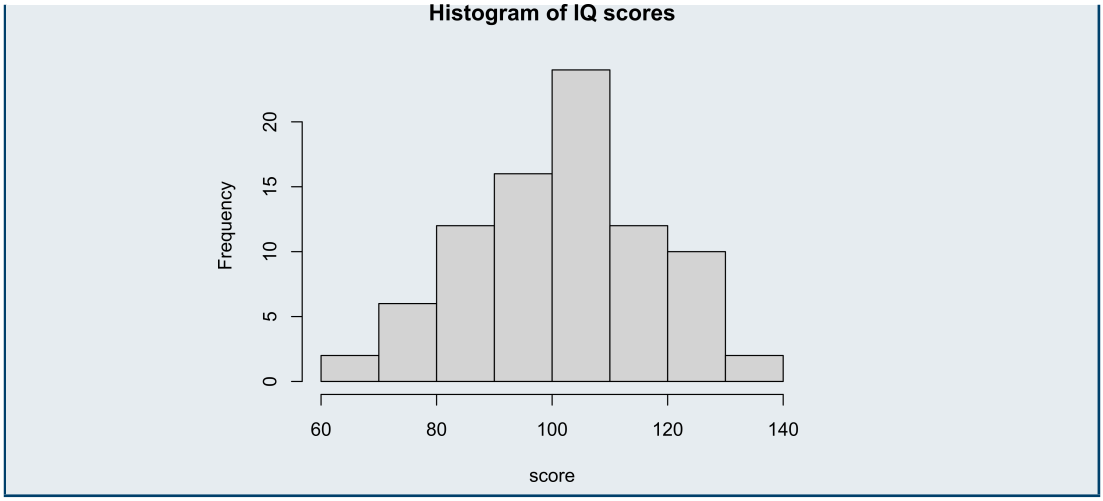
Bin	Number of scores
$60 \leq \text{IQ} < 70$	2
$70 \leq \text{IQ} < 80$	6
$80 \leq \text{IQ} < 90$	12
$90 \leq \text{IQ} < 100$	16
$100 \leq \text{IQ} < 110$	24
$110 \leq \text{IQ} < 120$	12
$120 \leq \text{IQ} < 130$	10
$130 \leq \text{IQ} < 140$	2

Download the IQ scores in R:

```
IQ <- read.csv("http://ryantgodwin.com/data/IQ.csv")
```

Create the histogram (with a title and label on the x-axis):

```
hist(IQ$scores, main = "Histogram of IQ scores", xlab = "score")
```

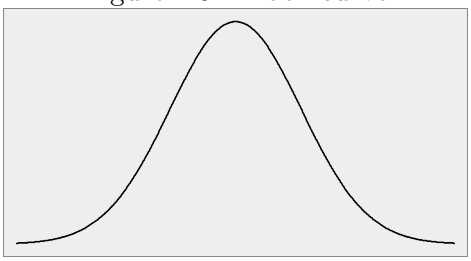


4.5.2 Describing distributions

Distribution. A distribution describes all the possible values that a random variable can take, assigning probability to all the possibilities.

A distribution defines the probabilities and possibilities of a random variable. It can sometimes be represented in a graph like in Figure 4.6 (except we would need numbers on the x- and y-axis). Some statisticians think that the data we observe is actually *created* by distributions. This would mean that it is crucially important to try to find the right distribution to describe a variable. Even if distributions do not generate data but merely help to describe randomness in the real world, finding the right distribution is an important task.

Figure 4.6: A bell curve.



There are hundreds of different statistical distributions. In this book we will limit ourselves to only a few. We will, however, develop terminology that is helpful in selecting the right distribution. For example, if a distribution is spread out or condensed, if it has skew or is multi-peaked, then the bell curve of Figure 4.6 would not be appropriate.

Shape, location, and spread

How would you describe the distribution for IQ scores (as it is portrayed by the histogram) from Example 4.4? Use words like *shape*, *location*, and *spread*.

- **Shape**: IQ scores appear to have a single peak, are not skewed, and follow a "bell" like shape.
- **Location**: The distribution is located at around 100. This appears to be the centre of the distribution (around where most values are located).

- **Spread**: The distribution is not particularly spread out, nor is it tightly packed. It matches the bell-curve nicely.

The "bell" curve that we mention is in reference to the important Normal distribution, which we discuss in a later chapter. It is a famous and important shape that you should already be familiar with: see Figure 4.6.

4.5.3 Skew

A non-symmetrical distribution is said to be skewed if it looks as if one of the "tails" of the curve has been stretched out. That is, a distribution is skewed if one of the tails is longer than the other. Skew is a descriptor that helps to characterize a distribution. Figure 4.7 illustrates two skewed distributions.[4]

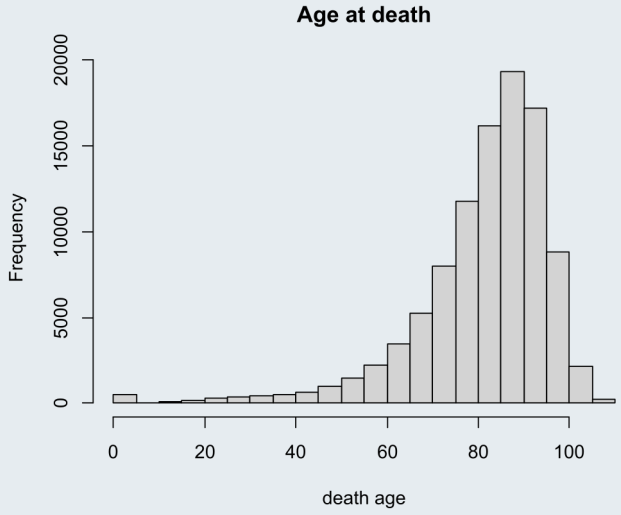
Skew. When one tail of the distribution is stretched out.

Example 4.5 — Left skew: age at death. The distribution of peoples ages at the time of their death is an example of left skew. Using Life Tables from Statistics Canada, download constructed data on the age-at-death of 99,976 individuals:

```
data <- read.csv("http://ryantgodwin.com/data/age-at-death.csv")
```

Create a histogram, give it a title, and label the x-axis:

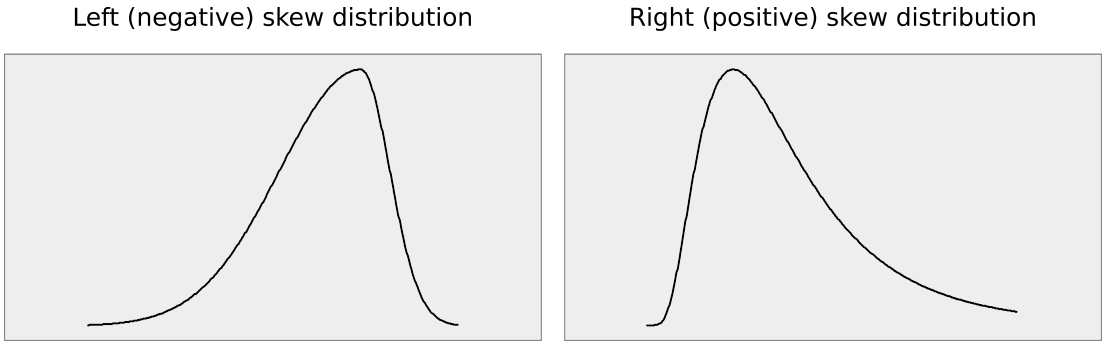
```
hist(data$death.age, main = "Age at death", xlab = "death age")
```



What do you see? This is a left skewed distribution. It peaks at age 85-90, with a bit of an extra "spike" at around age 0 (reflecting infant mortality).

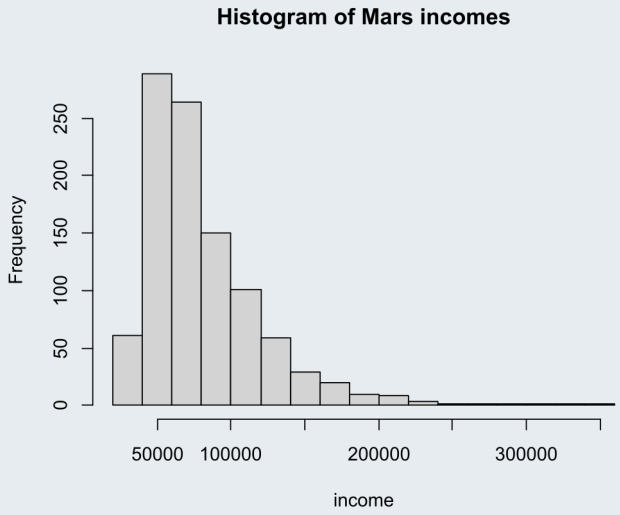
[4]The left skew distribution is from the "Skew Normal distribution" and the right skew is from the "Log-Normal distribution".

Figure 4.7: Skewed distributions.



Example 4.6 — Right skew: income. The incomes of individuals typically follow a right skewed distribution. For example, the majority of workers might be within the \$30,000 to \$100,000 range, with a small portion of workers making very large incomes. Let's draw a histogram of incomes from the sample of 1,000 employed Mars colonists:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
hist(mars$income, breaks = 16,
     main = "Histogram of Mars incomes", xlab = "income")
```



The option `breaks = 16` was used to control the number of bins in the histogram (try removing it and see what happens). We see that incomes on Mars appear to follow a right skewed distribution, with the majority making under \$100,000, and with some very large incomes in the sample.

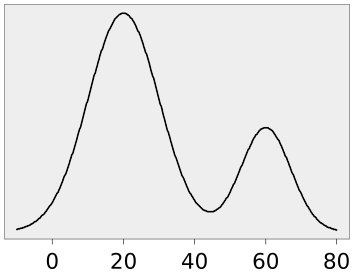
4.5.4 Multi-peaked distributions

All of the distributions we have seen so far have been single-peaked (think of the top of a mountain). This peak is the *mode* of the distribution (we will discuss mode later). Sometimes, however, we see distributions that are multi-peaked. These distributions

can arise for a variety of reasons, one of which being when two distributions are *mixed* together to create a single random variable. Figure 4.8 shows a multi-peaked (bi-modal) distribution.

For example, the percentage grades in university courses are often bi-modal (have two peaks): one peak around "C" grades and another peak around "B+". The number of years of education of individuals is often multi-peaked as well: for example one peak at 12 years (high school) and another peak at 16 (university degree).

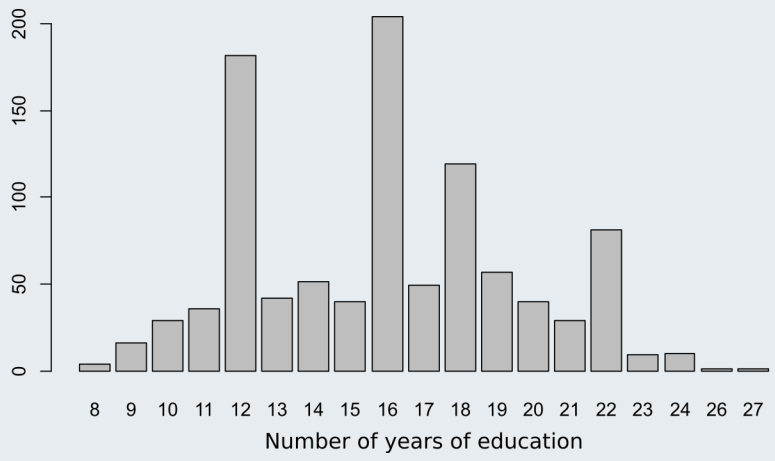
Figure 4.8: A multi-peaked distribution. $3/4$ of the values come from a bell curve located at 20, the other $1/4$ are located at 60.



Example 4.7 — A multi-peaked distribution for the number of years of education. We will use the sample of 1000 employed Mars colonists. Sometimes when we graph a *discrete* variable (see Section 4.3.2), we use a *bar plot* instead of a histogram. Discrete variables do not necessarily have to be binned! For education, we can count the number of people that have 8 years, that have 9 years,... there is no need to create a "bin" in order to cover a range of values, and using a bar plot avoids multiple integer values being binned together. Just like for the categorical data in Figure 4.5, the height of each bar is equal to the number of people for each integer value in the plot.

Download the Mars data, and create a bar plot for years of education:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
barplot(table(mars$years.education))
```



We see that the distribution has at least two-peaks: one for a high school degree, and one for a university degree. Note that barplots have spaces between the bars, whereas histograms do not.

4.5.5 Outliers

An important reason to graph data, using histograms and bar plots (and later scatter plots), is to detect the presence of *outliers*. Outliers are extreme values that differ significantly from the other observations in the data. An outlier might be sampled by chance, in which case the observation should usually remain in the data set.

Outlier. An extreme value that may indicate the presence of an error, in which case the observation should be removed from the sample.

If the outlier is due to an error, then it should be removed or corrected. Such errors may occur as the data is being measured or recorded. For example, an extra 0 might be typed when recording income, a single weight might be recorded in kilograms instead of pounds, or an economist may forget to convert Pesos to Dollars when examining trade.

Another possible source for outliers is if an observation comes from a *different* population. Remember that the sample is meant to represent the population. If we are interested in the income of employed Martian colonists, then the sample should not include a student, for example. Observing a small value for income might induce us to examine the observation more carefully and perhaps discover that the observation is indeed from the *wrong* population.

In Example 4.6 we see some outliers: some very high incomes in the right tail of the distribution. We should do our best to examine these observations to see if we can detect any data recording mistakes, or any indication that these observations do not belong in the sample.

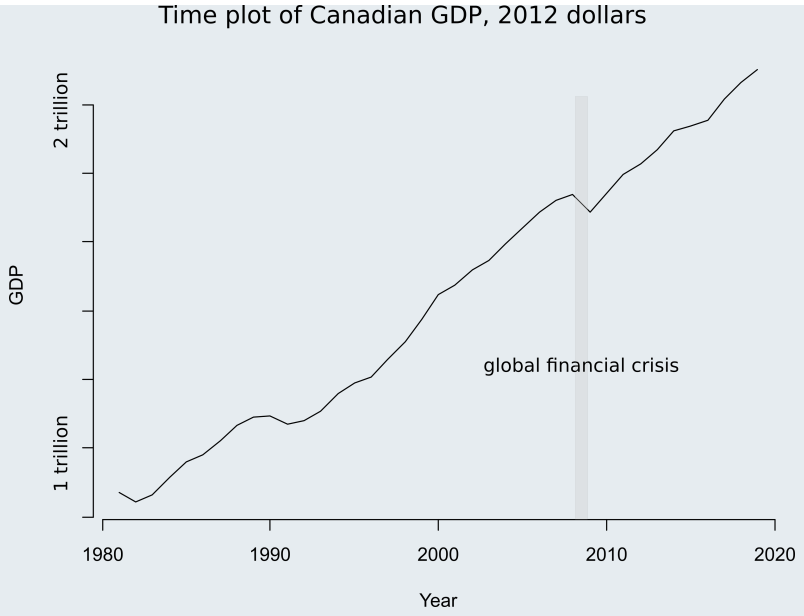
4.6 Time plots

A time-series is a single random variable that is measured repeatedly over time. A nations GDP is constantly changing over time, we might measure it quarterly or yearly. A households electricity consumption, CO_2 emissions, the price of a stock, the temperature in Winnipeg - these are all variables that could be recorded at different points over time.

Graphing these variables, with time on the x-axis, is called a *time plot*. Time plots can be useful to see how a variable evolves. Time plots can dramatically illustrate the effect of a timely event on a random variable, or can illustrate how quickly something is growing or declining.

Example 4.8 — Time plot of GDP. The data is from Statistics Canada[1], and contains GDP by year in millions of 2012 dollars:

```
gdp <- read.csv("http://ryantgodwin.com/data/canada-gdp.csv")
plot(x = gdp$Year, y = gdp$GDP,
     type = "l", xlab = "Year", ylab = "GDP")
```



We used the `plot` function, and chose:

- the x-axis to be the year (`x = gdp$Year`)
- the y-axis to be GDP (`y = gdp$GDP`)
- lines instead of dots (`type = "l"`)
- labels for the x- and y-axes (`xlab = "Year", ylab = "GDP"`)

Notice how each option in the `plot` function is separated by a comma.

4.6.1 Logarithms in time plots

The logarithm is the inverse to the exponent. If a variable is growing exponentially over *time* (for example), then taking the logarithm will *undo* the exponential growth and make the variable's relationship to time appear *linear*.

When a variable is said to grow in percentage terms, this implies *exponential growth*. For example, suppose Mars GDP grows by 6% on average, per year. This means GDP is growing exponentially. Starting at 11 billion in year 1, what is the expected GDP in year 10? In year 40?

$$GDP_{year=10} = 11 \times (1.06)^{10} = 19.7$$
$$GDP_{year=40} = 11 \times (1.06)^{40} = 113.1$$

GDP is really accelerating! Notice in the formula that the year is an *exponent*, so GDP is growing exponentially over time. If we take the *logarithm* of both sides of the equation:

$$\log(GDP_{year=40}) = \log(11) + 40 \times \log(1.06)$$

then $\log(GDP)$ (to any base) is growing *linearly* over time! (The 40 no longer appears as a "power"). This is extremely useful for graphing variables that grow exponentially[5]. To see this, load some data on Mars GDP:

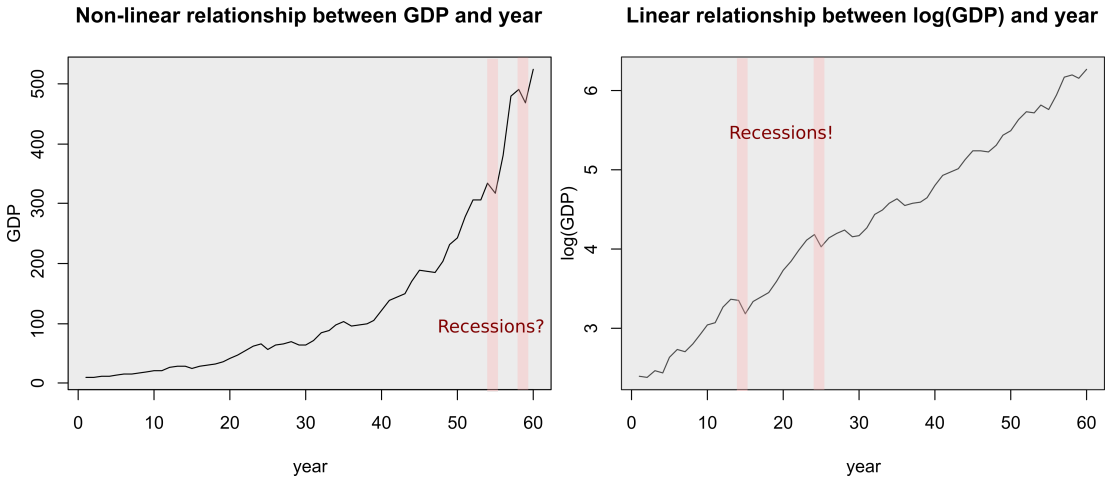
[5]It is also useful when trying to fit "straight line" models to non-linear relationships.

```
marsGDP <- read.csv("http://ryantgodwin.com/data/marsGDP.csv")
```

Create a time plot of GDP:

```
plot(marsGDP$time, marsGDP$GDP, type="l",
     main = "Non-linear relationship between GDP and year",
     xlab = "year", ylab = "GDP")
```

Figure 4.9: Mars GDP. It is difficult to locate the years in which recessions took place with taking GDP in logs.



In the left pane of Figure 4.9, it is difficult to see the values of GDP at the beginning of the time period. It looks like a smooth ride! This is because, by the end of the sample (year 60), the values for GDP are very large, making the scale of the y-axis unhelpful for seeing what is happening with GDP around year 10. Looking at the left pane of Figure 4.9, it appears that there were two recessions at the end of the sample. Let's now put the log of GDP on the y-axis instead (right pane of Figure 4.9):

```
plot(marsGDP$time, log(marsGDP$GDP), type = "l",
     main = "Linear relationship between log(GDP) and time",
     xlab = "year", ylab = "log(GDP)")
```

After graphing the *log* of GDP, we see a linear relationship. This shows that GDP is growing constantly over time. It also allows us to see that GDP at the beginning of the time period was actually quite tumultuous! The two major recessions occurred near the beginning of the time period, not the end. This was only visible after taking logs.

To summarize, if a variable is growing exponentially (or is growing with a constant percentage increase), then a common trick for visualizing such a variable is to take logs.

4.7 Scatter plots

A scatter plot can be used to visualize the *relationship* between *two* quantitative variables. Sometimes, one variable is suspected to *cause* or determine the other variable.

By looking at a scatter plot we can comment on the *strength*, *form*, and *direction* of the relationship between two variables.

In this section, we will:

- Define the dependent and explanatory variable.
- Describe the *strength*, *form*, and *direction* of a relationship when looking at a scatter plot.
- Graph scatter plots in R, and use a categorical variable to add colour.

4.7.1 Explanatory and dependent variables

Of the two variables in the scatter plot, the explanatory variable is the one that is suspected to *cause* or *determine* the dependent variable. Usually, the symbols "x" and "y" are used when referring to these two variables.

Explanatory and dependent variables.

- x - the explanatory variable. It is not caused or determined by y. The explanatory variable will appear on the x-axis of the scatter plot.
- y - the dependent variable, also called the response variable. It is thought to be caused, or is at least explained, by the x variable. The dependent variable appears on the y-axis of the scatter plot.

4.7.2 Points on a scatter plot

Figure 4.10: Happiness score and log GDP per capita.

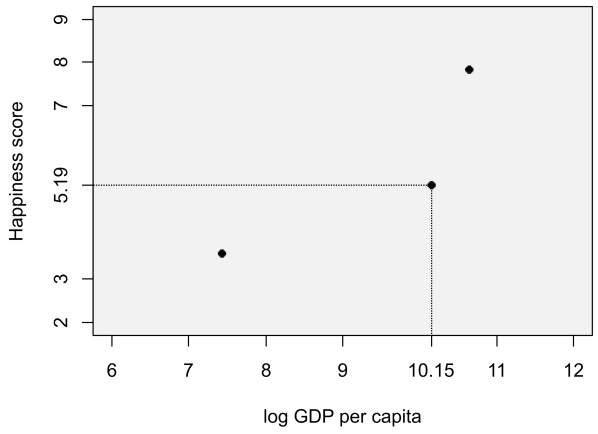


Table 4.1: Happiness score and log GDP per capita.

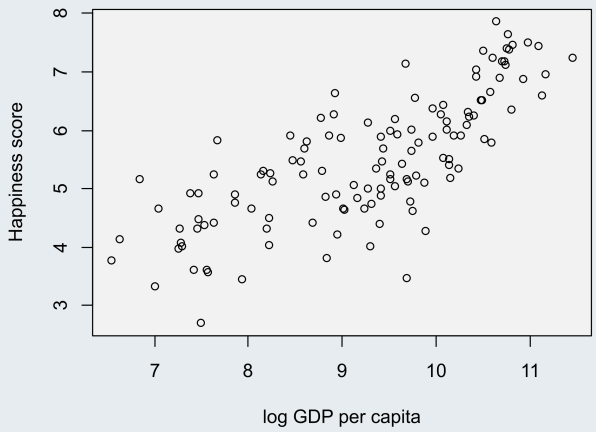
Country	Happiness score	log GDP per capita
Finland	7.86	10.64
Turkey	5.19	10.15
Haiti	3.62	7.42

Each point on the scatter plot represents a single observation (a row in the data set, see Section 4.2). The position on the plot is determined by the values of the dependent and explanatory variables; these values provide the *coordinates*.

Using the World Happiness report, Table 4.1 shows the average happiness score, and log GDP per capita, for a few countries. When dealing with GDP, it is common to use the *log*. Hypothesizing that GDP may cause happiness, we'll call "Happiness score" our dependent variable (the y variable) and "GDP per capita" our explanatory variable (the x variable). The three observations in Table 4.1 are plotted in Figure 4.10. (Make sure you can locate all the points!) By plotting all 127 countries in the data set, the scatter plot will show us whether there is a relationship between the two variables, and allow us to comment on the strength, form, and direction of the relationship.

Example 4.9 — Scatter plot for happiness and GDP per capita. Load the Happiness data, and create the scatter plot:

```
mydata <- read.csv("http://ryantgodwin.com/data/happiness.csv")
plot(x = mydata$Log.GDP.per.capita, y = mydata$Happiness.score,
     xlab = "log GDP per capita", ylab = "Happiness score")
```



Describe what you see using terms like *strength*, *form*, and *direction*.

- There is a fairly *strong* relationship between the two variables (the data points are quite tightly packed together, rather than being spread out).
- The *form* seems to be linear (rather than non-linear).
- There is a positive (direct) relationship between the two variables. When one variable increases, so does the other (rather than a negative or indirect relationship where the values move in opposite directions).

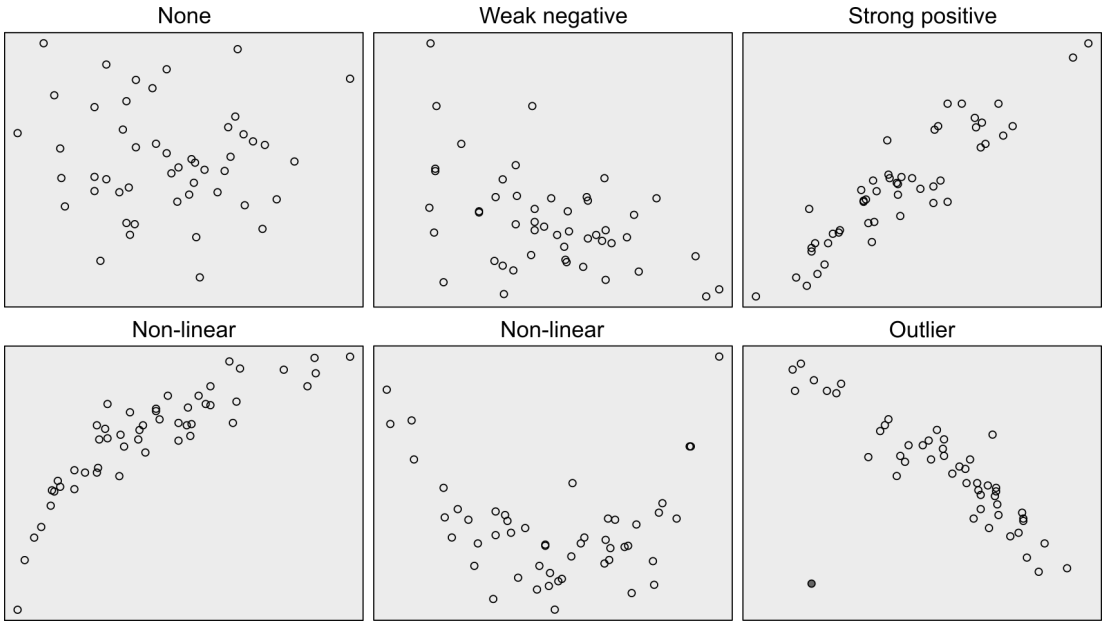
4.7.3 Scatter plots: types of relationships

From the scatter plot we can sometimes tell if:

- There is no relationship between the variables.
- The relationship is strong or weak.
- There is a positive or negative relationship.
- The relationship is linear or non-linear.
- There are outliers.

Figure 4.11 illustrates some of these possibilities.

Figure 4.11: Types of relationships between variables.



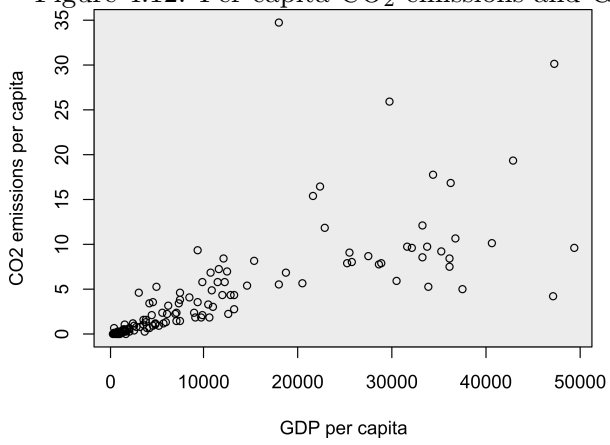
4.7.4 Categorical variables in scatter plots

Recall that categorical variables compartmentalize each observation into one of a few categories. Using colours to denote a category, the information contained in these variables can be made visible in a scatter plot. Colour-coding (or symbol-coding) each variable can reveal interesting patterns in the data.

Let's create a scatter plot of per capita CO_2 emissions, and GDP per capita (data is from 2007). We will hypothesize that CO_2 emissions is the *dependent* variable. Load the data, and create the plot:

```
co2 <- read.csv("http://ryantgodwin.com/data/co2.csv")
plot(co2$gdp.per.cap, co2$co2,
     ylab = "CO2 emissions per capita", xlab = "GDP per capita")
```

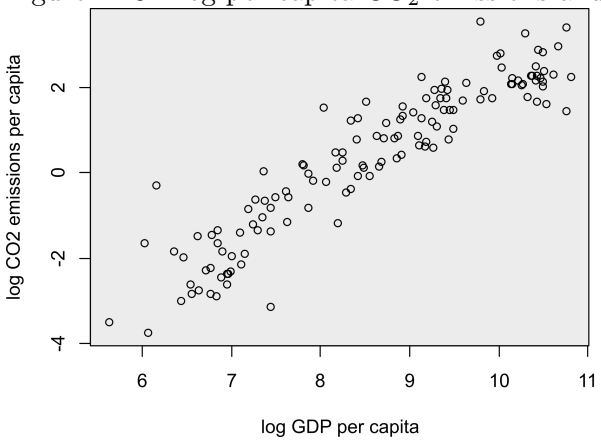
Figure 4.12: Per capita CO_2 emissions and GDP.



A problem with Figure 4.12 is that there are some very large values for CO_2 leading to a scale for the graph that makes it difficult to see what is happening for the majority of countries. As in Section 4.6.1, a trick for handling this is to take the *logs* of the variables[6]. We can do this easily in R:

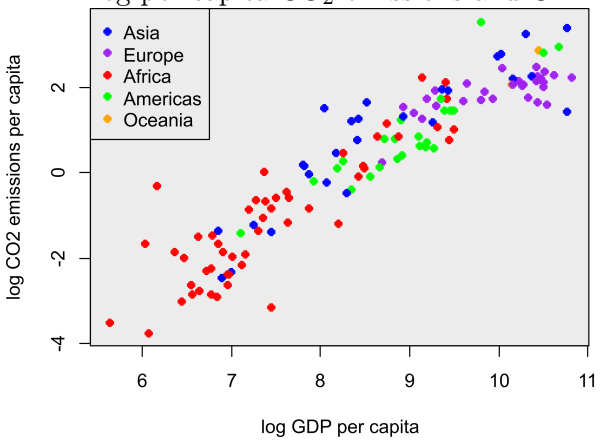
```
plot(log(co2$gdp.per.cap), log(co2$co2),
     ylab = "log CO2 emissions per capita", xlab = "log GDP per capita")
```

Figure 4.13: Log per capita CO_2 emissions and GDP.



In Figure 4.13, it is much easier to see that there is a *strong* and *positive* relationship between per capita CO_2 emissions and per capita GDP.

Figure 4.14: Log per capita CO_2 emissions and GDP by continent.



Finally, let's add colour to the scatter plot, by giving each point on the plot a different colour based on the country's *continent*. "Continent" is a qualitative variable in the data set that places each country in 1 of 5 categories. From Figure 4.14 we can now see that, compared to other countries with similar GDP, the Americas have fewer CO_2 emissions. These types of revelations occurs much more easily when colour

[6]Taking the logs of both variables leads to an approximate percentage change interpretation. That is, a percentage increase in GDP will be associated with a percentage increase in CO_2 emissions (approximately).

(or symbol) coding scatter plots using qualitative variables. The R code necessary for adding colour to the scatter plot is provided in Example 4.10.

Example 4.10 — Colour coding CO_2 emissions by continent. First, load the data:

```
co2 <- read.csv("http://ryantgodwin.com/data/co2.csv")
```

The first few observations in the data look like:

Country	Continent	CO_2	GDP per capita
Afghanistan	Asia	0.085	974.58
Albania	Europe	1.30	5937.03
Algeria	Africa	3.19	6223.37
Angola	Africa	1.20	4797.23

We need to create a `colour` variable that will control the colour of each data point. We begin this by initializing a colour variable:

```
colour <- character()
```

and then assigning it values based on "continent":

```
colour[co2$continent == "Africa"] <- "red"
colour[co2$continent == "Americas"] <- "green"
colour[co2$continent == "Asia"] <- "blue"
colour[co2$continent == "Europe"] <- "purple"
colour[co2$continent == "Oceania"] <- "orange"
```

Now we create the scatter plot, choosing the colour of each data point using the variable we have created:

```
plot(log(co2$gdp.per.cap), log(co2$co2),
     ylab = "log CO2 emissions per capita", xlab = "log GDP per capita",
     col = colour, pch = 16)
```

and then add a legend to explain what the colours mean:

```
legend("topleft",
       legend = unique(co2$continent),
       col = unique(colour), pch = 16)
```

This reproduces Figure 4.14. There are much easier ways to accomplish colour coding in R, for example by using the `ggplot2` downloadable extension for R. This example instead serves to illustrate the principle behind colour coding in a scatter plot: linking each possible value in a qualitative variable to a unique colour.

5. Describing distributions with statistics

A statistic is a numerical value that is a function of the sample data. When we say "function of the sample data," we mean a formula, algorithm, set of rules, etc. that uses the information in the data. Statistics can be used to *describe* a distribution. Some of the visual descriptors from the previous chapter, such as *location*, *spread*, and *skew*, can actually be measured using a numerical value.

Some statistics that we will cover in this chapter are:

- sample mean
- median
- interquartile range
- p^{th} percentile
- sample variance and standard deviation
- sample correlation

5.1 Sample mean

The sample mean (also called the sample average, arithmetic average, or average) is calculated by adding up all the values in the variable, and dividing by the sample size. The sample size (the number of rows in the data set, or the number of values in a variable) is usually denoted n.

If the variable y has values $y = \{6, 2, 5, 6, 1\}$, then the sample average is calculated by:

$$\bar{y} = \frac{6 + 2 + 5 + 6 + 1}{5} = 4$$

We divided by 5 because there are 5 observations in the variable ($n = 5$). The sample mean of the variable y is denoted $\bar{y}$.[1] The sample mean of a variable called *income* (for example) would be denoted $\overline{income}$.

[1] The symbol $\bar{y}$ is pronounced "y bar".

The general formula for calculating the sample mean is:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.1}$$

where y_i denotes the i^{th} observation, and where n denotes the sample size. The symbol Σ tells you to add, starting at the 1st observation ($i = 1$) and ending at the last (n). Equation 5.1 is a very common statistic, and should already be very familiar to you.

Example 5.1 — Sample mean in R. Load the variable $y = \{6, 2, 5, 6, 1\}$ into R using:

```
y <- c(6, 2, 5, 6, 1)
```

Calculate the sample mean of y using the `mean()` function:

```
mean(y)
```

```
[1] 4
```

Example 5.2 — Sample mean in R. Load some Mars data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

We can calculate the sample mean of "income" using the `mean()` function:

```
mean(mars$income)
```

```
[1] 80938.1
```

What does the sample mean $\bar{y}$ tell us? For one, it is an estimate of the *true population mean*. The true population mean is the "centre" of the distribution (e.g. the centre of a bell curve) that is generating the values for the variable. The *true population mean* of y is the value that we *expect* to observe for y.

The sample mean gives us an idea about the centre or location of the variable's distribution, and is called a "measure of central tendency." The sample mean is the "centre of mass" of the variable. That is, if the histogram of the variable were a physical object, the mean would be the location where we could balance the object on one finger along the x-axis.

The sample mean is one of the most important statistics, because it defines a very important feature of a distribution: its location.

5.2 Sample median

Another measure of "central tendency" is the *sample median*. The sample median is the "middle" observation. The sample median is the value for which half of the other values are smaller, and the other half are larger. That is, for a variable y, the median of y is where:

$$50\% \text{ of values} \leq \text{median}(y) \leq 50\% \text{ of values} \tag{5.2}$$

Whether the inequality is $<$ or $\leq$ in Equation 5.2 depends on whether the sample size n is *odd* or *even*. The algorithm for finding the median is as follows:

1. Order the observations in the sample from smallest to largest.
2. Label the smallest observation the 1st observation, the second smallest the 2nd observation, all the way to the nth observation.
3. Find the middle observation(s) in the ordered list.
    - If n is odd, the middle observation is $(n+1)/2$. This is the median.
    - If n is even, there are *two* middle observations. The median is the sample mean of these two middle observations. (The two middle observations are the $(n/2)^{th}$ and $[(n+1)/2]^{th}$.)

Example 5.3 — Sample median of y. Take the variable $y = \{6, 2, 5, 6, 1\}$ again. To calculate the median, we start by ordering the variable:

$y_{ordered} = \{1, 2, 5, 6, 6\}$

The sample size is odd ($n = 5$) so the middle observation is the 3rd observation ($(n+1)/2 = 3$). Finding the 3rd observation in the ordered variable gives us the median at 5.

Example 5.4 — Sample median of *income*. Load the Mars data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

We can calculate the sample median of "income" using the `median()` function:

```
median(mars$income)
```

```
[1] 70094
```

The sample median is important for similar reasons that the sample mean is important. It is a defining feature of the true underlying distribution that is generating or describing the data that we observe. In addition, it gives an idea about the "centre" or "middle" of the distribution of a random variable.

5.3 Comparing sample mean and median

The sample mean and sample median are both "measures of central tendency," but they have some pretty important differences.

For one, the sample mean is arguably more important than the sample median when characterizing the true underlying distribution. This is because many statistical distributions are defined by their mean. Later we will see that the mean (μ) is one of two parameters that define the bell curve.

Second, the median is perhaps a more intuitive concept and may be easier to understand. The median is the middle where 50% of values are below and the other 50% are above. This is easy to understand. In contrast, the mean is a bit more abstract, using concepts such as "centre of mass" or "expected value". However, when it comes to actually calculating the number, the mean is easier; just add them all up and divide by n.

Third, the sample median is unaffected by extreme values or outliers, whereas the mean is. For example, as long as 50% of the values are larger, the median will not change even if those 50% of values are located close to the median, or way far out in

the tail. That is, once the median has been found, all the values to the left (or right) of the median could be stretched out or rearranged and the median would be unchanged.

Example 5.5 — Resistance of the median to outliers. Take the ordered y variable from Example 5.2: $y = \{1, 2, 5, 6, 6\}$. The sample mean and median of y are:

$$\bar{y} = 4$$
$$\text{median}(y) = 5$$

Now, let's try changing the last value in the y variable so that it is an outlier, for example let:

$$y = \{1, 2, 5, 6, 100\}$$

Calculate the sample mean and median using R:

```
y <- c(1, 2, 5, 6, 100)
mean(y)
median(y)
```

```
> mean(y)
[1] 22.8
> median(y)
[1] 5
```

The sample mean has been drastically affected by this outlier (it went from 4 to 22.8), and the sample median has remained unchanged.

In a symmetrical distribution, the mean and median are always the same. In an asymmetrical distribution, they are always different. For example, in a right skewed distribution, the mean will always be greater than the median. If outliers are suspected to be in the data set, the median might be a safer measure of "central tendency" since the sample mean can be greatly swayed by extreme values.

5.4 Percentiles and quartiles

The median is the 50th percentile, and is also the 2nd quartile. The median divides the distribution into *two*. We could also divide the distribution into *four* and get *quartiles*, or we could divide the distribution 1% at a time and get *percentiles*(if we have enough observations). Percentiles and quartiles are a natural extension to the median; the median is just a special case.

5.4.1 Percentiles

To calculate a percentile, we again start by arranging the values of the variable in increasing order. Then, we count to the required percentage starting at the first observation. For example, the 20th percentile would be the $(0.2 \times n) + 1$ observation of the ordered variable. For a sample size of $n = 101$ for example, this would be the 21st largest value of the variable. 20% of the values would be smaller, 80% of the values would be larger.

Similar to the median, there may not be an exact correspondence between the desired percentile and the observation number in the ordered list. In this case, we

would take the sample mean of two values instead. For example, if $n = 100$, the 20th percentile would be the sample average of the values for the 20th and 21st observation.

Percentiles can be used to measure the *spread* of a variable. A 5% probability (a 1 in 20 chance) is a common value chosen in statistics for classifying an "extreme" event. We might wonder, in the extreme, what is the best and worst that could happen? The mean height of a person may be 1.65m, but that doesn't tell us anything about the *extremes* or *spread* of the distribution. What height marks the shortest 5%? At what income level are the top 5% of earners above?

Finally, *quantiles* are very similar to *percentiles*. A quantile is just expressed in different units (not in percentage points but as a real number between 0 and 1). For example, the 20th percentile is the 0.2 quantile.

Example 5.6 — Top and bottom 5% of Mars income earners. Load the Mars data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

To find the 5th percentile (the value for which approximately 5% of incomes are smaller):

```
quantile(mars$income, 0.05)
```

```
     5%
39092.15
```

So, 39092.15 is the 5th percentile of income. To find the income that marks the *top* 5% of earners, we can use:

```
quantile(mars$income, 0.95)
```

```
> quantile(mars$income, 0.95)
    95%
157921.2
```

5.4.2 Quartiles

Quartiles break up a distribution into four quarters. That is, one-quarter of the values will fall into each quartile. The 1st, 2nd, and 3rd quartiles correspond to the 25th, 50th, and 75th percentiles, respectively. The 2nd quartile is the same as the median. The first quartile is found by ordering the values, and then counting to the $(0.25 \times n) + 1$ observation. Again, two values might need to be averaged if $(0.25 \times n) + 1$ is not an *integer*. Similarly, the 3rd quartile is the $(0.75 \times n) + 1$ ordered value, and we already know how to find the 2nd quartile (the median).

Quartiles are more common than percentiles, and are a simple way to summarize the spread and shape of a distribution. When *summarizing* a variable, it is common to report the values for the 1st, 2nd, and 3rd quartiles, as well as the sample mean. The values for the quartiles tell us if the distribution is *skewed*, and in which direction, and can help to select the right distribution in order to characterize a variable.

Example 5.7 — Quartiles of income. Load the Mars data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

To find the quartiles of "income" we can ask for the 25th, 50th and 75th percentiles using the `quantile()` function:

```
quantile(mars$income, c(.25, .5, .75))
```

```
    25%     50%     75%
53516.25 70094.00 96815.00
```

So, the quartiles of income are $\{53516, 70094, 96815\}$. What does this tell us? Notice that the gap between the 1st quartile and the median (approximately 17k) is smaller than the gap between median and 3rd quartile (approximately 27k). In a symmetrical distribution, these gaps would be equal. The distribution is skewed to the right.

5.5 Min and Max

The minimum value and maximum value of a variable are sometimes reported. Similar to median, percentiles, and quartiles, we find the min by sorting the values of the variable in ascending order and selecting the 1st observation. Likewise, the max is the nth observation.

Example 5.8 — Min and max of Mars incomes. Load the Mars data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

The minimum income in the data is 24973, and the maximum is 358318. To find this in R use:

```
min(mars$income)
max(mars$income)
```

```
> min(mars$income)
[1] 24973
> max(mars$income)
[1] 358318
```

5.6 Summary of a variable

A variable is often "summarized" using some of the statistics that we have defined. In particular, the sample mean, quartiles, and min and max can be reported to provide a numerical characterization of the distribution of a variable.

Example 5.9 — Summary of Mars incomes. Load the Mars data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

Use the `summary()` command:

```
summary(mars$income)
```

```
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  24973   53516  70094  80938   96815 358318
```

Notice how several statistics from the previous few examples have all been calculated

under the `summary()` command.

5.7 Sample variance

The sample variance (and the closely related standard deviation) is a very common and important measure of the "spread" of a variable. It is very important for at least two reasons. (i) Along with the (population) mean, the (population) variance is a defining feature for most distributions. That is, if you know the mean *and* the variance, you can draw most statistical distributions in a plot (for example the bell curve). (ii) Sample variance provides a numerical measure of the average distance between each value and the centre of the distribution. Sample variance quantifies the chance of "extreme" values occurring.

The formula for calculating the sample variance of a variable y is:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5.3}$$

Similar to how we used the symbol $\bar{y}$ to denote the sample mean, we also use a symbol to denote the sample variance: s_y^2. If we were calculating the sample variance of *income* we would denote it s_{income}^2. The summation operator Σ is again telling us to add something up, starting at the first observation and ending at the last. This time, however, we are subtracting the sample mean from each observation, squaring that "distance", and then adding up all of these squared "distances".

Notice that sample variance is essentially a measure of *distance*.[2] Each value in the variable is compared to the sample mean $(y_i - \bar{y})$. This is measuring how far away the values tend to be from the "centre". However, we want to combine all of these distances into a single measure, so we add them up. But if we just added up all of the the $(y_i - \bar{y})$, negative distances would cancel out the positive distances!

To avoid this, we could take the *absolute values* of the distances: $|y_i - \bar{y}|$. This would lead to an alternative measure of the *spread* or *dispersion* of a variable, called the "Mean Absolute Deviation." The sample variance is a more popular measure of dispersion, and instead of taking *absolute* distances, we take squared distances: $(y_i - \bar{y})^2$. The squaring in the formula means that all distances are now positive, but that variance is very sensitive to large values in the data. As a value gets further and further away from the sample mean, the squared distance gets even further. Note that the "square" in the Equation 5.3 means that sample variance can *never* be negative. The smallest possible value for 5.3 is 0. A 0 can *only* occur when all of the values for y_i are identical (and hence there is no variation).

When we calculated the sample mean, we added everything up and then divided by n. Here we are instead dividing by $n-1$. Why? The reason is somewhat complicated, and we will not go into depth in this book. Instead, we will provide a cursory treatment of the topic of degrees of freedom in order to understand this $n-1$ in the formula for sample variance.

Degrees of freedom

Degrees of freedom can be thought to account for the number of *independent* pieces of information available when calculating a statistic. In Equation 5.3, notice that the

[2] In particular, sample variance involves the *squared Euclidean distance*.

formula for the statistic s_y^2 involves another statistic ($\bar{y}$)! Having the $\bar{y}$ on the right-hand-side of the formula for s_y^2 turns out to cause a distortion, and one degree of freedom is lost. Instead of n pieces of *sample* information, there are now only $n-1$ pieces of information available when calculating s_y^2.

This can be seen in a simple example. Take the variable $y = \{1, 3, ?\}$, and the sample mean of y at $\bar{y} = 3$. Can you figure out the missing y value? Good job! Together with $\bar{y}$, only 2 out of the 3 sample values ($n-1$) actually provide any unique information.

Example 5.10 — Sample variance of y. We'll use the variable $y = \{6, 2, 5, 6, 1\}$ again, and calculate the sample variance. First we need to calculate $\bar{y} = 4$. We take each of the values in the variable, subtract the mean, square the difference, and add al the squared differences:

i	y_i	$\bar{y}$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	6	4	2	4
2	2	4	-2	4
3	5	4	1	1
4	6	4	2	4
5	1	4	-3	9
				22

Finally, we divide the sum by by $n - 1 = 4$ to get the sample variance of $s_y^2 = 22/4 = 5.5$. We can also easily calculate this sample variance in R:

```
y <- c(6, 1, 2, 5, 6)
var(y)
```

```
> var(y)
[1] 5.5
```

Example 5.11 — Sample variance of Mars incomes. Load the Mars data and take the sample variance:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
var(mars$income)
```

```
> var(mars$income)
[1] 1605382317
```

This is quite a large number! What does it tell us? It is difficult to interpret this number, unless we compare it to some other distribution. For example, we could calculate the sample variance for Earth incomes, and see which distribution is more spread out.

5.8 Sample standard deviation

Standard deviation is the square root of variance. The formula to calculate the sample standard deviation for a variable y is:

$$s_y = \sqrt{s_y^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.4}$$

The standard deviation is obviously closely related to the sample variance, and often the two are used interchangeably. An important difference, however, is that s_y has the same units of measurement as y (whereas s_y^2 does not). Sometimes, the value of a variable is compared to the number of "standard deviations" it is away from the sample mean. This can provide an idea of how "extreme" a value is.

Example 5.12 — Standard deviation of Mars incomes. What is the standard deviation of Mars incomes? We know from Example 5.11 that:

$$s^2_{income} = 1,738,740,548$$

so that the standard deviation is:

$$s_y = \sqrt{s_y^2} = \sqrt{1738740548} = 41698.21$$

We can also get this number straight from R:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
sd(mars$income)
```

```
[1] 40067.22
```

5.9 Skewness and Kurtosis

When calculating $\bar{y}$ we added up the y_i values (they were to the power of 1). When we calculated s_y^2 we added up squared differences $(y_i - \bar{y})^2$ (they were to the power of 2). We could also add up cubed differences, and differences to the power of 4, or as high as we like! A statistic involving cubed differences is the skewness of the variable:

$$\text{skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} \tag{5.5}$$

We have already encountered the concept of skewness as a visual descriptor of a distribution. Equation 5.5 is a way of quantifying skewness. For example, a positive value for sample skewness means the right tail of the distribution is relatively more stretched out. The magnitude of sample skewness measures how stretched the tail is.

Similarly, kurtosis is a statistic that involves differences to the power of 4, but we do not report the formula here. Like variance and skewness, kurtosis is a measure of the shape of a distribution. Kurtosis measures whether the tails of the distribution are fat or slim. The higher the kurtosis number, the fatter the tails.

Skewness and kurtosis are more abstract and less intuitive than mean and variance. We introduce them here in order that we may later talk about the Jarque-Bera test for Normality, in Section 13.3.1.

5.10 Correlation

Correlation is a measure of the relationship between two variables. Correlation measures:

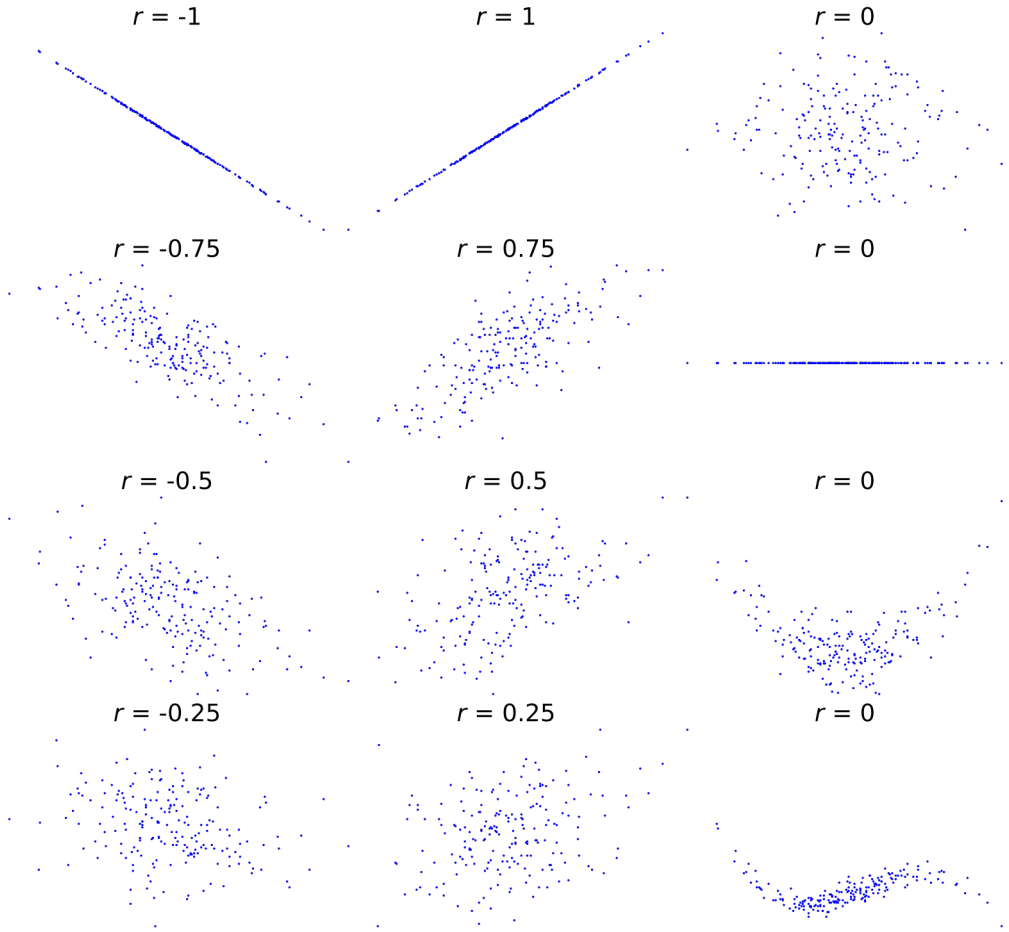
- how two variables *move* or *vary* in relation to each other.
- the *direction* of the relationship between two variables.
- the *strength* of the relationship between two variables.

The equation for the sample correlation between two variables (x and y for example) is:

$$r_{xy} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \tag{5.6}$$

Lower case "r" is used to denote the sample correlation coefficient. s_x and s_y are the sample standard deviation of x and y (see Section 5.8).

Figure 5.1: Scatterplots and sample correlations for x and y variables.



Correlation measures how often and how far two variables differ from their sample mean value (notice the $x_i - \bar{x}$ and $y_i - \bar{y}$ terms in Equation 5.6). If both variables tend to be larger than their mean at the same time, then correlation will be positive. If when one variable is larger than its mean, the other tends to be smaller than its mean, correlation will be negative. The larger the magnitude of the correlation number, the more often this statement holds true for specific pairs of values.

Correlation. Correlation is a measure of the direction (either negative or positive) and strength (between -1 and 1) of the association between two variables.

If the correlation is positive, then when one variable is larger (or smaller) than its mean, the other variable tends to be larger (or smaller) as well. The larger the magnitude of covariance, the more often this statement tends to be true. Covariance tells us about the direction and strength of the relationship between two variables.

Note the following properties of r_{xy}:

- r_{xy} is a measure of the *linear* relationship between x and y. Non-linear relationships cannot be quantified using correlation.
- $r_{xy} = 0$ implies that x and y are *linearly* independent.
- If x and y are *independent* (neither variable causes the other), then $r_{xy} = 0$. The converse is not necessarily true.
- Correlation is bound between -1 and 1. That is, $-1 \leq r_{xy} \leq 1$.
- A positive covariance means that the two variables tend to differ from their mean in the *same* direction.
- A negative covariance means that the two variables tend to differ from their mean in the *opposite* direction.
- $r_{xy} = 1$ means perfect positive linear association between x and y.
- $r_{xy} = -1$ means perfect negative linear association between x and y.

Correlation is a basic, and extremely common way to quantify the relationship between two variables. The correlation coefficient is almost always reported when discussing two variables that are thought to be associated.

Correlation is a way to quantify some of the association between variables that we can "see" in a scatterplot, and you may have recognized the same terms *direction* and *strength* being used in the section as were used in Section 4.7. Figure 5.1 shows several scatterplots along with the sample correlation between the variables being plotted.

Example 5.13 — Sample correlation. To calculate a sample correlation in R, use the `cor()` function:

```
cordata <- read.csv("http://ryantgodwin.com/data/cordata.csv")
cor(cordata$x, cordata$y)
```

```
[1] -0.7467756
```

The scatterplot for this data is shown in Figure 5.1. The sample correlation of -0.75 tells us that there is a negative or inverse relationship between x and y, and that the relationship is quite strong.

Using the Mars data, calculate the sample correlation between income and education:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
cor(mars$income, mars$years.education)
```

```
[1] 0.4552673
```

With a correlation of $r = 0.46$, there is a fairly strong and positive relationship between education and income. That is, when education tends to be higher (than the sample mean value) so does income.

6. Density curves

In this chapter, we introduce the concept of a probability distribution function (also called a density curve), and discuss several related topics. This chapter assumes a basic understanding of the concept of probability. In the next chapter, we go into greater detail on the meaning and rules of probability and randomness. In addition to using density curves when we discuss probability, we will also make heavy use of them later on for *hypothesis testing*, and in particular for p-values.

6.1 Probability distributions (densities)

A "probability distribution function", is also called a "density curve", or just "density". It is a mathematical way of modelling a variable's distribution. A probability function is an equation (it can also be a graph or table), which contains information about a random variable. The nature and properties of the randomness determines what type of equation is appropriate (for example a bell curve, or something else).

Density curve / probability distribution / probability function. The probability function accomplishes two things: (i) it lists all possible numerical values that the random variable can take, and (ii) assigns probabilities to ranges of values.

Areas under the distribution / density

A range of values on the x-axis corresponds to an area underneath the probability distribution. The area is the probability that the random variable will take on a value in the x-axis range. This means that, in the long run, an area under the density curve is equal to the proportion of values that fall in that region. Since the distribution / density function lists all possible numerical values that the variable can take, the area under the *entire* density curve must sum to 1.

6.2 Continuous uniform distribution

The continuous uniform distribution is a starting point in the illustration of probability distributions. The uniform distribution is defined by it's endpoints a and b. For

simplicity, let $a = 0$ and $b = 1$. If a variable y follows this distribution, we write:

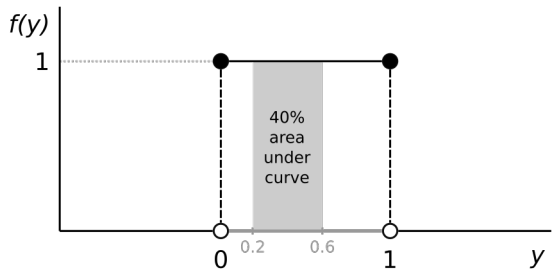
$$y \sim U_{[0,1]}$$

Such a variable has an *equal* probability of taking on any value in the interval [0,1]. The probability distribution can be written as:

$$f(y) = \begin{cases} 1 & \text{for } 0 \leq y \leq 1 \\ 0 & \text{for } y < 0 \text{ or } y > 1 \end{cases} \tag{6.1}$$

Equation 6.1 defines the height of the density curve for any value of y, and is depicted in Figure 6.1. To calculate the probability of y taking on a certain value in a range, we calculate the area under the density curve. For example, if we want to know the probability that y will be between 0.2 and 0.6, we calculate the area under the density curve, between 0.2 and 0.6. This area, and probability, is height $\times$ width $= 1 \times (0.6 - 0.2) = 0.4$.

Figure 6.1: Density curve for a uniform U(0,1) distribution. The area under the density curve represents the probability that y will be between 0.2 and 0.6.



6.3 Discrete uniform distribution

The discrete uniform distribution describes random variables that have an equal probability of taking on a finite number of values, for example a coin flip or a die roll. The following chapter will make extensive use of this distribution as a simple setting in which to explore and discuss various topics on probability and randomness. If y is the result of a die roll ($y = ⚀, ⚁, ⚂, ⚃, ⚄,$ or $⚅$), then y follows a discrete uniform distribution and we can write:

$$y \sim U\{1, 6\}$$

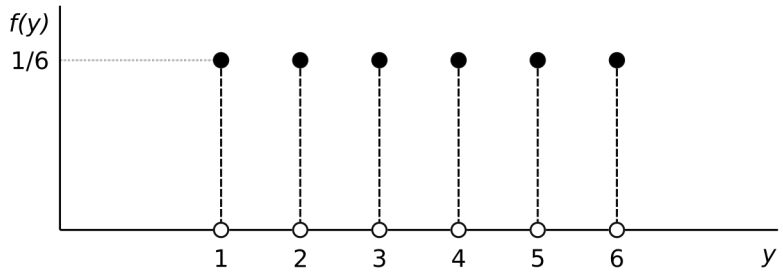
If the variable is *discrete*, then probabilities are determined by the *height* of the density, *not* the area under the density curve. Figure 6.2 shows the density function for a die roll that follows a discrete uniform distribution $U\{1, 6\}$.

6.4 The Normal distribution

The Normal distribution is an important probability distribution. It is important because it describes many different random variables. The probability function for a normally distributed random variable y is:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{6.2}$$

Figure 6.2: The discrete uniform distribution describes a die roll. For a discrete variable, the height of the distribution is the probability of y taking a value.



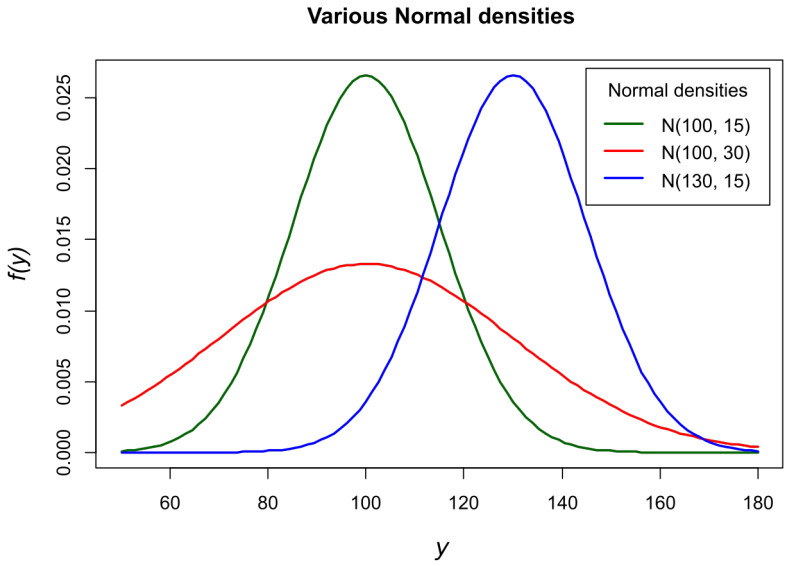
Equation 6.2 can look a little scary. But this is just the bell curve! If you plug in a y-value, you get a height on the Normal (bell) curve. If you plug in many y values into this equation, you can trace out the curve. Equation 6.2 has two *parameters*: μ (the mean) and σ (the standard deviation). These parameters control the location and shape of the curve.

If a variable y follows a Normal distribution we can write:[1]

$$y \sim N(\mu, \sigma)$$

Figure 6.3 shows Normal distributions for three different means and standard deviations: $N(100, 15)$, $N(100, 30)$, and $N(130, 15)$.

Figure 6.3: The mean (μ) controls the location of the normal distribution, and the standard deviation (σ) controls the shape.

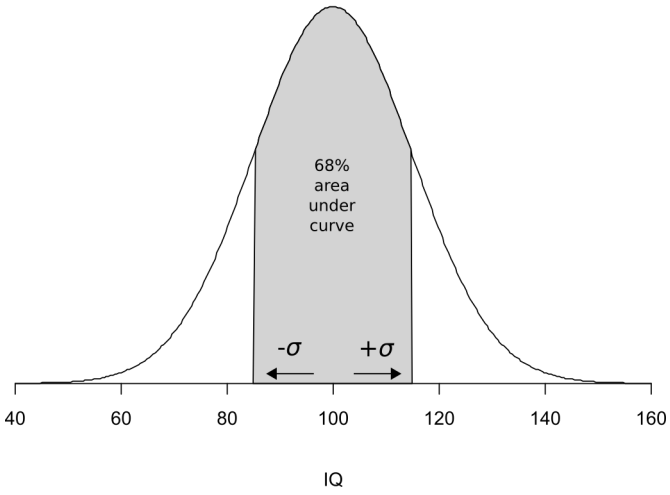


6.4.1 Areas under the Normal density

An area under a probability distribution function (density curve) is the probability of an observation lying in that range of values. Alternatively, it is the portion of times the

[1] It is also common to write the normal distribution in terms of its *variance* (σ^2) instead of its *standard deviation* (σ): $N(\mu, \sigma^2)$.

Figure 6.4: The probability of $85 \leq y \leq 115$ is an area under the Normal density.



variable is in the specified range. Calculating an area under the Normal distribution is trickier than, for example, the continuous uniform distribution, and requires integration (not covered in this book).

For example, suppose that $y \sim N(100, 15)$ and we wish to know the probability of y being between 85 and 115. This probability is the area under the $N(100, 15)$ curve shown in Figure 6.4. Note that, in this example, the range of values (85 to 115) happens to be plus-and-minus one standard deviation around the mean of the distribution: $\mu \pm \sigma = 100 \pm 15 = [85, 115]$.

6.4.2 68-95-99.7

The Normal distribution has an interesting property. No matter what the mean (μ) or variance (σ^2) of the Normal distribution, the area under the curve is always the same when the region of values is measured in *standard deviations*.

For example, if we take a range of $\pm\sigma$ around the centre of the distribution (μ), then the area under the curve in this region is always 0.68 (68%) (see Figure 6.4).[2] This holds true no matter what the values of μ and σ. 95% of the area is within 2 standard deviations of the mean ($\mu \pm 2\sigma$), and almost all of the area (99.7%) is within 3 standard deviations. Remember that area under the curve is probability. So this means that, for example, if a variable is Normally distributed then there is an approximate 95% probability that it will be within 2 standard deviations from its mean.

6.4.3 Standard Normal distribution N(0,1)

Due to the property that areas under the Normal curve are identical when regions are defined by standard deviations, any Normal distribution can be *transformed* so that the x-axis is measured in *standard deviations*. Putting standard deviations on the x-axis of the bell curve (instead of whatever units the variable was originally measured in), is called *standardizing*.

If y is a Normally distributed variable, then to standardize we subtract the mean (μ) form y, and divide by the standard deviation (σ). This creates a new random

[2] The values 68-95-99.7 are approximate.

variable (call it z):

$$z = \frac{(y - \mu)}{\sigma}$$

Standard Normal distribution. The Standard Normal distribution is a special case of the Normal distribution: it has mean $\mu = 0$ and standard deviation $\sigma = 1$, and is denoted $N(0, 1)$.

The original variable y has a Normal distribution with mean μ and variance σ^2 (we write this $N(\mu, \sigma^2)$). The z variable, which is created from y, has mean 0 and variance 1 ($N(0, 1)$). No matter what the values are for μ and σ^2, z will *always* be $N(0, 1)$. $N(0, 1)$ is a special case of the Normal distribution, and is called the *Standard Normal distribution*.

If we want to know the probability that y is within a range of values, we need to draw the Normal curve for y, and then calculate the area under the curve. Each situation presents different values for μ and σ, meaning that for each situation we have to draw a unique curve and calculate a unique area. This was historically problematic. Without computers, drawing these curves and calculating these areas was difficult.

Instead of drawing a curve and calculating an area for each unique situation, we can transform the situation such that it is characterized by the standard Normal distribution. This means we can have *one* curve, and we can calculate a bunch of areas under that curve *once*. These areas are reported in a *Standard Normal table.*

The benefit of "standardizing" a variable (subtracting its mean and dividing by its standard deviation) has somewhat been diminished along with advances in computing power. However, the topic is still worth studying. Standardization is ingrained in statistics, and some other concepts build upon or are analogous to it.

6.4.4 Testing for Normality

Often, we are unsure as to whether a variable is Normally distributed, or even approximately Normal. We can *test* to see if a variable is Normal, using some of the properties that Normal variables always have. We will revisit this topic later in Section 13.3.1.

Normal quantile-quantile plot

A Normal quantile-quantile plot (Q-Q plot) is where a variables quantiles (according to the Normal distribution) are plotted on the x-axis, and the variable itself is plotted on the y-axis. Remember that a quantile is very similar to a percentile (see Section 5.4.1).

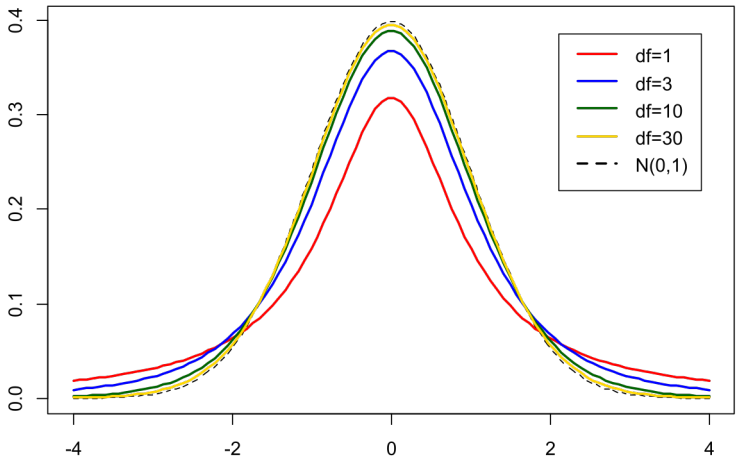
If the variable is Normally distributed, then the scatter plot of the quantiles vs. the values should approximately form a straight line. The Normal Q-Q plot is an informal visual aid for determining if a variable is Normally distributed (or distributed according to whatever distribution generates the quantiles).

Jarque-Bera test

We will only mention this test for now. One property of the Normal distribution that we have not discussed is that its skewness is always 0 (you might have guessed this already, since it is a symmetric distribution), and its kurtosis is 3. We can compare the *sample* skewness and kurtosis of our data to see if they are close to that which is required for a Normal variable.

6.5 t-distribution

Figure 6.5: Comparison of t-distributions with different degrees of freedom (df), and the Standard Normal $N(0, 1)$ distribution.



t-distribution. The t-distribution is used in hypothesis testing, and is similar to the $N(0, 1)$ distribution, becoming more and more Normal as the *degrees of freedom* (df) increases.

The t-distribution is most commonly used in hypothesis testing when the sample size n is small (see Section 11.2). The t-distribution is similar to the Standard Normal distribution, except that it has flatter tails. The t-distribution has only one parameter that controls its shape, called the *degrees of freedom* (df). As the degrees of freedom df increases, the t-distribution becomes closer and closer to the Standard Normal distribution $N(0, 1)$. If df is large enough, then the Standard Normal distribution can be used as an approximation to the t-distribution. Several t-distributions with varying df are drawn in Figure 6.5, as well as the $N(0, 1)$ for comparison.

7. Probability and Randomness

Probability is a way of providing structure to randomness. If there is uncertainty (randomness) surrounding a particular event, usually the best we can do is try to assign it a probability. In this chapter we discuss and define randomness and probability, and some related topics.

7.1 Randomness

Something is said to be random if its occurrence involves a degree of unpredictability or uncertainty. Outcomes that we cannot perfectly predict are random. Randomness represents a human failing, an inability to accurately predict what will happen. For example, if we roll two dice, the outcome is random because we are not skilled enough to predict what the roll will be. Things that we cannot, or do not want to predict (because it is too difficult), are random. We cannot know everything. However, we can attempt to model randomness mathematically.

The idea that randomness embodies a lack of information does not oppose a deterministic world view. While many things in our lives *appear* to be random, it is possible that all events are potentially predictable. In the dice example, it is not too far-fetched to believe that a camera connected to a computer could analyze hand movements and perfectly predict the result of a dice roll before the dice finish rolling!

Just because an outcome or event is random, doesn't mean that it is *completely* unpredictable, or that we can't at least try to guess what will happen. This is where *probability* comes in. Probability is a way of providing structure for things that are uncertain or random.

7.2 Sample space, outcomes, and events

Before we define *probability*, it is helpful to establish some terminology.

Random process. A process that results in some uncertain outcome.

Sample space. The sample space is the set of all possibilities (all outcomes) that can occur as a result of the random process.

Outcome. An outcome is a single point in the sample space. After the randomness resolves (is *realized*), the random process results in a single outcome.

Event. An event is a collection of outcomes. An event is a subset of the sample space.

Depending on the nature of the random process, the sample space may consist of integers or real numbers, qualities (for example ethnicity or gender), colours, locations, time, etc. The nature of the sample space, and the properties of the elements in the sample space, vary by random process. The sample space could be countably or uncountably infinite (e.g. the set of all integers or the set of all real numbers), could be bounded (e.g. between the number 0 and 1) or unbounded (e.g. between $-\infty$ and $+\infty$), and could take on a finite number of possibilities (e.g. $1, 2, 3, 4, 5, 6$).

Out of all the possibilities in the sample space, the *outcome* is where the random process arrives at. The outcome is a single element, point, or number, in the sample space.

An event is a collection of outcomes. There are three good reasons for caring to define *events*. (i) When we want to know the probability of *something* occuring, that *something* is usually a collection of outcomes. For example, what is the probability that someone is a millionaire? The event of interest consists of all the dollar outcomes that are greater than \$1 million. What is the probability that it will be cold tomorrow? "Cold" means below a certain temperature, not an exact temperature.

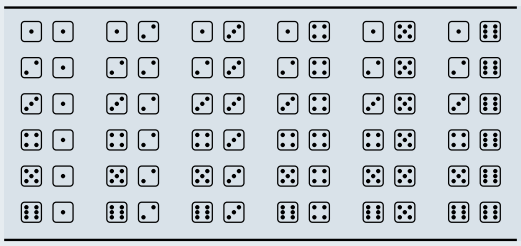
(ii) When the sample space has an infinite number of possibilities, as is the case for *any* continuous random variable (such as temperature or income), the probability of any one outcome occurring tends to zero. What is the probability that it will be $-20°C$? What about $-20.1°C$ What about $-20.000\,01°C$? Since there are infinite possibilities, the probability of any one of them occurring goes to 0. Instead, we must talk about *ranges* of values if we want to end up with non-zero probabilities. A range of values is just a collection of outcomes, or an event.

(iii) Finally, an event represents an *area* under the density curve (see Section 6.1). We will be able to calculate the probability of events using a density curve.

Example 7.1 — Rolling 2 dice. Consider the random process of rolling 2 dice.

1. What is the sample space?

   The sample space is the set of all possible outcomes that can result from rolling two dice. In this example, the sample space contains a finite number of outcomes:



2. Give an example of an outcome.

   An outcome is any one of the 36 entries in the table, for example ⚄ ⚂. After the dice are rolled, one of these outcomes will occur.

3. Give an example of an event.

   There are many events that we could consider. For example:

   - Rolling a 4. This event is a collection of the outcomes ⚀ ⚂, ⚁ ⚁, and ⚂ ⚀.
   - Rolling higher than 10. This event is a collection of the outcomes ⚄ ⚅, ⚅ ⚄, and ⚅ ⚅.
   - Rolling an even number. This a collection of half (18) of the outcomes in the table.

Example 7.2 — Percentage mark on a midterm. What percentage mark will you receive on your next midterm? If the professor told you that the midterm was out of 100 marks and no part-marks would be given, then the sample space would have exactly 101 values: $\{0\%, 1\%, 2\%, \ldots, 100\%\}$.

If the professor does not tell you the marking structure of the midterm, then your percentage score could be anything between 0% and 100%! Infinite possibilities. In this case, the sample space is written as $[0\%, 100\%]$. An outcome is any value in this range, and you will receive one of them for your score. An event, such as receiving an "A+", is the collection of outcomes $[93\%, 100\%]$, and is a subset of the sample space.

7.3 Probability

Probability can be defined several ways. There is a somewhat philosophical debate between "frequentists" and "Bayesians" on the definition and meaning of probability. I take the frequentist approach.

Probability. The probability of an event is the portion of times the event will occur, if the event could occur repeatedly.

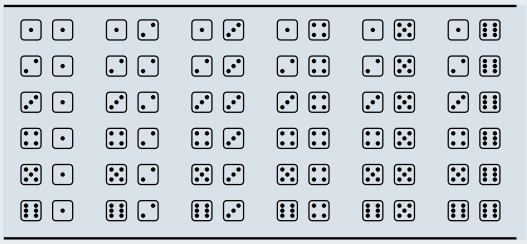
A probability is a number between 0 and 1 that is assigned to an event (sometimes expressed as a percentage). The probability of an event is the proportion of times it occurs in the long run. This definition is straightforward when we think about rolling dice or flipping a coin. The random process of flipping the coin can easily be repeated. If we imagine flipping the coin many (infinite) times, the proportion of times each event happens (heads or tails) is the probability. This definition of probability works because we can imagine repeating the random process many times under similar settings.

What about events that occur seldomly or only once? What is the probability that you will obtain an A+ in this course? What is the probability that Donald Trump will be president in 2025? For these examples, the former definition of probability takes a little bit more work and imagination. We need to imagine the random process being repeated many times under similar situation. For example, think of many parallel universes that are identical except for your performance in this course. In what proportion of those universes do you receive an "A+"?

A more general definition of probability is that it is a mathematical way of quan-

tifying uncertainty. For the Trump example, Bayesians would say that the probability of re-election is *subjective*. I may think the probability is 0.1, but someone else may assign a probability of 0.9. Which is right? These problems are better suited to a *Bayesian* framework, which is not discussed further in this book. The first definition of probability will be sufficient for the topics covered here.

Example 7.3 — Probabilities when rolling dice. When rolling 2 dice, what are the probabilities of various events? It turns out we can assign a probability to any event by making a simple assumption: there is an equal probability of the die[a] landing on any one of its six sides. That is, we assume that the die is "fair". This means that each of the outcomes in the sample space:



has an equal probability of occurring (a $1/36$ probability). To determine the probability of an event, we simply count the number of outcomes that satisfy the event, and add up the probabilities (this uses a rule of probability that we will discuss in Section 7.6). For example:

1. The probability of rolling a "7" is $1/6$. This is because there are 6 "ways" to roll a "7", out of the 36 possible outcomes: ⚀⚅, ⚁⚄, ⚂⚃, ⚃⚂, ⚄⚁, and ⚅⚀. So, $\Pr[Y = 7] = 6/36 = 1/6$. Here, "Pr" stands for probability, the event is written in the square brackets [ ], and Y needs to be defined as the sum of 2 die rolls.
2. The probability of rolling a "2" is $1/36$. Only 1 outcome satisfies the event: ⚀⚀.
3. The probability of rolling a "1" is 0, since there are no outcomes in the sample space that can satisfy the event.
4. The probability that the roll is higher than "10" is $3/36$: 3 outcomes out of 36 satisfy the event.

[a] "Die" is the singular of "dice".

7.4 Random variables

Random variable. A random variable assigns a unique numerical value to each of the possible outcomes in the random process.

A random variable is when outcomes are translated into numerical values. For example, a die roll only has numerical meaning because someone has etched numbers onto the sides of a cube. A random variable is a human-made construct, and the choice of numerical values can be arbitrary. Different choices can lead to different properties of the random variable. For example, I could measure temperature in Celsius, Fahrenheit, Kelvin or something new (degrees Ryans).

7.4.1 Discrete and continuous random variables

Random variables can be *discrete* or *continuous* (see Section 4.3). A discrete random variable takes on a countable number of values, e.g. {0, 1, 2, ...}. The result of the dice roll is a discrete random variable. Number of years of education, ethnicity, gender, are all examples of discrete random variables.

In contrast, a continuous random variable takes on a continuum of possible values (an uncountably infinite number of possibilities). Some examples of continuous random variables that we have mentioned so far are temperature, income, GDP, happiness score, etc.

Although a continuous random variable may have lower and upper bounds, there are still infinite possibilities. The temperature tomorrow is a continuous random variable, that may be bound between -50°C and 50°C, but there are still infinite possibilities. What is the probability that it is 20°C? What about 20.1°C? What about 20.0001°C? We could keep adding 0s after the decimal.[1] In fact, the probability of the temperature taking on any one value (outcome) approaches 0. For continuous random variables, instead of considering individual outcomes, we must consider *ranges* of outcomes (recall that a range of outcomes is defined as an *event*). For example, we could consider the probability that the temperature will be above 20°C.

7.4.2 Realization of a random variable

Finally, we make note of the difference between a *random variable* and the *realization of a random variable*. Before we roll the die, the outcome is random. After we roll the die and get a ⚄ (for example), the "4" is just a number - a *realization* of a random variable. It might be confusing to see a spreadsheet full of numbers in R and call them random variables, but the idea is that the numbers we see are the *realizations* or results of a random process.

7.4.3 Key points

Some of the key points we have discussed in this section are:

- A random variable can take on different values (or ranges of values), with different probabilities.
- It is sometimes helpful to differentiate between discrete and continuous random variables.
- Continuous random variables can take on an infinite number of possible values, so we can only assign probabilities to *ranges* of values (events).
- We can assign probabilities to all possible values (outcomes) for a discrete random variable, because we can count all the outcomes that can occur.
- When randomness resolves, we see the outcome as a *realization* of the random process. It is now just a number.

7.5 Independence

Often, we consider two or more random processes simultaneously. In economics, we frequently want to know if one variable is "associated" with, or "causes" another. For example, how does a change in inflation effect GDP or employment? How does the number of years of education of a worker influence their wages? There are elements of

[1] We have already discussed this idea in Sections 4.3.2 and 7.2.

randomness in all of these variables. When considering two or more potentially random processes together, a key consideration is whether or not the processes are *independent*.

If two random variables are independent, then the outcome of one variable does not influence the outcome of the other variable. Observing the value (outcome) of one variable does not give any clues about what the other variable will be. Finding out that two random variables are independent is very important in statistical analyses.

Implications of independence. If two variables are independent, then:

- The outcome of one variable can't influence or affect the other.
- One variable is useless for predicting the outcome of the other variable.
- Neither variable can cause the other.

For example, finding out that *education* and *income* are independent would be a shocking discovery. It would mean that education could not cause an improvement in wages.

Example 7.4 — The gambler's fallacy. The gambler's fallacy occurs when the independence of events is ignored. It is an incorrect belief that if independent events occur more or less than usual, then that must mean that the event is more (or less) likely to occur in the future.

For example, if a slot machine has been "cold" all night (has not paid out any jackpots), then that must imply that the probability of a jackpot on the next pull is somehow effected (different gamblers may avoid or seek out such a machine). This is an incorrect belief. The probability of a jackpot is the same for each pull. Each pull is *independent* from the last - the past events do not change the probabilities of future events.

What is the probability of rolling a "7" with 2 dice? From Example 7.3 we know this to be $1/6$. What if we had just rolled a "7" three times in a row? As long as the dice are fair, and there is no magical being interfering with the dice, the probability of rolling a "7" is still $1/6$. Each roll of the dice are independent from each other. Knowing past events does not help predict future events that are independent.

7.6 Rules of probability

Probabilities must follow several rules. These rules not only help to solidify our understanding of probability, but also have various uses. Below are five of the rules that probabilities must follow.

Rule 1: Probabilities are always between 0 and 1

The probability of an event (call it "A" for example) must be between 0 and 1:

$$0 \leq \mathrm{P}(A) \leq 1$$

The probability of an impossible event is 0, and the probability of a certain event is 1. Everything else must be between these two "extremes".

Rule 2: The probability of something happening is 1

The probability of some outcome occurring in the sample space is 1. Something must happen. If the sample space is truly exhaustive (describes everything that can possibly happen in the random process), then one of these outcomes must occur. To express

this mathematically, we define an event called "S" which is comprised of all outcomes in the sample space. Then:

$$\mathrm{P}(S) = 1$$

Rule 3: Complements

If event A does not occur, then the event "not A" must occur. A^c is the event "not A", and is called the *complement* of event A. The probability of A^c occurring is:

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$$

Rule 4: Addition

The probability of *either* event "A" or event "B" occurring can be determined by adding and subtracting probabilities depending on whether the two events are *mutually exclusive* or not.

(a) If "A" and "B" are mutually exclusive (meaning that they do not have any outcomes in common) then:

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

(b) If "A" and "B" are *not* mutually exclusive (they have some outcomes in common), then:

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B)$$

This is so we don't "double-count" probabilities.

This rule extends to more than two events, for example: $\mathrm{P}(A \text{ or } B \text{or } C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C)$ (in the case of mutually exclusive events).

Rule 5: Multiplication

The probability of *both* event "A" and "B" occurring can be determined through multiplication.

(a) If events A and B are *independent* (neither event influences or affects the probability of the other occurring) then:

$$\mathrm{P}(A \text{ and } B) = \mathrm{Prob}(A) \times \mathrm{Prob}(B)$$

(b) If events A and B are *dependent* then we must *condition* on one of the events occurring:

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B|A) = \mathrm{P}(B) \times \mathrm{P}(A|B)$$

The vertical line | means "conditional" or given. $\mathrm{P}(B|A)$ means the probability of event B, given that A has already occurred.

This rule also extends to more than two events, for example: $\mathrm{P}(A \text{ and } B \text{and } C) = \mathrm{P}(A) \times \mathrm{P}(B) \times \mathrm{P}(C)$ (in the case of independent events).

Example 7.5 — Snow storm and a cancelled midterm. What is the probability of *both* a snow storm and a cancelled midterm occurring? Suppose that in *good weather* the probability of a midterm being cancelled tomorrow is only 1%. However, if there is a *snow storm*, then the probability of a *cancelled midterm* is:

$$\mathrm{P}(\textit{cancelled mid} \mid \textit{snow}) = 50\%$$

Suppose further that there is a risk of a snow storm tomorrow and the weather forecast gives it a 20% chance:

$$\mathrm{P}(\textit{snow}) = 20\%$$

Using the multiplication rule, the probability that both a snow storm and a cancelled midterm occurs is:

$$\mathrm{P}(\textit{snow} \text{ and } \textit{cancelled mid}) = \mathrm{P}(\textit{snow}) \times \mathrm{P}(\textit{cancelled mid} \mid \textit{snow}) = 0.2 \times 0.5 = 0.1$$

7.7 Mean and variance from a probability distribution

Recall that a probability distribution lists all the possible outcomes that can occur for the random variable, and assigns a probability to each outcome (or ranges of outcomes in the case of a continuous random variable). Sometimes, we can use our intuition to completely describe a probability distribution (for example, in the case of flipping a coin or rolling dice). In other cases, we must use statistics to *estimate* some unknown parts of the probability distribution (for example, the location and shape of the bell curve). In this section, we will consider the former case, where the probability distribution is known. When the probability distribution is completely known, we can calculate the true population mean and variance directly. We begin this section with simple examples of probability distributions, and then use them to calculate mean and variance.

Example 7.6 — Probability distribution for a coin flip. What is the probability distribution for a coin flip? We begin by describing all the possible outcomes that can occur. We can either get "tails" (T) or "heads" (H). So, the sample space for the coin flip (call it Y) is $\{T, H\}$. Next, we need to assign a probability to each possible outcome. If the coin is fair (not weighted), then the probability of each outcome is equal. Putting this all together, we can write the probability distribution as:

$$\mathrm{P}(Y = T) = 0.5$$
$$\mathrm{P}(Y = H) = 0.5$$

Example 7.7 — Probability distribution for a die roll. What is the probability distribution for a die roll? The sample space (all the outcomes that can occur) is: $S = \{1, 2, 3, 4, 5, 6\}$. If the die is fair (not weighted), then the probability of each outcome is equal. Denoting the result of the die roll as Y, we can write the proba-

bility distribution as:

$$\begin{aligned}
&\mathrm{P}(Y = 1) = 1/6 \\
&\mathrm{P}(Y = 2) = 1/6 \\
&\mathrm{P}(Y = 3) = 1/6 \\
&\mathrm{P}(Y = 4) = 1/6 \\
&\mathrm{P}(Y = 5) = 1/6 \\
&\mathrm{P}(Y = 6) = 1/6
\end{aligned}$$

Probability distributions can be written in alternate ways. The important points are that (i) all of the possible outcomes are defined, and (ii) probabilities are assigned to each outcome. We can rewrite the above probability function as:

$$\mathrm{P}(Y = k) = 1/6 \quad ; \quad k = 1, \ldots, 6$$

7.7.1 Mean / expected value

The *mean* or *expected value* of a random variable is the value that is expected, or the value that occurs on average through repeated realizations of the random process. The mean of a random variable can be determined from its probability function. The probability function contains all possible information we could hope to have about the random variable, so it's no surprise that if we want to know the mean we can use the probability function. The mean (and variance, etc.) is just summarized information derived from the probability function.

Sample versus population means. Caution! There is a confusing but important distinction between the *sample mean* and the *true population mean*. Here, we are discussing the *true population mean*. The true population mean is determined from the probability function. The sample mean is determined by adding up sample values and dividing by the sample size. The two are related: in general if you don't know the true population mean, you can instead use the sample mean to "guess" or estimate the truth. This distinction is a key point in statistics, and we will try to highlight its importance in the coming chapters.

Let Y be a discrete random variable, for example the result of a die roll. Notation for the mean of Y or expectation of Y is μ_Y or $\mathrm{E}[Y]$. As mentioned above, $\mathrm{E}[Y]$ can be determined from its probability distribution.

Mean of a discrete random variable. For discrete random variables, the mean is determined by taking a weighted average of all possible outcomes, where the weights are the probabilities of each outcome occurring. The equation for the mean of discrete random variable Y is:

$$\mathrm{E}[Y] = \sum_{k=1}^{K} P_k Y_k \tag{7.1}$$

where P_k is the probability of the k^{th} event, Y_k is the numerical value of the k^{th}

outcome, and K is the total number of outcomes[a].

[a] K can be infinite!

Example 7.8 — Mean of a die roll. Let Y be the result of a die roll. What is $\mathrm{E}[Y]$? We will use Equation 7.1, and from the probability function for the die roll (see Example 7.7) we know that $K = 6$ and each $P_k = 1/6$, so:

$$\mathrm{E}[Y] = \sum_{k=1}^{K} P_k Y_k = \frac{1}{6} \times (1) + \frac{1}{6} \times (2) + ... + \frac{1}{6} \times (6) = 3.5$$

Notice that the mean of 3.5 is not a number that we can possibly roll on the die! However, it is still the *expected* result.

Mean for a continuous random variable

Equation 7.1 is valid for any discrete random variable. Calculating the mean of a continuous random variable is analogous, but more difficult. Again, the mean is determined from the probability function, but instead of *summing* across all possible outcomes we have to *integrate* (since the random variable can take on a continuum of possibilities).

Let y be a continuous random variable. The mean of y is

$$\mathrm{E}[y] = \int y f(y)\, \mathrm{d}y$$

If y is normally distributed, then $f(y)$ is equation (6.2), and the mean of y turns out to by μ. You do not need to integrate for this course, but you should have some idea about how the mean of a continuous random variable is determined from its probability function.

7.7.2 Rules of the mean / expected value

Some rules of the mean / expected value are discussed in this section. These rules can help in the understanding of the concept of the mean, and can be useful in real situations.

Rule 1: Mean of a constant

If the random variable is not random at all, but is a constant c, then $\mathrm{E}[c] = c$.

Rule 2: Addition

The mean of the sum of two (or more) random variables is equal to the sum of the means:

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

where X and Y are random variables. Similarly, the mean of the sum of a constant and a random variable is:

$$\mathrm{E}[c + Y] = c + \mathrm{E}[Y]$$

Rule 3: Multiplication by a constant

The expected value of the product of a constant and a random variable is equal to the product of the means:

$$\mathrm{E}[cY] = c \times \mathrm{E}[Y]$$

Note that, in general, the expected value is not multiplicative. $\mathrm{E}[XY] \neq \mathrm{E}[X]\,\mathrm{E}[Y]$. Only if X and Y are *independent* is the expected value multiplicative. Note that a constant c and a random variable Y are always independent!

Example 7.9 — Changing the numbers on a die. Suppose that we create our own custom die. A typical die has sides that read ⚀, ⚁, ⚂, ⚃, ⚄, ⚅. On our custom die, we instead make the sides read $\{3, 4, 5, 6, 7, 8\}$. What is the expected value (mean) of the custom die?

Instead of defining the probability distribution for this custom die, and using Equation 7.1, we can instead use the rules of the mean. Let Y represent the typical die, and X represent the custom die. What is the relationship between Y and X? We can get the custom die by adding 2 to each side of the typical die, so:

$$X = 2 + Y$$

and using the rules of means for constants and addition we have that:

$$\mathrm{E}[X] = 2 + \mathrm{E}[Y] = 2 + 3.5 = 5.5$$

Example 7.8 shows where "3.5" comes from.

Example 7.10 — The sum of two dice. Often, in games, players roll two dice and take the sum (backgammon, craps, Monopoly, Catan, Dungeons and Dragons, etc.). In Monopoly, you move forward a number of spaces equal to the number that you roll with two dice. How many spaces forward do you *expect* to move? That is, what is the mean value of the sum of two dice?

To answer this question, we could either determine the probability function for the sum of two dice and use Equation 7.1, or we could use the rules of means. Let X be the result of one of the die rolls, and Y the result of the other. The number of spaces the game piece moves forward is equal to $X + Y$. The mean dice roll, using the rules of the mean, is:

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 3.5 + 3.5 = 7$$

You can expect to move forward 7 spaces. This is an important thing to know if you gamble or play board games!

7.7.3 Variance

Sample versus population variances. Again, be aware of the difference between the *true population variance* (which we are discussing here), and the *sample variance* (see Equation 5.3 for sample variance). In this section, the probability functions are completely known, so that we can use them to determine the variance of the random variable directly. In cases where there is some question as to the *shape* of the probability function, we can use the sample variance to guess or estimate the true population variance.

The variance of a random variable is a measure of its *spread* or *dispersion*. It tells us how far away the numerical outcomes tend to be, relative to the *mean*. A higher variance means that there is a higher probability that the random variable will take on

values that are far away from the mean or expected value.

Variance of a discrete random variable. Variance is the expected squared difference of the random variable from its mean. For a discrete random variable Y, the variance (denoted by σ_Y^2 or $\text{Var}[Y]$) is:

$$\text{Var}[Y] = \text{E}[(Y - \text{E}[Y])^2] \tag{7.2}$$

As long as Y is a discrete random variable[a], equation (7.2) becomes

$$\text{Var}[Y] = \sum_{k=1}^{K} P_k \times (Y_k - \text{E}[Y_k])^2 \tag{7.3}$$

where P_k is the probability of outcome k occurring, Y_k is the numerical value of outcome k, and K is the total number of (possibly infinite) outcomes. Note that equation 7.3 is a weighted averaged of squared distances. The variance is measuring how far, on average, the variable is from its mean. The higher the variance, the higher the probability that the random variable will be far away from its expected value.

[a]When the random variable is continuous, equation (7.2) becomes:

$$\text{Var}(y) = \int (y - \text{E}[y])^2 f(y)\, dy$$

but you don't need to know this for the course.

Example 7.11 — Variance of a die roll. What is the variance of a die roll, Y? We already know that $\text{E}[Y] = 3.5$, and using Equation 7.3, we have:

$$\begin{aligned}
\text{Var}[Y] &= \sum_{k=1}^{K} P_k \times (Y_k - \text{E}[Y_k])^2 \\
&= \frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \cdots + \frac{1}{6}(6 - 3.5)^2 \\
&= \frac{1}{6}(6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) \\
&= \frac{17.5}{6} \approx 2.92
\end{aligned}$$

This is telling us that we *expect* the squared distance between the die roll and the mean value of 3.5 to be equal to 2.92. This is a measure of the *dispersion* of Y.

7.7.4 Rules of variance

Variance must follow several rules, which are both illuminating and useful in practice.

Rule 1: Variance of a constant

The variance of a constant, c, is zero:

$$\text{Var}[c] = 0$$

A constant is always the same (it never varies). The distance of the constant from its mean is always zero.

Rule 2: Addition

The variance of the sum of two random variables is equal to the sum of the variances, plus a *covariance* term (covariance defined later):

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \times \text{Cov}[X, Y]$$

If the two random variables are *independent* (the outcome of one does not influence or affect the outcome of the other), then the covariance[2] between them is 0, and:

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

Since the variance of a constant is zero, adding constants to random variables does not change the variance:

$$\text{Var}[c + Y] = \text{Var}[Y]$$

Rule 3: Multiplication

The variance of a random variable multiplied by a constant is equal to the *square* of the constant multiplied by the variance of the random variable:

$$\text{Var}[cY] = c^2\text{Var}[Y]$$

The variance rules for the product of two random variables are more complicated and are not used here.

Rule 4: Non-negativity

Variance cannot be a negative number. Note the "square" in the formula for variance (Equation 7.2). Since distance from the mean is being squared, we can never get a negative variance for a random variable Y: $\text{Var}[Y] \geq 0$.

Example 7.12 — Variance of a custom die. Consider the custom die from Example 7.9, with sides $\{3, 4, 5, 6, 7, 8\}$. Call the result of the custom die roll random variable X. What is $\text{Var}[X]$? It's the same as the standard die! That is, $\text{Var}[X] \approx 2.92$. This makes sense: the distance between each consecutive outcome is 1, whether looking at the custom die or the standard die.

We can verify this intuition either using Equation 7.3, or by using the rules of variance. Again, the relationship between the custom die X, and a standard die (call it Y) is: $X = 2 + Y$. Using the rules of variance:

$$\text{Var}[c + Y] = \text{Var}[Y] = 2.92$$

Adding and subtracting a constant does not affect the variance of a random variable.

Example 7.13 — Variance of the sum of two dice. Take the situation in Example 7.10. What is the variance of the sum of two dice? Call one die X and the other Y. From the rules of variance:

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \times \text{Cov}[X, Y]$$

but the dice are independent (the result of one roll cannot influence the other), so

[2] Covariance is very similar to correlation (see Section 5.10.

that the covariance between the two dice is 0. So:

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \approx 2.92 + 2.92 \approx 5.83$$

Example 7.14 — Another custom die. Let's create another custom die. The sides of the die will be equal to 2 times the sides of a regular die, so that the six sides read: $\{2, 4, 6, 8, 10, 12\}$. Call the result of this new custom die Z. What is the variance of Z?

Again, we could use Equation 7.3, or we could make our lives simpler by using the rules of variance. The relationship between this new custom die, and a standard die, is: $Z = 2 \times Y$. Using the rules of variance, we have:

$$\text{Var}[cY] = c^2\text{Var}[Y] \approx 2^2 \times 2.92 \approx 11.7$$

The values on the traditional die were multiplied by 2, the variance increases by 4.

8. Statistical Inference

Statistical inference. Statistical inference is when a statistic (for example the sample mean) is used to *infer* (i.e. "guess" or "estimate") something about the population.

In this chapter, we put some of what we have learned about statistics and probability together. This introduction quickly outlines what statistical inference means, and some key points. We will spend the remainder of the chapter dissecting and explaining the statements made in this overview. The key point in this chapter is that the *sample mean* is a random variable!

You have seen the equation for the sample mean before (see Section 5.1). The sample mean is found by adding up all values of a variable in the sample, and dividing by the sample size:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{8.1}$$

where y_i denotes the i^{th} observation, and where n denotes the sample size. The sample mean is a very popular method for *inferring* an unknown population mean. Suppose there is a random variable (call it y), and we want to know the true population mean of y. The true population mean value for y is denoted μ_y, and is an unknown *parameter*. We can *infer*[1] the value of μ_y by randomly drawing a sample from the population, and calculating the sample average. This process is called *statistical inference*.

Since $\bar{y}$ is calculated from a randomly selected sample, $\bar{y}$ is itself a random variable. $\bar{y}$ turns out to be Normally distributed, thanks to the *central limit theorem* (we will cover the central limit theorem in Section 8.4.1). The mean of $\bar{y}$ turns out to be the true population mean! This partly explains why $\bar{y}$ is such a popular estimator.

We end this introduction with a *simulation experiment*. We already know from Example 7.8 that the true population mean of a die roll is 3.5. Let's pretend, however, that we don't know this true population mean. How could we *estimate* it? We can use

[1] Instead of using the word "infer", we could also say "guess" or "estimate". In fact, $\bar{y}$ is an "estimator" for μ_y.

the *sample mean*! First we must collect a sample. We could sit at our desks and roll a die repeatedly, recording each result. Suppose we only have enough time to collect a sample of $n = 20$. Then, we take the sample average of all recorded die rolls. You can use your own die to accomplish this, or use the following R code to simulate rolling a die 20 times:

```
dierolls <- sample(1:6, 20, replace=TRUE)
```

Take a look at the die rolls (yours will be different):

```
dierolls
```

```
[1] 4 5 3 4 3 6 6 3 2 1 2 1 2 6 2 1 6 5 5 5
```

and calculate the sample average:

```
mean(dierolls)
```

```
[1] 3.6
```

If we didn't know that the true mean die roll is 3.5, we could collect a sample and use the sample mean to come up with a pretty good guess!

8.1 Parameter versus statistic

One main purpose for calculating a statistic is to represent, guess, or *estimate* some feature or characteristic of the population. Suppose we would like to know some true feature of the population, but it is something that we cannot observe directly (perhaps because the population is too large!). We could be interested in many different things: mean, mode, minimum or maximum value, variance, a percentile, etc.

In this chapter, we will focus on the mean of a population, as it is important and commonly sought after. For example, we may want to know:

- The mean[2] income, or the mean years of education, of all Martian colonists.
- The mean height of a human being.
- The mean quantity demanded of Mars diamonds, given prices.
- The mean number of doctor visits for individuals with health insurance.
- The mean sales for Fortune 500 companies.
- The mean temperature and CO_2 emissions by country.

Population parameter. A population parameter is a fixed number, often unknown, that governs the location and shape of a distribution.

There are many other examples, and many reasons for wanting to know the mean of a population. In each example, we could calculate a sample mean, and use that sample mean to *infer* the true population mean. The true expected or mean wage of a Mars colonist could be *estimated* using the sample average. It is important to note the distinction between the true population mean, and the sample mean. The

[2]Instead of "mean" we could use the word "average" for these examples, but we want to stress that we are talking about the *true* mean of the population, and not a "sample average".

true population mean income of Mars colonists is μ_{income}, a parameter that determines the entire population distribution of incomes. It can be considered a fixed number, unknown to us, but that we desire to discover. The sample mean, $\bar{income}$ is a statistic that can be used in place of the unknown μ_{income}.

Statistic. A statistic is calculated from a sample of data, and can be used to estimate an unknown parameter.

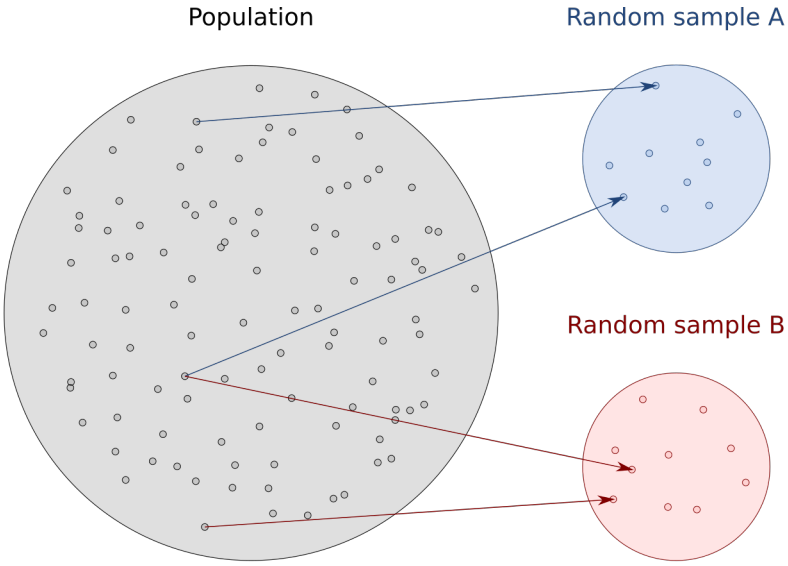
In each of the above examples, we could collect a sample, and calculate a sample mean in order to infer the unknown population mean. Measuring and averaging the heights of a sample of humans would allow us to guess at the expected human height. Observing a few diamond sales allows us to guess at the true quantity demanded for diamonds. Recording the number of doctor visits made by some randomly selected individuals would allow us to guess at the overall demand for healthcare.

Population parameter versus statistic. The "things" in the population that we might wish to know, such as mean, variance, median etc., are determined by *parameters* (often denoted with Greek letters, like μ and σ). Parameters are fixed numbers that govern or characterize the population distribution. In the case of the Normal distribution, the two parameters μ and σ control the location and shape of the bell-curve. If we could know the true parameters for any population, we would have incredible knowledge of the random process that is creating the data that we observe. Typically however, these population parameters are unknown and must be *estimated*. This is where a *statistic* comes in: a statistic can be used in place of the true unknown population parameter.

8.2 Population versus sample

In the examples above, it is not feasible to observe the entire population, so that the true population mean can never be known! How can we *estimate* the population mean in situations like these? One method is to collect a *random* sample from the population, and then use the information in that sample to *estimate* the population mean (or whatever feature of the population we wish to know). Using a statistic (calculated from a sample) in order to guess or estimate a parameter (from the population), is called statistical inference.

The difference between populations and samples has already been explained (see Section 3.2). The sample is a *subset* of the population, and the individuals or units that comprise the sample are *randomly selected*.



Each circle ◦ represents an individual in the population, and has a chance to be randomly selected into sample A ◦ or sample B ◦. Samples A and B are just two of millions of possible samples of size $n = 10$ that could be drawn from the population. In reality, we only have one of these possible samples to work with.

Statistics are random variables!. Since statistics are calculated from the sample, and the sample is randomly selected from the population, statistics are themselves random variables! This idea is key to understanding many of the concepts that follow.

8.2.1 Collecting a random sample

Consider the process of obtaining a random sample for some of the examples listed above. How might we select the individuals to be included in our sample? If we had everyone's social insurance number, we could use a random number generator[3] (RNG) to select as many individuals as we can afford to interview. A good way to obtain a random sample is by having a list of everyone in the population, and using RNG to select members for the sample. RNG can be accomplished by flipping a coin, rolling dice, drawing numbers from a hat, or by using more sophisticated tools like a computer.

Mars colonists

Pretend that we have a list of all 620,136 working-aged Mars colonists. According to our budget, we can afford to contact 100 individuals. We assign a number to each colonist, and then use RNG to generate 100 numbers. The selected colonists are contacted, interviewed, and entered into the sample.

The RNG determined the sample, and any statistics calculated from that sample! If we were to (hypothetically) repeat the process, we would get a different sample and different corresponding statistics. Statistics calculated from the sample are random variables, because the whole process began by randomly selecting a sample through RNG.

[3]The closest we can actually get are called *pseudo* random numbers. We start with a *seed*, and apply a complicated process to obtain an unpredictable result.

Example 8.1 — Random sample of Mars colonists. Begin by downloading a data set containing information on all 620,136 Mars colonists aged 18 and older (give it a few minutes, it's a large data set):

```
mars18 <- read.csv("http://ryantgodwin.com/data/mars18.csv")
```

Now, pretend that we *don't* have this entire data set. This is the entire population - if we had information on the entire population we would not need *statistical inference*. We will *simulate* sampling from this population. Let's pretend that our budget allows us to interview 100 individuals. Draw a random sample of 100 individuals from the population:

```
msample <- mars18[sample(1:620136, 100), ]
```

Take a look at the people in your sample:

```
View(msample)
```

Let's calculate the sample mean income from the randomly drawn sample:

```
mean(msample$income)
```

```
[1] 51686.45
```

The sample mean value is $\bar{income} = 51,687$. You will get a different sample mean! This is because your random sample will consist of different colonists. Try the following lines of code many times:

```
msample <- mars18[sample(1:620136, 100), ]
mean(msample$income)
```

Each time you will get a different value for the sample mean of income. We have just conducted a *simulation experiment*. In reality, we will only have one sample of size $n = 100$ to work with. In this experiment, we are drawing many samples in order to imagine what else we could possibly calculate for the sample mean. Since this is an experiment, we also know the population mean:

```
mean(mars18$income)
```

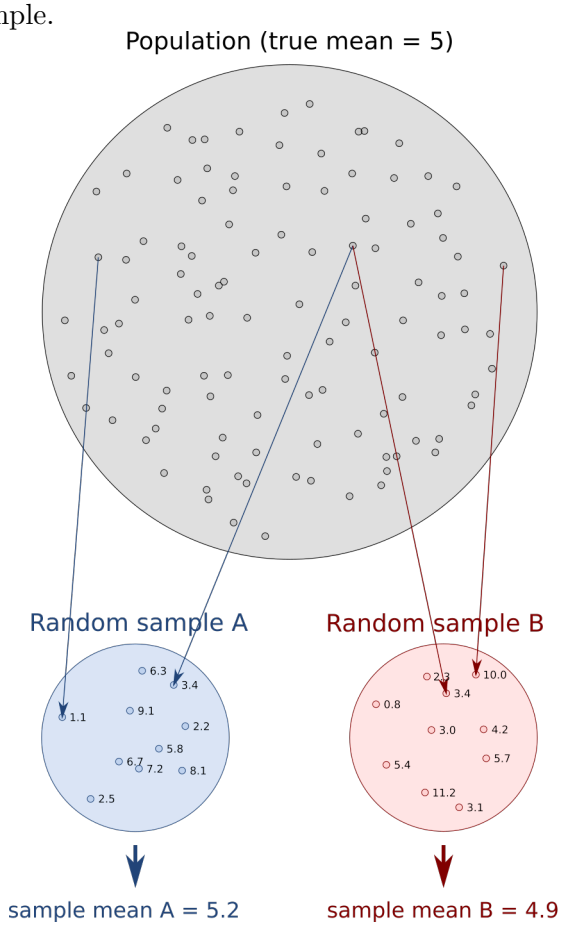
```
[1] 51737.09
```

How close were the sample averages to the true population mean? It is important to keep in mind that this is an experiment: in reality we only have one sample, and we do not know the true population mean.

Height of a human

We could go out into the street at 2 pm and record people's heights, and obtain a sample. Since we don't know who we're going to meet, the sample is random. But what if we had decided to go out at 3 pm instead? We would have recorded a different sample of heights. Any statistics calculated from the two hypothetical samples (2 pm and 3 pm) would differ, even though the true population height remains unchanged and is a fixed parameter.

Figure 8.1: The true population mean is 5 ($\mu = 5$). Each possible random sample y that we could draw from the population gives us a different sample average ($\bar{y}^A = 5.2$ and $\bar{y}^B = 4.9$ for example). $\bar{y}$ is a random variable because it is calculated from a randomly drawn sample.



8.3 The sample mean is a random variable

In this section, we want to emphasize again that the sample mean is a random variable, along with any other statistic that we might compute using a random sample.

A popular choice for estimating the population mean (denoted by $\mathrm{E}[y]$ or μ_y) is by using the *sample mean* (or *sample average*, or just *average*). The sample mean of y is usually denoted by $\bar{y}$. You have seen the equation for the sample mean before (see Section 5.1):

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

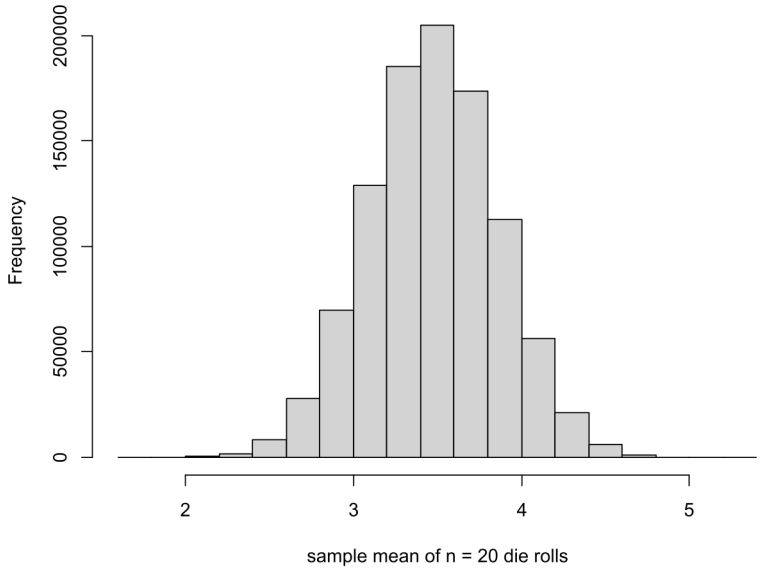
where y_i denotes the i^{th} observation, and where n denotes the sample size.

To reinforce the idea that the sample mean is random, consider the following situation. You will roll a die 20 times, collect the results, and calculate the sample average, $\bar{die rolls}$. What will be the number that you calculate for $\bar{die rolls}$? You might guess that it will be close to 3.5, but you can't completely predict the result. It is random, because the sample values $\{1, 2, 3, 4, 5, 6\}$ are randomly collected.

When we think about sampling individuals from a population, remember that they are chosen randomly. Imagine what would happen if we got a different sample, if we were in a parallel universe, if we collected the sample on a Tuesday instead of a Monday, etc. The sample values *could* be different, meaning that anything that is calculated from the sample *could* be different. See Figure 8.1 for a visualization of this idea.

8.4 Distribution of the sample mean

Figure 8.2: Histogram of sample means: simulated sampling distribution for the sample mean of 20 die rolls.



An important question is: how good is the estimator? That is, how good of a job is the estimator doing at "guessing" the true unobservable thing in the population? In our specific example: how good is the sample mean at estimating the true population mean of heights? This is an importannt question, because there are many ways that we could use the information in the sample to try to estimate the true mean. Why is equation (8.1) so popular?

The fact that a statistic is a random variable has important implications for statistical inference. If we are using the sample mean to estimate the population mean, we might wonder: "how well does the sample mean represent the true population mean?" One way to answer this question is to consider the *distribution* of the sample mean. It's a random variable after all, and it has a distribution!

Let's start from the fact that the sample mean, $\bar{y}$, is random. What is the distribution for $\bar{y}$? Remember that the probability distribution for a random variable accomplishes two things: (i) it lists all possible numerical values that the random variable can take, and (ii) assigns probabilities to ranges of values. So, what are the possible values that $\bar{y}$ can take? How likely is $\bar{y}$ to take on certain values? Ideally, we would like to know the exact *location* and *shape* of the probability distribution for $\bar{y}$.

Before we proceed, let's define the term *sampling distribution*. When the random variable is an *estimator* (such as the sample mean), then its probability distribution gets a special name - *sampling distribution*. That is, a *sampling distribution* is just a fancy name for the probability function of an estimator.

The sampling distribution is a hypothetical construct. It describes the probability of all outcomes for $\bar{y}$, but in the real world we only get one sample and one estimate $\bar{y}$.

Sampling distribution. Imagine that you could draw all possible random samples of size n from the population, calculate $\bar{y}$ each time, and construct a relative frequency diagram (a histogram) for all of the $\bar{y}$s. This relative frequency diagram would be the sampling distribution of the estimator $\bar{y}$ for sample size n.

This definition of the sampling distribution can be approximated using a computer. Instead of "all possible samples" we can use a computer to draw many many samples of size n from a population, and calculate and record $\bar{y}$ each time. Let's return to the die rolling experiment. Let y be the result of a die roll. We collected a sample of $n = 20$ die rolls, giving us one sample mean $\bar{y} = 3.6$. The *sampling distribution* for $\bar{y}$ can be simulated by collected a sample of $n = 20$ die rolls 1 million times (or as many times as you like), calculating and recording $\bar{y}$ each repetition:

```
nrep <- 1000000
allmeans <- numeric(nrep)
for(i in 1:nrep) {
  dierolls <- sample(1:6, 20, replace=TRUE)
  allmeans[i] <- mean(dierolls)
}
hist(allmeans)
```

The histogram from this R code is shown in Figure 8.2. Notice that there were a few "weird" samples drawn, where the sample mean was calculated to be very low or high, but this happens rarely. Most of the sample means from the experiment tend to be centered between 3 and 4. What is the exact *location* of this distribution? In fact, we can find this location by taking the sample mean of all 1 million $\bar{y}$:[4]

```
mean(allmeans)
```

```
[1] 3.498985
```

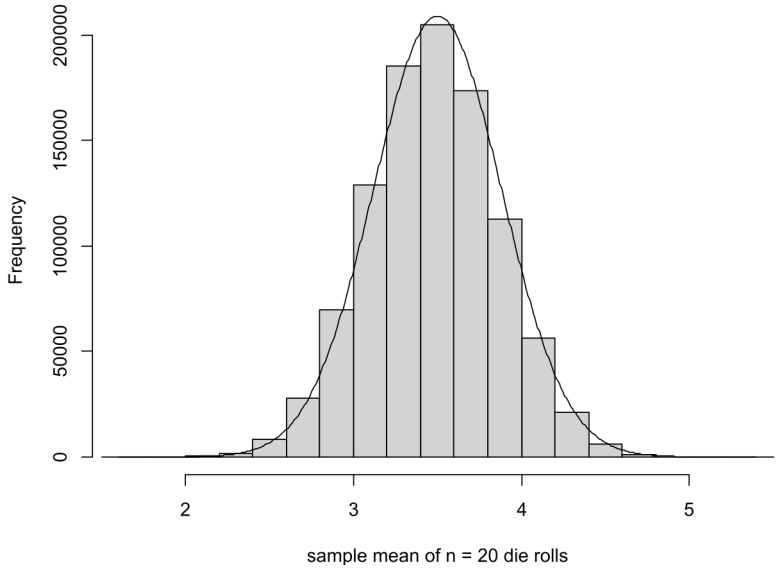
Wow, the value of 3.499 is very close to the true population mean die roll of 3.5! So, even though the possible values for $\bar{y}$ can be all over the place, on average they give the correct answer! In this example, the expected value of the sample mean is exactly equal to the true population mean: $\mathrm{E}[\bar{y} = \mu_y$, which is part of the reason why $\bar{y}$ is a popular statistic.

Unbiased estimator. When the expected value of the estimator is equal to the true population parameter intended to be estimated, the estimator is said to be "unbiased." The sample mean, $\bar{y}$ is an unbiased estimator (under certain assumptions).

Returning to Figure 8.2, what shape characterizes the histogram? It is the familiar Normal distribution, or bell curve! Figure 8.3 shows a Normal distribution superimposed onto the histogram of $\bar{y}$ for die rolls. In fact, the sample average $\bar{y}$ always (approximately) follows a Normal distribution, regardless of the distribution of the variables in the sample! This is due to the *central limit theorem*.

[4] It may be confusing to take the sample mean of sample means. Just focus on the fact that $\bar{y}$ is a random variable. It is natural to try to find the mean and variance of a random variable.

Figure 8.3: Normal distribution with $\mu = 3.5$ and $\sigma^2 = 0.145$, and histogram simulating the sampling distribution for the sample mean of 20 die rolls.



8.4.1 The central limit theorem

No matter what distribution the random process follows, when we start adding up random variables, the resulting sum is Normally distributed. We can even add different types of random variables. It only matters that we add up enough. If the random outcomes that we seek to model are the results of many random factors all added together, then the central limit theorem applies. This is a casual explanation of the CLT; there are several conditions required for it to hold, and several versions.

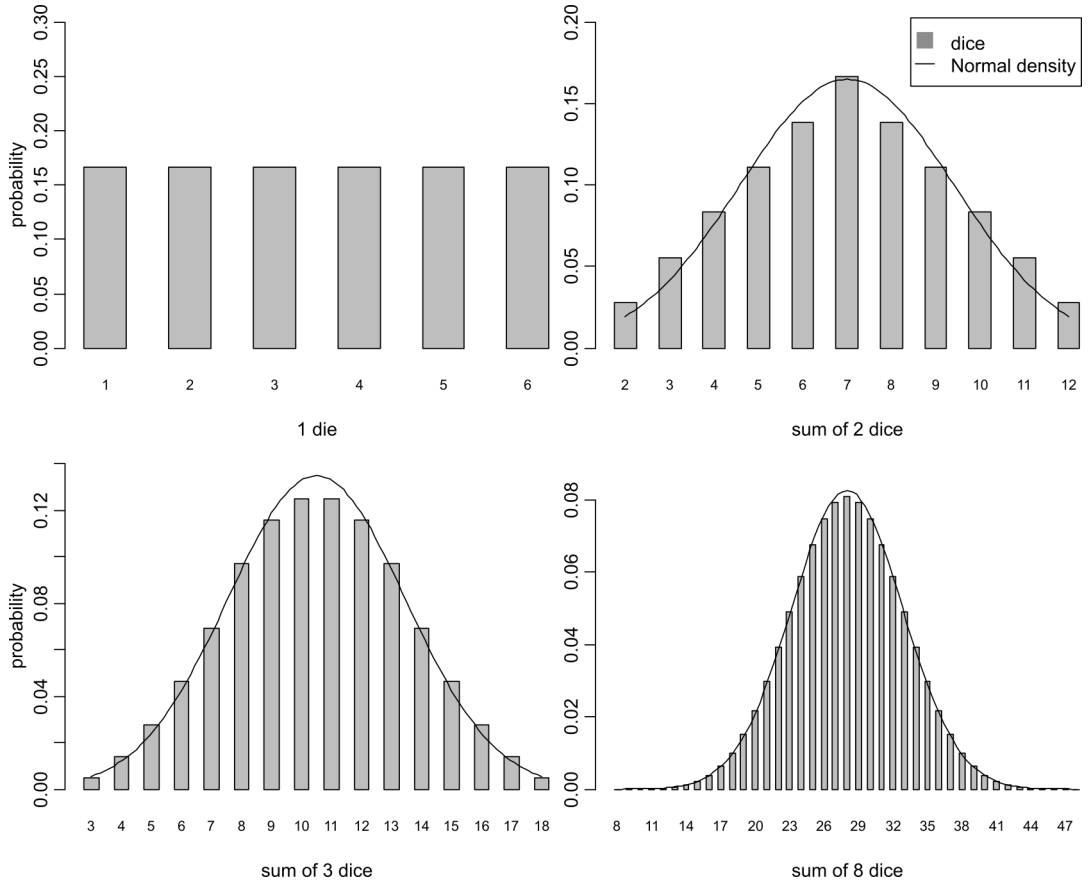
Central limit theorem. Loosely speaking, the central limit theorem (CLT) says that the sums of random variables are Normally distributed.

To illustrate the CLT, let's again use dice for an example. We know that a single die roll is *uniformly distributed* (equal probability of each number coming up). But what if we start *adding* the results of dice rolls? Figure 8.4 shows the probability function for the sum of two dice. It's no longer flat (uniform)! It even seems to have a bit of a curve to it.

Now, let's add a third die, and see if the probability function looks more normal. Let Y = the sum of *three* dice. It turns out the mean of Y is 10.5 and the variance is 8.75. The probability function for Y is shown in Figure (8.4). Also in Figure (8.4) is the probability function for a Normal distribution with $\mu = 10.5$ and $\sigma^2 = 8.75$. Notice the similarity between the two probability functions.

The CLT says that if we add up the result of *enough* dice, the resulting probability function should become Normal. Finally, we add up *eight* dice, and show the probability function for both the dice and the Normal distribution in Figure(8.4), where the mean and variance of the normal probability function has been set equal to that of the sum of the dice.

Figure 8.4: Probability function for the sums of dice, with Normal density functions superimposed. As the number of random variables that we sum increases, the distribution of the sum becomes Normal. This is due to the central limit theorem (CLT).



CLT and $\bar{y}$. So, what does the CLT have to do with the sample mean? Look at Equation 8.1 again. Notice the summation Σ operator. Taking a sample average involves adding up random variables, so the CLT means that $\bar{y}$ is randomly distributed.

9. Confidence intervals

This chapter discusses how to construct and interpret confidence intervals. Confidence intervals are very easy to calculate but very difficult to understand, and are commonly misinterpreted.

One of the uses of a confidence interval is to quantity the uncertainty surrounding the estimate $\bar{y}$. Confidence intervals can be calculated along with the calculation of $\bar{y}$, provided a measure of how "close" $\bar{y}$ might be to the true population mean μ_y.

The first part of the chapter lays down the groundwork necessary to understand confidence intervals. Sampling distributions, estimators, and the variance of estimators, are some of the required concepts that we begin with.

9.0.1 Simplifying assumptions

In statistics textbooks, it is customary to begin discussion of confidence intervals, hypothesis tests, and test statistics, by assuming that the *population variance is known*. This is a simplifying, but very unrealistic assumption. It is unrealistic because confidence intervals and hypothesis tests are used when the *population mean is unknown*. If the population mean (μ) is unknown, then the population variance (σ^2) is usually unknown as well.

In a later chapter, we tackle the more realistic situation that σ^2 is unknown. In this more realistic case, confidence intervals and hypothesis tests are altered slightly, but the overall principle and interpretation remains the same.

In this chapter, we will be using the result that the sample average follows a Normal distribution: $\bar{y} \sim N$. Usually, this Normal distribution is only an *approximation* to the true distribution of $\bar{y}$, and the approximation only works well when the sample size n is *large*. In some of the examples, we will only have a sample size of $n = 10$. This is too small for the Normal approximation to work well in practice, but we still use a small sample for simplicity in some the examples. Be aware: the Normal approximation only works well for large n!

Assumptions.

- The population mean μ_y is unknown and must be estimated using $\bar{y}$.
- The population variance σ_y^2 is assumed to be known (unrealistic, but simplifying assumption).
- n should be large for the Normal approximation to work well. We use examples with $n = 10$ for simplicity, but $n = 10$ is too small in reality.

9.1 Sample mean and population mean

In previous chapters, we have learned that we can use the *sample mean* to estimate an unknown *population mean*. Below we summarize some important differences between sample mean and population mean.

Sample mean $\bar{y}$	**Population mean μ_y**
Can be calculated from a sample of data.	Is usually unknown (except for things like coins or dice, see Example 7.8).
Is a random variable.	Is a fixed parameter (just a number).
Has an approximate Normal probability distribution. We use "sampling distribution" instead of "probability distribution" when talking about a statistic.	Doesn't have a probability distribution (because it's a constant).

Since the sample mean is a random variable, we can consider probabilities of $\bar{y}$ taking on certain values. For example, we could try to determine: $P(\bar{y} > 4)$ for a sample of dice rolls, or $P(80k \leq \bar{y} \leq 90k)$ for a sample of Mars incomes. As long as we know the *distribution* for $\bar{y}$, we can determine these probabilities by taking the area under the probability distribution (density) curve. So, what is the exact sampling distribution (probability distribution) for $\bar{y}$?

9.2 Exact sampling distribution of $\bar{y}$

In Section 8.4 we introduced the idea that since the sample mean is a random variable it has a probability distribution (renamed *sampling distribution* since $\bar{y}$ is a statistic). In Section 8.4.1 we said that $\bar{y}$ follows the Normal distribution due to the central limit theorem. In the current section, we use the sampling distribution of $\bar{y}$ to calculate the probability of getting an "extreme" sample average. An "extreme" sample average is one that is far away from the true population mean[1]. This idea is related to confidence intervals, and also leads to hypothesis testing (which is covered in the next chapter).

The sampling distribution of $\bar{y}$.

$$\bar{y} \sim N\left(\mu_y\ ,\ \frac{\sigma_y^2}{n}\right)$$

[1]What constitutes "extreme" and "far away from the truth" is subjective.

This says that $\bar{y}$ is Normally distributed with mean μ_y and with variance σ_y^2/n. The mean of $\bar{y}$ is the same as the mean of y. The variance of $\bar{y}$ is whatever the variance of y is, divided by n.

This sampling distribution of $N(\mu_y, \sigma_y^2/n)$ is only valid in certain situations. The sample size n has to be large enough for the central limit theorem to provide a Normal distribution, and the y data must be *identically and independently distributed* (which is assured if the y data was collected by simple random sampling).

Taking the distribution of $\bar{y}$ as $N(\mu_y, \sigma_y^2/n)$, it is easy to calculate the probability of getting various values for $\bar{y}$, provided μ_y and σ_y^2 are known!

Example 9.1 — Probability of getting a $\bar{y} > 4$. Suppose that you are about to roll 10 dice, and take the sample average $\bar{y}$. We know that the sample average "should" give us an answer that is close to 3.5 (the true mean of a die roll). What is $P(\bar{y} > 4)$? That is, what is the probability that we get some "extreme" value for the sample average? We now know that the sample average (approximately) follows the Normal distribution with mean μ_y and variance σ_y^2/n. From Example 7.8 we know that the mean of a die roll is 3.5:

$$\mu_y = 3.5$$

From Example 7.11 we know that the variance of a die roll is:

$$\sigma_y^2 = \frac{35}{12} \approx 2.92$$

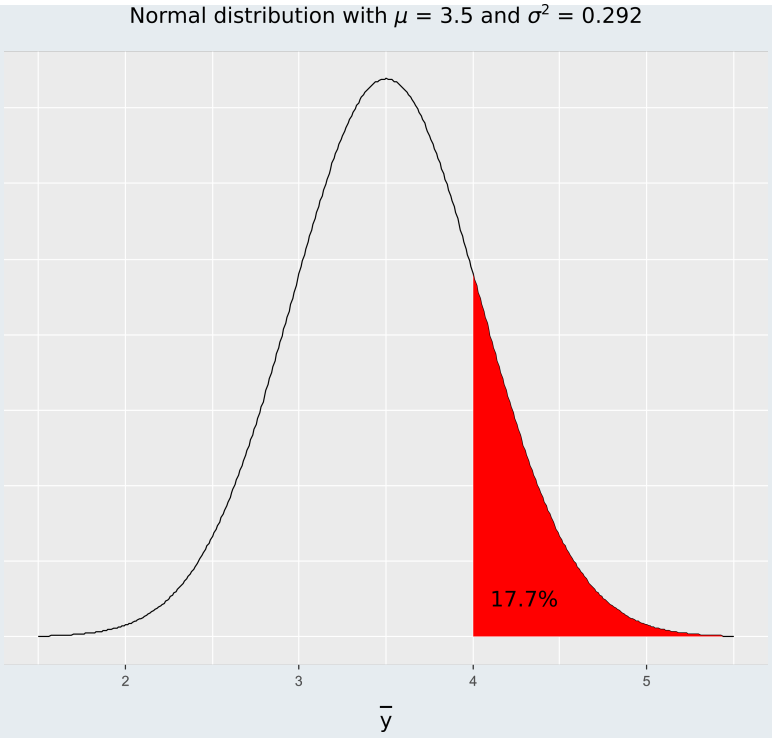
The sample size is going to be $n = 10$, so the variance of $\bar{y}$ is:

$$\frac{\sigma_y^2}{n} = \frac{35/12}{10} \approx 0.292$$

Putting this together, we have that the sampling distribution for the sample average of 10 die rolls is:

$$N(3.5, 0.292)$$

We can now get R to draw this Normal distribution, and calculate the area under the curve to the right of 4. This area tells us the probability of getting a $\bar{y}$ that is "extreme", or greater than 4. The R code for calculating this probability is:




```
pnorm(4, mean = 3.5, sd = sqrt(0.292), lower.tail = FALSE)
```

```
[1] 0.1774071
```

The `pnorm()` function calculates an area (a probability) under the Normal curve. The first argument in the function is `4`: we want $P(\bar{y} > 4)$. Next we give the function the mean and standard deviation[a] so that we draw the correct curve: `mean = 3.5` and `sd = sqrt(0.292)`. Finally, we tell the function that we want the "upper tail" (the area to the right of 4), so we set `lower.tail = FALSE`.

So, there is only a 17.7% probability of getting a $\bar{y} > 4$ when we average 10 dice!

[a] Remember that standard deviation is just the square root of the variance.

Example 9.2 — Number of times getting a $\bar{y} > 4$. In the previous example (Example 9.1) we found that if we were to roll 10 dice, and take the sample average, that:

$$P(\bar{y} > 4) = 0.177$$

One way of interpreting this probability of 0.177 is that, of all the samples of $n = 10$ die rolls that we could obtain, 17.7% will give a sample average higher than 4. This can easily be verified! Roll 10 dice, take the average. Repeat this many times. 17.7% of sample averages calculated should be above 4. Instead of actually rolling dice, we can use R:

```
roll <- sample(1:6, 10, replace=TRUE)
mean(roll)
```

```
[1] 3.9
```

Repeat the above code many times, and you will see that roughly 17.7% $\bar{y}$s are above 4!

9.3 Accuracy of $\bar{y}$ increases with n

The variance of $\bar{y}$ is:

$$\text{var}(\bar{y}) = \frac{\sigma_y^2}{n}$$

Variance measures the "spread" of a random variable. The formula shows that as n gets larger, the variance of $\bar{y}$ decreases. $\bar{y}$ gets more accurate with a bigger n! This is one of the reasons we want the sample size n to be as large as possible. As we collect more information in the sample, the sample average gets "better". This is true for many other statistics as well, such as the median or mode.

Example 9.3 — Variance of $\bar{y}$ for increasing n. The sample average $\bar{y}$ is a random variable that (approximately) follows the Normal distribution, with mean μ_y, and variance σ_y^2/n. Notice the n in the denominator. This means that as the sample size grows, the sample average gets more accurate. In other words, a larger sample size reduces the probability of getting an "extreme" value for the sample average.

For different sample sizes let's calculate the variance of $\bar{y}$, and the probability of getting a $\bar{y} > 4$. In Example 9.1 we found that if we were to take the sample average $\bar{y}$ of 10 dice, the variance of $\bar{y}$ would be:

$$\frac{\sigma_y^2}{n} = \frac{35/12}{10} \approx 0.292$$

Where $35/12$ is the variance of a single die roll. If we were to instead roll 20 dice and take the average, the variance of $\bar{y}$ would be:

$$\frac{\sigma_y^2}{n} = \frac{35/12}{20} \approx 0.146$$

This gives us the sampling distribution for $\bar{y}$ when $n = 20$: $\bar{y} \sim N(3.5, 0.146)$. The probability of getting a $\bar{y} > 4$ is similarly found by calculating the area under the $N(3.5, 0.146)$ curve, to the right of $\bar{y} = 4$. R can do this for us:

```
pnorm(4, mean = 3.5, sd = sqrt(0.146), lower.tail = FALSE)
```

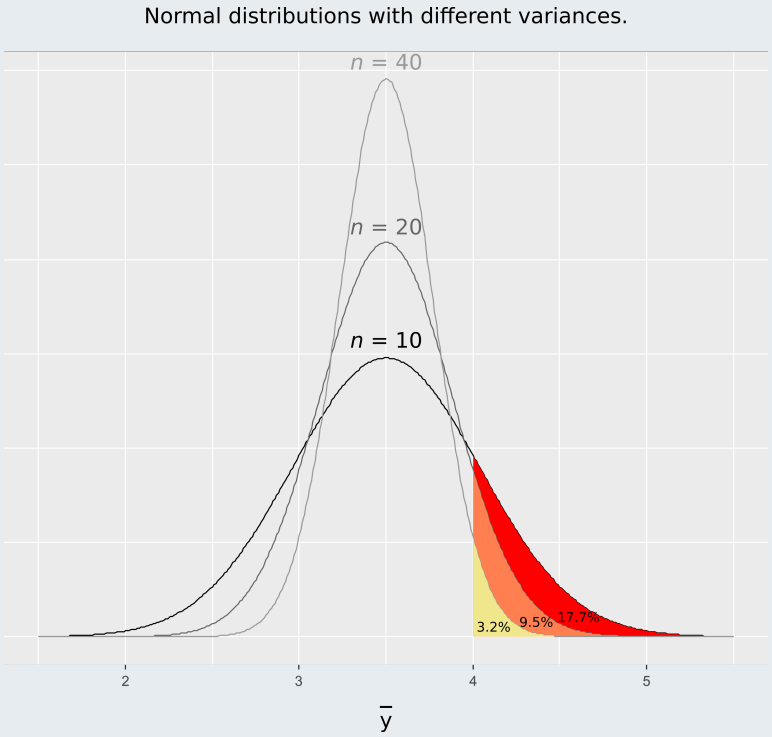
```
[1] 0.09534175
```

The probability of getting a $\bar{y}$ that is "far away" from the true mean of 3.5 is getting smaller as n increases! Let's try one more time for 40 dice.

$$\frac{\sigma_y^2}{n} = \frac{35/12}{40} \approx 0.073$$

```
pnorm(4, mean = 3.5, sd = sqrt(0.073), lower.tail = FALSE)
```

```
[1] 0.03211478
```



sample size	var($\bar{y}$)	P($\bar{y} > 4$)
10	0.292	17.7%
20	0.146	9.5%
40	0.073	3.2%

The Normal curves, and $P(\bar{y} > 4)$, for $n = 10, 20, 40$, are shown above.

9.4 Sampling distribution of $\bar{y}$ with unknown μ

The sampling distribution of $\bar{y}$ is:

$$\bar{y} \sim N\left(\mu_y \ , \ \frac{\sigma_y^2}{n}\right)$$

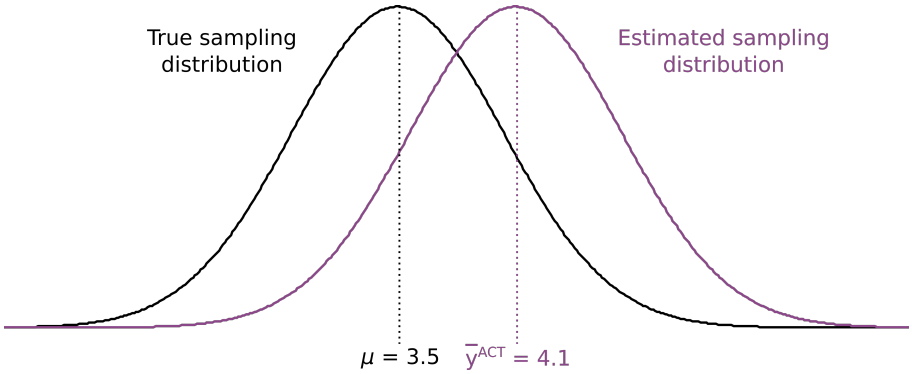
The problem is, in reality μ_y is usually not known![2] We need to calculate $\bar{y}$ as a way of estimating the unknown true population mean μ_y. So, in reality we don't know where the sampling distribution is "centered". That is:

$$\bar{y} \sim N\left(? \ , \ \frac{\sigma_y^2}{n}\right)$$

How can we calculate probabilities involving $\bar{y}$? We need to locate the sampling distribution before we can do so.

[2] Our dice examples are an exception.

Figure 9.1: An "actually" calculated value for the sample mean $\bar{y}^{ACT}$ is used to locate the sampling distribution, since the true location μ_y is typically unknown.



What is our best guess for the unknown population mean μ_y? It's $\bar{y}$! We can estimate the sampling distribution for $\bar{y}$ by replacing the unknown μ_y with a value that we "actually" calculate for the sample average. Call this value $\bar{y}^{ACT}$, where ACT stands for a number that we actually calculate:

$$\bar{y} \sim N\left(\bar{y}^{ACT}\ ,\ \frac{\sigma_y^2}{n}\right)$$

The idea of replacing the unknown parts of the sampling distribution with estimated numbers is the beginning step in constructing confidence intervals, and performing hypothesis tests.

Example 9.4 — An estimated sampling distribution. Suppose that we do not know that the true mean of a die roll is $\mu_y = 3.5$! We can estimate this unknown mean using the sample average, and use it for confidence intervals and hypothesis tests. A typical way of proceeding is to:

1. Collect a sample.
2. Calculate $\bar{y}$ to use as an estimate for μ_y.
3. Use $\bar{y}$ in place of μ_y in the sampling distribution, in order to calculate confidence intervals and hypothesis tests.

Begin by collecting a sample. Roll 10 dice:

```
set.seed(2040)
roll <- sample(1:6, 10, replace=TRUE)
roll
```

```
[1] 1 4 5 2 6 6 6 4 1 6
```

Here we used `set.seed(2040)` so that all the randomly generated numbers will be the same no matter who runs the code! Next, use this sample to calculate $\bar{y}^{ACT}$:

Random seed. Random number generation begins with a "seed". Complicated formulas are applied to the seed, so complicated that we can't predict the result. This generates "pseudo" random numbers. If we do not choose a seed, the computer default is to use the system time.

```
mean(roll)
```

```
[1] 4.1
```

Lastly, we can calculate probabilities of getting different values for $\bar{y}$, by using the $N\left(\bar{y}^{ACT}, \sigma_y^2/n\right)$ distribution. In Example 9.1 we calculated the probability of getting an "extreme" $\bar{y}$. Let's calculate the probability that, if we were to draw another sample of size $n = 10$, that the $\bar{y}$ we calculate from this sample is within ± 1 of $\bar{y}^{ACT}$. That is, we want: $\text{P}(3.1 \leq \bar{y} \leq 5.1)$. To get this probability, we use a Normal distribution with $\mu = 4.1$ and $\sigma_y^2/n = 2.92/10 = 0.292$.

Notice that $\text{P}(3.1 \leq \bar{y} \leq 5.1) = \text{P}(\bar{y} \leq 5.1) - \text{P}(\bar{y} \leq 3.1)$. We calculate two probabilities in R, and subtract:

```
pnorm(5.1, mean = 4.1, sd = sqrt(0.292))
  - pnorm(3.1, mean = 4.1, sd = sqrt(0.292))
```

```
[1] 0.9357704
```

This tells us that, if the true population mean were 4.1, there would be a 93.6% chance of calculating a $\bar{y}$ between 3.1 and 5.1 with a new sample of size $n = 10$. These types of probability statements, involving what would happen if we could hypothetically recalculate $\bar{y}$ with a new sample, forms the basis for confidence intervals and hypothesis testing.

9.5 Confidence intervals

Statistical inference typically includes some measure of the uncertainty surrounding the actual estimate. A confidence interval accomplishes this task. It is an interval surrounding an estimate (such as $\bar{y}$), that reflects "accuracy". The wider the interval, the less confident we are about how well $\bar{y}$ represents or is "close" to the true μ_y.

Measuring uncertainty surrounding an estimate, such as by using a confidence interval, relies on knowing the distribution of the estimator. Luckily, in Section 9.4, we determined the distribution of $\bar{y}$ to be:

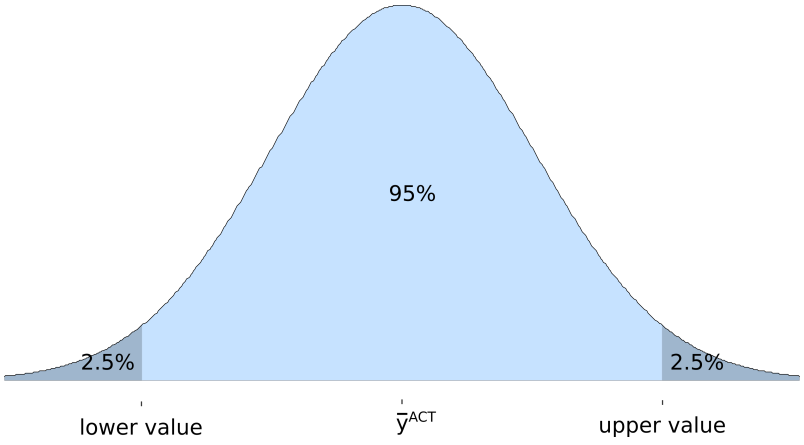
$$\bar{y} \sim N\left(\mu_y \ , \ \frac{\sigma_y^2}{n}\right)$$

and said that we can replace the unknown μ_y with an actual value for $\bar{y}$:

$$\bar{y} \sim N\left(\bar{y}^{ACT} \ , \ \frac{\sigma_y^2}{n}\right)$$

Now, consider the following question:

Figure 9.2: Solving for "lower value" and "upper value" provide the 95% confidence interval around $\bar{y}^{ACT}$.



> Which interval around $\bar{y}^{ACT}$ has a 95% probability of containing a new $\bar{y}$?

To answer this question, we could draw the $N\left(\bar{y}^{ACT}, \frac{\sigma_y^2}{n}\right)$ distribution, put 95% of the area in the middle, and figure out the lower and upper bounds. See Figure 9.2.

Example 9.5 — Confidence intervals using R. Returning to an earlier dice example (Example 9.4):

$$n = 10$$
$$\bar{y} = 4.1$$
$$\sigma_y^2 = 2.92$$
$$\sigma_y^2/n = 0.292$$
$$\sqrt{\sigma_y^2/n} = 0.54$$

The estimated sampling distribution for $\bar{y}$ is $N(4.1, 0.292)$. We can use R to find the lower value, that puts 2.5% of the area under the curve to the left:

```
qnorm(.025, mean = 4.1, sd = 0.54)
```

```
[1] 3.040894
```

Find the value that puts 2.5% of the area under the curve to the right:

```
qnorm(.025, mean = 4.1, sd = 0.54, lower.tail = FALSE)
```

```
[1] 5.159106
```

These two values define the confidence interval around $\bar{y}^{ACT}$: [3.04 , 5.16].

In addition to using R (Example 9.5), we can also find the 95% confidence interval

using:

$$\text{lower value} = \bar{y} - 1.96 \times \sqrt{\frac{\sigma_y^2}{n}}$$
$$\text{upper value} = \bar{y} + 1.96 \times \sqrt{\frac{\sigma_y^2}{n}}$$

or:

95% confidence interval.

$$\left[\bar{y} - 1.96 \times \sqrt{\sigma_y^2/n} \quad , \quad \bar{y} + 1.96 \times \sqrt{\sigma_y^2/n}\right] \tag{9.1}$$

The number 1.96 in Equation 9.1 is coming from the Standard Normal distribution: $\text{N}(0, 1)$ (see Section 6.4.3). In a Standard Normal distribution, 2.5% of the area in the "tails" is located outside of the values -1.96 and 1.96.

Instead of drawing out the $N\left(\bar{y}^{ACT}, \sigma_y^2/n\right)$ distribution and calculating areas, Equation 9.1 uses the values $\bar{y}^{ACT}$ and σ_y^2/n in order to transform the distribution to the Standard Normal distribution $N(0, 1)$, where the number 1.96 is well known. Essentially, we are "standardizing"[3] $\bar{y}$: creating a different variable that instead follows $N(0, 1)$, and using what we know about $N(0, 1)$ (that ± 1.96 puts 95% area in the middle).

Example 9.6 — Confidence intervals using Equation 9.1. Returning to earlier dice examples (Examples 9.4 and 9.5):

$$n = 10$$
$$\bar{y} = 4.1$$
$$\sigma_y^2 = 2.92$$
$$\sigma_y^2/n = 0.292$$
$$\sqrt{\sigma_y^2/n} = 0.54$$

Calculate the 95% confidence interval using Equation 9.1:

$$\begin{aligned} 95\%\ \text{CI} &= \left[\bar{y} - 1.96 \times \sqrt{\sigma_y^2/n}\ ,\ \bar{y} + 1.96 \times \sqrt{\sigma_y^2/n}\right] \\ &= 4.1 \pm (1.96 \times 0.54) \\ &= 4.1 \pm 1.06 \\ &= [3.04\ ,\ 5.16] \end{aligned}$$

This is the same confidence interval from Example 9.5!

9.5.1 Standard error

The "standard error" is the standard deviation of an estimator. The special name "standard error" is used instead of "standard deviation" when referring to an estimator in particular, rather than just any random variable. The sample standard error is used to calculate confidence intervals and perform hypothesis tests. The standard error of $\bar{y}$, for example, is often abbreviated $\text{s.e.}(\bar{y})$.

[3] Standardization will be covered in detail in the next chapter.

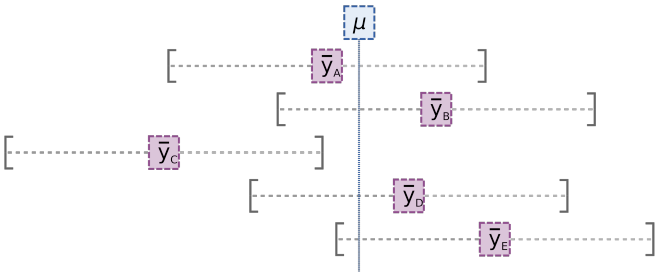
Standard error. $\sqrt{\sigma^2/n}$ is called the "standard error" of $\bar{y}$, abbreviated s.e.($\bar{y}$).

9.5.2 Interpreting confidence intervals

So we have calculated an interval around $\bar{y}$, but what does it all mean? There are several ways to interpret a 95% confidence interval:

- There is a 95% probability that a 95% confidence interval will contain the true μ_y.
- 95% of such intervals constructed in this way will contain the true μ_y.
- We are confident that 95% of the time, the interval contains the truth.

Figure 9.3: Each hypothetical sample of size n that we could draw (sample A, sample B, etc.) provides a 95% confidence interval that has a 95% probability of containing the true population mean μ_y. In reality, we will only draw one sample from the population, and calculate one sample mean and interval. The confidence interval provides a measure of the uncertainty surrounding $\bar{y}$.



The confidence interval is itself a random interval. The randomness all begins with the random sample. From the random sample we get $\bar{y}$, which is a random variable. From $\bar{y}$ we get the confidence interval. Since $\bar{y}$ is random, so must be the confidence interval.

It turns out that there is a 95% probability that we will draw a sample that leads to a 95% confidence interval containing μ_y. Of all the possible samples that we could draw from the population, 95% of them will produce 95% confidence intervals that contain the truth.

How not to interpret a confidence interval

Some misconceptions on how to interpret confidence intervals persist. The following interpretations are *wrong*:

- There's a 95% probability that the true μ_y lies inside the 95% confidence interval.
- The 95% confidence interval contains the true μ_y 95% of the time.

These interpretations are subtly wrong. The reason is that the interval is random, and μ_y is a fixed parameter, not the other way around.

Margin of error

The distance in the confidence interval, on either side of $\bar{y}$, is sometimes called the *margin of error*. That is:

$$95\% \text{ margin of error} = 1.96 \times \sqrt{\sigma^2/n}$$

The term "error" is in keeping with the idea that the confidence interval is measuring the uncertainty surround an estimate.

9.5.3 The width of a confidence interval

From the equation for the 95% confidence interval:

$$\bar{y} \pm 1.96 \times \sqrt{\sigma_y^2/n}$$

we can see several factors that will effect the *width* or size of the interval:

- The number "1.96". This number is associated with a 95% *confidence level*. A 90% confidence level will make the interval narrower, and a 99% confidence level will make the interval wider.
- The sample size n. As n grows, $\bar{y}$ becomes more accurate, and confidence intervals narrow.
- The population variance σ_y^2. If there is more variance in the population, then $\bar{y}$ becomes less accurate and confidence intervals widen.

Notice that $\bar{y}$ only determines the location of the interval, not it's width.

Example 9.7 — Effect of a larger n. In Example 9.6 we calculated a 95% confidence interval from a sample of size $n = 10$. Let's take another sample from the same population, but with $n = 20$. From Section 9.3 and Example 9.3, we know that the accuracy of $\bar{y}$ increases for larger n. This should be reflected in a *narrower* confidence interval when $n = 20$.

In fact, the width of the confidence interval is determined by the margin of error, not the actual value for $\bar{y}$. The width of the interval is:

$$2 \times \left(1.96 \times \sqrt{\sigma_y^2/n}\right)$$

As long as we are sampling from the same population so that σ_y^2 is constant, then the width of the confidence interval only depends on n. From the dice examples, where the true variance of a die roll is $\sigma_y^2 = {}^{35}/_{12}$, the margin of error from a 95% confidence interval using $n = 20$ will be:

$$\left(1.96 \times \sqrt{(35/12)/20}\right) = 0.749$$

so that any 95% confidence interval calculated by sampling $n = 20$ from this population will just be:

$$\bar{y} \pm 0.749$$

Comparing this to the 95% confidence interval from when $n = 10$:

$$\bar{y} \pm 1.06$$

we can see that doubling the sample size decreases the width of the confidence interval by a factor of $\sqrt{2}$.

9.5.4 Confidence level

Common confidence levels are 90%, 95%, and 99%, but any confidence level may be chosen. A confidence level determines the probability that a confidence interval will contain the true population parameter (μ_y). We can use R to find the "critical values" that correspond to these confidence levels. For a 99% confidence interval:

```
qnorm(.005, mean = 0, sd = 1)
```

```
[1] -2.575829
```

In the `qnorm()` function we chose the area `.005`. This calculates the value for a Standard Normal variable that puts 0.5% area in the left tail. This gives 1% area in both tails, hence 99% area in the middle. Similarly, for the 95% critical value, we put 2.5% area in the left tail of the distribution:

```
qnorm(.025, mean = 0, sd = 1)
```

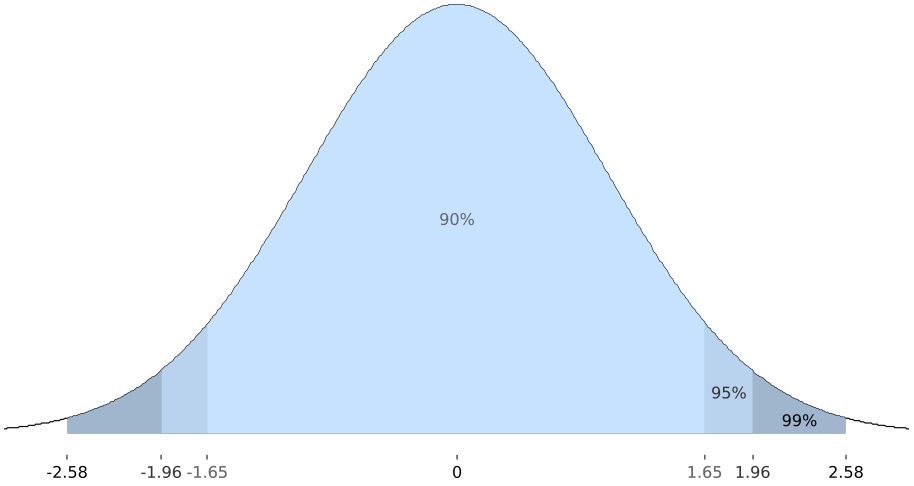
```
[1] -1.959964
```

and finally for the 90% confidence interval:

```
qnorm(.05, mean = 0, sd = 1)
```

```
[1] -1.644854
```

Figure 9.4: Standard Normal distribution. "Critical values" of ±2.58, ±1.96, and ±1.65 are used to construct 99%, 95%, 90% confidence intervals (respectively). These numbers can be used when $\bar{y}$ (at least approximately) follows a Normal distribution.



These critical values are depicted in Figure 9.4. Using these values, we can construct confidence intervals with varying levels of confidence:

$$99\%\ \text{CI} = \bar{y} \pm 2.58 \times \sqrt{\sigma_y^2/n}$$
$$95\%\ \text{CI} = \bar{y} \pm 1.96 \times \sqrt{\sigma_y^2/n}$$
$$90\%\ \text{CI} = \bar{y} \pm 1.65 \times \sqrt{\sigma_y^2/n}$$

Example 9.8 — Confidence intervals for Mars incomes. Let's calculate 99%, 95% and 90% confidence intervals using the sample of Mars colonists. In this chapter, we are operating under the unrealistic assumption that the population variance is *known*. Since the Mars data is fake (I generated it), in this example we can know what the

true population variance is. Load up the entire population of 720,720 colonists:

```
wholepop <- read.csv("http://ryantgodwin.com/data/marsregistry.csv")
```

This will take awhile since the file is large! Now that we can unrealistically "see" the entire population, we can get the population variance of income for "employed" individuals:

```
var(wholepop$income[wholepop$occupation == "employed"])
```

```
[1] 1452833175
```

So, the true population variance is $\sigma^2_{income} = 1.4$ billion. Now, pretend that we do not have any other information on the population other than $\sigma^2_{income} = 1.4$ billion. Let's use the sample of $n = 1000$ colonists, calculate the sample mean income $\bar{income}$, and construct the confidence intervals.

```
sample <- read.csv("http://ryantgodwin.com/data/mars.csv")
mean(sample$income)
```

```
[1] 80938.1
```

Let's calculate the "standard error" first, to make things easier:

$$s.e.(\bar{y}) = \sqrt{\sigma^2/n} = \sqrt{1452833175/1000} = 1205.335$$

and now for the confidence intervals:

$$99\%\ \text{CI} = 80938.1 \pm 2.58 \times 1205 = [77829\ ,\ 84047]$$
$$95\%\ \text{CI} = 80938.1 \pm 1.96 \times 1205 = [78576\ ,\ 83299]$$
$$90\%\ \text{CI} = 80938.1 \pm 1.65 \times 1205 = [78950\ ,\ 82926]$$

10. Hypothesis testing

In this chapter, we introduce hypothesis testing. Hypothesis testing involves assessing statements made about unknown population parameters. One of the unknown population parameters that we have been focusing on in this book is the population mean μ_y. For example, we might hypothesize that the true population mean height of U of M students is 173 cm, that the mean income of Mars colonists is 82000, or that GDP growth is 2% per year. A hypothesis test uses the information in the sample to assess the plausibility of such statements. Having knowledge of the sampling distribution of $\bar{y}$, which is our estimator for the unknown part of the hypothesis test μ_y, is the key to conducting a hypothesis test.

In this chapter, we are maintaining the unrealistic assumption that the population variance σ_y^2 is known. We relax this assumption in the next chapter.

10.1 Null and alternative hypotheses

In general, a hypothesis test begins with a null hypothesis, and an alternative hypothesis:

$$H_0 : \mu_y = \mu_{y,0}$$
$$H_A : \mu_y \neq \mu_{y,0}$$

H_0 is the null hypothesis. The null hypothesis is "choosing" a value for the unknown population mean, μ_y. The hypothesized value of the population mean is denoted $\mu_{y,0}$. The alternative hypothesis is denoted by H_A. One of the two situations must occur. This is called a "two-sided" hypothesis test: the null hypothesis is wrong if the population mean (μ_y) is either "too small" or "too big" relative to the hypothesized value.

The hypothesis test concludes with either: (i) "reject" H_0 in favour of H_A, or (ii) "fail to reject" H_0. We should never say that we "accept" either of the hypotheses: we either have evidence to reject H_0, or we do not have enough evidence to reject H_0.

The decision to "reject" or "fail to reject" H_0 may begin by the researcher *subjectively* deciding on a *significance level* and then doing one or more of the following:

- Calculating a (p-value) and comparing it to the significance level.
- Seeing whether or not $\mu_{y,0}$ is contained in a confidence interval.
- Calculating a test statistic and seeing if it exceeds a critical value.

We'll use the example of the incomes of Mars colonists in order to illustrate hypothesis testing. Suppose that the Mars government claims that the population mean income of employed Mars colonists is 82,000. Let's begin by formally stating the null and alternative hypotheses:

$$\begin{aligned} H_0 &: \mu_{income} = 82000 \\ H_A &: \mu_{income} \neq 82000 \end{aligned} \tag{10.1}$$

If we think the government is lying, we will reject their claim. This is a "two sided" hypothesis test; the government is lying if the true income is either greater than or less than 82000.[1]

10.2 Distribution of $\bar{y}$ assuming H_0 is true

"Classical" hypothesis testing (as we are doing here), begins by making the assumption:

Hypothesis Testing Assumption 1. The null hypothesis is correct (H_0 is true).

The hypothesis test concludes by re-evaluating this assumption. If this assumption appears to be incorrect, we "reject" the null hypothesis. We make an additional assumption for this chapter:

Hypothesis Testing Assumption 2. The true population variance is known. For the Mars income example, this means that σ^2_{income} is assumed to be known.

Now, consider this important question:

> What is the distribution of the sample average, assuming that the null hypothesis H_0 is true?

The sample average, $\bar{income}$, has a Normal distribution with mean equal to μ_{income}, and variance equal to σ^2_{income}/n. That is:

$$\bar{income} \sim N\left(\mu_{income}\ ,\ \frac{\sigma^2_{income}}{n}\right)$$

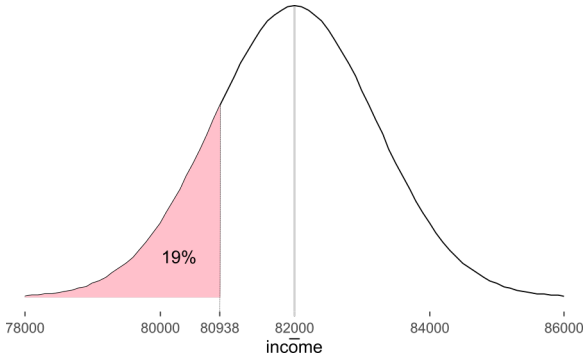
If H_0 is true, then $\mu_{income} = \mu_{income,0}$. In our example, this would mean that $\mu_{income} = 82000$ and:

$$\bar{income} \sim N\left(82000\ ,\ \frac{\sigma^2_{income}}{n}\right)$$

Using our unrealistic assumption that the population variance is known, we also have that $\sigma^2_{income}/n = 1452833$. We can draw the sampling distribution for the sample mean, assuming that the null hypothesis H_0 is true. See Figure 10.1. We can also use this sampling distribution to calculate the probability of drawing a sample that leads to an "extreme" or "weird" sample average, like we did in Example 9.1.

[1] We cover one-sided hypothesis tests in Section 10.6.

Figure 10.1: Sampling distribution of the sample average income $\bar{income}$ if H_0 : $\mu_{income} = 82000$ is correct. $2 \times 19\% = 38\%$ is the probability of getting a "worse" sample average, and is called the p-value.



Example 10.1 — Probability of an extreme sample average assuming H_0 true. Assume that H_0 : $\mu_{income} = 82000$ is true. What is the probability of getting $\bar{income} < 80000$? We need to calculate the area under the $N(82000\ ,\ 1605382)$ curve, to the left of 80,000:

```
pnorm(80000, mean = 82000, sd = sqrt(1452833))
```

```
[1] 0.04852874
```

If the population mean income is truly 82000, then there is only a 4.9% chance that $\bar{income} < 80000$.

10.3 p-values

After stating H_0 and H_A, the next step is to actually estimate the parameter (μ_y for example) that the hypothesis is about. A p-value can then tell us whether the difference between the what we hypothesis (H_0 : $\mu_{y,0}$) and what we actually observe from the sample ($\bar{y}$) is "large" enough to warrant rejection of the hypothesis.

For example, to test whether $\mu_{income} = 82000$ or not, we proceed by estimating the unknown μ_{income}. In several examples using a sample of $n = 1000$ employed Mars colonists we have calculated that:

$$\bar{income} = 80938$$

Notice that our estimate of 80938 is clearly different from our hypothesis that the true population mean is 82000. The difference between what we actually estimated from the sample, and our null hypothesis, is $82000 - 80938 = 1062$. Just because there is a difference does not imply we should reject H_0 outright. We need to assess whether this difference is "large". Assessing whether the difference is large can be accomplished using a p-value. We will only reject H_0 if the probability of getting an $\bar{income}$ (from another hypothetical sample drawn from the population) further away than 1062, is small. This probability is called a p-value.

p-value. A p-value is the probability of getting a new (hypothetical) estimate that is more adverse to the null hypothesis than the estimate just calculated, assuming the null hypothesis is true.

By "more adverse" we mean a difference $\bar{y} - \mu_{y,0}$ that is even larger than the difference calculated with our given sample. If H_0 is actually true, then the probability of calculating a sample average that is more "extreme" than the one we just calculated is two times[2] the probability that $\bar{income} < 80938$, using the $N(82000, 1605382)$ curve. From R, this probability is:

```
2 * pnorm(80938, mean = 82000, sd = sqrt(1452833))
```

```
[1] 0.3782731
```

This is the p-value for our example hypothesis test. It tells us that, if H_0 is true, there is a 38% chance of getting a sample that would lead to an $\bar{income}$ that is further away from 82000 than the sample average of 80883 that was just calculated. That is, out of all the hypothetical samples of $n = 1000$ that we could draw from an $N(82000, 1452833)$ distribution, 38% would give a sample average further than $82000 - 80883 = 1117$ from H_0.

All that remains is to decide whether the p-value of 38% is "large" or "small". This decision is subjective. With a p-value of 38%, most researchers would decide to "fail to reject" the null hypothesis.

Hypothesis testing decision rule.

- If the p-value is large, we "fail to reject" H_0. A large p-value indicates that there are many worse estimates (e.g. $\bar{y}s$) that we could calculate, relative to H_0, if H_0 is actually true. A large p-value indicates that the sample average is "close" to H_0.
- If the p-value is small, we "reject" H_0 in favour of H_A. A small p-value indicates that there are few sample averages that we could observe that are more extrem, if H_0 is actually true. A small p-value indicates that the sample average is "far" from H_0.
- If the p-value is greater than the *significance level*, then the p-value is considered large. That is, if p-value $< \alpha$, reject H_0.

10.4 Significance of a test (α)

At what point should we decide that the p-value is too small, such that we should reject the null hypothesis? The choice is somewhat arbitrary, and is up to the researcher (you). Standard choices are 10%, 5%, and 1%. A pre-decided maximum p-value under which H_0 will be rejected is called the *significance level* of the test. It is sometimes denoted by α. In the Mars income example, we would fail to reject the null at the 10% significance level. Note that failing to reject at the 10% level implies that we also fail to reject H_0 at the 5% and 1% significance levels.

[2] We multiply by 2 because it is a two-sided hypothesis test.

10.4.1 Type I error

Take another look at Figure 10.1. Suppose that the null hypothesis is true and Figure 10.1 is the correct sampling distribution for $\bar{income}$. We could still randomly draw a weird sample that makes H_0 appear to be "wrong". That is, even when the null is true, some of the hypothetical samples we could draw would give an $\bar{income}$ that is far from the truth. In these cases, we will erroneously reject the null. If the null hypothesis is falsely rejected, it is called a *type I error*. Type I error is the probability that H_0 is rejected when the null is true:

$$\text{P(type I error)} = \text{P(reject } H_0 \mid H_0 \text{ is true)} = \alpha$$

How do we determine what this type I error will be? As soon as we pick the significance of the test, it has been determined. That is, type I error $= \alpha$. If we always pick a 5% significance level, we will make a type I error in 5% of hypothesis tests. That is, if we conduct thousands of scientific studies where we always use $\alpha = 5\%$, in 5% of those studies where we reject the null, we will be doing so falsely.

In reality, we do not know the population values, so we will never know if we have made a type I error or not. Type I error tells us nothing about the particular sample that we are working with. It only tells us something about what happens through repeated applications of our testing procedure.

10.4.2 Type II error

There is another type of error we can make. There are two possibilities for H_0: either it is true or false. In type I error, we considered that H_0 is actually true. If we consider that H_0 is actually false, then we make a *type II error* when we *fail to reject*. The probability of a type II error is:

$$\text{Pr(type II error)} = \text{Pr(fail to reject } H_0 \mid H_0 \text{ is false)}$$

If H_0 is actually false, one of two things can happen: we "reject" or we "fail to reject". The probabilities of both of these events must sum to 1 (something must happen). So:

$$\text{Pr}(1 - \text{type II error}) = \text{Pr(reject } H_0 \mid H_0 \text{ is false)} \tag{10.2}$$

Equation (10.2) is called the *power* of the test. We want the power to be as high as possible. That is, we do not want to make a type II error, and we want the probability of rejection to be as high as possible when H_0 is actually false.

Determining the type II error (and power) of a test is difficult or impossible. This is because power depends on knowing the unobservable population. The concept is useful, however, when we are trying to find the "best" test available (which is not covered in this book).

10.4.3 Trade-off between type I and II errors

We choose the significance (α) of the test (e.g. either 1%, 5%, or 10%). Type I error is equal to α. So why don't we just choose α to be really small, in order to minimize our type I error? The answer is that there is a trade-off between type I error and type II error. Generally, as type I error decreases, type II increases. A small α means that the sample mean can be far away from the null hypothesis before it is rejected. Confidence intervals get wider. It becomes more difficult to reject the null hypothesis, whether it is true or false.

10.5 The z test statistic

A *test statistic* is a convenient way of measuring the difference between the null hypothesis and what is actually estimated. Test statistics provide an alternative way of obtaining a p-value. If we want to use the above testing procedure in different situations, we would have to "graph" a different Normal curve (similar to the one in Figure 10.1) each time, in order to calculate the area under the curve to get the p-value. Historically, calculating an area under the Normal curve was difficult (now it is easily done in R). Consequently, a method was devised so that every testing problem could use the same distribution: the *Standard Normal* (see Section 6.4.3. Historically, areas under the Standard Normal curve were tabulated to provide p-values without the need for integration or a computer. One of these tables is reproduced in Table 10.1.

The z-statistic is obtained by *standardizing*. To standardize a variable, we subtract its mean and divide by its standard deviation. If the variable we begin with is Normal, then this process creates a new Normal random variable from the old one, which follows the $N(0,1)$ distribution. For example, let $y \sim N(\mu_y, \sigma_y^2)$. We standardize by creating a new variable z where:

$$z = \frac{y - \mu_y}{\sigma_y}$$

Now, z is still Normally distributed, but has mean 0 and variance 1 since:

$$\mathrm{E}[z] = \mathrm{E}[y - \mu_y] = \mathrm{E}[y] - \mu_y = \mu_y - \mu_y = 0$$

and:

$$\mathrm{Var}[z] = \mathrm{Var}\left[\frac{y}{\sigma_y}\right] = \frac{\mathrm{Var}[y]}{\sigma_y^2} = \frac{\sigma_y^2}{\sigma_y^2} = 1$$

(refer to the rules of mean and variance in Sections 7.7.2 and 7.7.4).

How is standardization helpful for hypothesis testing? The sampling distribution of $\bar{y}$ under the null hypothesis is $\bar{y} \sim N(\mu_{y,0}, \sigma_y^2/n)$. Create a new variable z. Subtract $\mu_{y,0}$ (the mean of $\bar{y}$ if the null is true) from $\bar{y}$. z has mean 0 (if the null is actually true). Divide by the standard error (standard error = the standard deviation of an estimator) of $\bar{y}$, and z has variance of 1. That is:

$$z = \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma_y^2/n}} \sim N(0,1)$$

This is the "z-test statistic" for the null hypothesis that $\mu_y = \mu_{y,0}$. If the null is true, then z should be "close" to 0. The probability of observing a $\bar{y}$ further away from H_0 than what we just observed from the sample is obtained by plugging $\bar{y}$ and $\mu_{y,0}$ into the z statistic formula, and calculating a probability using the Standard Normal distribution. From our Mars incomes example, the z statistic is:

$$z = \frac{80938 - 82000}{\sqrt{\frac{1452833175}{1000}}} = -0.881$$

Now, the question:

> "What is the probability of getting further away than 80938 from the null hypothesis of 82000?"

has just been translated to:

> "What is the probability of an $N(0, 1)$ variable being less than -0.881, or greater than 0.881?"

Get this probability from R:

```
2 * pnorm(-0.881, mean = 0, sd = 1)
```

```
[1] 0.3783178
```

It is the same p-value that we obtained in Section 10.3! We only need to calculate the area under the curve for several possible z values. These were tabulated long ago, and are reproduced in Table 10.1.

Example 10.2 — Hypothesis on Mars incomes - z test.

State the null and alternative hypotheses.

$$H_0 : \mu_{income} = 82000$$
$$H_A : \mu_{income} \neq 82000$$

Choose the significance level.

Let's choose $\alpha = 5\%$.

Collect a sample.

Load the sample of 1000 employed Mars colonists:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

Estimate the population parameter in the hypothesis.

We calculate $\bar{income}$, which is an estimator for the unknown μ_{income}:

```
mean(mars$income)
```

```
[1] 80938.1
```

Calculate the z test statistic.

$$z = \frac{80938.1 - 82000}{\sqrt{\frac{1452833175}{1000}}} = -0.881$$

Calculate the p-value.

The Normal distribution is symmetric, so we can look up the value 0.881 (instead of -0.881) in Table 10.1. The number in the table is 0.1894. Multiplying by 2 (since it's a two sided test), gives a p-value of 0.3788. Instead of the table we can get the p-value from Standard Normal distribution in R:

```
2 * pnorm(-0.881, mean = 0, sd = 1)
```

```
[1] 0.3783178
```

Make a decision.
Since the p-value is greater than the significance level, we fail to reject H_0. There is insufficient evidence to reject the claim that $\mu_{income} = 82000$. Note that we also reject the null at the 10% significance level.

10.5.1 Critical values

Critical values are the most extreme values allowable for the test statistic, before the null hypothesis is rejected. Suppose that we choose an $\alpha = 5\%$ significance level for our test. This means that if we receive a p-value that is less than 0.0250 in Table 10.1, we should reject the null hypothesis (since $2.5\% \times 2 = 5\%$). If we use Table 10.1 to find the z statistic that corresponds to a significance level, we are finding the critical value for the test. According to Table 10.1, we see that a p-value of 0.0250 corresponds to a z statistic of 1.96. This is the 5% critical value. We know that if the z statistic that we calculate for our test ends up being greater than 1.96 or less than -1.96, we will get a p-value that is less than 0.05, and we will reject the test.

A rejection rule. In a two-sided hypothesis test, H_0 is rejected at the 5% significance level if $|z| > 1.96$.

10.5.2 Confidence intervals again

Given a significance level α, the $1 - \alpha\%$ confidence interval may be used to "reject" or "fail to reject" a null hypothesis. For example, a 95% confidence interval can be used to conduct a hypothesis test at the 5% significance level. An alternative interpretation of the 95% confidence interval is:

Confidence interval. The 95% confidence interval contains all of the values for $\mu_{y,0}$ (all values for the null hypothesis) that will not be rejected at 5% significance.

If the null hypothesis is in the $1 - \alpha\%$ confidence interval, we will "fail to reject" that null hypothesis at the $\alpha\%$ significance level.

10.6 Two-sided vs. one-sided hypothesis tests

So far, we have only considered two-sided hypothesis tests. An alternative is a one-sided hypothesis test, which takes the form:

$$H_0 : \mu_y > \mu_{y,0}$$
$$H_A : \mu_y \leq \mu_{y,0}$$

or:

$$H_0 : \mu_y < \mu_{y,0}$$
$$H_A : \mu_y \geq \mu_{y,0}$$

On one side of $\mu_{y,0}$ the null is true, on the other side it's false. For a one-sided test, we only calculate the area under the curve on *one* side of the Normal distribution. In many instances throughout the chapter we multiplied the area under the Normal curve by 2: this is because we were conducting a *two* sided hypothesis test. For a *one* sided test, we do not multiply by 2.

Table 10.1: Area under the Standard Normal curve, to the right of z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

11. Hypothesis testing with unknown σ^2

In this chapter, we tackle the situation where the population variance σ^2 is unknown. If we want to perform hypothesis testing, we need to estimate this variance. So far our confidence intervals, test statistics, and p-values all rely on the unknown value σ^2:

confidence interval for $\bar{y}$	$\bar{y} \pm z_c \times \sqrt{\frac{\sigma_y^2}{n}}$
z test statistic	$(\bar{y} - \mu_{y,0}) / \sqrt{\frac{\sigma_y^2}{n}}$
p-value	Calculated by using the distribution of z (and z requires σ_y^2)

Note that we have written the confidence interval using z_c, instead of 1.96 (for example). When σ_y^2 is known, $z_c = 1.96$ for the 95% confidence interval. This "1.96" is coming from the Standard Normal distribution. In this chapter, z_c will change to t_c, once we estimate σ_y^2.

We will replace the unknown σ^2 with an *estimator*, s^2. We will discuss how confidence intervals, test statistics, and p-values are altered slightly when we substitute the unknown σ^2 with the estimator s^2.

11.1 Estimating σ^2

So far we have assumed that σ_y^2 is known. After calculating $\bar{y}$, we needed this σ_y^2 for confidence intervals, test statistics, and p-values. Hypothesis testing relies on knowing the population variance σ^2.

If we have to estimate μ_y, it is unlikely that we would know σ_y^2. That is, if the population mean is unknown, it is likely that the population variance is unknown as well. Equation 11.1 provides a way of estimating the unknown σ_y^2.

Sample variance of y.

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{11.1}$$

A discussion of this formula (including where the $n-1$ comes from), along with examples, were presented in Section 5.7 (review this section now). Review Example 5.11 to see how to use R to calculate the sample variance.

All instances where we used σ^2 can use s^2 instead, with some minor modifications:

confidence interval for $\bar{y}$	$\bar{y} \pm t_c \times \sqrt{\frac{s_y^2}{n}}$
t test statistic	$(\bar{y} - \mu_{y,0}) / \sqrt{\frac{s_y^2}{n}}$
p-value	Calculated by taking an area under the t-distribution

Replacing σ^2 with s^2. The reason why confidence intervals and hypothesis testing changes is because we are replacing a *parameter* (σ^2) with a *random estimator* (s^2). s^2 has its own sampling distribution. Introducing another element of randomness into confidence intervals and hypothesis testing has the effect of changing the relevant underlying distributions slightly.

Where we used the Standard Normal distribution before for confidence intervals and hypothesis testing, we should now use the t-distribution.

11.2 t-distribution

The t-distribution (in place of the Standard Normal distribution) can be used in the calculation of confidence intervals, p-values, and for hypothesis testing in general. See Section 6.5 for a review of the t-distribution. It is the appropriate distribution when σ^2 is replaced by the estimator s^2 in the formula for the z-statistic. Whereas the z-statistic follows the Standard Normal distribution:

$$z = \frac{\bar{y} - \mu_y}{\sqrt{\frac{\sigma_y^2}{n}}} \sim N(0,1)$$

the t-statistic follows the t-distribution:

$$t = \frac{\bar{y} - \mu_y}{\sqrt{\frac{s_y^2}{n}}} \sim t_{(n-k)}$$

The only difference is that σ^2 has been replaced with s^2, but introducing the random estimator s^2 into the equation changes the distribution of z. The t-distribution is denoted $t_{(n-k)}$. It is very similar to the $N(0,1)$ distribution, but it has fatter tails. It is symmetric and centered at 0. The shape of the t-distribution depends on the degrees of freedom $(n-k)$, where n is our sample size, and k (in this case) is 1.

Relationship between the t-distribution and Standard Normal. As the sample size n grows, the t-distribution becomes identical to the Standard Normal distribution. The developments in this chapter can essentially be ignored when the sample size n is large enough. That is, the Standard Normal distribution is an approximation to the t-distribution, and the approximation gets better as n increases.

11.3 Confidence intervals using s^2

When we use s^2 instead of σ^2, the "critical value" used in the confidence interval can no longer come from the Standard Normal distribution. Instead of using the critical value for a z score (z_c), we must instead use a critical value from the t-distribution (t_c):

$$\bar{y} \pm t_c \times \sqrt{\frac{s_y^2}{n}}$$

In the previous chapters, recall that if we wanted a 95% confidence interval the critical value (z_c) was found by finding the values on the x-axis that put 2.5% area in each tail of the $N(0, 1)$ distribution (see Figure 9.4). The 95% critical value using a Standard Normal distribution can be found in R using:

```
qnorm(.025)
```

```
[1] -1.959964
```

This is where the number "1.96" in the confidence interval formula comes from. The 95% critical value using the t-distribution can be found in R using:

```
qt(.025, 19)
```

```
[1] -2.093024
```

Try increasing the number "19" in the command `qt(.025, 19)`. You will see that the critical value produced approaches 1.96. This number is the "degrees of freedom" $(n - k)$ for the t-distribution. The "19" would correspond to a sample size of $n = 20$. For a sample this size, we can see that the confidence interval will be quite a bit wider under the t-distribution compared to the Standard Normal distribution. This is always the case: confidence intervals using the t-distribution are always wider than those using the Standard Normal.

Example 11.1 — Confidence intervals for Mars incomes - unknown s^2. This example mimics Example 9.8, but here we use s^2 instead of σ^2 and the t-distribution instead of the Standard Normal distribution. Let's calculate the 95% confidence intervals around the sample mean income of Mars colonists. In this chapter, we are operating under the realistic assumption that the population variance is *unknown*.

Using the sample of $n = 1000$ colonists, calculate the sample mean income ($\bar{income}$):

```
sample <- read.csv("http://ryantgodwin.com/data/mars.csv")
mean(sample$income)
```

```
[1] 80938.1
```

Let's calculate the "standard error" first, to make things easier:

```
sqrt(var(sample$income) / 1000)
```

```
[1] 1267.037
```

So, $\sqrt{s_i^2 ncome/n} = 1319$. Next we need the critical value from the t_{999} distribution:

```
qt(.025, 999)
```

```
[1] -1.962341
```

The critical value of 1.96 is a value that we have become accustomed to. This critical value from the t-distribution is nearly identical to that from the Standard Normal. This is because the sample size of $n = 1000$ is large enough that the t-distribution is approximately Normal. Finally, we can calculate the confidence interval:

$$95\%\ \text{CI} = 80938.1 \pm 1.96 \times 1267 = [78455\ ,\ 83421]$$

We will see in Example 11.2 that the confidence interval can be generated automatically in R using the `t.test()` function.

11.4 The t-test

Now that we know how to estimate σ_y^2, we can estimate the variance of the sample average using:

$$\text{Estimated variance of } \bar{y} = \frac{s_y^2}{n}$$

We can implement hypothesis testing by replacing the unknown σ_y^2 with its estimator s_y^2. The z test statistic now becomes:

$$t = \frac{\bar{y} - \mu_{y,0}}{\sqrt{\frac{s_y^2}{n}}}$$

This is the t statistic. Because we have replaced σ_y^2 with s_y^2 (a random estimator) in the z statistic formula, the form of the randomness of z has changed. The t statistic is no longer a standard normal variable. It follows its own probability distribution, called the t-distribution. When performing a t test, the p-values are different than in Table 10.1 (those obtained from the Standard Normal distribution). However, as the sample size grows, the t-distribution becomes the standard normal distribution. This means that, for sample sizes of approximately $n > 100$, using the standard normal distribution (Table 10.1) instead of the t distribution, makes very little difference.

Example 11.2 — Hypothesis on Mars incomes - t test. This example mimics Example 10.2, in which σ^2 was assumed to be known.

State the null and alternative hypotheses.

$$H_0 : \mu_{income} = 82000$$
$$H_A : \mu_{income} \neq 82000$$

Choose the significance level.

Let's choose $\alpha = 5\%$.

Collect a sample.

Load the sample of 1000 employed Mars colonists:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

Estimate the population parameter in the hypothesis.

We calculate $\bar{income}$, which is an estimator for the unknown μ_{income}:

```
mean(mars$income)
```

```
[1] 80938.1
```

Estimate the population variance.

We calculate s^2_{income}, which is an estimator for the unknown σ^2_{income}:

```
var(mars$income)
```

```
[1] 1605382317
```

Calculate the t-test statistic.

$$t = \frac{80938 - 82000}{\sqrt{\frac{1605382317}{1000}}} = -0.838$$

This is somewhat close to the value of the z test statistic in Example 10.2 (-0.881).

Calculate the p-value.

We can look up the value 0.838 in Table 10.1 to get an *approximate* p-value. The number in the table is 0.2005. Multiplying by 2 (since it's a two sided test), gives a p-value of 0.4010 (compared to 0.378 in Example 10.2). Get a p-value from the t-distribution using R:

```
2 * pt(-0.838, 999)
```

```
[1] 0.4022311
```

The p-value for this test is 0.402. The "999" in `pt(-0.838, 999)` is the degrees of freedom ($n - k = 1000 - 1$).

Using the R function `t.test()`

Use R to accomplish all of the above, in one command:

```
t.test(mars$income, mu=82000)
```

```
       One Sample t-test

data: mars$income
t = -0.8381, df = 999, p-value = 0.4022
alternative hypothesis: true mean is not equal to 82000
95 percent confidence interval:
 78451.74 83424.45
sample estimates:
mean of x
  80938.1
```

The same p-value of 0.402 was found above. Notice that `t.test()` also provides the 95% confidence interval. The null hypothesis is inside this confidence interval, so we will end up failing to reject H_0 at the 5% significance level (at least).

Make a decision.

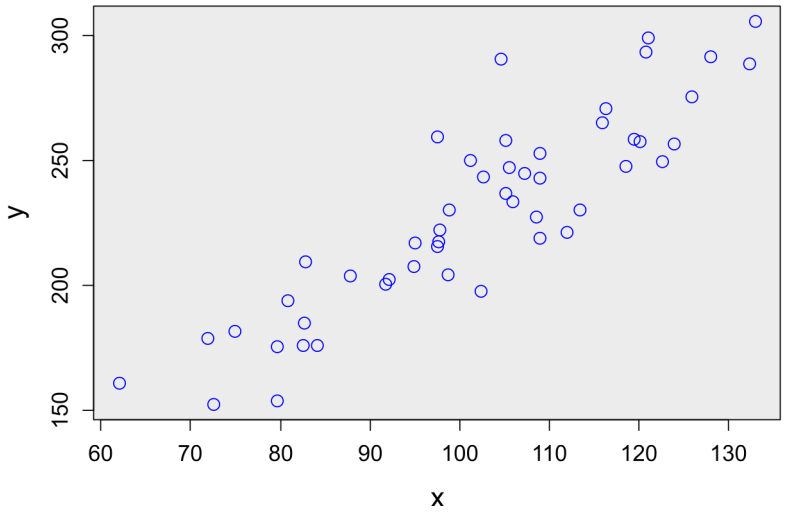
Since the p-value is greater than the significance level, we fail to reject H_0. There is insufficient evidence to reject the claim that $\mu_{income} = 82000$. Note that we also reject the null at the 10% significance level.

12. Least-squares regression

This chapter introduces *least-squares regression*, which is a way of modelling and quantifying the relationship between *two* or more variables. In the preceding chapters, we have mostly considered methods of analysis that deal with only *one* variable. In many cases in economics (and other subjects), we want to know how much a change in one variable might be associated with, or *cause*, a change in another variable.

12.1 The least-squares regression line

Figure 12.1: Scatter plot of x and y.



How might you *quantify* the relationship between two variables? The relationship between an x and a y variable is depicted in Figure 12.1 (see Section 4.7 for a review of scatter plots). Download the data from Figure 12.1, and reproduce the scatter plot yourself:

```
data <- read.csv("http://ryantgodwin.com/data/ls1.csv")
plot(data$x, data$y)
```

Looking at Figure 12.1, there appears to be a *strong*, *positive*, and *linear* relationship between x and y. One way to *quantify* this linear relationship is by using the correlation coefficient:

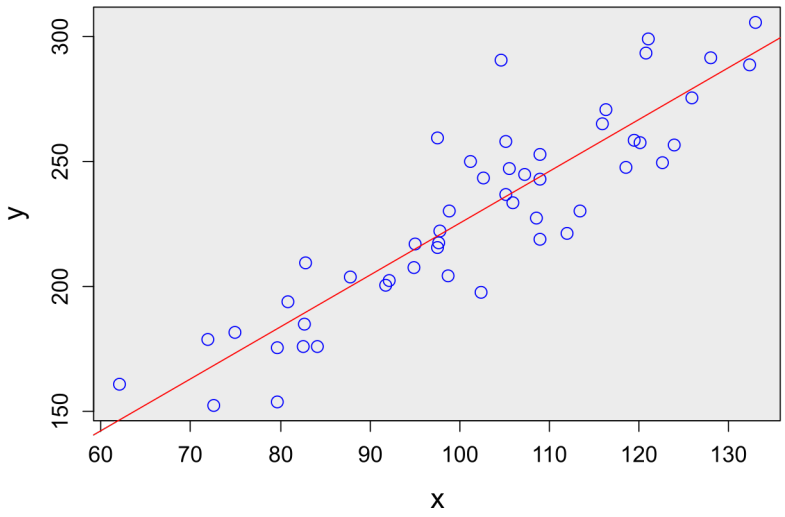
```
cor(data$x, data$y)
```

```
[1] 0.8919703
```

Another way to quantify, or *model*, the relationship between x and y is by drawing or "fitting" a straight line through the scatter plot (see Figure 12.2):

```
plot(data$x, data$y)
abline(lm(data$y ~ data$x))
```

Figure 12.2: A least-squares line has been "fitted" through the scatter plot of x and y.



The line in Figure 12.2 is called a least-squares *regression* line. The term "regression" refers to taking the information from all of the data points and "regressing" or reducing it to a single line. If the vertical distances between the data points and the line are just random "noise", then the relationship between x and y can be represented by the regression line (provided we make a few other assumptions, for example that the relationship is *linear*).

12.2 Equation of the least-squares regression line

The least-squares line from Figure 12.2 is defined by an *intercept* and a *slope*. A common way to write the equation of a line is:

$$y = a + bx$$

where y and x are variables, a is the "intercept" of the line, and b is the "slope" of the line. Often in econometrics we instead use the symbol b_0 for the intercept, and b_1

for the slope[1]. Since there is some randomness involved such that the data points in Figure 12.2 are scattered around the regression line, we will write the equation of the line in terms of $\hat{y}$ instead of y:

$$\hat{y} = b_0 + b_1 x$$

The variable $\hat{y}$ is the least-squares predicted y value, which we will explain in Section 12.5. The values for the intercept b_0 and the slope b_1 can be calculated using the `lm()`[2] function in R:

```
lm(data$y ~ data$x)
```

```
Call:
lm(formula = data$y ~ data$x)

Coefficients:
(Intercept)      data$x
     18.264       2.071
```

From the R output, the intercept for the line in Figure 12.2 is $b_0 = 18.3$. The slope of the line is $b_1 = 2.1$. The line in Figure 12.2 is written as:

$$\hat{y} = 18.3 + 2.1x$$

12.3 Interpreting the least-squares regression line

The least-squares regression line is defined by its intercept b_0 and its slope b_1. The intercept b_0 is the value for y when $x = 0$. Often, the intercept is not very interesting. The slope, however, is the change in y associated with a change in x of 1 unit[3]. The slope b_1 is sometimes called the estimated *marginal effect* of x on y:

$$\frac{\Delta \hat{y}}{\Delta x} = b_1$$

If the x variable is thought to *cause* the y variable, then b_1 is the estimated causal effect. It tells us how much y will change if we change x by 1 unit. This is very powerful knowledge, and very useful if we have control over the x variable, but we can never know if an x variable causes a y variable using statistics. *Causation* is a very strong statement and very difficult to determine. At the very least, we can say that b_1 represents the change in y *associated* with a change in x. So, in our example above, for $b_1 = 2.1$, we can say that an increase in x of 1 unit is associated with an increase in y of 2.1 units. Similarly, a decrease in x of 1 is associated with a decrease in y of 2.1.

In most of the examples and figures in this book, the least-squares regression line happens to have a *positive* slope, but the line can just as easily have a negative slope. If the value for b_1 is negative, then an increase in x is associated with a decrease in y, and vice versa.

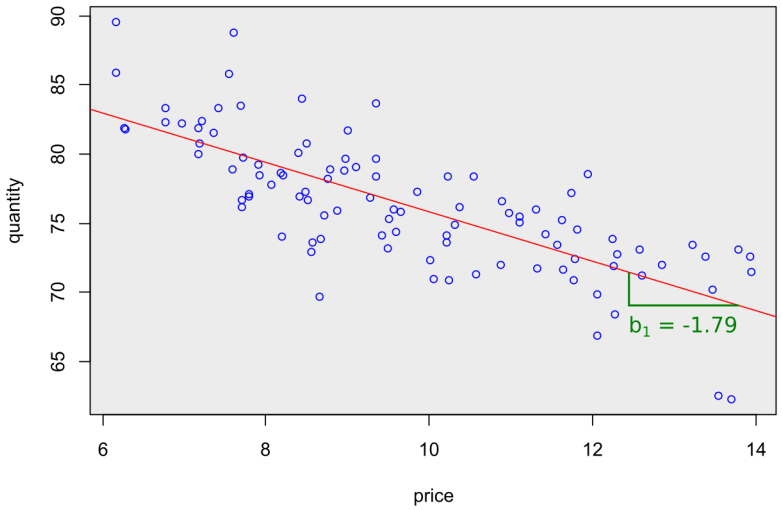
Figure 12.3 shows a situation where the least-squares regression line has a negative slope. The value $b_1 = -1.79$ means that for a 1 unit increase in price, quantity demanded decreases by 1.79 on average.

[1] In the next chapter we will introduce a *population* intercept and slope, and will denote them using β_0 and β_1.

[2] "`lm`" stands for "linear model".

[3] This interpretation is only valid when x is a continuous variable, and not, for example, a dummy variable.

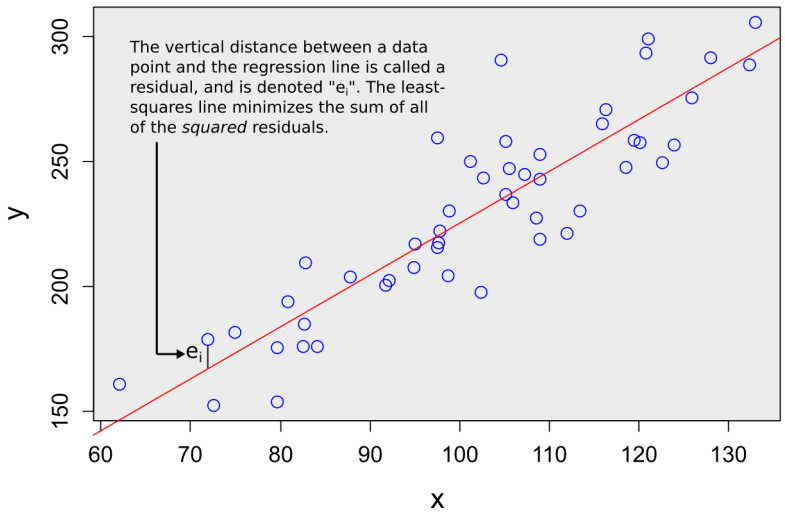
Figure 12.3: Mars completely controls the price of alcohol, and has experimented with different prices to see how colonists respond with their quantity of drinks demanded. The slope of $b_1 = -1.79$ is the average *decrease* in drinks when Mars government *increases* the price of alcohol by 1.



12.4 Formula for the intercept and slope of the regression line

How are the least-squares regression lines in Figure 12.2 and 12.3 "fitted"? That is, how are b_0 and b_1 chosen? As the name implies, the "least-squares" regression line has something to do with minimizing squared values. In particular, the line is chosen such that the sum of all of the *squared* vertical distances between the regression line and data points is *minimized*. "Least" refers to minimizing, and "squares" refers to squaring the distance between the points and the regression line.

Figure 12.4: Least-squares residuals.



Each of these vertical distances is called a "residual". There is one residuals for each data point, and we refer to a single residual as e_i. See Figure 12.4. The formulas for b_0 and b_1 can be found by solving a calculus minimization problem (which is beyond

this course):

$$b_1 = \frac{\sum_{i=1}^{n} \left[(y_i - \bar{y}) (x_i - \bar{x}) \right]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1 \bar{x} \tag{12.1}$$

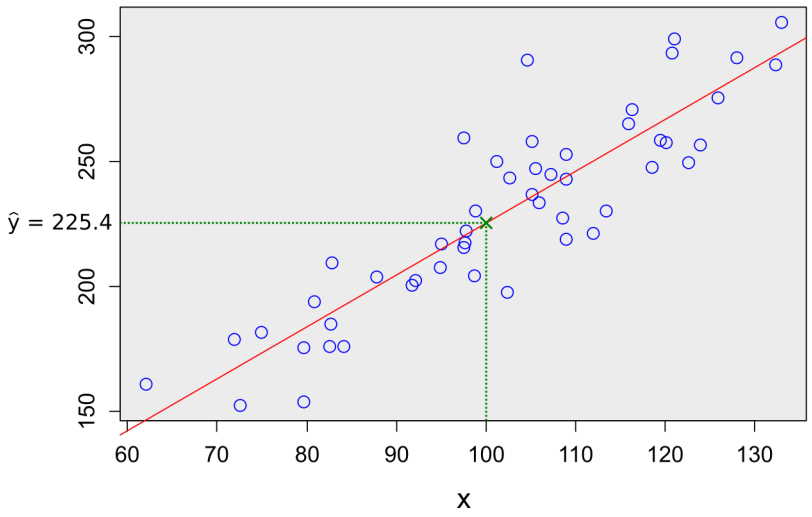
Equation 12.1 tells us how to use the x and y data in such a way to pick the intercept and slope of a line that passes closely through the data points, by minimizing the sum of all of the squared vertical distances.

12.5 Predicted values and residuals

A least-squares predicted value is the value for y that we get if we "plug" in a value for x into the equation $y = b_0 + b_1 x$. For example, say that we substitute $x = 100$ into the fitted line in Figures 12.2-12.4. Substituting $x = 100$ into the least-squares regression line we get:

$$\hat{y} = b_0 + b_1 x = 18.3 + 2.1x$$
$$\hat{y} = 18.3 + 2.1(100) = 225.4$$

Figure 12.5: Least-squares predicted value for when $x = 100$.



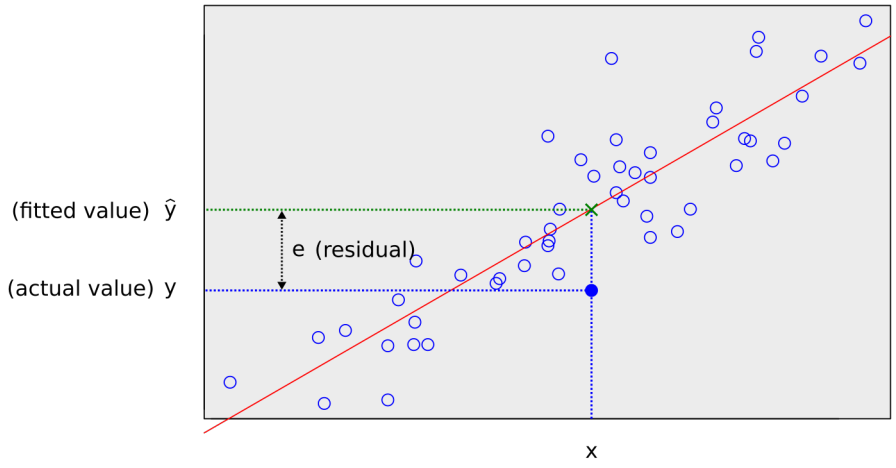
Least-squares predicted values are denoted $\hat{y}$ (pronounced "y hat"). A predicted value will always lie on the least-squares regression line. See Figure 12.5.

We can also take all of the original values for the x variable, plug all of them into the least-squares regression line, and get n "in sample" predictions. These predicted values are called the least-squares fitted values. There is one fitted value for each data point.

The fitted values $\hat{y}$, for when we use the original x data, don't quite coincide with the actual y values. See Figure 12.6. The differences between the two are the least-squares *residuals*, which we have already mentioned above. The least-squares residuals are *prediction errors*, and can be calculated by the equation:

$$e = y - \hat{y}$$

Figure 12.6: Actual value, fitted value, and residual.



The sum of all of these squared residuals is the very thing that the least-squares regression line minimizes.

Now that we have defined the least-squares residuals, we can write a new equation for the y variable:

$$y = b_0 + b_1 x + e \tag{12.2}$$

Equation 12.2 says that each y value has a predictable part $(b_0 + b_1 x)$, and an unpredictable part that cannot be explained (e)

12.6 R-squared

R-squared (R^2) is a "measure of fit" of the least-squares regression line. It is a number between 0 and 1[4]. R^2 indicates how close the data points are to the regression line. R-squared is the portion of sample variance in the y variable that can be explained by variation in the x variable.

The assumption is that changes in x are associated with or are leading to changes in y. But, changes in x are not the only reason, or explanation, for changes in y. There are *unobservable* variables that are leading to changes in y, otherwise all of the data points in the scatter plot would line up exactly in a straight line. R^2 helps answer the question: how much of the change in y is coming from x? Some equivalent ways of interpreting R^2 are:

- How well the estimated model explains the y variable.
- How well changes in x explain changes in y.
- How well the estimated regression line "fits" the data.
- The portion of the sample variance in y that can be explained using the estimated model.

[4] $0 \leq R^2 \leq 1$ as long as there is an intercept b_0.

To get the R^2 using R[5], we need to put the `lm()` function inside of the `summary()` function in order to get some more information about the "fitted" regression line:

```
summary(lm(data$y ~ data$x))
```

```
Call:
lm(formula = data$y ~ data$x)

Residuals:
    Min      1Q  Median      3Q     Max
-32.811 -12.356  -1.758   9.887  55.437

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.2636   15.6668   1.166    0.249
data$x       2.0712    0.1515 13.669   <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.01 on 48 degrees of freedom
Multiple R-squared: 0.7956,  Adjusted R-squared: 0.7914
F-statistic: 186.8 on 1 and 48 DF, p-value: < 2.2e-16
```

In this R output, make sure you can find the intercept and slope ($b_0 = 18.2636$, $b_1 = 2.0712$). There is quite a bit of information provided, including $R^2 = 0.7956$[6]. This value can be interpreted as: 79.6% of the variation in y can be explained using the x variable.

R-squared provides a measure of the predictive power of the fitted least-squares regression line. The higher the value of R^2, the more power x has for explaining or predicted values of y. A low value for R^2 does not mean that x is *insignificant*, however. Whether or not x is $significant$ in explaining changes in y is better left to a hypothesis test (covered in the next chapter).

12.6.1 R^2 and correlation

The value of R-squared from a least-squares line for x and y is closely related to the correlation between x and y: R^2 is the square of the correlation. The interpretation of R^2 is thus very similar to that of correlation; both are measuring the strength of the association between two variables. Note that due to the "squaring", R^2 does not tell us the *direction* of the relationship between the two variables (i.e. either positive or negative).

R^2 becomes more important when we add additional variables to the model, that is, when we are working with more than just two variables. In such a case, the correlation coefficient is not useful in assessing the overall "fit" or predictive power of the least-squares model.

[5] The measure of fit R-squared (R^2) and the R statistical environment (the application we are using to perform econometrics) are completely unrelated.

[6] "Multiple R-squared" can be used when we only have one x variable. When there is more than one x variable in the model, we should use the "Adjusted R-squared".

12.7 Three algebraic facts of least-squares regression

Three algebraic facts of least-squares regression emerge, simply as a consequence of minimizing the squared vertical distances between the data points and the regression line[7].

1. The least-squares residuals sum to zero.
2. The regression line passes through the sample means of the data (the line passes through $\bar{x}$ and $\bar{y}$).
3. The sample mean of the least-squares predicted values is equal to the sample mean of the actual y data ($\bar{\hat{y}} = \bar{y}$).

These facts become important when we delve into more advanced topic in econometrics. These three results can be proven mathematically, but in this text we only illustrate that they are true via Example 12.1.

Example 12.1 — Three least-squares facts illustrated. We'll illustrate the three results using the data from Figures 12.1 - 12.6. Load the data into R, and in this example we'll suppress scientific notation:

```
data <- read.csv("http://ryantgodwin.com/data/ls1.csv")
options(scipen = 999)
```

Next, estimate the least-squares line and save it as "`ls.line`":

```
ls.line <- lm(data$y ~ data$x)
```

From the saved object `ls.line` we can extract the *residuals* and the in-sample predicted or *fitted values*. To illustrate result (1), save the residuals and check that they sum to zero (or are at least very close to zero):

```
resids <- residuals(ls.line)
options(scipen = 999)
sum(resids)
```

```
[1] 0.0000000000001425526
```

To illustrate fact number (2), get the predicted or fitted values, and check that the mean of the fitted and actual y values are equal ($\bar{\hat{y}} = \bar{y}$):

```
yhat <- predict(ls.line)
mean(yhat)
mean(data$y)
mean(yhat)
```

```
[1] 229.5679
```

```
yhat <- predict(ls.line)
mean(data$y)
```

```
[1] 229.5679
```

To verify fact (3), we can plot the point $(\bar{x}, \bar{y})$, draw the least-squares regression

[7] There are rare situations where the intercept b_0 is excluded from the model. In this case, these properties do not necessarily hold.

line, and see the line passing through the point:

```
plot(mean(data$x), mean(data$y))
abline(ls.line)
```

12.8 Least-squares regression example

Load the per country, per capita CO_2 and GDP data that was used in Figure 4.12:

```
data <- read.csv("http://ryantgodwin.com/data/co2.csv")
```

The idea here is that GDP per capita is associated with per capita CO_2 emissions. We'll fit a least-squares line through the data, which will give us b_1. Then, b_1 will tell us how an increase in GDP is associated with an increase in CO_2 emissions. We don't know if GDP *causes* CO_2 emissions (or vice versa); such causal statements are beyond the scope of this statistical analysis.

To make the interpretation of b_1 easier, measure GDP per capita in 1000s of dollars:

```
data$gdp <- data$gdp.per.cap / 1000
```

Now, estimate the least-squares regression line, and get some "summary" information:

```
summary(lm(data$co2 ~ data$gdp))
```

```
Call:
lm(formula = data$co2 ~ data$gdp)

Residuals:
    Min      1Q  Median      3Q     Max
-11.8964 -1.0479 -0.6367  0.0841 28.2401

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.49730    0.45799   1.086     0.28
data$gdp    0.33110    0.02675  12.380   <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.945 on 132 degrees of freedom
Multiple R-squared: 0.5373,  Adjusted R-squared: 0.5338
F-statistic: 153.3 on 1 and 132 DF, p-value: < 2.2e-16
```

Plot the data, and add the least-squares regression line to the plot (see Figure 12.7):

```
plot(data$gdp, data$co2, col = "blue",
     ylab = "CO2 emissions per capita", xlab = "GDP per capita")
abline(lm(data$co2 ~ data$gdp), col = "red")
```

The slope of the least-squares regression line is $b_1 = 0.33$. This can be interpreted as: an increase in per capita GDP of \$1000 is associated with an increase in per capita CO_2 emissions of 0.33. The R-squared of 0.54 tells us that per capita GDP explains 54% of the variation in per capita CO_2 emissions between countries.

Note that it would be better to fit a regression line through the *logs* of per capita

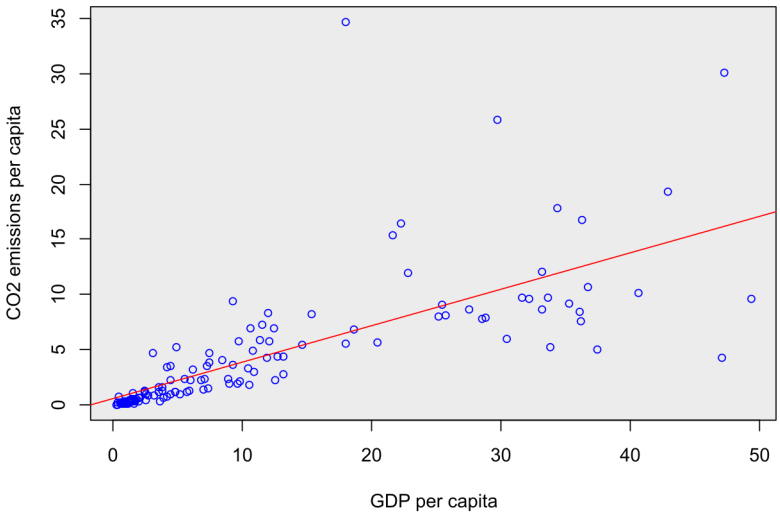


Figure 12.7: Per capita CO_2 and GDP, with fitted least-squares regression line

GDP and CO_2 emissions (see Figure 4.13). Doing so would give b_1 a *percentage change* interpretation. Using logs in a least-squares regression is a more advanced topic that is left out of this text.

13. Least-squares continued

13.1 The linear population model

The least-squares estimators (b_0 and b_1) of the previous chapter are *estimators* for unknown parameters in a linear population model. The linear population model expresses a possible *true* (and ultimately unobservable) relationship between y and x. The reasons for thinking in terms of a true population model are the same as for the true population mean and population variance. In earlier chapters, we differentiated the sample mean ($\bar{y}$) and sample variance (s^2) from the true unknown population values: μ and σ^2. This allows us to view the sample mean and sample variance as estimators for something that is unknown, and is a required concept for the view that estimators (like $\bar{y}$) are random variables that have sampling distributions. In turn, the sampling distributions allow us to conduct hypothesis testing (and derive statistical properties of the estimators, which is beyond this book).

The linear population model is often written as:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{13.1}$$

Once again Greek letters (β) are being used to denote unknown *population* parameters. β_0 and β_1 are expressing a *true* linear relationship between y and x. The least-squares intercept (b_0) and slope (b_1) from the previous chapter are just *estimators* for β_0 and β_1. The terms β_0, β_1, and ϵ are unobservable components of the model. Typically, the goal is to use the y and x data in order to *estimate* β_0 and β_1, which can be tricky given the random "noise" introduced through ϵ.

An *observable* counterpart to the linear population model has already been shown in Section 12.5:

$$y = b_0 + b_1 x + e$$

The least-squares regression model "replaces" the unobservable components of the linear population model: $\beta_0 \to b_0$, $\beta_1 \to b_1$, and $\epsilon \to e$

In the linear population model, the *slope* β_1 has a very important interpretation. It represents the *true* change in y associated with a 1 unit change in x. If there is a causal

relationship between y and x, then β_1 is the true marginal effect of x on y. Much of applied economics is focused on estimating the effect of one variable on another. Statistical methods for estimating β_1 are vital.

13.2 The random error term ϵ

An important component of the linear population model is the random error term ϵ (Greek letter epsilon). Equation 13.1 says that y is determined, or generated, by a value for x. The quantity $\beta_0 + \beta_1 x$ is the deterministic (non-random) part of the model; ϵ is the *random* part.

It may be helpful to envision the unknown *data generating process* for y (the way in which a y value is created). For example, suppose a value for x is chosen: this determines the quantity $\beta_0 + \beta_1 x$. Then, a random number (ϵ) is drawn, and added to $\beta_0 + \beta_1 x$ to determine the value for y. The same x value, when plugged into the model, could generate different y values depending on the randomly drawn *epsilon*. However, by making certain assumptions about the random ϵ, we can say that the population line $\beta_0 + \beta_1 x$ provides the mean, or expected value, of y.

Sometimes the random error term reflects *measurement error*, but usually in econometrics the random error term (ϵ) is thought to encompass all variables (besides x) that determine or cause the y variable. When we think of examples of y and x variables it is likely that we can think of many factors that determine y other than just x. There are potentially hundreds or thousands of other variables that might be associated with y. If we cannot observe these other variables, they are contained in the random error term. Thus, the random error term is the sum of all of the effects on y that other variables might have. Some examples of variable effects that might be contained in ϵ are shown in Table 13.1.

Table 13.1: Examples of variable effects contained within ϵ.

y **variable**	x **variable**	**other factors that determine** y **(represented in** ϵ**)**
wage	years of education	age, work experience, race, gender, IQ, economic and social factors, minimum wage laws, family characteristics, personality
happiness score	GDP per capita	social support, life expectancy, freedom, corruption, trust in neighbours, government ideologies
CO_2 emissions	GDP per capita	industrialization, urbanization, technological progress, population density, temperature

Random error term ϵ. The linear population model contains a random error term ϵ. The error term encapsulates all variables that determine y, besides x. ϵ contains the effects of many variables acting on y, all summed together into one term.

13.3 Five least-squares assumptions

The successfulness of the least-squares estimators (b_0 and b_1) at estimating the true population model (β_0 and β_1) relies on the properties of the random error term (except property 5). In this section, we identify some of the conditions that are often required in order for the least-squares method to be considered suitable. In particular, for least-squares to work well, we need:

1. ϵ to be uncorrelated with x
2. ϵ to be identically and independently distributed (i.i.d.)
3. ϵ to be Normally distributed
4. ϵ is mean zero ($\mathrm{E}[\epsilon] = 0$)
5. the relationship between y and x should be *linear* in β_0 and β_1

The first requirement, that ϵ and x be uncorrelated, is the most important. Correlation between x and ϵ has serious implications for estimating β_1. To complicate matters, it is difficult or impossible to tell if there is a relationship between the *unobservable* ϵ and x. Requirements (2)-(4) are usually considered to be less important. If (2) or (3) are not satisfied then alternative models or refinements to the basic least-squares method may be used. Requirement (4) only affects estimation of β_0, which is typically not of interest. If property (5) is untrue, then least-squares cannot be used to estimate the relationship between y and x.

Requirements (1) - (5) are collectively known as the "least-squares assumptions". These assumptions are mentioned in this book because they will become of greater import in more advanced econometrics courses. Some of these assumptions can be checked, the easiest of which being assumption (3) - the Normality of the error term ϵ. One strategy for testing assumptions about ϵ is to examine the least-squares residuls (e), since the properties of e should mimic those of the unobservable ϵ.

13.3.1 Testing the Normality of the error term

The assumption that the random error term follows a Normal distribution is required in order for the t-test (and other tests not covered) to be valid. If the random error term ϵ is Normally distributed, then so should be the residuals e from the least-squares estimation of the population model. In this section, without going into detail we look at two ways to test the Normality of the error term by examining the residuals; the Jarque-Bera test and a Normal Q-Q plot (quantile-quantile plot).

Jarque-Bera test

In the Jarque-Bera test, the null and alternative hypotheses are:

H_0 : ϵ is Normal

H_A : ϵ is not Normal

The test is based on comparing the sample skewness and kurtosis of a sample variable, to the true skewness and kurtosis of a Normally distributed variable (which should be 0 and 3 respectively). Similar to Example 12.1, we can begin by getting the residuals from a least squares model:

```
data <- read.csv("http://ryantgodwin.com/data/ls1.csv")
ls.model <- lm(data$y ~ data$x)
resids <- residuals(ls.model)
```

Next, we need to install and load the `tseries` package into R, which contains the Jarque-Bera test:

```
install.packages("tseries")
library("tseries")
```

Finally, we apply the Jarque-Bera test to the residuals from our least-squares regression:

```
jarque.bera.test(resids)
```

```
       Jarque Bera Test

data:  resids
X-squared = 3.7476, df = 2, p-value = 0.1535
```

Since the p-value is greater than 0.1, we fail to reject the null hypothesis at the 10% significance level. In this model, the assumption that the random error term is Normally distributed seems plausible.

Q-Q plot

A quantile-quantile plot compares the sample quantiles[1] of a variable to the theoretical quantiles of the Normal distribution distribution (or any other distribution). The Q-Q plot is a visual diagnostic tool. If the sample quantiles, plotted against the theoretical quantiles, appear to follow a straight line, then the variable is thought to be Normally distributed. To generate a Q-Q Normal plot in R, we can use the `qqnorm` function, along with `qqline` in order to draw a straight line through the plot.

```
qqnorm(resids)
qqline(resids)
```

The plot generated by the above R code is shown in Figure 13.1. Most researchers would conclude that the Q-Q plot supports the conclusion of the Jarque-Bera test: the residuals appear to be Normally distributed, indicating that the Normality assumption for ϵ is reasonable.

13.4 Hypothesis testing and confidence intervals

In Chapters 10 and 11 we made hypotheses regarding the unobservable population mean μ, and used the sample mean $\bar{y}$ to test these hypotheses. In this section, we make hypotheses about β_1, and test the hypotheses using b_1. The general framework of hypothesis testing, and the interpretation of p-values, remains unchanged.

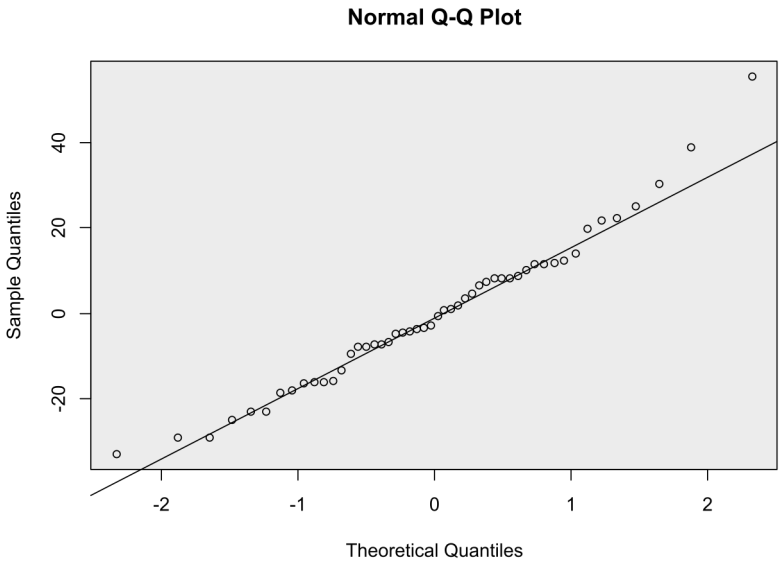
Once again, we begin the hypothesis test by stating the null and alternative hypotheses:

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_A : \beta_1 \neq \beta_{1,0}$$

The 0 subscript in $\beta_{1,0}$ denotes the value for β_1 under the null hypothesis. We could also make hypotheses about β_0, but usually the focus is on β_1. We can use the t-test

[1] (See Section 5.4 on quartiles and percentiles, which are similar to quantiles.

Figure 13.1: Q-Q Normal plot. If the variable is Normally distributed, then the sample quantiles should "line up" with the theoretical quantiles from the Normal distribution.



to perform hypothesis tests involving the β in the linear population model. In Section 11.4, the t-test statistic used for hypotheses on the population mean μ was:

$$t = \frac{\bar{y} - \mu_{y,0}}{\sqrt{\frac{s_y^2}{n}}}$$

There are three components to this formula: an estimator, a hypothesized value, and the standard error of the estimator. In the context of the linear population model, these components respectively become: the least-squares estimator (b_1), the value for β_1 under the null hypothesis, and the standard error of b_1 (denoted s.e.(b_1)[2]). A generic formula for the t test statistic is:

$$t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}} \tag{13.2}$$

See Table 13.2 for a comparison of notation for the t test statistic, between hypotheses involving the population mean μ, and the slope β_1 in the linear population model.

Table 13.2: Examples of variable effects contained within ϵ

	population mean μ	population slope β_1
estimate	$\bar{y}$	b_1
hypothesized value	$\mu_{y,0}$	$\beta_{1,0}$
standard error	s.e.$(\bar{y}) = \sqrt{\frac{s_y^2}{n}}$	s.e.(b_1)

[2]The formula for the standard error of $\bar{y}$ is $\sqrt{\frac{s_y^2}{n}}$, but the formula for the standard error of b_1 is more complicated and not covered in this book.

Applying the generic formula (Equation 13.2) to the least-squares situation gives us t-test statistic for testing hypotheses about β_1:

$$t = \frac{b_1 - \beta_{1,0}}{s.e.(b_1)}$$

This t-statistic follows the same t-distribution (see Section 11.2) whether it is used to test a population mean μ or a population slope β_1. We can obtain p-values from the t-distribution the same way in which we did in Chapter 11. If the sample size n is large, then this t-statistic is approximately Standard Normal $N(0, 1)$, and we can use Table 10.1 to obtain p-values.

The p-value is compared to a significance level (see Section 10.4). As before, if the p-value exceeds the significance level (if p-value $> \alpha$), we *fail to reject* the null hypothesis.

Suppose that we want to test:

$$H_0 : \beta_1 = 2$$
$$H_A : \beta_1 \neq 2$$

To calculate the t-test statistic using R, we need the least-squares estimate b_1, and the s.e.(b_1). Load data, estimate the model, and use the `summary()` function:

```
mydata <- read.csv("http://ryantgodwin.com/data/ls1.csv")
ls.model <- lm(y ~ x, data = mydata)
summary(ls.model)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.2636   15.6668  1.166   0.249
data$x       2.0712    0.1515 13.669  <2e-16 ***
```

From the output, we see that $b_1 = 2.07$, and s.e.$(b_1) = 0.15$. The t-statistic for testing $H_0 : \beta_1 = 2$ is:

$$t = \frac{2.07 - 2}{0.15} = 0.47$$

This t-statistic follows the t-distribution, whose shape is determined by the degrees of freedom, $n - k$ (see Section 11.2). In the present context, $n - k = 50 - 2 = 48$ (because the sample size is 50 and *two* least-squares estimates have been calculated, b_0 and b_1). The p-value is:

```
pt(0.47, 48, lower.tail = FALSE)
```

```
[1] 0.3202417
```

If the null is correct, then the expected value of the t-statistic is 0. That is, the difference between what we estimate and hypothesize $(b_1 - \beta_{1,0})$ should be zero on average. The p-value for this hypothesis test comes from the area in the t-distribution, to the *right* of 0.47. This gives us the probability of calculating a b_1 that is more "extreme" than the value of 2.07 that we just calculated. Thus, we need to set `lower.tail = FALSE` in the above R code. If the sample size n is large enough, then the t-distribution is well approximated by the $N(0, 1)$ distribution. Using Table 10.1, we obtain the same value of 0.32.

Finally, the alternative hypothesis is *two-sided*, so we need to multiply the area under the curve by 2: p-value $= 0.32 \times 2 = 0.64$. Since the p-value > 0.1, we fail to reject the nullhypothesis that the true $\beta_1 = 2$. There is a 64% chance that, if we had drawn a different sample from the same population generating y and x, that it would produce a b_1 further away from the null hypothesis, compared to what we just witnessed (a distance of 2.07 - 2 = 0.07 away from H_0).

13.5 Tests of "significance"

For the linear population model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

A special, and common hypothesis test is:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

If $\beta_1 = 0$ then x does not have a *linear* effect on y. That is, a change in x does not lead to a change in y. The marginal effect of x on y is zero. If $H_0 : \beta_1 = 0$ is *rejected*, then x is said to be "significant". If we *fail to reject* $H_0 : \beta_1 = 0$, then x is said to be "insignificant".

13.6 Confidence intervals

Confidence intervals around b_1 (or b_0) are calculated and interpreted in the same way that they were in Section 9.5. The formula for a confidence interval around b_1 (for example) is:

95% confidence interval around b_1.

$$[b_1 - 1.96 \times \text{s.e.}(b_1) \quad , \quad b_1 + 1.96 \times \text{s.e.}(b_1)] \tag{13.3}$$

In equation 13.3, the value of 1.96 comes from the 95% confidence level (-1.96 and 1.96 put 2.5% area in each tail of the Standard Normal distribution). A 90% or 99% confidence interval would use 1.65 or 2.58 respectively, instead of 1.96. We can obtain these critical values in R using:

```
qnorm(.05)
qnorm(.025)
qnorm(.005)
```

```
[1] -1.644854
[1] -1.959964
[1] -2.575829
```

For small samples, we could instead use the *critical values* from the t-distribution, to get more accurate confidence intervals (the critical values of 1.65, 1.96, and 2.58 are approximate values for when the sample is large). Using a degrees of freedom of $n - k = 48$ (for example), we get critical values of 1.68, 2.01, and 2.68 for the 90%, 95% and 99% confidence levels:

```
qt(.05, 48)
qt(.025, 48)
qt(.005, 48)
```

```
[1] -1.677224
[1] -2.010635
[1] -2.682204
```

The standard error of 0.15 is displayed in the `summary()` command:

```
data <- read.csv("http://ryantgodwin.com/data/ls1.csv")
ls.model <- lm(data$y ~ data$x)
summary(ls.model)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.2636   15.6668  1.166   0.249
data$x       2.0712    0.1515 13.669  <2e-16 ***
```

With $b_1 = 2.07$, a critical value of 2.01 (from the t-distribution), and s.e.$(b_1) = 0.15$, the 95% confidence interval for the above example is:

$$95\%\ \text{CI} = 2.07 \pm 2.01 \times 0.15 = [1.77\ ,\ 2.37]$$

Notice that this confidence interval contains the value $\beta_1 = 2$ from the null hypothesis in the previous section. The 95% confidence interval around b_1 contains all null hypotheses for β_1 that will be rejected at the 5% significance level.

13.7 Least-squares regression analysis

This section goes through some typical aspects of a least-squares regression analysis. It does not include other aspects that should accompany a proper analysis, such as cleaning the data, checking for outliers, calculating summary statistics (minimum, sample mean, sample variance, etc.), or plotting the data (in histograms, scatterplots, etc.).

- State the population model.
- Estimate the model using least-squares.
- Interpret the estimates.
- Comment on R^2.
- Report a confidence interval.
- Conduct any hypothesis tests of interest.

The following examples go through these steps using the Mars data.

Example 13.1 — Mars income and number of years on Earth.

State the population model

A population model to investigate the relationship between the number of years spent on Earth, and income earned on Mars, is:

$$income = \beta_0 + \beta_1 years.on.earth + \epsilon \tag{13.4}$$

β_1 is the true effect on income of an additional year spent living on Earth. ϵ contains

all the factors that affect income, besides years on Earth.

Estimate the model using least-squares

Load the sample of 1000 employed Mars colonists:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

To estimate β_0 and β_1 in this model, use the `lm()` command in R, and the `summary()` command to see the results:

```
model1 <- lm(income ~ years.on.earth, data = mars)
summary(model1)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  78287.77   2338.59 33.476  <2e-16 ***
years.on.earth 123.45     91.91  1.343    0.18
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41680 on 998 degrees of freedom
Multiple R-squared: 0.001805, Adjusted R-squared: 0.0008044
F-statistic: 1.804 on 1 and 998 DF, p-value: 0.1795
```

Interpret the estimates

The estimated intercept is $b_0 = 78287.77$. For a Mars colonist who has lived 0 years on Earth, their expected income is 78287.77. The estimated slope is $b_1 = 123.45$. On average, for each additional year having been spent on Earth, income is expected to increase by 123.45.

Comment on R^2

The R-square from this regression is $R^2 = 0.0018$. This is quite low, meaning that variation in the number of years spent on Earth explains less than 1% of the variation in income.

Report a confidence interval

The 95% confidence interval around b_1 is:

$$95\%\ \text{CI} = 123.45 \pm 1.96 \times 91.91 = [-56.69\ ,\ 303.59]$$

Because the sample size is large ($n = 1000$), it makes little difference whether we get the critical value (1.96) from the Standard Normal distribution, or from the t-distribution.

Conduct a hypothesis test

Test the hypothesis that years spent on Earth has no effect on income:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

The t-statistic for this hypothesis is:

$$t = \frac{123.45 - 0}{91.91} = 1.34$$

The p-value, either from Table 10.1 or from R is:

```
2 * pnorm(1.34, lower.tail = FALSE)
```

```
[1] 0.1802453
```

With a p-value of 18%, we fail to reject the null hypothesis. It appears that years on Earth has *no* effect on income. Note that:

(i) This is a test of *signficance* (see Section 13.5).
(ii) The `summary()` command has already calculated the t-statistic and p-value for $H_0 : \beta 1 = 0$.
(iii) The variable `years.on.earth` is said to be "insignificant".

Concerning item ((ii)): in the output from `summary(model1` above, make sure that you can find the values for the t-statistic of 1.343 and p-value of 0.18.

14. Multiple regression

In this final chapter, we briefly introduce *multiple regression* (in contrast to *single* variable regression in the previous two chapters). Multiple regression refers to least-squares estimation of population models that include more than one "x" variable. In practice, the vast majority of models estimated by researchers contain multiple regressors (x variables), and in reality there are likely many factors that determine the value for y. A linear population model with multiple x variables can be written:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon \tag{14.1}$$

k is the total number of x variables in the model, all of which may determine the value for y. β_1 is the effect of x_1 on y, holding all other x variables constant. Similarly, β_2 is the effect of x_2 on y, etc.

There are several reasons for including more than one x variable in the population model, three of which are:

(i) We may be interested in the effect that several x variables have on y.
(ii) To improve our ability to predict a value for y.
(iii) There may be other variables that are correlated to both our main x variable of interest, that y.

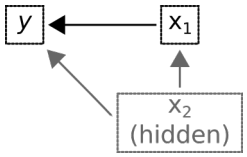
We will briefly investigate the importance of item (iii) as it relates to *causal inference*.

Estimation, interpretation of the estimates, and hypothesis testing and confidence intervals, all remain largely unchanged in the multiple regression model compared to the single variable regression model.

14.1 Lurking or confounding variables

A *lurking*, or *confounding* variable is one that threatens our ability to correctly estimate the effect that an x variable has on a y variable. Lurking variables are a major issue in analyses of *causal inference*, and are of tremendous import in many areas, not just economics.

Figure 14.1: A hidden x_2 variable that determines both y and x_1 will make estimation of the effect of x_1 on y difficult (or impossible).



The situation depicted in Figure 14.1, where x_2 is a determinant of both x_1 and y, implies that the effect of x_1 on y cannot be measured in the single variable linear population model. That is, the estimated β_1 (b_1) is *wrong* in the population model:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

The reason that b_1 gives the *wrong* answer for the true effect of x_1 on y is that:

- A change in x_2 is associated with a change in both x_1 and y.
- When we "see" x_1 changing, we know x_2 is also changing.
- Attributing changes in y due to changes in x_1 alone becomes impossible, since we don't know how much of the change in y came from x_2.

The solution to the problem is to include the x_2 variable in the model! If we can't actually observe x_2 then we must use clever strategies and more advanced methods to attempt to estimate the effect of x_1 on y.

14.2 Estimating the multiple regression model

A multiple regression model is estimated in R by including all of the desired x variables on the right-hand-side of the `lm()` command, each separated by a `+`. For example, suppose we wish to use the Mars data to estimate the population model:

$$\text{income} = \beta_0 + \beta_1 \text{years.education} + \beta_2 \text{years.on.earth} + \epsilon$$

In R:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
model <- lm(income ~ years.education + years.on.earth, data = mars)
summary(model)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6746.60   5313.59  1.270  0.2045
years.education 4740.39   315.68 15.016  <2e-16 ***
years.on.earth -141.89     74.33 -1.909  0.0566 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35150 on 997 degrees of freedom
Multiple R-squared: 0.1851,  Adjusted R-squared: 0.1835
F-statistic: 113.3 on 2 and 997 DF, p-value: < 2.2e-16
```

Note that the R output `<2e-16` is using scientific notation for the p-value, and means that the p-value is less than 2×10^{-16}, and is the smallest decimal that R can represent (the p-value is essentially zero).

14.2.1 Interpreting the estimation results

The estimated β values (b_0, b_1, b_2)

The estimated β values have a similar interpretation as before, but with one very important addition. For example, the estimated value of 4740.39 means that an additional year of education is associated with an increase in income of 4740.39, on average. The important addition to the interpretation of this b_1 value is that this estimated effect is *ceteris paribus*. That is, holding all else equal. The estimated increase in income due to education is while holding years on Earth constant. The value of -141.89 means that each additional year lived on Earth is associated with a *decrease* in income of 141.89, holding years of education *constant*.

In the multiple regression model, the effect of each x variable on y is while *controlling* for all the other variables. This is a major advantage of the multiple regression model. Lurking or confounding variables, when included in the model, are *controlled* for. The relationships between multiple x variables can be accounted for as long as they are included in the model.

Adjusted R^2

In the multiple regression model, we look at adjusted R-squared (not the multiple R-squared as in previous chapters). Adjusted R^2 has a similar interpretation as before: it is the ratio of the sample variance in y that can be explained using *all* of the x variables in the model. For example, the value of 0.1835 above means that `years.education` and `years.on.earth` can together explain 18.35% of the changes in income.

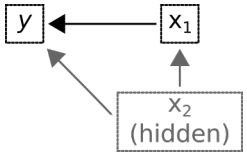
Hypothesis testing

The types of hypothesis tests that we have already discussed are conducted identically in the multiple regression model as they were in the single variable model. Confidence intervals, t-tests and p-values, are all calculated and interpreted the same as before. In the multiple regression model, however, there is the opportunity to formulate hypothesis that involve more than just one of the βs. Multiple hypothesis testing is more complicated and not covered in this book.

14.3 Lurking or confounding variables

A *lurking*, or *confounding* variable is one that threatens our ability to correctly estimate the effect that an x variable has on a y variable. These variables are a major issue in analyses of *causal inference*, and are of tremendous import in many areas, not just economics.

Figure 14.2: A hidden x_2 variable that determines both y and x_1 will make estimation of the effect of x_1 on y difficult (or impossible).



The situation depicted in Figure 14.2, where x_2 is a determinant of both x_1 and y, implies that the effect of x_1 on y cannot be measured in the single variable linear

population model. That is, the estimated β_1 (b_1) is *wrong* in the population model:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

The reason that b_1 gives the *wrong* answer for the true effect of x_1 on y is that:

- A change in x_2 is associated with a change in both x_1 and y.
- When we "see" x_1 changing, we know x_2 is also changing.
- Attributing changes in y due to changes in x_1 alone becomes impossible, since we don't know how much of the change in y came from x_2.

The solution to the problem is to include the x_2 variable in the model! If we can't actually observe x_2 then we must use clever strategies and more advanced methods to attempt to estimate the effect of x_1 on y.

To illustrate the issue, consider the population model:

$$\text{income} = \beta_0 + \beta_1 \text{age} + \epsilon$$

We might guess that *age* has a positive effect on income, as we tend to see people making more money the older they are. Let's try estimating this model in R using the Mars data:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
model1 <- lm(income ~ age, data = mars)
summary(model1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 71634.84  4161.95 17.212  <2e-16 ***
age           223.49     95.26  2.346  0.0192 *
```

The estimation results suggest that each additional year of *age* is associated with an increase in income of 223.49, on average. Now, test the hypothesis that *age* has zero effect on income (that age does not determine or is not associated with income). This is a test of the "significance" of the variable *age*:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

R has already calculated the t-statistic (2.346) and p-value (0.0192) for this hypothesis test. Since the p-value is less than 0.05, we reject the null hypothesis at the 5% significance level, and conclude that *age* is a "significant" determinant of *income*.

Given our recent discussion for the need for the multiple regression model, can you think of any *lurking* variables? We should be thinking about other variables that are correlated or related with *age*, and also determine *income*. What about *education*? The older a person is, the more likely they have more education. A worker who is 18 cannot have more than 12 years of education. The idea here is that *age* might just be indicating (acting as a proxy for) years of education. Consider the following population model instead:

$$\text{income} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{years.education} \epsilon$$

In this model, we can examine the effect of *age* on *income* while controlling for *education*. It is as if we can compare workers who all have the same *education*, but differ only in their *age*. Estimate this model in R:

```
model2 <- lm(income ~ age + years.education, data = mars)
summary(model2)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -607.56   5861.52 -0.104   0.917
age              42.88     85.83  0.500   0.617
years.education 5010.40   314.32 15.940  <2e-16 ***
```

Notice that the effect of *age* on *income* has reduced, and is no longer significant! After controlling for *education*, we now conclude that *age* is not a determining factor of *income*. The positive relationship between *age* and *education* (older people tend to have more education) resulted in the overestimation of the effect of *age*, when *education* was omitted from the model.

14.4 Multiple regression model for Mars incomes

We conclude this text by estimating a multiple regression model using the Mars data. We'll try to determine the effect that education has on income. Note that economics studies that involve *income* typically use the logarithm of income, an aspect that we ignore for simplicity. Start with a single variable population model:

$$\text{income} = \beta_0 + \beta_1 \text{years.education} + \epsilon$$

Estimate this model in R:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
ls1 <- lm(income ~ years.education, data = mars)
summary(ls1)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      847.5    5084.9  0.167   0.868
years.education 5031.1     311.5 16.154  <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35690 on 998 degrees of freedom
Multiple R-squared: 0.2073,  Adjusted R-squared: 0.2065
F-statistic: 260.9 on 1 and 998 DF, p-value: < 2.2e-16
```

The interpretation of b_1 is that each additional year of education leads to an average increase in income of 5031.1. Education is statistically significant, with a p-value of essentially zero (`<2e-16`) for $H_0 : \beta_1 = 0$. The 95% confidence interval is $b_1 \pm 1.96 \times 311.5 = [4421, 5642]$. Years of education explains 21% of the variation in incomes.

Take a look at the variables in the `mars` dataset. Which other variables might be important to include in a multiple regression model. In fact, there is very little to lose by including *all* of the variables in the data as a general practice, but many are dummy variables which we won't address in this book. In particular, we are looking for variables that might be correlated with *education*, and also determine *income*. Some possibilities might be *age*, or *years on Earth*. In particular, we should be looking at the individual's *IQ* score. It is very likely that *IQ* causes both *years of education* and *income*. Education might be higher in individuals who score well in IQ tests because

they have an easier time obtaining an education (it is less costly). A higher IQ may also lead to higher incomes.

Let's include these variables in a population model:

$$\text{income} = \beta_0 + \beta_1 \text{years.education} + \beta_2 \text{IQ} + \beta_3 \text{age} + \beta_4 \text{years.on.earth} + \epsilon$$

and estimate the model in R:

```
ls2 <- lm(income ~ years.education + IQ + age + years.on.earth, data = mars)
summary(ls2)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -38238.81   8804.80 -4.343 1.55e-05 ***
years.education 2447.18    508.11  4.816 1.69e-06 ***
IQ               788.99    124.13  6.356 3.14e-10 ***
age               31.55     91.04  0.347   0.729
years.on.earth    35.19    120.39  0.292   0.770
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35040 on 995 degrees of freedom
Multiple R-squared: 0.2384,  Adjusted R-squared: 0.2353
F-statistic: 77.87 on 4 and 995 DF, p-value: < 2.2e-16
```

Notice that:

- The model explains 24% of the variation in *income*.
- After including *IQ* in the model, the estimated effect of an additional year of *education* is to increase income by 2447.18 on average (compared to 5031.1 from the single variable model).
- In the single variable model, IQ was a lurking variable that was causing both *income* and *education*. Without including *IQ* in the model, it is difficult to get the correct estimate for the effect of *education* on *income*.
- Both *education* and *IQ* are "significant", *age* and *years on Earth* are not significant.

14.4.1 The future

The multiple regression model is a widely used tool, and if the various assumptions and requirements of the data are satisfied, can be extremely useful. However, the real world is rarely simple enough that this basice multiple regression model is appropriate! In your future econometrics studies you will use the multiple regression model with variables that are non-continuous (dummy variables), approximate non-linear relationships between variables using logarithms and polynomials, test hypotheses that involve multiple variables, examine the properties of the least-squares estimators and the assumptions that back them, deal with things like heteroskedasticity and instrumental variables, and so much more!

References

[1] https://doi.org/10.25318/3610022201-eng. "Table 36-10-0222-01 Gross domestic product, expenditure-based, provincial and territorial, annual (x 1,000,000)". In: () (cited on page 41).

[2] Courtney Kennedy et al. "An evaluation of the 2016 election polls in the United States". In: *Public Opinion Quarterly* 82.1 (2018), pages 1–33 (cited on page 26).

[3] Mitchell Langbert, Anthony J Quain, Daniel B Klein, et al. "Faculty voter registration in economics, history, journalism, law, and psychology". In: *Econ Journal Watch* 13.3 (2016), pages 422–451 (cited on page 25).

[4] Dominic Lusinchi. ""President" Landon and the 1936 Literary Digest poll: Were automobile and telephone owners to blame?" eng. In: *Social science history* 36.1 (2012), pages 23–54. ISSN: 0145-5532 (cited on page 26).

[5] Bruce D Meyer, Nikolas Mittag, and Robert M George. "Errors in survey reporting and imputation and their effects on estimates of food stamp program participation". In: *Journal of Human Resources* (2020), 0818–9704R2 (cited on page 26).