Quantitative Methods in Economics

Ryan T. Godwin

First printing, July 2021

Contents

Part One

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, if not all of the material in the book should be covered in a single semester course.

1.2 Quantitative Methods

Some questions that need answering...

- What are quantitative methods?
- What is the relationship between statistics, quantitative methods, and econometrics?
- What are these methods used for?
- What are the limitations of quantitative methods?

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
- Understand behaviour and conditions of counts and proportions
- Draw conclusions about a population or process from sample data
- Model a response based on an explanatory variables

- Perform quantitative analysis using *R*

1.4 Format of this Book

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

Definitions in the text. Important definitions that take up a lot of space will appear in the main text.

Shorter definitions. Shorter definitions will appear in the margins.

> "How do I know when something is a quote?" asked the student.

Exercise 1.1 This is a place to write exercises.

Example 1.1 This is an example of the examples you will see in these 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. Janelle Mann for help with outline, content, and edits. University of Manitoba for providing financial support. All images from NASA. Statistics performed using *R* and *RStudio*.

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 organisations. *R* is great for any job or task that uses data.

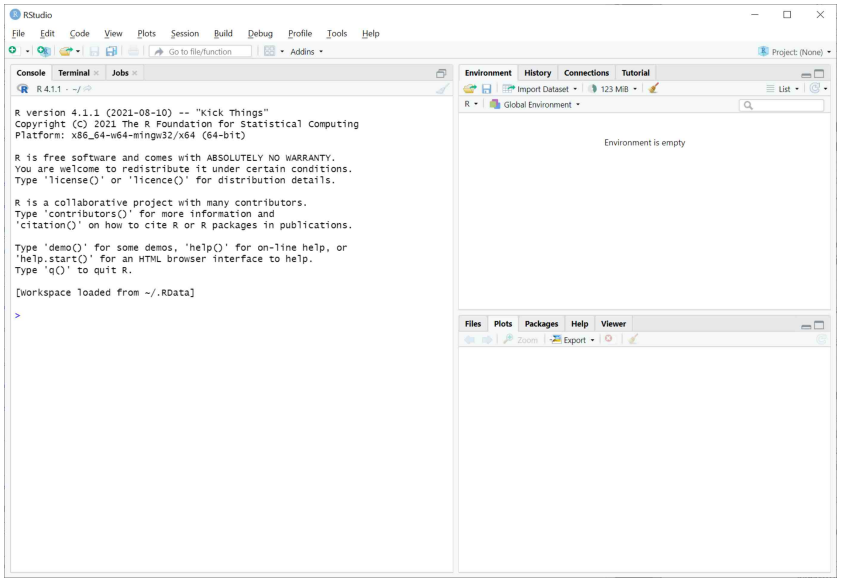
2.2 Where to get *R*

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*

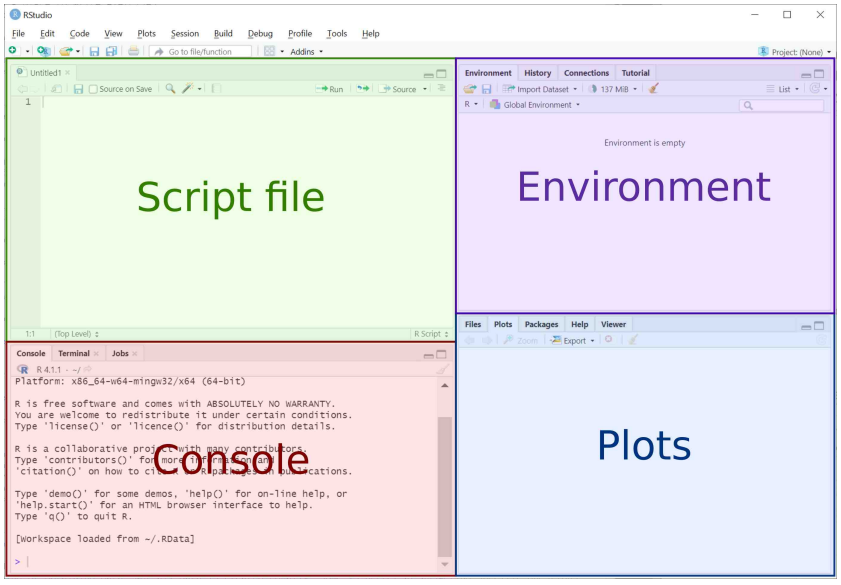
2.3.1 Open *RStudio*

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".



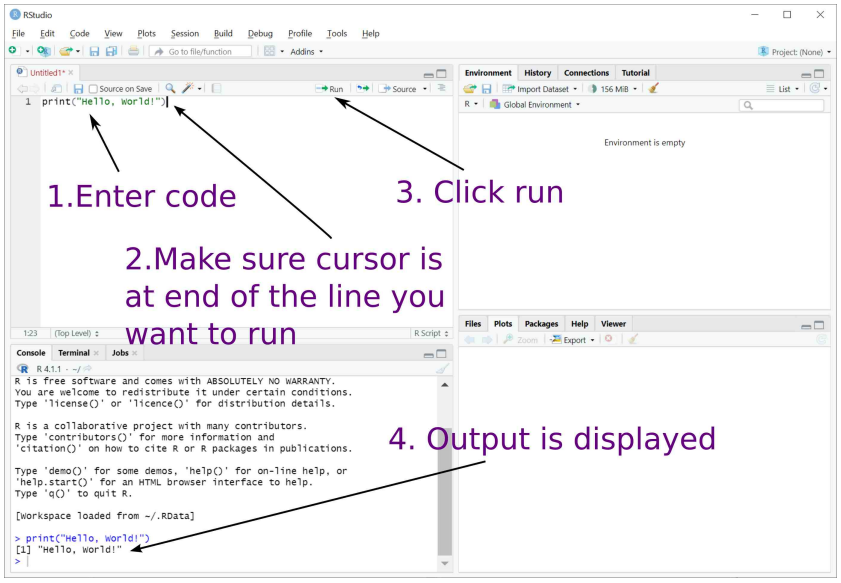
- In the top left is your Script file. R commands can be run from the R Script file, and saved at any time.
- In the bottom left is the Console window. Output is displayed here. R commands can be run from the Console, but not saved.
- In the top right is the Environment. Data and variables will be visible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

2.3.3 Your first program

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).



Sometimes in this book, 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

Exercise 2.1 Use *R* to perform the following arithmetic operations:

1. $3+5$
2. $12-4$
3. 2×4
4. $^{16}/_{2}$
5. 2^3
6. $\frac{10+6}{2}$

Answer: For question 6 you need to enter:

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

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, it doesn't have to be `a` and `b`):

```
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*

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:

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

Example 2.1 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.

Exercise 2.2 Try all of the functions in the table on `myvector`.

2.7 Logical operators

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

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

for men and women, 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 females 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*

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).

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.

2.8.1 Directly from the internet

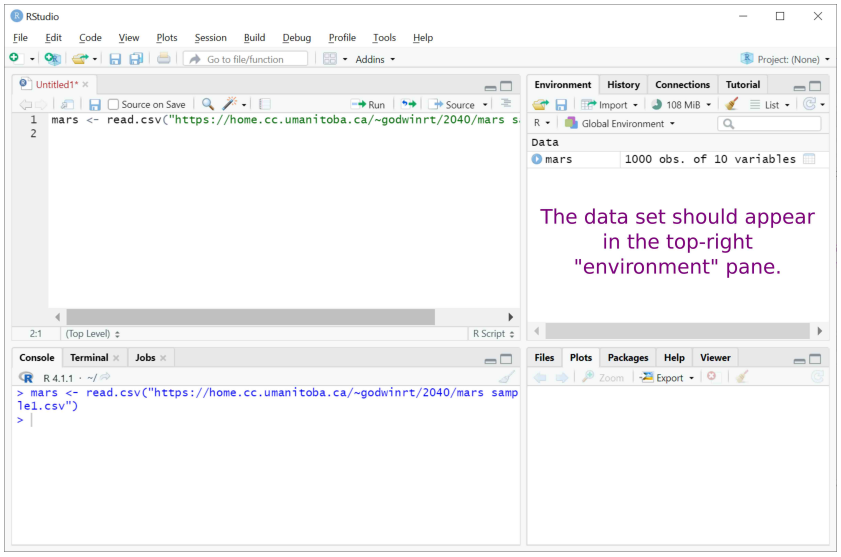
We can use the R code:

```
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("https://home.cc.umanitoba.ca/~godwinrt/2040/mars sample1.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` to your computer, you can use the `read.csv()` command to load the file from a location on your computer. For example:

```
mars <- read.csv("c:/data/mars sample1.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)
```

This allows you to view your data in spreadsheet form. See Figure 2.1.

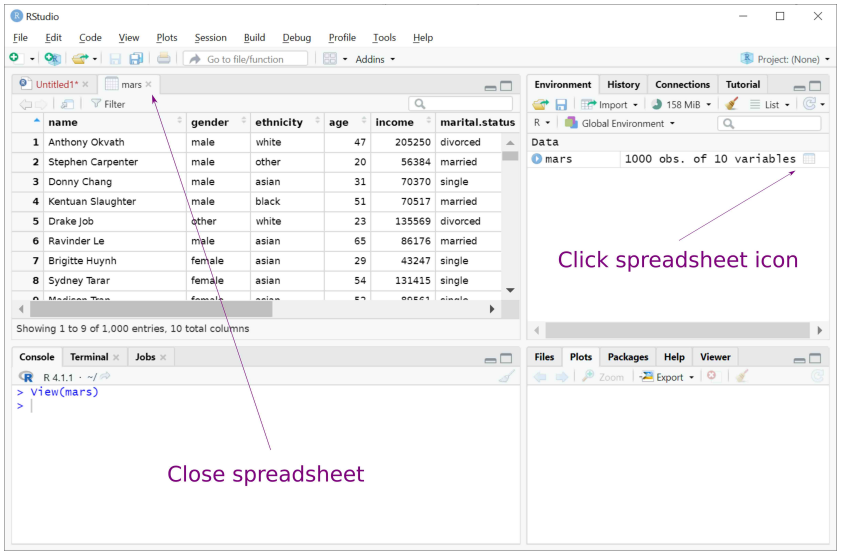


Figure 2.1: View your data in spreadsheet form.

3. Collecting Data

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 *biases*. In this chapter, we will discuss situations in which the data set might not represent the population in a way that creates misleading statistical conclusions.

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.

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*.

3.1.1 Anecdotal evidence

Other people's stories or accounts of experience, 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.

Anecdotal evidence. Anecdotal evidence is based on individual experiences or stories.

An anecdote is a story. For example:

> "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:

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

Table 3.1: Data set from anecdotal evidence

There is a perfect negative correlation (-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!

In addition to not providing a large enough sample size, anecdotal evidence may only exist because the stories 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*.

"Treatment" can be defined quite broadly. Traditionally it meant treatment with a new drug or medical procedure, but treatment could refer to education, a labour training program, health insurance, etc. The treatment group are those individuals that receive the "treatment"; the control group does not receive treatment (in a drug study they might receive a "placebo"). The effect of the treatment can then be determined by comparing the *outcomes* of the two groups. The outcome could be cholesterol (due to the new drug), an increase in the employment rate among those receiving job training, the wages

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 *bias* 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.

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.

Random assignment in an experiment is sometimes said to provide "controls". 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 figure out 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 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 is recorded without being able to apply any *control* over whether the individuals in the data are in the treatment group, or 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[1].

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

The lack of random assignment or 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. Instead, we must just observe an individual's wage, and whether they have a university education or not. We have no control over which individuals obtain a university education. Consider that the individuals who were enrolled in the previous experiment were instead allow 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?

[1] *Natural experiments* are one exception.

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 11.8%. Then, individuals with a higher earning potential will be more attracted to a university education. They have more to gain.

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* (11.8%). 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 obtained an education (their self assessed earning potential).

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) using observational data. In this textbook we will not tackle such issues, but will be working primarily with observational data. We need to be aware of the limitations of observational data, being very careful about inferring causality when using it.

Endogeneity. In economics, endogeneity usually refers to a situation where an individual's anticipation of an outcome influences the choices that they make.

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 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 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

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. If we are interested in the effect of education on wage, the population is 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 or member is an individual 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 in the discussions are valid whether we are talking about individuals, businesses, schools, institutions, countries, etc.

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

3.2.2 Sample

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.

Sample. A sample collects data on a subset of members from 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[2].

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 only useful in describing the population if it is a fair and unbiased representation of the population. Bad sample data can result 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 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*.

Surveys, or polls, are used for many purposes and are an important source of data. It is usually better to observe information about the members in the sample directly, rather than ask the members to report the information. In some following examples, we will discuss polls. A poll, or a survey, is a certain way of collecting a sample of data. It is when an individual is asked to respond to questions. The alternative is to observe the data directly. 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.

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

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 predictive that the next government will not be conservative? Are these 30 individuals a fair representation of the voting

[2] When comparing economic indicators such as GDP, usually the entire population (all countries) are used, since the population consists of at most 195 members.

population? Probably not. Professors in social science overwhelming vote on the left[2], and so do students. While collecting this sample might be *convenient* for us, a university campus is not a fair subset of the political views of the population at large. 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. The *Digest's* mistake was attributed to using "telephone books and automobile owners."[3] 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, 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 leads to a sample not reflecting 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 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%[1]. 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 has been challenged[3]. *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[4], 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

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. This 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 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. 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%. 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.

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.

Figure here.

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 learn the appropriate type of graph for a certain type of variable, and how to create that graph in *R*.

By graphing a *quantitative* variable, 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* 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 reperesents 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.

variables

observations

name	gender	ethnicity	age	income	marital.status
Simon Tran	male	asian	7	0	single
Abdul Fattaah el-Basher	male	arabic	40	95074	single
Jerry Yen	male	asian	21	20000	married
Cody Ironshield	male	indigenous	50	86844	married
Mikio Mulholland	male	asian	12	0	single
Nadiyah Williams	female	black	37	201435	single
Samih Esquibel	male	black	47	147290	married
Nhi Lumpkin	female	asian	65	33803	married
Haifaaa al-Ally	female	arabic	49	20000	married
Tori Abeyta	female	hispanic	33	208805	single
Jawan Fortune	male	black	63	20000	married
Javonte Ali	male	black	24	23070	married
Chalese Mcgowan	female	black	38	131892	married

Figure 4.1: Data set on Mars colonists

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!

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 the first 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 them-

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

selves 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.

	Country	Happiness.score	GDP.per.capita
1	Finland	7.769	1.340
2	Denmark	7.600	1.383
3	Norway	7.554	1.488
4	Iceland	7.494	1.380
5	Netherlands	7.488	1.396
6	Switzerland	7.480	1.452
7	Sweden	7.343	1.387

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 tells you the *observation type*.

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 individuals 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 categorical 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.

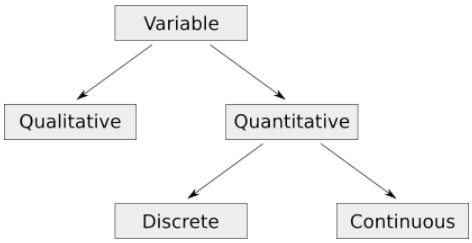


Figure 4.3: Types of variables

4.3.1 Categorical variables

A categorical variable is one that takes on two or more possible qualitative values (categorical variables are also called qualitative 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.

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

Figure 4.4: Statistics Canada ethnic categories.

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 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 thought to belong to one of the two categories. In contemporary statistical analyses, researchers try to recognize a broader spectrum of gender categories, such as non-binary, trans, and possibly dozens others. The more inclusive the better, and it has been suggested to at least allow for non-cisgender individuals to fall into a broadly defined "other" category. With more than two categories, gender would no longer be a "dummy" variable.

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.

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

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 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.

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

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.

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!

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

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, and income is a *continuous* variable[1].

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, are:

- Times a customer might visit a store.

[1] 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" variable.

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

- 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[2]. By contrast, *continuous* variables can take on an *uncountable* number of values!

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 `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.

Example: identify the observation type, variables, and values that the variables can take.

Exercise.

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.

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

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

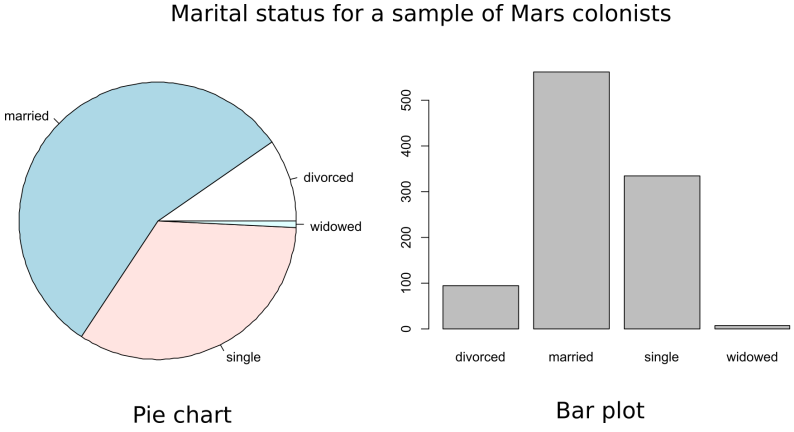


Figure 4.5: Example pie chart and bar plot

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 (the code must all be on a **single line**):

```
mars <- read.csv("https://home.cc.umanitoba.ca/
       ~godwinrt/2040/mars sample1.csv")
```

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

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

```
divorced married  single  widowed
      96     561     335       8
```

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	96	561	335	8

4.5 Graphing quantitative data

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

By counting how many observations are in each "bin", we graph the height of each bin, and obtai

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

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.

Shape, location, and spread

Skew

Multi-peak

Outliers

4.5.2 Time plots

4.5.3 Scatterplots

Bibliography

Articles

[1] 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 25).

[2] 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).

[3] 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 25).

[4] 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).

Books