Elementary Statistics

Rebekah Robinson and Homer White

Version: December 31, 2014

Contents

Chapter 1

Introduction

1.1 What are R and RStudio?

R is a computer program that can do anything from basic arithmetic to complicated statistical analysis and graphs. Learning the language will take some time, but the best way to learn how to use R is by using it. So, let's get started.

RStudio is an integrated development environment (IDE) that facilitates the use of R. In short, RStudio makes using R easier and more fun! You'll notice that you have four panels in the RStudio window.

1.1.1 Panels and Tabs

The **top left panel** is called the Source. You will primarily be working with RScript (.R) and RMarkdown (.Rmd) files. To create a new file, select *File*, then *New File*, then select what type of file you want. This will open a new file with a template document to help you get started.

The **bottom left panel** is called the Console. This panel is the 'brain' of RStudio. Anything entered in this panel will be executed by R. Also, R only 'knows' what is entered into the Console.

The **top right panel** has two tabs - Environment and History. As commands are entered into the Console, they are simultaneously stored in the History tab. This gives you a running history of the work you do. You can Search through the History to locate previous commands. You can also highlight a previous command and send it back *To Console* or *To Source*. The Workspace tab shows the objects that the Console 'knows' - datasets, functions, etc.

The **bottom right panel** has four tabs that we will work with in this class - Files, Plots, Packages, and Help. The Files tab is exactly what it sounds like - it shows you a list of your available files and allows you to upload new files. The Plots tab displays the plots that you create in the Console. The Packages tab allows you to install and load necessary packages. The Help tab is where you will view R help files.

This will all start to make more sense as we go along, so don't worry if it seems overwhelming at first. This information will be more useful as a reference tool as you start getting used to RStudio.

1.1.2 Differences Between RScript and RMarkdown Files

These descriptions of different types of files is not intended to overwhelm you on the first day of class! It should be treated as a resource for you to use as different points throughout the semester as you get used to using RStudio.

The most basic way to get RStudio to perform a command is to type the command directly into the Console. This tells R what you want it to do and it does it. The drawback to this is that you do not have a saved, and editable, copy of your commands. (Everything you type into the Console is automatically saved in the History tab. However, you cannot edit commands in the History tab.) This becomes an inconvenience if you are typing long commands. It is advantageous to store all of your commands in a file called an RScript. This way, you can go back and edit as you please. Once a command is typed into an RScript you can run it through the console a couple of different ways:

- You can select the entire command with your mouse. Then copy and paste it into the console.
- You can place your cursor on the line that contains your command and press the `Run` button at the top right of the Source window.

Another type of file that you will use in this class is an RMarkdown file. It differs slightly from an RScript file. While it is likely that you will primarily read your course notes from the printed version you purchased from the bookstore, you will also have access to the RMarkdown version of these notes and your homework assignments. Other occasions will also arise throughout the semester where you will need to create RMarkdown documents.

RMarkdown documents integrate text with code into one document that can be compiled, or knit, into a readable HTML. The code in an RMarkdown document is offset by a 'chunk'. This way, when you knit your HTML, the chunks will all be run through the console as Rcode and any output will be put in your HTML file.

While you are working in your RMarkdown document, you can run the code in a chunk several different ways:

- You can use your cursor and mouse to copy the code and paste it in the console.
- You can place your cursor anywhere in the chunk and select the 'Run Current Chunk' option in the 'Chunks' dropdown button at the top right of this window.
- You can place your cursor on the line you want to run and hit 'Run' at the top right of this window.

1.1.3 Basic R

In this class, anytime we use R, we will always start by loading two necessary packages.

The package `tigerstats` is the main package for our class. It includes all of the datasets, functions, and interactive apps that we will use throughout the semester. Anytime you do any work in RStudio for this class, you will want to `require(tigerstats)`.

The `mosaic` package is a package that contains functions that we will use for some of our work this semester. For this reason, anytime you do any work in RStudio, you will also want to `require(mosaic)`.

You will encounter various apps throughout this class that are designed to let you interactively explore concepts that we will be discussing in class. You should always take the time to tinker with these apps, as they will improve your understanding of difficult concepts. The package that must be loaded before you can use these apps is the package `manipulate`. Anytime you want to play with one of the apps in `tigerstats`, you must `require(manipulate)`.

Let's start by loading the necessary packages.

```r
require(tigerstats)
require(mosaic)
require(manipulate) #if you want to play with an app in tigerstats
```

Note: Sometimes when you're typing code into either a RScript or an RMarkdown, it's nice to add a note to remind yourself what that particular line does. Anything that is typed after a # will not be run as Rcode in the console. It's simply a note for yourself.

There are lots of cool features of R, but let's begin by using R in the most basic way - like a calculator to do simple calculations.

```
5+4
```

```
## [1] 9
```

```
24*3.7
```

```
## [1] 88.8
```

```
18/3
```

```
## [1] 6
```

```
sqrt(81) #square root function.
```

```
## [1] 9
```

It's that easy! R will do any arithmetic that you want. Just remember to use parentheses where they are appropriate to preserve order of operations.

1.1.3.1 Help

Now, suppose that you did not know what the command `sqrt()` did. You could get the help file for this command by typing `help(sqrt)` into the Console.

The Help tab will pop up and display the R Documentation for this command. It will give you:

- a description of what the command does,
- how it is entered into the Console,
- the arguments the command takes,
- some other details and references, and
- most importantly, examples of how to use it

Shortcut Help: If you type in the first couple letters of a command into the Console and then hit the *Tab* key on your keyboard, a Help window will pop up with possible matches and allow you to select exactly which one you want. If you still need more information, you can then hit *F1* on your keyboard to open the Help tab to read more about it.

1.1.3.2 Assigning Values

If you want to assign a value to a place, you use the `<-` sign. For example,

```
mysum <- 5+6
```

This computes the value of 5+6 and assigns it to the object named `mysum`. This will not show you what the value is; it simply stores it. If you want to see the value of `mysum`, you need to call it by typing it into the console (or a code chunk in an RMarkdown file).

```r
mysum
```

```
## [1] 11
```

```r
myproduct<-5*6
myproduct
```

```
## [1] 30
```

Assigning values to an object allows you to access them later by their name. For example, suppose that I wanted to add 10 to whatever value I had stored in `mysum` earlier in my work.

```r
mynewsum <- mysum+10
mynewsum
```

```
## [1] 21
```

Notice that you can choose any name that you wish for the object (the name to the left of the `<-`). However, you must be careful with what you type to the right of the `<-`. This must be recognizable by R. It should either by a number, a function, or an object that you have already assigned a value to.

While it does not really matter what you choose to name objects, you should be a little bit careful.

- It is helpful to choose a name that is descriptive. Notice that we chose the name `mysum` and `mynewsum` for the sums that we computed above.

- You do not want to name an object something that is already used as a function name or the name of a dataset. This is both confusing and can cause problems for you later.
- You can use some symbols in your names, such as a period (.) and an underscore (_). However, you cannot include other symbols or spaces in your object names.

1.1.3.3 Several Important Functions

1.1.3.3.1 Concatenation Function Let's talk about a function that you will see and use alot this semester. The concatenation, `c()` function combines values into a list. Creating a list of the numbers 1, 3, and 5 can be accomplished by:

```r
mylist <- c(1,3,5)
mylist
```

```
## [1] 1 3 5
```

If you want to combine letters (or words), you must put them in quotes. You can remember this by thinking of R as a really fancy calculator. R operates with numbers and functions, so letters and words have to be treated differently by placing them in quotes.

```r
mygrades <- c("A", "B", "C", "D", "F")
mygrades
```

```
## [1] "A" "B" "C" "D" "F"
```

1.1.3.3.2 Replicate Function Suppose we want to create a list of numbers that are all the same. For example, say we want to make the list 1,1,1,1,1. We could accomplish this with the concatenation function by

```
myreps <- c(1,1,1,1,1)
myreps
```

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

This will work, but there is an easier way to do this, using the replicate function, `rep()`.

```
myreps <- rep(1,5)
myreps
```

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

The replication function requires two inputs. The first input is the value that we want to be replicated and the second is the number of times we want that value replicated.

You can also combine the concatenation and replication functions to create a list of repeated letters or words. Just make sure that you use quotes for the letters.

```
myletters<-c(rep("A",3),rep("B",10),rep("C",7))
myletters
```

```
##  [1] "A" "A" "A" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "C" "C" "C" "C"
## [18] "C" "C" "C"
```

1.1.3.3.3 Sequence Function Suppose we want to create a list of numbers that appear in a certain sequence. For example, say we want to make the list 1, 2, 3, 4, 5, 6. Again, we could accomplish this with the concatenation function.

```
myseq <- c(1,2,3,4,5)
myseq
```

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

The sequence function, `seq()`, makes this task easier.

```
myseq <- seq(from=1, to=5, by=1)
myseq
```

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

The sequence function requires three pieces of information.
Where the sequence should start, `from=` * Where the sequence should end, `to=` * The increment, `by=`

Suppose we want to create the sequence 55, 57, 59, 61, 63, 65, 67, 69, 71.

```
seq(from=55, to=71, by=2)
```

```
## [1] 55 57 59 61 63 65 67 69 71
```

Notice that we did not store the sequence above in a object. The sequence was still created, but we will not be able to access it later.

That should be enough to get you started (and overwhelmed). Remember, treat this chapter as a reference guide to help you throughout the semester.

1.2 Let's play cards!

1.2.1 The Game

We are going to introduce an important statistical concept that will be a continuing theme throughout this class by playing a game with a standard deck of cards. In a standard 52 card deck, there are 4 suits (Hearts, Diamonds, Spades, Clubs). Each suit has an Ace, King, Queen, Jack, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Half of the cards are red (Hearts and Diamonds) and half of the cards are black (Spades and Clubs).

Here's the game: A volunteer draws 10 cards from our classroom deck of cards, with replacement. The cards will be shuffled between each draw. The volunteer wins a point for each red card drawn and the dealer wins a point for each black card drawn.

Drawing the cards **with replacement** means that the volunteer will draw one card, record the color, and then put it back in the deck before drawing another card. Doing it this way assures that each color has the same chance of being drawn each time.

Let's think about a couple of questions before playing. *If* we are playing with a standard deck,

- What is the probability that the volunteer will pull out a red card?
- Of the 10 cards drawn, how many do you *expect* to be red?

This is our best guess (hypothesis), based on the information we have at hand.

- Do we think that the volunteer will draw **exactly** the hypothesized number of red cards?

There is randomness at play when we draw cards from a deck.

Okay, enough thinking. Let's play!

1.2.2 The Results

Suppose that our class drew 9 red cards (out of 10 cards drawn). Consider the following questions:

- Do these results seem consistent with how many we *expected* to be red? Or do they seem strange?
- Do you still believe your hypothesized probability of drawing a red card? In other words, do you still believe that we are playing with a standard deck?

Although these results may seem strange, this was only one game. We agreed that there is randomness at play here, so our results may have just been a fluke. Let's test this by playing the game again. Better yet, we can speed things up by simulating the game in RStudio. This way, we can play *lots* of games quickly to see if the results we got in our classroom game are really that surprising.

We can easily create a virtual deck of cards using the concatenation function. Since we only care about the color of the card (not the suit or the face value), let's make it easier to count by using a virtual deck that tells us only the color. We'll need 26 red cards and 26 black cards.

```
mycards<-c(rep("Red",26),rep("Black",26))
mycards
```

```
##  [1] "Red"   "Red"   "Red"   "Red"   "Red"   "Red"   "Red"   "Red"
##  [9] "Red"   "Red"   "Red"   "Red"   "Red"   "Red"   "Red"   "Red"
## [17] "Red"   "Red"   "Red"   "Red"   "Red"   "Red"   "Red"   "Red"
## [25] "Red"   "Red"   "Black" "Black" "Black" "Black" "Black" "Black"
## [33] "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black"
## [41] "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black"
## [49] "Black" "Black" "Black" "Black"
```

Let's again start by shuffling the deck.

```
shuffle(mycards)
```

```
##  [1] "Black" "Black" "Black" "Red"   "Red"   "Red"   "Black" "Red"
##  [9] "Red"   "Red"   "Black" "Black" "Black" "Black" "Red"   "Red"
## [17] "Black" "Black" "Black" "Red"   "Red"   "Black" "Black" "Red"
## [25] "Red"   "Black" "Black" "Black" "Black" "Red"   "Red"   "Red"
## [33] "Black" "Red"   "Black" "Black" "Black" "Black" "Red"   "Red"
## [41] "Red"   "Red"   "Black" "Red"   "Black" "Black" "Red"   "Red"
## [49] "Red"   "Red"   "Black" "Red"
```

We will randomly deal 10 cards from this new deck using the `sample` function. The `sample` function takes three arguments. You need to tell is where to sample from, how many to sample, and whether or not to sample with replacement.

```
sample(mycards, size=10, replace=TRUE)
```

```
##  [1] "Black" "Red"   "Red"   "Red"   "Black" "Black" "Red"   "Red"
##  [9] "Black" "Red"
```

Let's create a table to count the number of red cards and black cards that were in our hand of 10.

```
table(sample(mycards, size=10, replace=TRUE))
```

```
##
## Black   Red
##     4     6
```

Again, this was just one game. We would like to repeat this lots of times to see if our class results were really that unusual. We are looking to answer the questions:

- How *often* do we see our in-class results when we're playing with a standard deck? * What are the *chances* (or *probability*) of seeing our in-class results when we're playing with a standard deck?

Let's repeat the game three times.

```
do(3)*table(sample(mycards, size=10, replace=TRUE))
```

```
## Loading required package: parallel
```

```
##   Black Red
## 1     3   7
## 2     6   4
## 3     8   2
```

The first row of this table represents the first game in which 3 black cards were drawn and 7 red cards were drawn. The second row of this table represents the second game in which 6 black cards were drawn and 4 red cards were drawn. The third row of this table represents the third game in which 8 black cards were drawn and 2 red cards were drawn. If you wanted to simulate more games, you could change the 3 in the line of code above. For example, suppose you wanted to simulate 20 games. You would simply type:

```
do(20)*table(sample(mycards, size=10, replace=TRUE))
```

Let's look at this another way by creating a table that keeps track of how many of our games result in a certain number of red cards drawn.

```
## Red
##  0  1  2  3  4  5  6  7  8  9 10
##  0  0  1  0  1  0  0  1  0  0  0
```

The first row of this table represents the number of red cards drawn. The second row gives the number of games (out of the three that we played) that resulted in drawing that many red cards. We had one game where 7 red cards were drawn, one game where 4 red cards were drawn, and one game where 2 red cards were drawn.

We could get an even better idea of what's going on if we could simulate the game many times, say 1000 times! We will create the same type of table we just did to keep track of these games.

```
## Red
##   0   1   2   3   4   5   6   7   8   9  10
##   0  10  38 116 209 234 216 114  57   4   2
```

At this point, it's starting to seem that drawing a high number of red cards doesn't happen very often. It seems unlikely, but it sure would be nice to have an idea of just how unlikely it is.

Perhaps having this table in terms of percents is more useful:

```
##
## Red 0 1    2    3    4    5    6    7   8   9  10 Total
##     0 1  3.8 11.6 20.9 23.4 21.6 11.4 5.7 0.4 0.2   100
```

Consider the following questions based on our simulation:

If we draw 10 cards from a standard deck,

- What is the estimated *chance* of drawing 5 red cards?

Answer: There is a 23.4% chance of drawing 5 red cards.

- What is the estimated *probability* of drawing 1 red card?

Answer: There is a 1% probability of drawing 1 red card.

- How *likely* is it that our volunteer drew their original hand, based on our simulations?

Answer: Assuming that our volunteer drew 9 red cards, the likelihood that our volunteer drew their original hand (based on our simulations) is 0.4%.

Based on this likelihood, does it seem reasonable to believe that the deck we used in class was a standard deck?

Let's check out how RStudio can help us to visualize the data from our 1000 simulated games graphically, in the form of a histogram. We will be learning more about histograms in Chapter 2.

Here is the table again followed by the graphical representation (histogram). See Figure[Histogram].

```
##
## Red 0 1   2    3    4    5    6    7   8   9  10 Total
##     0 1 3.8 11.6 20.9 23.4 21.6 11.4 5.7 0.4 0.2   100
```

```
## Warning in mean.default(evalF$right[, 1], ...): argument is not numeric or
## logical: returning NA
```

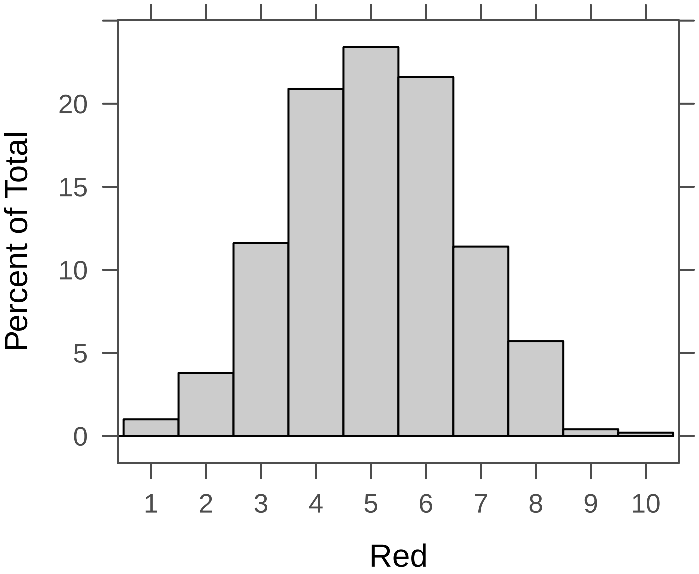


Figure 1.1: Histogram: Graphical representation of the table of counts for the 1000 simulated games.

The horizontal axis gives us the number of red cards drawn in a hand of 10 cards from a standard deck. The vertical axis gives us the percent of times (out of the 1000 simulated games) that a particular number of red cards was drawn.

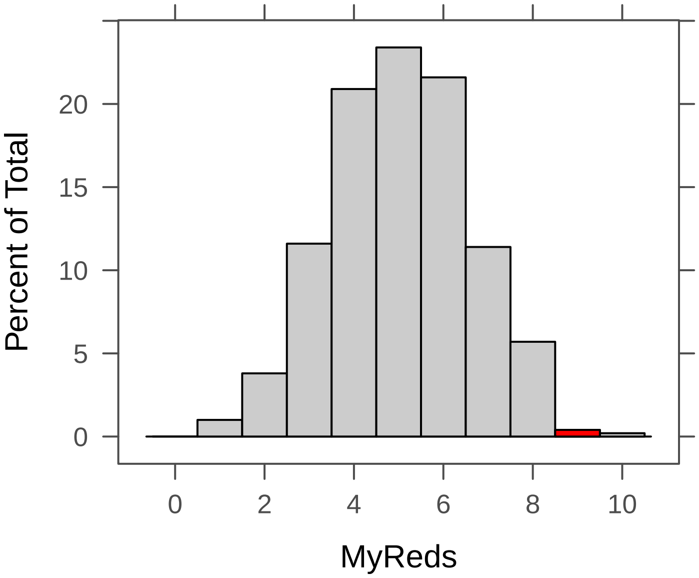


Figure 1.2: Class Probability: The shaded rectangle in the histogram represents the probability that the volunteer would draw 9 cards if they were drawing from a standard deck of playing cards.

Notice that the width of each bar in the histogram is 1 and the height of each bar is equal to the percent of the 1000 games that resulted in that number of red cards.

We can color the bar in the histogram to mark how many red cards our volunteer drew in the class game. See Figure[Class Probablity].

This colored part of the histogram represents the estimated *chance* of drawing our particular draw from a standard deck.

We can think about how *likely* it is that our class game resulted in such a high number of red cards (or higher). This *likelihood* is called a **p-value**. We can compute the p-value numerically and view it graphically on the histogram.

Numerically, the p-value is:

```
##     9
## 0.006
```

The p-value is the estimated *probability* of drawing as many red cards or more (out of 10) as our volunteer drew if we are drawing from a standard deck of playing cards.

Graphically, the P-value is the total area in the histogram that lies to the right of the vertical line in the graph. See Figure[P-Value].

1.2.3 The Conclusion

Summing this up:

1. We started with a hypothesis. We assumed that we were playing with a standard deck.
2. We gathered data from a real-world experiment to test our hypothesis. We played one real game.
3. Our results seemed strange, so we wondered if our hypothesis was true. We questioned: How likely was it to draw the hand that we did if we drew from a standard deck?

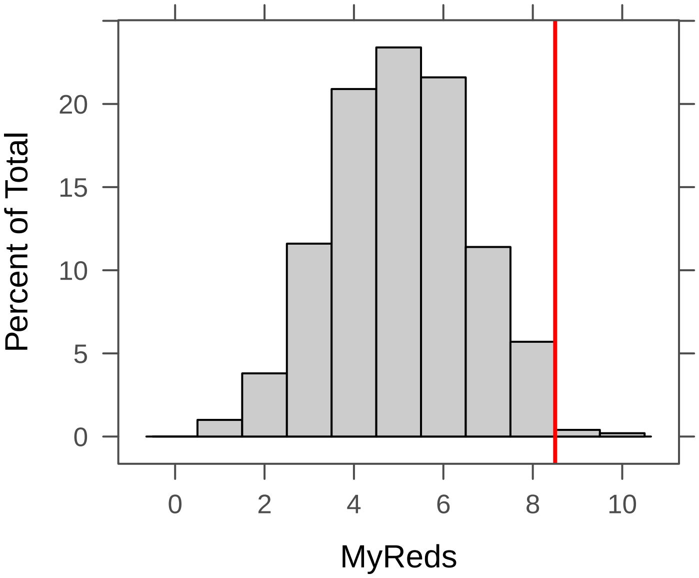


Figure 1.3: P-Value: The area of the histogram that lies to the right of the vertical line is the p-value.

4. We tested the hypothesis by simulating 1000 games using R. We counted up the number of games that gave us the result we got in class.
5. We calculated a P-value, the probability of getting results as extreme as ours (or more so!) from a standard deck.
6. Finally, we need to draw a conclusion.

Conclusion: If we assume that our volunteer drew 10 cards from a standard deck of cards, there is about a 0.006 (0.6 %) chance of drawing 9 red cards. Since the probability of drawing 9 (or more) red cards is so small, this should cause you to question whether the volunteer was really drawing from a standard deck.

1.3 Statistics

The ultimate goal of **statistics** is to translate data into knowledge and understanding of the world around us. It's the art and science of learning from data!

The card game we played above is a perfect example of the three aspects of statistics.

Design - asking the right questions and collecting useful data.

After we played the card game in class, we questioned whether we were playing with a standard deck of cards. We came up with a logical way to test whether our results were really as unlikely as they seemed. We used the power and speed of RStudio to quickly simulate 1000 games. This was our *data*.

Description - summarizing and analyzing data.

Once we had the raw data, we summarized it in the form of a table and a histogram. These summaries allowed us to analyze the data by calculating a p-value. A p-value is the probability of obtaining our results (or more extreme results) if our original hypothesis is true.

Inference - making decisions, generalizations, and turning data into new knowledge.

This is where the *art* of statistics comes into play. Based on your analysis, how convinced are you that the deck was stacked? How much evidence is enough? Were you convinced after 1 game, 3 games, 1000 games? Are you convinced now or do you think we should gather more data?

These are all ideas that we will investigate more thoroughly throughout the semester! As we continue to talk about these things, keep the Card Game in mind as a reference example.

1.4 Thoughts on R

Important R commands:

- `a+b` basic addition, “a plus b”
- `a-b` basic subtraction, “a minus b”
- `a*b` basic multiplication, “a times b”
- `a/b` basic division, “a divided by b”

Know how to use these functions:

- `sqrt()` square root function
- `help()` help function will pull up the help file for any function for which you need more information
- `c()` concatenation function combines values, letters, or words into a list
- `rep()` replication function creates a list of repeated values, letters, or words.
- `seq()` sequence function creates a sequence of values starting and ending at a specified spot with a specified increment.

Always remember to **`require()`** the needed packages.

- `require(tigerstats)` anytime that you open a new Rscript or RMarkdown file.
- `require(knitr)` anytime that you are working in a RMarkdown file.

- `require(manipulate)` anytime you want to run one of the class apps, available in `tigerstats`.

Chapter 2

Describing Patterns in Data

Before you begin work, it's a good idea to check that the necessary packages are loaded:

```r
require(mosaic)
require(tigerstats)
```

2.1 Data Basics

2.1.1 Getting Data in R

To start, let's work with the `m111survey` data. It is always present when you load `tigerstats`, and can be used right away. However, when a dataset is new to you, you should first take a few steps to become familiar with it. We recommend the following procedure:

First, put `m111survey` into your Global Environment:

```r
data(m111survey)
```

Next, take a quick look at it:

```r
View(m111survey)
```

Finally, learn more about the dataset: who collected it and why, and what the variables mean. You can do this quickly with:

```r
help(m111survey)
```

This is *sample data*: we took a sample from the population of all GC students. It comes in the form of a *data frame*. Each row corresponds to an *individual* (sometimes called an *observation*). Each column corresponds to a *variable*. A variable is something that you measure on an individual. The *value* of the variable can change from one individual to another.

2.1.2 Variable Types

- *Categorical* (called “factor” in R“). Values are not numbers. Example: **sex** in `m111survey`. The values of **sex** are:”female" and “male”). Some categorical variables come in a natural order, and so are called *ordinal* variables. Example: **seat** in `m111survey`.
- *Quantitative* (called “numeric” in R“). Values are numbers. There are two sub-types:
  - *Discrete* (called “integer” in R). Values are whole numbers. Example: how many brothers you have. Possibles values are 0,1,2 . . . .
  - *Continuous* (officially called “double” in R, but usually just listed as “numeric”). Values lie in a range of real numbers. Example: **height** in `m111survey`.

R will classify the variables in a data frame for you, if you call the **str** (structure) function:

```
str(m111survey)
```

```
## 'data.frame':    71 obs. of  12 variables:
##  $ height        : num  76 74 64 62 72 70.8 70 79 59 67 ...
##  $ ideal_ht      : num  78 76 NA 65 72 NA 72 76 61 67 ...
##  $ sleep         : num  9.5 7 9 7 8 10 4 6 7 7 ...
##  $ fastest       : int  119 110 85 100 95 100 85 160 90 90 ...
##  $ weight_feel   : Factor w/ 3 levels "1_underweight",..: 1 2 2 1 1 3 2 2 2 3 ...
##  $ love_first    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ extra_life    : Factor w/ 2 levels "no","yes": 2 2 1 1 2 1 2 2 2 1 ...
##  $ seat          : Factor w/ 3 levels "1_front","2_middle",..: 1 2 2 1 3 1 1 3 3 2 ...
##  $ GPA           : num  3.56 2.5 3.8 3.5 3.2 3.1 3.68 2.7 2.8 NA ...
##  $ enough_Sleep  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 2 1 2 ...
##  $ sex           : Factor w/ 2 levels "female","male": 2 2 1 1 2 2 2 2 1 1 ...
##  $ diff.ideal.act.: num  2 2 NA 3 0 NA 2 -3 2 0 ...
```

Structure guides interpretation! We will see this in the next section: the types of the variables you are interested in determine the methods you will use to study them and to describe them to other people.

2.1.3 Descriptive Statistics

Reading Data for Interpretaion: When you look at sample data, you may wish to examine it to see what patterns it has, and you may wish to summarize the data and describe these patterns to others. Such practices are called *descriptive statistics*. The tools for both of these jobs are the same, but the specific tools you use depend on the kind *Research Question* you have about the data.

2.2 Outline

There are many tools in statistics to describe patterns in data, and in the next few chapters we will introduce you to some of them. Since the tools you use will depend on the research Question you have in mind, we will organize the discussion around the different types of Research Questions, and introduce the tools as we need them.

In this chapter we will cover the basic tools needed to address Research Questions involving:

- One factor variable
- One numeric variable

- Relationship between two factor variables
- Relationship between a factor variable and a numeric variable

Questions about relationships between variables are especially interesting, so in the next chapter, we will look more deeply into the relationship between two factor variables, and in the chapter after that we will consider the relationship between two numeric variables.

The `m111survey` data frame will be used in most of the examples in this chapter, but from time to time we will introduce additional data frames.

2.3 One Factor Variable

Say that we are interested in the following:

> **Research Question**: What percentage of the sample are females?

This Research Question is about **one factor variable**, namely the variable **sex**. We can describe it in two ways, *graphically* and *numerically*. Let's first try a numerical approach.

2.3.1 Tables

We can make a tally of males and females:

```
xtabs(~sex,data=m111survey)
```

```
## sex
## female   male
##     40     31
```

This gives us counts, but we would like to see percents, so we try:

```
rowPerc(xtabs(~sex,data=m111survey))
```

```
##
## sex female  male Total
##      56.34 43.66   100
```

We see that a majority (56.34%) of the students in the sample are female.

Note: If you want to see both the counts and the percentages without having to do too much typing, then you should first store the table in a new object, with a good descriptive name of your own choosing, one that will help you remember what the object contains. Here, we will use the name `tablesex`:

```
tablesex <- xtabs(~sex,data=m111survey)
```

Then you can print the table to the console simply by typing:

```
tablesex
```

```
## sex
## female   male
##     40     31
```

You can also get the row percents with:

```
rowPerc(tablesex)
```

```
##
## sex female  male Total
##      56.34 43.66   100
```

Here is another example:

> **Research Questions**: *What is the distribution of seating preference in the sample, and what percentage of students prefer to sit in the front of a classroom?*

The *distribution* of a categorical variable is simply a statement of the percentage of the time it takes on each of its possible values, so we would compute:

```
rowPerc(xtabs(~seat,data=m111survey))
```

```
##
## seat 1_front 2_middle 3_back Total
##        38.03    45.07   16.9   100
```

Apparently the students in the sample tend to prefer front and middle (38% and 45% respectively) more than they prefer the back (only about 17%).

2.3.2 Barcharts.

Barcharts convey the same information as `xtabs`, but in a graphical way.

To make a barchart, send your table to the **barchart** function (see Figure [Barchart]).

```
barchartGC(~sex,data=m111survey,
           main="Distribution of Sex",
           type="percent")
```

2.4 Two Factor Variables

Now let's say that we are interested in the following:

> **Research Question**: *Who is more likely to sit in the front: a guy or a gal?*

This Research Question is about the relationship between two factor variables, namely **sex** and **seat**. We wonder whether knowing a person's sex might help us predict where the person prefers to sit, so we think of **sex** as the *explanatory* variable and **seat** as the *response* variable.

Important Idea: When we are studying the relationship between two variables X and Y, and we think that X might help to cause or explain Y, or if we simply wish to use X to help predict Y, then we call X the *explanatory* variable and Y the *response* variable.

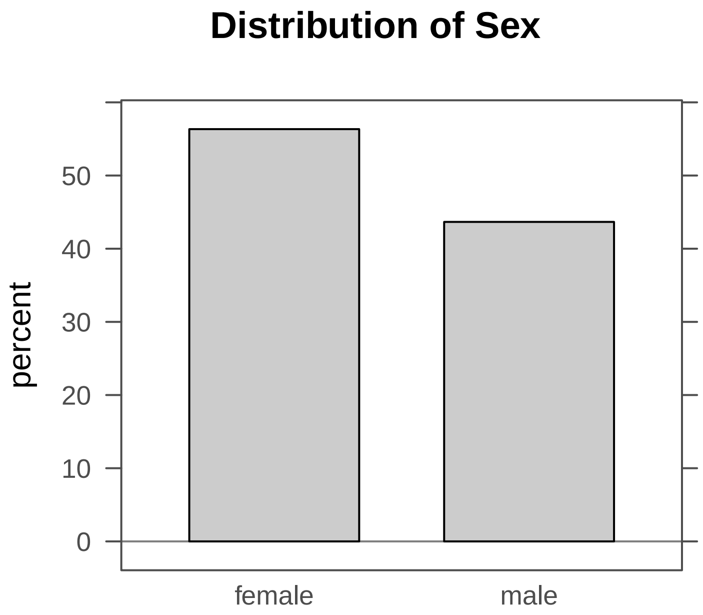


Figure 2.1: Barchart

2.4.1 Two-Way Tables

These are also called *cross-tables* or *contingency tables*. You can make them using `xtabs()`:

```
tabsexseat <- xtabs(~sex+seat,data=m111survey)
tabsexseat
```

```
##         seat
## sex       1_front 2_middle 3_back
##   female       19       16      5
##   male          8       16      7
```

In the formula for `xtabs()` above, the variable you put first goes along the rows of the table. As a rule, we like to put the explanatory variable first.

The counts in the two-way table don't do much to help us figure out if guys and gals differ with regard to seating preference. That's because there are more gals than guys in the sample, in the first place, so just because there are more gals in the sample who prefer the front than guys in the sample who prefer the front (19 vs. 8) doesn't tell you much. What we really want to do is to compare percentages:

```
rowPerc(tabsexseat)
```

```
##         seat
## sex       1_front 2_middle 3_back  Total
##   female    47.50    40.00  12.50 100.00
##   male      25.81    51.61  22.58 100.00
```

We see that 47.5% of the women prefer the front, whereas only 25.81% of the men prefer the front. Looking at all the row percentages, it appears that as far as the sample is concerned the guys tend to sit more towards the back, as compared to gals.

2.4.2 Barcharts Again

Again, a barchart will convey the information in a table, but in graphical form (see Figure [A Two-Way Barchart]).

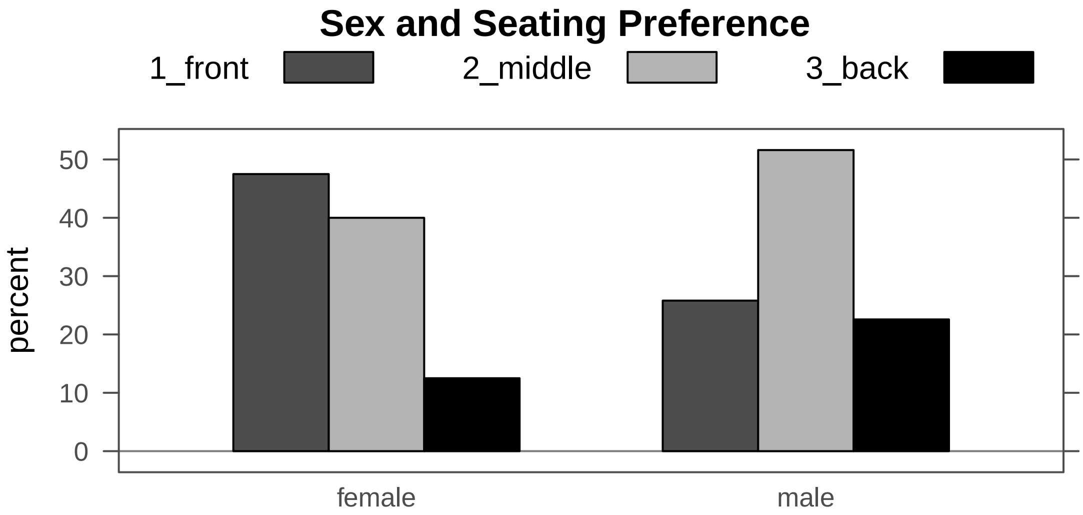


Figure 2.2: A Two-Way Barchart

Some people like to make “flat” barchaerts, using the optional `flat` parameter:

```
barchartGC(~sex+seat,
           data=m111survey,
           type="percent",
           main="Sex and Seating Preference",
           flat=TRUE)
```

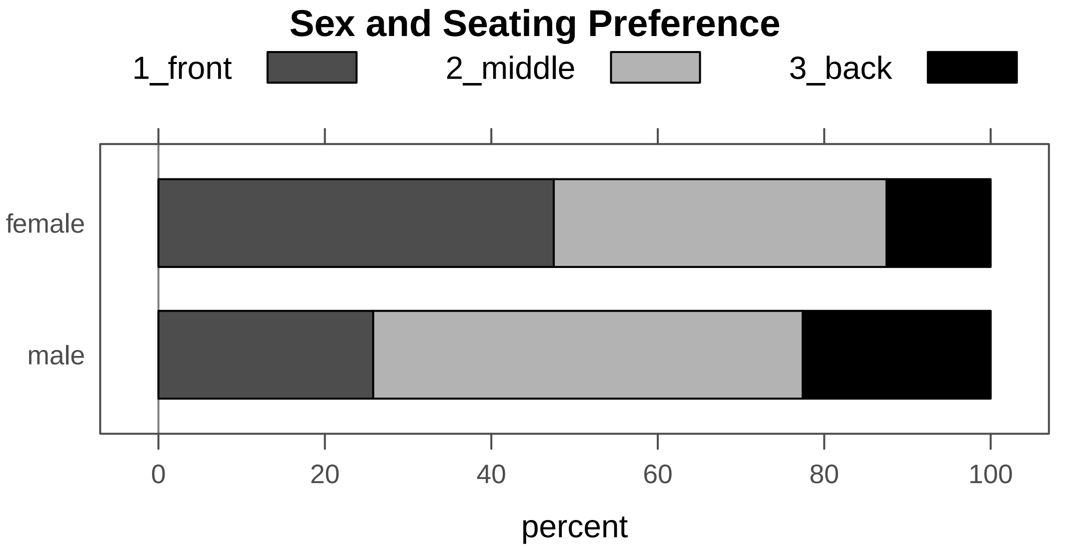


Figure 2.3: A Flat Two-Way Barchart

The lure of flat barcharts (like Figure [Flat Two-Way Barchart]) is that they are a visual “match” for a two-way table of row percentages.

Here is another example:

> **Research Question**: *Who is more likely to believe in love at first sight: a guy or a gal?*

To address the question numerically, we make a two-way table:

```
sexlove <- xtabs(~sex+love_first,data=m111survey)
sexlove
```

```
##         love_first
## sex       no yes
##   female 22  18
##   male   23   8
```

The row percents are:

```
rowPerc(sexlove)
```

```
##         love_first
## sex           no    yes  Total
##   female  55.00  45.00 100.00
##   male    74.19  25.81 100.00
```

In the sample, females appear to be more likely to believe in love at first sight (45% vs about 26% for the guys).

Figure [Barchart Sex and Love] shows the same thing graphically.

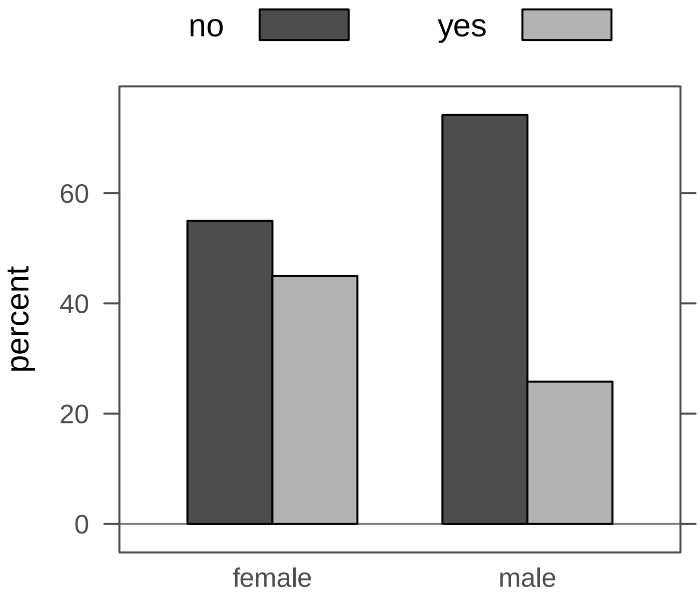


Figure 2.4: Barchart Sex and Love: belief in love at first sight in the m111survey data, broken down by sex of respondent.

2.5 One Numerical Variable

Suppose we are interested in the following:

> **Research Question**: *How fast do GC students drive, when they drive their fastest?*

This Research Question deals with **one numerical variable.** As usual, it can be explored both graphically and numerically. There is a lot to say about numerical variables, and many tools have been developed to study them.

Describing the distribution of a single factor variable is pretty straightforward: one simply states the percentages of the time each value of the variable occurs in the data. When we describe a single numerical variable, there is more to say. Specifically, we want to describe:

- The Center
- The Spread
- The Shape

We will learn more about these terms as we go along.

To help with describing center and spread, we will learn about several numerical measurements:

- the median and other percentiles
- the five-Number Summary
- the mean
- the standard deviation
- the Interquartile Range

To help with describing shape, we will consider three graphical tools:

- Histogram
- Density Plot
- Boxplot

2.5.1 Numerical Measures

A convenient way to obtain some important numerical summaries of a dataset is provided by the R-function `favstats`. For the current Research Question, we would invoke:

```r
favstats(~fastest,data=m111survey)
```

```
##  min   Q1 median    Q3 max     mean      sd  n missing
##   60 90.5    102 119.5 190 105.9014 20.8773 71       0
```

In this section we will discuss each of the measurements provided by `favstats`.

2.5.1.1 The Mean

The mean of a list of numbers is the sum of the numbers, divided by how many there are. When the list of numbers is a sample from a population, then the mean is called the *sample mean* and is written $\bar{x}$. The mean of a population is called the the *population mean* and it is often written as μ.

Sometimes you see the formula for the sample mean written like this:

$$\bar{x} = \frac{\sum x_i}{n},$$

where:

- $\sum$ means summing
- x_i denotes the individual values to be summed
- n denotes the number of values in the list.

Most of the time we will get the mean of a variable through favstats, but if we ever want the mean alone we can get it with the R-function `mean`:

```
FakeData <- c(2,4,7,9,10)
mean(FakeData)
```

```
## [1] 6.4
```

Sometimes it's just easiest to compute the mean "by hand", or use R as a calculator to compute the mean:

```
(2+4+7+9+10)/5
```

```
## [1] 6.4
```

For many data sets, the mean appears to be about in the "middle" of the data, so it is often used as a measure of the center of the distribution of a variable.

2.5.1.2 The Standard Deviation

The standard deviation (or SD for short, or even just s) indicates how much a typical data value differs from the mean of the data.

Here is the mathematical formula for the SD of a sample:

$$s = \sqrt{(\sum (x_i - \bar{x})^2)/(n-1)}.$$

This is a symbolic way of saying:

- Find the mean of the numbers.
- Subtract the mean from each number x_i (the results are called the *deviations*), and then square the deviations.
- Add up the squared deviations.
- Average them, almost, by dividing the sum by how many there are MINUS ONE. (What's up with *that*?? Consult the GeekNotes to find out!)

- Take the square root of this “almost-average.”

The bigger the SD, the more spread out the data are, and so the SD is often used as a measure of the spread of the distribution of a variable. Most of the time, a majority of the numbers in a dataset lie within one standard of the mean, and nearly all of them lie within two SDs of the mean. We’ll learn more about this later, when we come to the 68-95 Rule.

2.5.1.3 The Median

Suppose that you have some data, sorted in order from lowest to highest:

```
FakeData <- c(2,4,7,9,10)
```

The *median* of the data is the number that is right in the middle:

```
median(FakeData)
```

```
## [1] 7
```

If there are an even number of data points, then the median is the average of the two points closest to the middle:

```
FakeData2 <- c(2,4,7,9,10,15)
median(FakeData2)
```

```
## [1] 8
```

Note that in `FakeData2` the values 7 and 9 were closest to the middle, and the median is the average of these two values:

```
(7+9)/2
```

```
## [1] 8
```

About 50% of the data values will lie below the the median of the data. Like the mean, it is often used as a measure of the center of a distribution.

2.5.1.4 Quantiles and the IQR

Since the median lies above about 50% of the data, it is often called the 50th *percentile* of the data, or the 50th *quantile* of the data. For every percentage from 0% to 100% there is a corresponding quantile. You can get as many of them as you like by using R’s `quantile` function. For example, if you want the 20th, 50th, 80th and 90th quantiles of the variable **fastest** in the `mat111survey` data, try

```
with(m111survey,
     quantile(fastest,probs=c(0.2,0.5,0.8,0.9))
     )
```

```
## 20% 50% 80% 90%
##  90 102 120 130
```

Here's you interpret the quantiles provided above:

- About 20% of the students in the survey drove slower than 90 mph.
- About 50% of the students in the survey drove slower than 102 mph. (This is the median.)
- About 80% of the students in the survey drove slower than 120 mph.
- About 90% of the students in the survey drove slower than 130 mph.

Of course you could also say:

- About 80% of the students in the survey drove *faster* than 90 mp.
- About 50% of the students in the survey drove *faster* than 102 mph.,

and so on!

Two important quantiles are:

- the 25th quantile (also called the *first quartile*)
- the 75th quantile (also called the *third quartile*)

They are provided in `favstats` as Q1 and Q3 respectively:

```
favstats(~fastest,data=m111survey)
```

```
##  min   Q1 median    Q3 max     mean      sd  n missing
##   60 90.5    102 119.5 190 105.9014 20.8773 71       0
```

The difference between them is called the *interquartile range*, or IQR for short:

$$IQR = Q3 - Q1 = 119.5 - 90.5 = 29.$$

The IQR tells you the range of the middle 50% of the data, so it is often used as a measure of spread: the bigger the IQR, the more spread out your data are.

The *five-number summary* gives you a good quick impression of the distribution of a variable: The five numbers are:

- the minimum value of the data
- the first quartile Q1
- the median
- the third quartile Q3
- the maximum value of the data

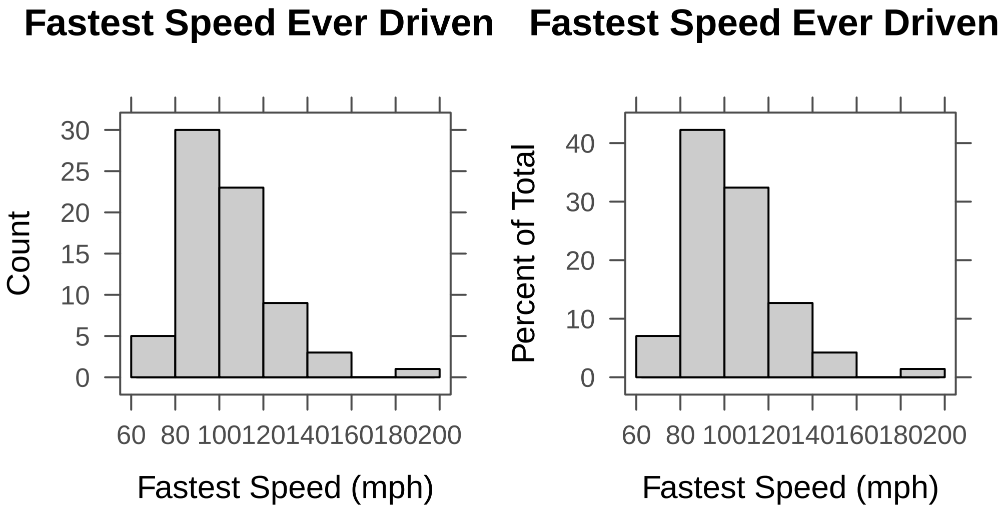
Note that all of these measurements are provided in `favstats`.

2.5.2 Graphical Tools

Numerical measurements aren't very satisfying on their own. To understand fully what they are telling you, and to be able to say something about the *shape* of a distribution as well, you need to combine them with graphical tools for describing the variable.

2.5.2.1 The Histogram

The histogram is a popular device for representing the shape of the distribution of a numerical variable. Figure [Two Histograms] shows two common types: a *frequency* histogram on the left, and a *percentage* or *relative frequency* histogram on the right.



Here's the difference:

- In a frequency histogram, the height of a rectangle gives the number of observations falling within the boundaries indicated by the rectangle's base. For example, in the frequency histogram on the left, the rectangle from 60 to 80 mph on the the x-axis is 5 units high: this means that five of the students drove between 60 and 80 mph, when driving their fastest. (The speed 60 mph is included in this rectangle, whereas a speed of 80 mph would be included in the next rectangle.)
- In a relative frequency histogram, the height of a rectangle gives the *percentage* of the observations falling within the boundaries indicated by the rectangle's base. For example, in the relative frequency histogram on the right, the rectangle from 80 to 100 mph on the the x-axis is about 42 units high: this means that about 42% of the students drove between 60 and 80 mph, when driving their fastest.

Most of the time, we will deal with a third type of histogram known as a *density* histogram. An example is given in Figure [Density Histogram].

In a density histogram:

- the *area* of a rectangle gives you the proportion of data values that lie within the left and right-hand endpoints of the rectangle;
- the total area of all of the rectangles equals 1.

For example, in the rectangle from 100 to 120 mph, the width is 20 and the height is about 0.016, so the area is

$$20 * 0.016 = 0.32,$$

which says that 0.32 is the proportion of the students who drove between 100 and 120 mph. In more familiar terms, we would say that about 32% of the students drove between 100 and 120 mph.

Density histograms seem to be a round-about way to get percentages: why not just look at relative frequency histograms instead? The answer will become more clear when we study the next graphical device: the density plot.

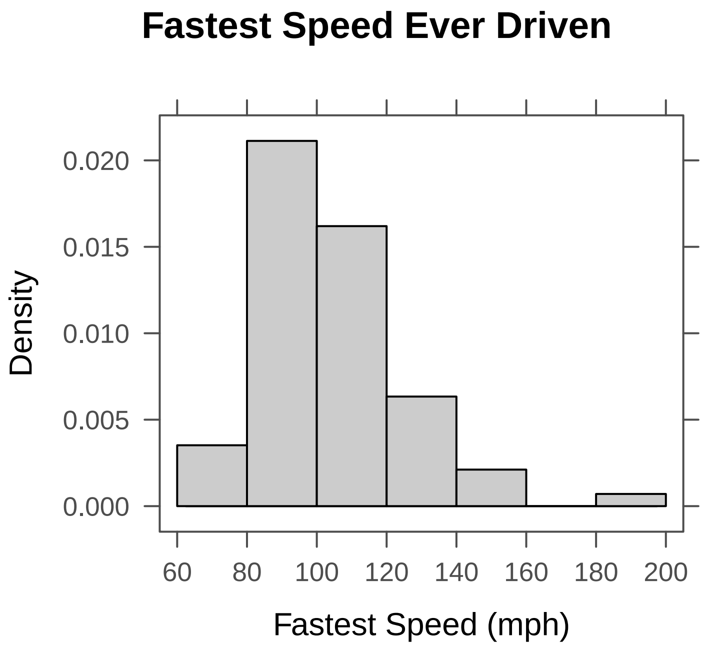


Figure 2.5: Density Histogram. The area of each rectangle gives the proportion of observations that fall within the boundaries indicated at the base of the rectangle.

2.5.2.2 Density Plots

From time to time we will work with an imaginary population contained in the data frame `imagpop`:

```
data(imagpop)
```

There are 10,000 people in this population.

Figure [Income Histogram] shows a density histogram of the annual incomes in this population. The histogram consists of approximately 100 rectangles.

When the number of rectangles is very large, don't you just *see*, in your mind, a smooth curve that the tops of the rectangles seem to follow? R can draw that curve for you if you call the function `densityplot`:

```
densityplot(~income,data=imagpop,
            main="Distribution of Income\n(Density Plot)",
            xlab="Annual Income (dollars)",
            plot.points=FALSE)
```

The results are shown in Figure [Income Density Plot]. Areas under a density curve give you proportions, so the total are under a density curve is 1.

Density plots aren't just for whole populations. When you are dealing with sample data from a population the density plot gives you an estimate–based on your sample–of what the distribution of the population might look like. Figure [Fastest Density Plot] shows a density plot of the fastest speed ever driven for the subjects in `m111survey` sample.

```
densityplot(~fastest,data=m111survey,
            main="Fastest Speed Ever Driven",
            xlab="speed (mph)",
            plot.points=TRUE)
```

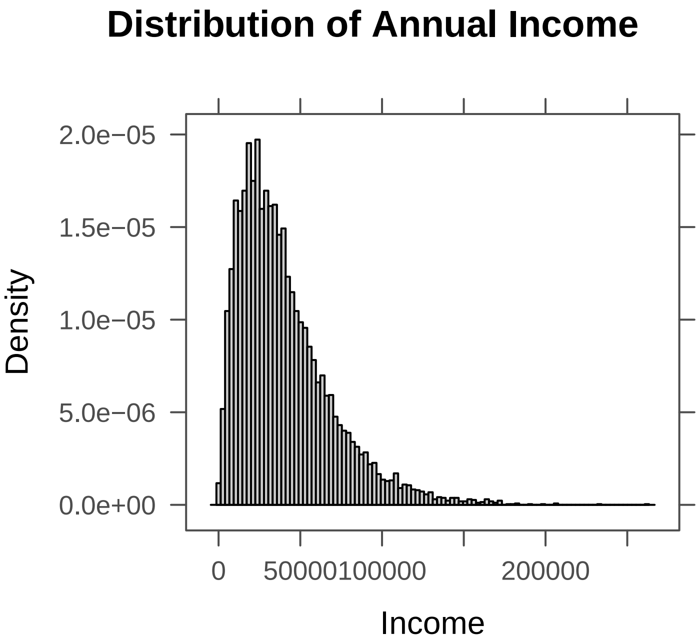


Figure 2.6: Income Histogram. The number of rectangles has been chosen to be very large.

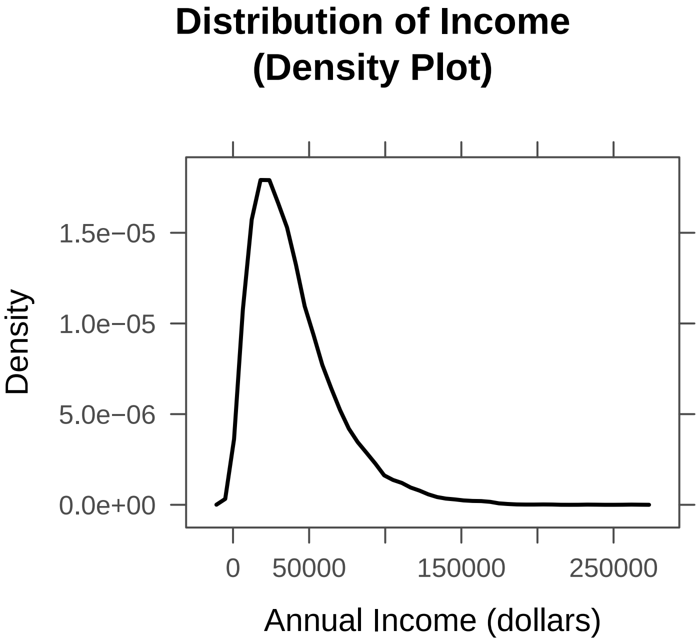


Figure 2.7: Income Density Plot. The plot shows the general shape of the distribution.

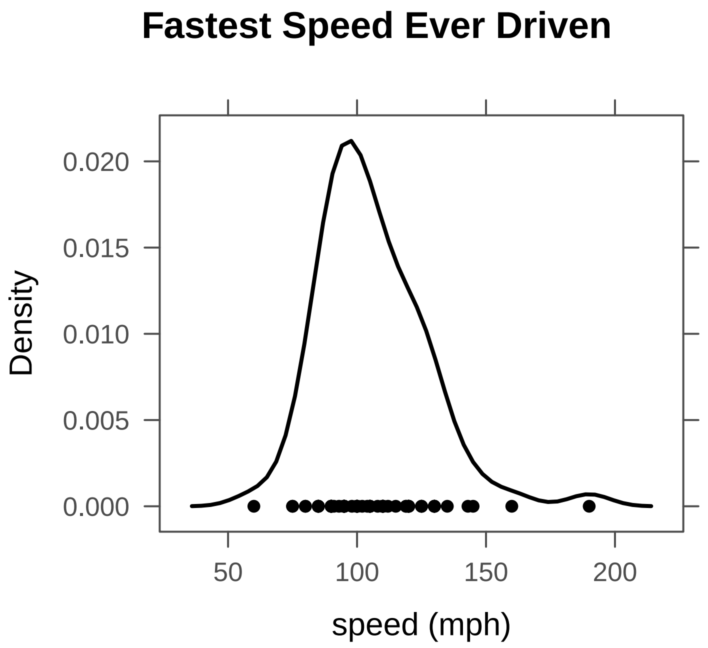


Figure 2.8: Fastest Density Plot. When the number of data points is not too large, it is good to produce a "rug" of individual points by setting plot.points to TRUE.

2.5.2.3 Describing Shape

When it comes to describing a distribution, recall that we want to describe center, spread and shape. So far we have measures for:

- Center (mean, median)
- Spread (standard deviation, interquartile range)

We use our graphical tools to describe shape. Some important features to look for when describing the shape are:

- skewness (left-skewed, right skewed)
- symmetry
- number of modes (unimodal, bimodal)

A *symmetric* distribution looks like a mirror image of itself around some central vertical line. A *skewed* distribution has a *tail* running off to the left (smaller values) or to the right (larger values). The *modes* of a distribution correspond to "humps" in the density curve: values that occur quite commonly in the data. Look at Figure [Shapes] for standard examples of these concepts.

Here are some examples:

- Looking back at Figure [Fastest Density Plot], we see that the distribution of the variable **fastest** is unimodal, and skewed a bit to the right.
- Figure [Kim Kardashian Temperature] shows a density histogram of ratings given by people in `imagpop` to the celebrity Kim Kardashian. The ratings are on a scale of 0 to 100, where 0 indicates that one doesn't like her, and 100 indicates a very high level of liking. The ratings distribution is symmetric, but bimodal: there are "humps" near 0 and near 100.

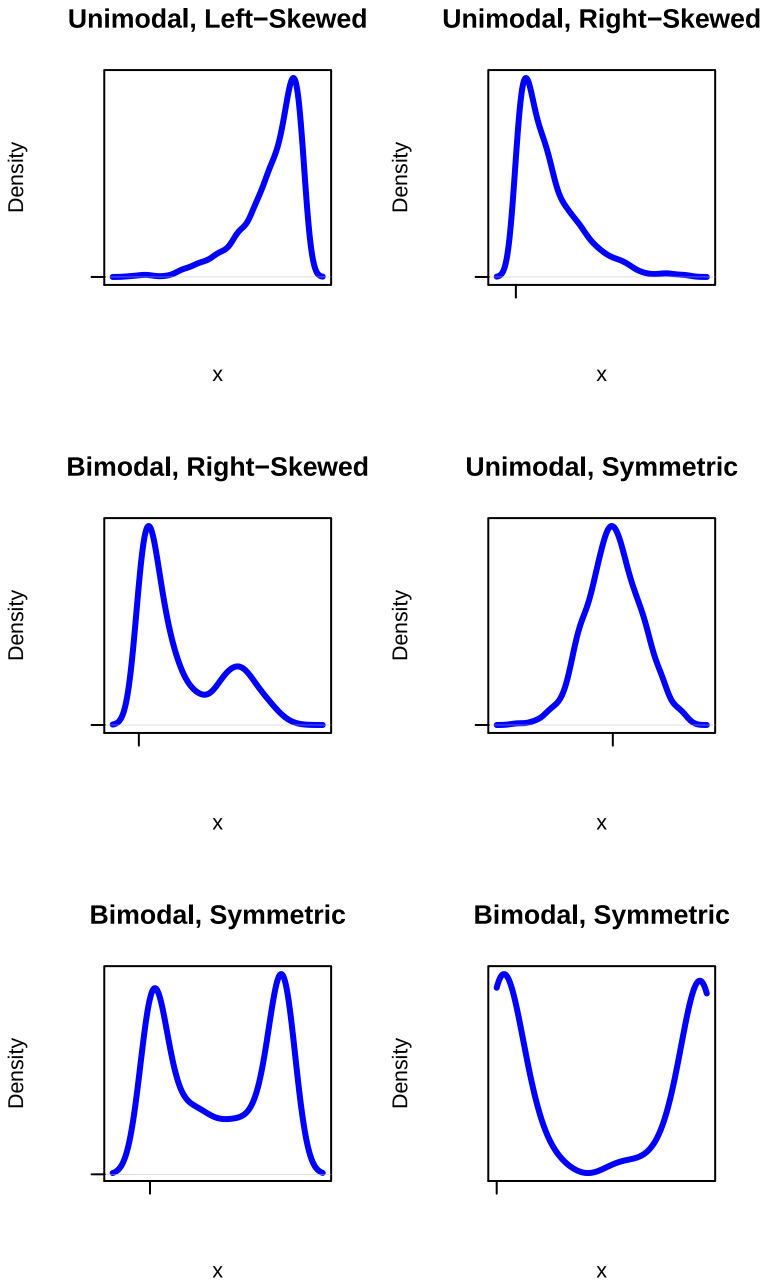


Figure 2.9: Shapes. Terminology for describing the shape of a distribution.

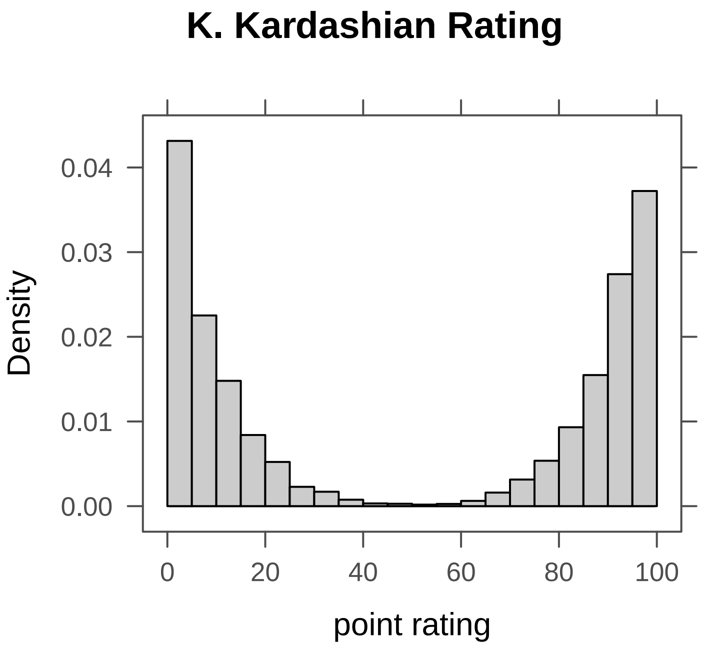


Figure 2.10: Kim Kardashian Temperature. People either love her or hate her!

2.5.2.4 Box-and-Whisker Plots

The five-number summary is the basis for a very useful graphical tool known as the *box-and-whisker plot*, or just "boxplot" for short. Figure [Boxplot] shows a boxplot of some imaginary data:

```
ImaginaryData <- c(7.1,7.3,7.5,8.2,8.5,9.1,9.5,
                   9.8,9,9,9.9,10,10.5,10,9)
bwplot(~ImaginaryData,xlab="x",main="Example Boxplot")
```

Here are the basic features of a boxplot:

- The dot is at the median of the data.
- The box starts at Q1 and goes to Q3. Therefore the length of the box is the IQR of the data. The middle 50% of the data lie within the box.
- The lower hinge is at the minimum of the data. From the lower hinge to Q1, we see the lower "whisker", indicated by a dotted horizontal line. The lowest 25% of the data is in this whisker.
- The upper hinge is at the maximum of the data. The upper whisker extends from Q3 to the upper hinge, and contains the largest 25% of the data.

Figure [Boxplot Height] shows a boxplot of the heights of the students in `m111survey`. It was produced as follows:

```
bwplot(~height,data=m111survey,
       main="Height at GC",
       xlab="height (inches)")
```

Note that this time the lower hinge does not extend all of the way down to the minimum value: there are two heights so small that R decided that you might consider them to be outliers. An *outlier* in a dataset is a

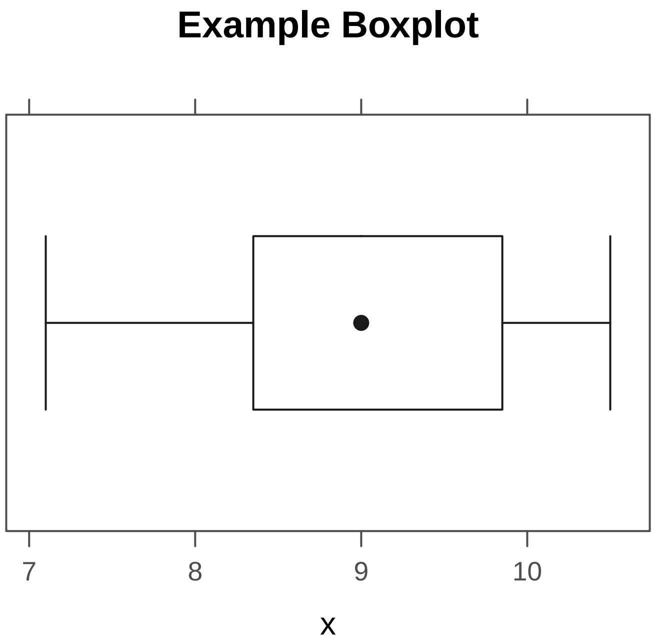


Figure 2.11: Boxplot. Box-and-whisker plot of some imaginary data.

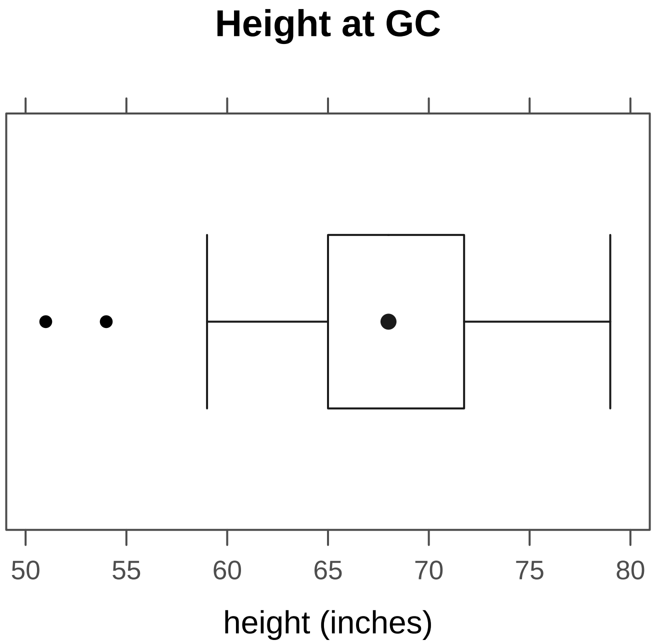


Figure 2.12: Boxplot Height. Note the two outliers.

value that lies far above or far below most of the other values. When R detects possible outliers, they are plotted as individual points.

When R plots outliers on a boxplot, it extends the whisker to the most extreme value that it did NOT consider to be an outlier. Thus in the heights data, there was at least one person who was 59 inches tall (lower hinge at 59), but this value was not deemed an outlier.

Boxplots are not only useful for detecting possible outliers: they can also detect skewness easily. Look at Figure [Boxplot Fastest], which shows the fastest speeds driven by the `m111survey` participants. As we have seen previously, this distribution is a bit skewed to the right, and the boxplot shows the skewness: the upper whisker is somewhat longer than the lower whisker, indicating that there is longer “tail” toward the higher values of speed. We also see that there is an outlier at 190 miles per hour.

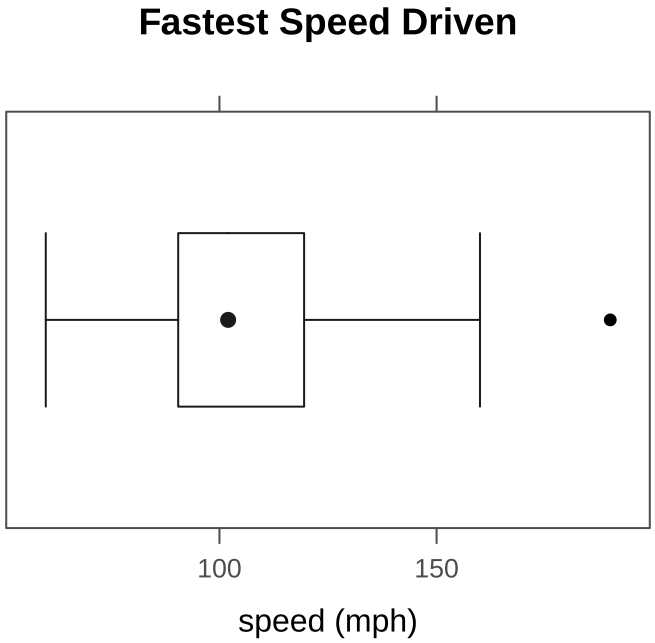


Figure 2.13: Boxplot Fastest. The distribution is a little bit skewed to the right, and there is an oulier at about 190 mph.

Although boxplots excel at detecting skewness and outliers, they fail miserably at detecting modes, so if you are interested in modes you should look at a density plot or histogram as well. Figure [Kardashian Violin] shows a combination of a violin plot and a boxplot for the Kim Kardashian ratings in the `imagpop` population. A violin plot is nothing more than a density plot combined with a mirror image of itself. Places where the violin is “thick” are regions where data values are relatively crowded together; in “thin” regions, data values are more widely separated from one another. Note that the boxplot correctly indicates that the distribution is symmetric, but entirely misses the bimodality.

2.6 Factor and Numerical Variable

Suppose that we are interested in the following:

> **Research Question**: *Who tends to drive faster: GC guys or GC gals?*

This research question concerns the relationship between two variables in the `m111survey` data: **fastest** and **sex**. **fastest** is numerical, and **sex** is a factor. Since we are inclined to think that one’s sex might,

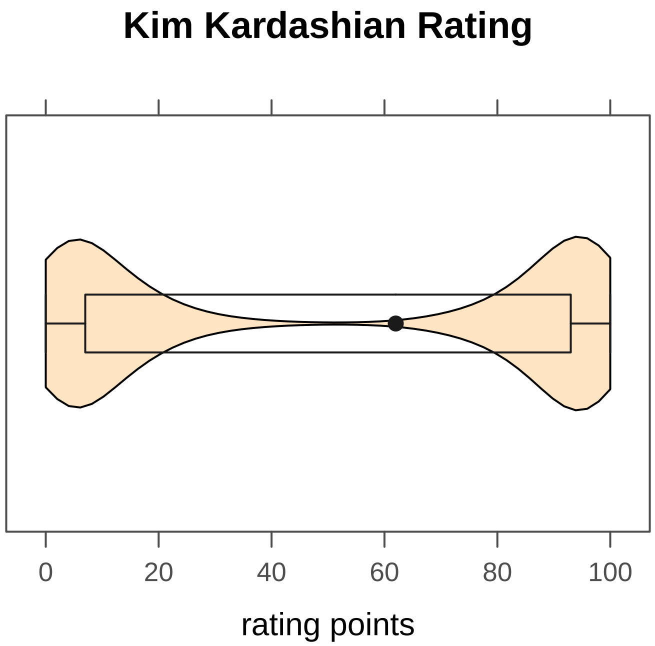


Figure 2.14: Kardashian Violin. The violin plot supplements the boxplot, indicating that the data are clumped around 0 and around 100, and are quite sparse in the middle.

through cultural conditioning, have some effect on how fast one likes to drive, we shall consider **sex** to be the explanatory variable and take **fastest** to be the response variable.

Let's investigate this question both numerically and graphically.

2.6.1 Numerical Tools

We use the same numerical measurements as when we are studying one numerical variable, but we have to compute them separately for each of the groups that go along with the different possible values of the factor variable. In the current Research Question, the factor variable **sex** has two values ("female" and "male"), so we need to separate the two sexes and compute numerical statistics for each of the two groups thus formed.

The R-function `favstats` can do this easily. For formula-data input, we use `m111survey` as the data, just as usual, but the formula must look like

$$numerical \sim factor$$

Thus we compute:

```
favstats(fastest~sex,data=m111survey)
```

```
##   .group min Q1 median    Q3 max     mean       sd  n missing
## 1 female  60 90     95 110.0 145 100.0500 17.60966 40       0
## 2   male  85 99    110 122.5 190 113.4516 22.56818 31       0
```

The next step is to decide what numbers to compare. The minima and the maxima differ a lot for the two groups, but each of these numbers might be based on only one individual. It's better, therefore, to compare measures of center. The mean for the males is 113.5 mph, considerably higher than the female mean of 100 mph. Therefore it seems that the males tend to drive faster, on average than the females in the sample drive. (One could also compare medians and arrive at the same conclusion.)

2.6.2 Graphical Tools

Again, we can use histograms, density plot and boxplots, but whichever tool we choose we must "break down" the numerical data into groups determined by the values of the factor variable.

2.6.2.1 Parallel Boxplots

For boxplots, use the same *numerical* ∼ *factor* formula as for `favstats`. Figure [Boxplot of Fastest by Sex] shows that guys tend to drive faster: the median for the guys is somewhat higher than the median for the gals. We might also point out that "box" for the guys, which represents the middle 50% of the guys' speeds, is higher than the middle 50% of the gals' speeds.

```
bwplot(fastest~sex,data=m111survey,
       main="Fastest Speed Driven, by Sex",
       xlab="Sex",
       ylab="speed (mph)")
```

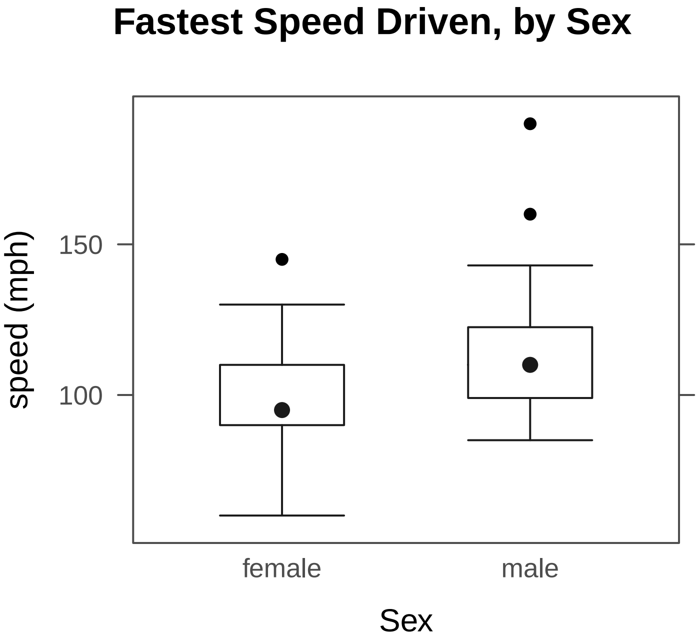


Figure 2.15: Boxplot of Fastest by Sex.

Here is another example:

> **Research Question**: *Who tends to have higher GPAs? People who prefer sitting in the front, the middle or the back?*

We will investigate the question graphically and numerically. For a numerical approach, we use `favstats`. The variable **GPA** is numerical, and **seat** is a factor, This tells us how to call the `favstats` function:

```
favstats(GPA~seat,data=m111survey)
```

```
##     .group min   Q1 median  Q3 max     mean        sd  n missing
## 1  1_front 2.1 3.05  3.500 3.7 4.0 3.337667 0.4822950 27       0
## 2 2_middle 1.9 2.90  3.200 3.5 3.9 3.110645 0.5344276 31       1
## 3   3_back 2.2 2.80  3.057 3.5 3.7 3.092333 0.4473687 12       0
```

Front-sitters seem to have a little bit higher GPA, on average, than other folks do (3.34 for front-sitters, vs. about 3.1 for the other two groups). It's not clear whether this is a really big or important difference, though.

For a graphical approach, let's look at Figure [GPA by Seat]. We see that the front-sitters have a higher median than the other two groups do.

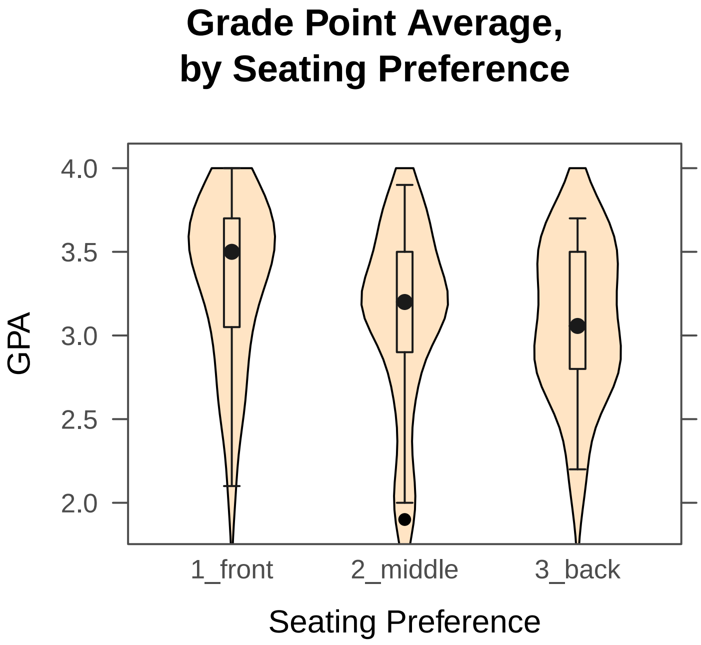


Figure 2.16: GPA by Seat.

2.6.2.2 Side-by-Side Histograms

In order to make a histogram for each group determined by the values of a factor variable, you need to "condition" the numerical variable the factor variable in your formula. This is accomplished using the vertical line "|" on your keyboard (look above the backslash character). In the Research Question about the relationship between **fastest** and **sex**, we would invoke:

```
histogram(~fastest|sex,data=m111survey,
       type="density",
       main="Fastest Speed Driven, by Sex",
       xlab="Fastest Speed, in mph")
```

The results appear in Figure [Speed by Sex], which shows that the guys tend to drive faster: the male histogram is shifted somewhat toward higher speeds, as compared to the female histogram.

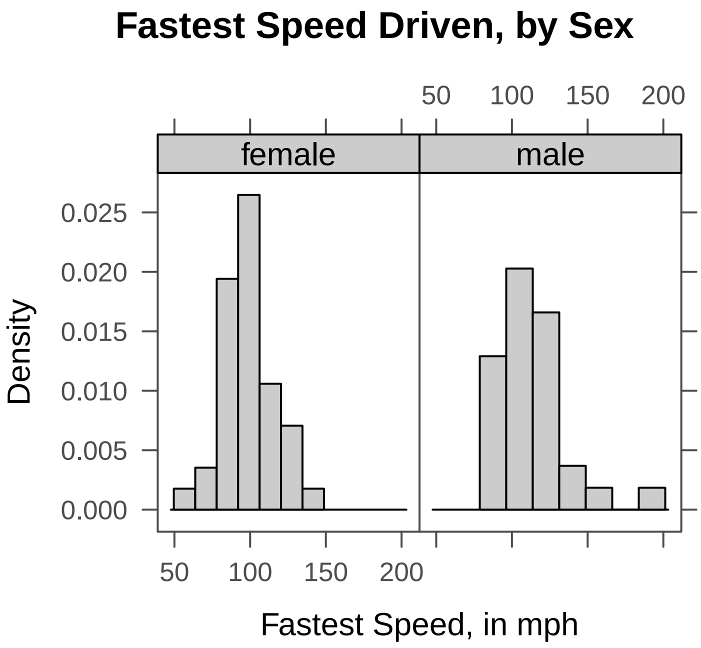


Figure 2.17: Speed by Sex. Histogram for female and male speeds appear in separate panels.

2.6.2.3 Verical Layout

It's not always easy to compare histograms side-by-side. You may prefer to use the arrange the histograms vertically using the `layout` option:

```
histogram(~fastest|sex,data=m111survey,
       type="density",
       main="Fastest Speed Driven, by Sex",
       xlab="Fastest Speed, in mph",
       layout=c(1,2))
```

In the graph [Speed by Sex (2)], the variable **sex** has two values, so you want the vertical layout to have two rows. The option `layout=c(1,2)` in the code for the graph specifies one column and two rows.

2.6.2.4 Grouped Density Plots

Overlaying two density plots – one for the guys and one for the gals – can make the difference between the two distributions quite clear. Overlaying can be accomplished using the *groups* argument:

```
densityplot(~fastest,data=m111survey,
       groups=sex,
       main="Fastest Speed Driven, by Sex",
       xlab="speed (mph)",
       auto.key=TRUE)
```

Figure Grouped Density Plots shows the results. Indeed, the guys drive faster: their density plot is shifted a bit to the right, in comparison to that of the gals.

If you would rather see the plots in different panles, you can do so in the same way you did with histrograms. Again, you might want to include the `layout` argument. Figure [fastestsexdensity2] shows the results

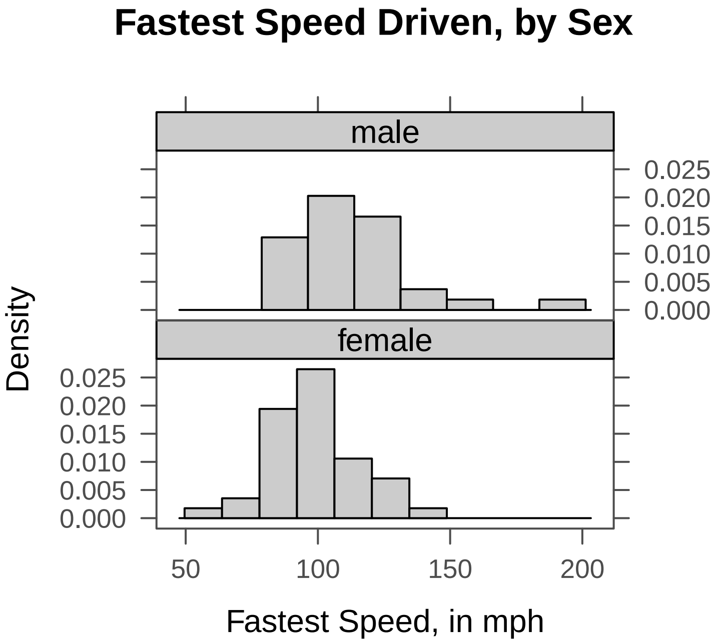


Figure 2.18: Speed by Sex (2). Histogram for female and male speeds appear in separate panels, laid out in one column.

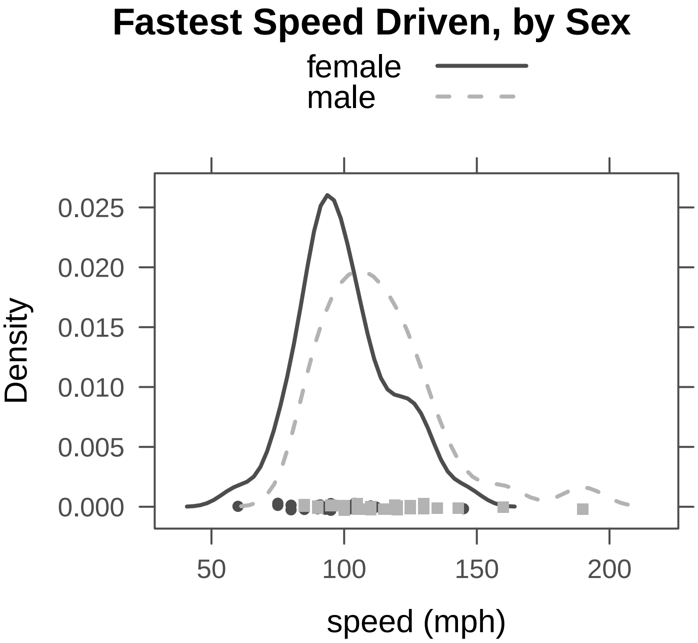


Figure 2.19: Grouped Density Plots. It is very easy to see that the mode for the males is greater than the mode for the females.

```
densityplot(~fastest|sex,data=m111survey,
        main="Fastest Speed Driven, by Sex",
        xlab="speed (mph)",
        auto.key=TRUE,
        layout=c(1,2))
```

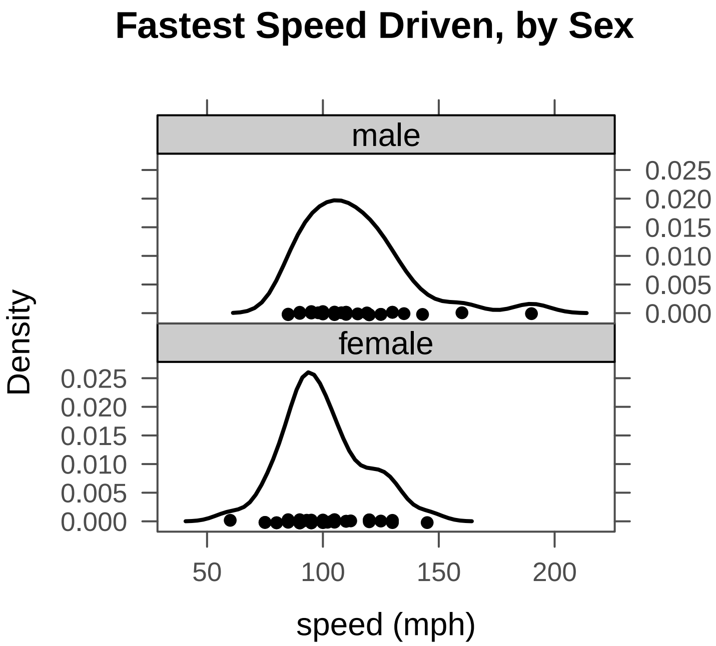


Figure 2.20: Density Plots arranged vertically. Again it is easy to see that the mode for the males is greater than the mode for the females.

2.7 Choice of Measures

2.7.1 Mean, Median and Skewness

If you are looking at a histogram or density plot of some numerical data, then:

- about half of the area of the plot will lie below the median (this is because about 50% of the data values are less than the median);
- the mean of the data will be approximately the place where you would put your finger, if you wanted the histogram or density plot to "balance"" if you held it up supported only by your finger.

Accordingly, when the distribution is symmetric the mean and the median will be about the same. But what happens if the distribution is skewed? In order to make the plot balance, the mean will have to be closer to the tail of the data than the median is. Figure [Symmetry and Skewness] illustrates what happens.

You can investigate these ideas with following app, too:

```
require(manipulate)
Skewer()
```

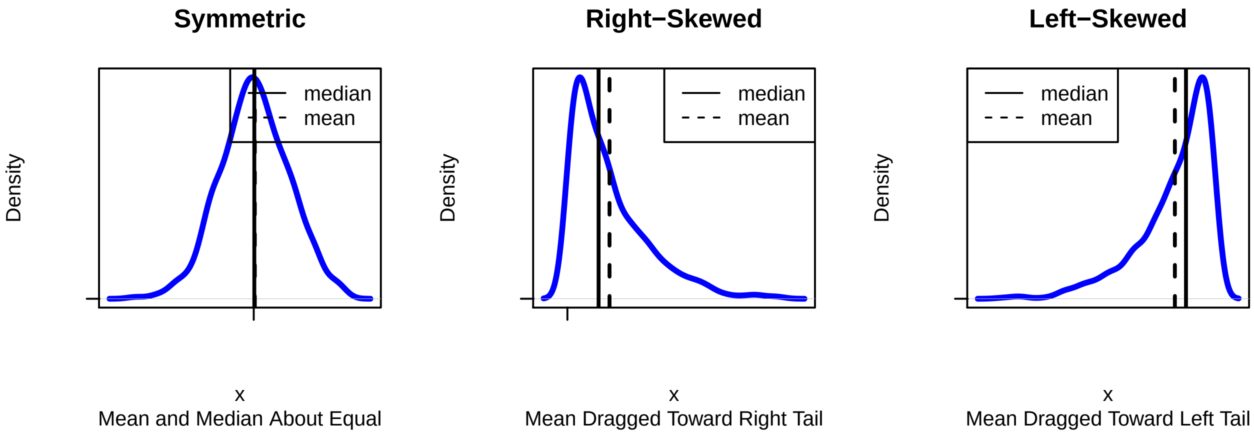


Figure 2.21: Symmetry and Skewness. The mean and the median are about the same when the distribution is symmetric. For right-skewed distributions the mean is bigger than the median, and for left-skewed distributions the reverse is true.

2.7.2 Mean, Median and Outliers

Not only is the mean "dragged" toward the tail of a skewed distribution – it is also dragged toward outliers. the median, on the other hand, is not much affected by outliers at all. For example, consider the small dataset:

```
SmallDataset
```

```
## [1] 32 45 47 47 49 56 56 56 57
```

SmallDataset has just nine values. A call to `favstats` gives:

```
favstats(~SmallDataset)
```

```
##  min Q1 median Q3 max     mean       sd n missing
##   32 47     49 56  57 49.44444 8.079466 9       0
```

The mean is about 49.4, and the median is the fifth value in the dataset: the 49. Now let's add just one extreme value to the data, say the number 200:

```
NewData <- c(SmallDataset,200)
NewData
```

```
##  [1]  32  45  47  47  49  56  56  56  57 200
```

Then call up `favstats` again:

```
favstats(~NewData)
```

```
##  min Q1 median Q3 max mean       sd  n missing
##   32 47   52.5 56 200 64.5 48.21537 10       0
```

The median is now the average of the fifth and sixth data values: $(49 + 56)/2 = 52.5$, so it has gone up just a bit. But the mean has increased markedly, from 49.4 to 64.5. In fact, the mean is now considerably larger than every data value except for the outlier at 200 – it no longer serves well to indicate what a "typical" data value might be.

Note also that the SD is affected by the outlier, increasing from about 8 (SmallDataset) to about 48 (NewData). No longer does it serve well as a measure of how "spread out" most of the data is. The IQR, on the other hand, is scarcely affected by the outlier.

2.7.3 Mean/SD vs. Median/IQR

We have two sets of measures for the center and the spread of a distribution:

- the mean (center) and the SD (spread);
- the median (center) and the IQR (spread).

In many circumstances either one of these pairs serves well to describe center and spread. However:

- when a distribution is STRONGLY skewed, or
- when it has SEVERE outliers in one direction but not the other,

the preferred measures of center and spread are the median and the IQR, rather than the mean and the SD. The graphs and calculations from the previous section back up this criterion.

2.8 Bell-Shaped Distributions

2.8.1 The 68-95 Rule

People like to use the mean as a measure of center and the SD as a measure of spread, because of the following rule of thumb:

The 68-95 Rule (also known as the *Empirical Rule*): If the distribution of sample data or of a population resembles a unimodal symmetric ("bell-shaped")" curve, then

- About 68% of the values lie within one SD of the mean.
- About 95% of the values lies within two SDs of the mean.
- About 99.7% of the values lie within three SDs of the mean.

The rule works surprisingly well, even when the data are somewhat skewed. The following manipulate app illustrates something of the scope and the limitations of the 68-95 Rule:

```r
require(manipulate)
EmpRule()
```

Here is an example of the use of the 68-95 Rule:

> A certain population consists of people whose heights have a roughly-bell-shaped distribution, with a mean of 70 inches and a standard deviation of 3 inches.

1. About what percentage of the people in the population are between 67 and 73 inches tall?

2. About what percentage are more than 73 inches tall?
3. About what percentage are less than 64 inches tall?

For the first question, we note than 67 and 73 are respectively one SD below and one SD above the mean of 70. Hence by the "68" part of the 68-95 Rule, **about 68% of the population should be between 67 and 73 inches tall**.

For the second question, note that 64 is two SDs below the mean of 70. By the "95" part of the 68-95 Rule, we know that about 95% of the population is between 64 and 76 inches tall (76 is two SD *above* the mean of 70). The remaining 5% lies outside of this range, and since bell-shaped distribution is symmetric, about half of this amount should lie below 64 and about half should lie above 76. Half of 5% is 2.5%, so **about 2.5% of the poulation is less than 64 inches tall**.

It is good to have graphs in mind when you think about the 68-95 Rule. The app `EmpRuleGC` provides a graphical way to use the 68-95 Rule. In order to solve the problems in the previous example, just try:

```
require(manipulate)
EmpRuleGC(mean=70,sd=3,xlab="height (inches)")
```

2.8.2 z-scores

When we use the 68-95 Rule, we think about how many SDs a number is away from the mean of the data. In general, even when we aren't thinking about the 68-95 Rule, we can measure how "unusual" a data value is by figuring out how many SDs it is above or below the mean of all of the data. If x is some value, then we compute:

$$z = \frac{x - \bar{x}}{s},$$

where:

- x is the actual value
- $\bar{x}$ is the mean of the data
- s is the standard deviation of the data.

z is called the *z-score* for x.

For example, suppose that Linda is 72 inches tall. How does she compare with the other GC students in the `m111survey` data? Is she unusually tall, unusually short, or rather typical? We can use her z-score to judge.

To find Linda's z-score, we need the mean and the SD of the heights in the `m111survey` data, so we call:

```
favstats(~height,data=m111survey)
```

```
##  min Q1 median    Q3 max     mean       sd  n missing
##   51 65     68 71.75  79 67.98662 5.296414 71       0
```

The mean is about 67.987 inches, and the SD is about 5.296 inches. Hence Linda's z-score is:

```
(72-67.987)/5.296
```

```
## [1] 0.7577417
```

The z score is about 0.76, which means that Linda is only about three-fourths of a standard deviation above the mean height. She is not that unusual. If she were a full standard deviation above the mean, then the 68-95 Rule would tell us that about 16% of the students in the `m111survey` data are taller than her (think about why this is so), but since she is not quite as tall as that we know that probably MORE than 16% of the students are taller than her. So Linda is taller than average, but not unusually tall.

Of course, we might wonder whether Linda is unusually tall, *for a female.* To find her z-score relative to females, we need the mean and the SD for females in `m111survey`, so we call:

```
favstats(height~sex,data=m111survey)
```

```
##   .group min Q1 median Q3 max     mean       sd  n missing
## 1 female  51 63     65 68  78 64.93750 4.621837 40       0
## 2   male  65 70     72 74  79 71.92097 3.048545 31       0
```

We find that the mean for the female sis about 64.938 inches, and the SD is about 4.622 inches. Linda's z-score relative to the females is:

```
(72-64.838)/4.622
```

```
## [1] 1.549546
```

So Linda is about 1.55 SDs above the mean female height. That's more impressive, but still not terribly unusual.

Let's adopt the following convention:

> *A value shall be considered* ***unusual*** *if its z-score is less than -2 or more than 2.*

This convention is most useful when the distribution is roughly bell-shaped, but it makes sense for other distributions, too, as long as they are not too strongly skewed, and don't have extreme outliers.

Here is another example, in which we use z-scores to compare individuals.

> **Example**. George comes from a school where the mean GPA is 3.4, with a SD of 0.3. Linda comes from a school where the mean GPA is 2.8, with a SD of 0.4. George's GPA is 3.6, and Linda's GPA is 3.3. Although the schools have different grading patterns, the students at both schools are believed to be equally strong, on the whole.
>
> 1. Compute George's z-score.
> 2. Compute Linda's z-score.
> 3. Based on the z-scores, who do you think is the stronger student?
> 4. Harold is from the same school as George. His z-score is -0.5. Is Harold above or below average? What is Harold's actual GPA?

For Question 1, we simply compute George's z-score as follows, using the mean and SD for Harold's school:

```
(3.6-3.4)/0.3
```

```
## [1] 0.6666667
```

For Question 2, we compute Linda's z-score, using the mean and SD for her school:

```
(3.3-2.8)/0.4
```

```
## [1] 1.25
```

For Question 3, we check to see who has the higher z-score. Linda is 1.25 SDs above the mean for her school, whereas George is only 0.67 SDs above the mean for his school. Since both schools are thought to be equally strong in terms of the academic profile of their students, we conclude that Linda is the more outstanding student.

For Question 4, we note that Harold's z-score is -0.5, putting him half of a standard deviation below average for his school. Since the mean is 3.4 and the SD is 0.3, Harold's actual GPA must be:

```
3.4-0.5*0.3
```

```
## [1] 3.25
```

So Harold's GPA is 3.25.

Here is one last example:

> **Example**. Belinda comes from a school where the mean GPA is 3.2, with a SD of 0.25. At this school you win a small scholarship if oyur GPA is higher than all but 97.5% of the other students. Approximately what is the minimum GPA that will secure Belinda a scholarship?

To answer this question, we recall that about 95% of the students will have GPAs within 2 SDs of the mean. Two SDs is $2 \times 0.25 = 0.5$, so 95% are between $3.2 - 0.5 = 2.7$ and $3.2 + 0.5 = 3.7$. Half of the remaining 5% (that's 2.5%) will be bigger than 3.7, so the rest (97.5%) will be BELOW 3.7. This is our answer: Belinda needs to secure a GPA of at least 3.7 in order to get the scholarship.

2.9 Reading in Statistics

Reading is a primary skill that is developed during your first year at Georgetown College (think about your FDN 111 class). IN FDN 111, you learn that the Read Skill has four components:

- Reading in context
- Reading for structure
- Reading to interpret
- Reading in a spirit of critical engagement

In this course, we not only read the course text, we also read data and we read tables and graphs. Even in this "quantitative" type of reading, all of the four components of the Read Skill come into play:

- We read data for **structure**: we note rows (individuals) and columns (variables). We look at the structure of the data, because the type of a variable determines how we explore and describe it.
- We read data to **interpret** it. Once we know the type of a variable, we know what descriptive techniques might help us to summarize and describe it.
- We read tables and graphs **for structure** and **to interpret** them. Tables and graphs have their own structures. Tables have rows, columns, cells, and sometimes marginal totals. Graphs have axes, scales on the axes, axis labels, titles, legends, etc. The parts work together to guide us to a good summary of the data.

- We read data **in context**. For example, it is important to recall the Help file on **mat111**: the Help file said that the survey was a survey of MAT 111 students at Georgetown College, and it told us what the variable names meant, what units they were measured in, etc. In more involved situations, such as data analysis in science, reading in context also means learning about the scientific problem that motivated the collection of the data.
- We read in a spirit of **critical engagement.** For example, when we learn that the students in the `m111survey` data are all from MAT 111, we might wonder whether this sample is very much like a random sample. If not, it might be trustworthy as a guide to how the GC population looks. Also, even when the sample is random we always wonder: "Are the patterns we see in the data also present in the population, or are they just due to chance?" As we proceed in the course, we'll learn how to answer this sort of question.

As we go along in the course, we will find that the Argue and Write skills are also very much in play when we practice statistics.

2.10 Thoughts on R

2.10.1 New R Functions

Know how to use these functions:

- `xtabs`
- `rowPerc`
- `barchartGC`
- `histogram`, `densityplot`, `bwplot`
- `quantile`
- `favstats`

2.10.2 Those Pesky Formulas!

The formula-data input format can be confusing at first. However, there are some patterns that will make life easier for you:

The symbol "~" appears somewhere in every formula. This is what alerts R to the presence of a formula.

When you deal with just one variable x, the format is:

```
goal(~x,data = MyData)
```

When you deal with the relationship between a numerical value y and a factor variable x, the format is:

```
goal(y~x,data = MyData)
```

In the formula above y is usually the response variable and x is the explanatory, but R doesn't know anything about that. R just puts the values of the first variable along the y axis, and the values of the second variable along the x-axis.

When you deal with the relationship between two factor variables, the format is:

```
goal(~x+y, data = MyData)
```

If there is an explanatory variable, we will try to remember to put it first, and the response variable second. (But R neither knows nor cares about explanatory vs. response.)

If your variables come from a data frame, don't forget to supply the name of the data frame, using the `data` argument!

Chapter 3

Two Factor Variables

3.1 Introduction

In Chapter Two we looked quickly at how to investigate the relationship between two factor variables. In the present chapter we will go into more depth on this topic.

First, a look at a Research Question from the `m111survey` data:

> *What is the relationship at Georgetown College between sex and how one feels about one's weight?*

This question, like most Research Questions, actually has two aspects:

- **Descriptive Aspect**: What is the relationship, *in the sample data*, between sex and how one feels about one's weight?
- **Inferential Aspect**: Supposing we see a relationship in the data, how much evidence does the data provide for a relationship *in the GC population at large* between sex and how one feels about one's weight? Does the data provide lots of evidence for a relationship in the population, or could the relationship we see in the data be due just to chance variation in the sampling process that yielded the data?

The previous chapter dealt with the descriptive aspect of Research Questions; in this chapter we will develop the descriptive aspect a bit more, and then turn to the inferential aspect.

3.2 The Descriptive Aspect

The current Research Question involves two factor variables from the `m111survey` data frame:

- **sex**. Its values are "male" and "female"
- **weight_feel**. Its values are:
  - 1_underweight
  - 2_about_right
  - 3_overweight

From Chapter Two know that in order to study a relationship between two factor variables we begin with a two-way table:

```r
SexWt <- xtabs(~sex+weight_feel,data=m111survey)
SexWt
```

```
##         weight_feel
## sex      1_underweight 2_about_right 3_overweight
##   female             1            11           28
##   male               8            14            9
```

We put **sex** first because in this study it is natural to consider it to be the explanatory variable: we think that one's sex might, through cultural conditioning, affect how one feels about one's weight.

3.2.1 Terminology for Two-Way Tables

The two-way table is called "two-way" because it has two dimensions: it has rows and columns.

- There are two rows, because the first variable **sex** has two values.
- There are three columns because the second variable **weight_feel** has three values.

Because there are two rows and three columns, the table has

$$2 \times 3 = 6$$

cells. In each cell, there is an *observed count*. For example, in the cell for males who feel underweight, the observed count is 8.

You can add up the observed counts in the rows to get *row totals*. For example, the row total for the first row is

$$1 + 11 + 28 = 40,$$

and this gives the total number of females in the study. If you add up the observed counts in the second row, you get 31, the total number of males in the study.

The sum of the row totals is called the *grand total*. It gives the total number of individuals in the study.

You can add up columns to get *column totals*, and if you add up column totals you will also get the grand total.

Below is the same two-way table, with an extra row and column for the totals:

```
##        1_underweight 2_about_right 3_overweight Total
## female             1            11           28    40
## male               8            14            9    31
## Total              9            25           37    71
```

The numbers in the `Total` column on the right of the table, excluding the Grand Total of 71, should be familiar: they are the tallies for the **sex** variable that we studied in chapter 2. Together they describe the distribution of **sex**. Because they occur in a margin of the two-way table, they are called the *marginal* distribution of **sex**. Similarly, the totals along the bottom give the marginal distribution of **weight_feel**.

However, when we are studying the relationship between **sex** and **weight_feel**, it doesn't help much to know the marginal distributions. For example, the marginal distribution of **sex** tells us that a majority (40 out of 71) of the people in the study are female, and the marginal distribution of **weight_feel** tells us that a majority of the students in the study (37 out of 71) felt overweight. But these two facts don't say anything about whether males and females *differ* in how they feel about their weight: they say nothing about the relationship between **sex** and **weight_feel**.

In order to address the relationship question, we have to find the distribution of **weight_feel** among the females, and the distribution of **weight_feel** among the males.

These two important distributions have special names:

- the distribution of **weight_feel** among the females is called the *conditional distribution* of **weight_feel**, given that **sex** is female.
- the distribution of **weight_feel** among the males is called the *conditional distribution* of **weight_feel**, given that **sex** is male.

If these two conditional distributions differ, then we will know that **sex** and **weight_feel** are related, in this sample of students.

Since the two sexes occur on two different rows of the two-way table, we can get them by computing row percentages:

```
rowPerc(SexWt)
```

```
##         weight_feel
## sex      1_underweight 2_about_right 3_overweight  Total
##   female          2.50         27.50        70.00 100.00
##   male           25.81         45.16        29.03 100.00
```

Let's check a couple of the figures. Out of 40 females in the study, 28 thought they were overweight. Compute

```
28/40*100
```

```
## [1] 70
```

Sure enough, we get the 70% figure that we see in the table of row percents. The two-way table also tells us that 9 of the 31 males in the study thought they were overweight; as a percentage, that is:

```
9/31*100
```

```
## [1] 29.03226
```

Again, this agrees with the figure in the table of row percentages.

You can also get column percents (observed counts divided by column totals):

```
colPerc(SexWt)
```

```
##         weight_feel
## sex      1_underweight 2_about_right 3_overweight
##   female         11.11            44        75.68
##   male           88.89            56        24.32
##   Total         100.00           100       100.00
```

The column percents give you three conditional distributions:

- the conditional distribution of **sex** given that **Weight_feel** is "underweight" (the first column);
- the conditional distribution of **sex** given that **Weight_feel** is "about right" (the second column);
- the conditional distribution of **sex** given that **Weight_feel** is "overweight" (the third column).

> **Practice** If we want to know the percentage of all men who feel that they are underweight, are we looking for a row percentage or a column percentage? What is the percentage?

The relevant fact here is that out of all 31 men, 8 think that they are underweight. These two numbers occur along a single row, so we are looking for a row percentage. From the row-percent table we see that this is 25.81%.

> **Practice** If we want to know the percentage of men among all students who feel that they are overweight, are we looking for a row percentage or a column percentage? What is the percentage?

This time we focus on all students who feel they are overweight (37 total), and we see that 9 of them are men. This time the two numbers occur in a single column, so we want the column percentage, which is 24.32%.

> **Practice** Suppose we want the percentage of all people who are men and who feel that they are overweight: is this a row percentage, a column percentage, or neither?

There are 9 people who are men and who think that they are overweight. The grand total is 71 people, so the percentage of all people who are men and who feel they are overweight is

```
9/71*100
```

```
## [1] 12.67606
```

This is neither a row percentage nor a column percentage.

3.2.1.1 Barchart Reminder

Remember that you can use barcharts to investigate graphically the relationship between two factor variables. Flat barcharts can be especially useful (see Figure [Sex and feeling about weight]):

```
barchartGC(SexWt,
           type="percent",
           ylab="Sex",flat=TRUE)
```

3.2.2 Detecting and Describing Relationships

3.2.2.1 Detection

Let's focus back on the conditional distributions of **weight_feel**, given **sex**:

```
rowPerc(SexWt)
```

```
##         weight_feel
## sex       1_underweight 2_about_right 3_overweight  Total
##   female           2.50         27.50        70.00 100.00
##   male            25.81         45.16        29.03 100.00
```

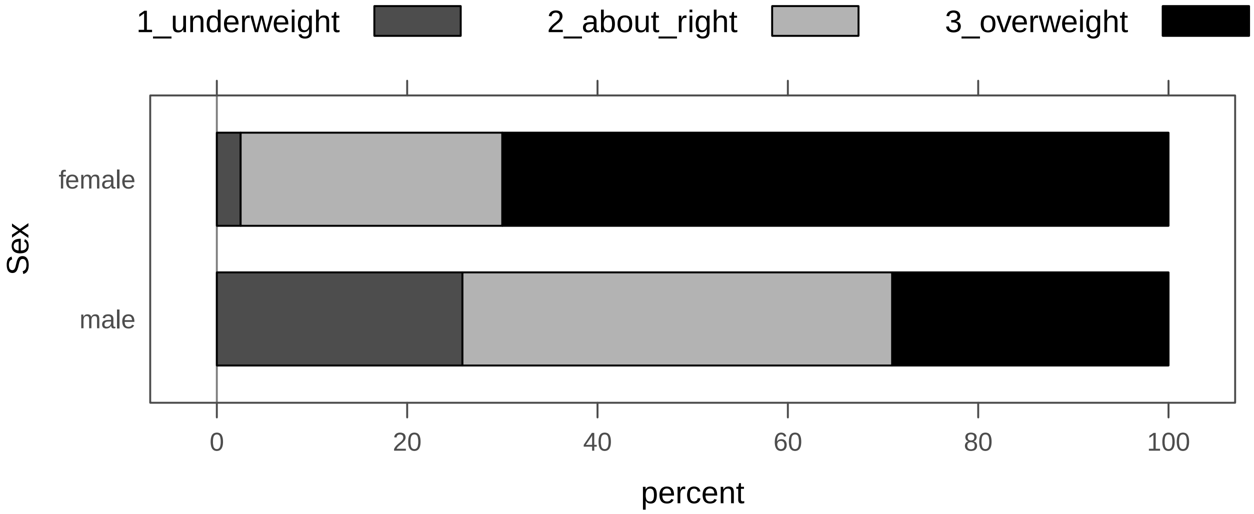


Figure 3.1: Sex and feeling about weight at GC

These two distributions obviously differ. For example, 70% of women feel overweight, but only 29.03% of the men feel overweight. This tells us that the women in the sample are more likely to feel overweight than the men in the sample are. We have just discovered that **sex** and **weight_feel** are related, in this sample!

This suggest a general procedure for detecting a relationship between two factor variables:

1. Make a two-way table. If one of the variables is clearly the explanatory variable, put it along the rows.
2. Make the table of row percents.
3. Compare row percents down columns. If you find a column where the row percents differ substantially, this indicates a relationship between the variables. The bigger the difference, the *stronger* the relationship.
4. If, in every column, the row percents are about the same, then there is little or no relationship between the two variables.

If you want, you could compute column percentages, and follow a similar procedure for them:

- When two variables are unrelated in a sample or in a population, then for every row in a two-way table, the column percentages do not change as you go across the row.
- If there is at least one row where the column percentages are not all the same, then there is some relationship between the two variables. The bigger the differences, the stronger the relationship.

When the explanatory variable is along the rows, as in our SexWt table, we usually look just at row percentages.

3.2.2.2 Description

How do we *describe* the relationship that we have detected? Here is one way:

> "Sex and feeling about weight are related, in our sample data. For example, the males were more likely than the females to think that they were underweight (25.81% as compared to 2.50% for the females)."

The key is to communicate specifically the features of the data that allowed you to detect the relationship. This helps convince your reader that you are right. Be sure to use specific numbers from the table to back up your assertions.

Here is a check-list for detecting and describing a relationship between two factor variables.

- Compare row percents down columns.
- The bigger the differences down a column, the stronger the relationship.
- Your description should incorporate at least two row percents from the same column.
- Focus on columns with an important-looking difference. This could be a big difference in percentages, or where the percentages may not differ so much but are based on high cell counts.
- Don't incorporate too many percents in your description. The reader can always look back at the table for more info.

3.2.2.3 Warning

Consider the following made-up data. Suppose that George plants 100 seeds in plot A, and 200 seeds in plot B. One week later, he finds that 70 in plot A have sprouted, and that 140 in plot B have sprouted. He makes a two-way table:

```
##        sprouted not.sprouted
## Plot.A       70           30
## Plot.B      140           60
```

George wants to see if there is a relationship, in the data, between type of plot and whether or not a seed sprouts in the first week. He computes row percentages

```
##        sprouted not.sprouted Total
## Plot.A       70           30   100
## Plot.B       70           30   100
```

George describes the relationship as follows: "There is a strong relationship between type of plot and sprouting: in both plots, a considerable majority of seeds sprouted within one week (70% sprouting vs. 30% not sprouting)."

George is quite mistaken, however. In fact there is no relationship at all between type of plot and whether a seed sprouts: the conditional distributions of sprouting, given plot, are both identical! (The row percents are the same, as you go down any single column.) Intuitively, this tells you that plot-type *makes no difference* in whether or not a seed sprouts.

3.3 The Idea of Inference

So maybe we have examined a two-way table based on sample data, and have detected a relationship between the two factor variables under study. Great! We've found a pattern in our data. But:

- does that pattern in our data exist in the population at large, or
- could it be that there is no pattern in the population, and that the pattern in the data is the product solely of chance variation in the data-collection process?

A quick-and-dirty way to ask the question is as follows:

- Is the data-pattern *real*, or
- just due to *chance*?

The question is actually quite complex. For our first time through, it's best to work with a small dataset.

```
data(ledgejump)
View(ledgejump)
help(ledgejump)
```

As we learn from `help()`, this data frame is constructed from 21 recorded incidents in England, in which a suicidal person was contemplating jumping from a ledge or other high structure and a crowd gathered to watch. The weather at the time of the incident is recorded, along with the behavior of the crowd.

Our Research Question is:

> *Does the weather affect the way the crowd behaves?*

In other words, is there a relationship between **weather** and **crowd.behavior**?

First, we will detect and describe the relationship in the data. We make the two-way table:

```
WeBe <- xtabs(~weather+crowd.behavior,data=ledgejump)
WeBe
```

```
##        crowd.behavior
## weather baiting polite
##    cool       2      7
##    warm       8      4
```

Next we make row percents:

```
rowPerc(WeBe)
```

```
##        crowd.behavior
## weather baiting polite  Total
##    cool   22.22  77.78 100.00
##    warm   66.67  33.33 100.00
```

In the data there appears to be a strong relationship between weather and crowd behavior: when the weather was warm, the crowd was far more likely to bait the would-be jumper than when the weather was cool (66.67% baiting in warm weather vs. 22.22% baiting in cool weather).

But is this data-pattern real, or just due to chance? After all, there are many other factors besides weather that can influence a crowd's behavior, for example:

- the size of the crowd (the larger the crowd, the more likely it is to be unruly);
- how long the crowd has to wait for the incident to be resolved;
- the personality/behavior of the would-be jumper, etc.

We model these other factors as "chance." If we could go back into time and watch the same 21 incidents, they would play out differently because of the other chance factors. We would probably not get exactly the same two-way table.

A natural question arises: if there is no relationship between weather and crowd behavior, what two-way table would we EXPECT to see? In other words: what cell counts would we expect to see?

We can estimate these expected cell counts. Here's the chain of reasoning that will produce our estimates:

There was a grand total of 21 incidents. Of these 21 incidents, the crowd baited 10 times. That's

$$10/21 * 100 = 47.6$$

percent. The crowd was polite 11 times, or

$$11/21 * 100 = 52.4$$

percent of the time. So if there is no relationship, our best guess is that the crowd baits 47.6% of the time and is polite 52.4% of the time. So if we could run the study all over again and there is no relationship, then our best guess is that out of the 9 warm-weather incidents, the crowd would bait in 47.6% of them: that's

$$\frac{10}{21} \times 9 = 4.29$$

times. Hence 4.29 is our estimate of the *expected cell count* for the warm-baiting cell of the two-way table.

Also, if we could run the study all over again and there is no relationship, then our best guess is that out of the 9 warm-weather incidents, the crowd would be polite in 52.4% of them: that's

$$\frac{11}{21} \times 9 = 4.71$$

times. Hence 4.71 will be our estimate of the *expected cell count* for the warm-polite cell of the two-way table.

There is a pattern in the above reasoning: to estimate an expected cell count if there is no relationship, compute

$$\frac{colSum}{GrandTotal} \times rowSum$$

for that cell.

Using this formula, we can compute expected cell counts for the remaining cells. For the cool-baiting cell, the column sum is $8 + 2 = 10$ and the row sum is $8 + 4 = 12$. The grand total is 21, so the expected cell count is:

```
10/21*12
```

```
## [1] 5.714286
```

which is about 5.71, if we round to two decimal places. For the cool-polite cell, we get:

```
11/21*12
```

```
## [1] 6.285714
```

which is about 6.29. The table of expected cell counts is:

```
##      baiting polite
## cool    4.29   4.71
## warm    5.71   6.29
```

That's the table you would expect to see, if there is no relationship between weather and crowd behavior.

Well, not exactly, of course:

- There is no such thing as 4.29 crowds!
- Also, when chance is involved, you don't expect to see *exactly* what you expect!

The idea is that someone who believes that there is no relationship between weather and crowd behavior would expect to see about 4.29 in the warm-baiting cell, *give or take a bit for chance error* in the data-collection process.

There certainly are some differences between what one would expect, and what we actually saw. The following table shows what you get if you subtract the expected cell count from the actual cell count, in each cell:

```
##        crowd.behavior
## weather baiting polite
##    cool   -2.29   2.29
##    warm    2.29  -2.29
```

There are four cells in our table, so there are four differences. We would like to find *one number* that provides an overall measure of the difference between the observed table and the expected one.

One well-known measure is called the *chi-square statistic*. To find it:

- Square the differences between observed and expected counts
- Divide each squared difference by the expected count
- Add them all up

Let's try it out on our table:

```
(2-4.29)^2/4.29+(7-4.71)^2/4.71+(8-5.71)^2/5.71+(4-6.29)^2/6.29
```

```
## [1] 4.087924
```

We get about 4.09. (The actual value is closer to 4.07: our-round-offs along the way led to some error.)

Here are some facts about the chi-square statistic:

- It is always at least 0.
- The only time it is 0 is when you observe *exactly* what one would expect if there is no relationship.
- The bigger the statistic, the bigger the difference between observed and expected.

Now, back to our big question: "Is the data-pattern real, or is it just due to chance?" We can rephrase the question as follows: "We got a chi-square of 4.07. How likely is to get 4.07 or more, just by chance?"

Well, we find probabilities by repeating an experiment many times. The following app allows us to explore what would happen if we could repeat the study several times on the same 21 incidents, under the assumption that there is no relationship between weather and crowd behavior.

```
require(manipulate)
ChisqSimSlow(~weather+crowd.behavior,
             data=ledgejump, effects="fixed")
```

We find that if there is no relationship between **weather** and **crowd.behavior**, then there is a bit more than a 5% chance of getting a chi-square statistic of 4.07 or more, as we did in the actual study.

This probability–the chance of getting a chi-square statistic at least as big as the one we actually got, if there is no relationship–is called a *P-value.* The smaller the P-value, the more evidence the data provides *against* the idea that there is no relationship. The smaller the P-value, the less reasonable it is for someone to claim that there is no relationship, that the results are due just to chance variation.

How small should the P-value be, in order to rule out the claim of no relationship? There is no one natural "cut-off", but as a convention we say that it is 0.05 (or 5%).

This time, the P-value appears to be a bit more than 5%. We might be suspicious that weather and crowd behavior are related, but with the amount of data on hand we cannot quite rule out the idea that our results are due solely to chance variation. (Maybe we should study more ledge-jumping incidents!)

3.4 Tests of Significance

3.4.1 Five Step Procedure

The chain of reasoning we followed in the ledge-jump study is so common in inferential statistics that it has a name: it is called a *test of signifiance*, or a *test of hypothesis*. It will be repeated so many times that we should learn to break it down into steps, and to assign special names to important components of the argument.

Step One. State Null and Alternative Hypotheses.

The Null Hypothesis always involves the claim that there is no particular pattern in the population, or in the process that resulted in the data under study. When we are investigating the relationship between two variables, it claims "no relationship."

H_0 : There is no relationship between weather and crowd behavior.

This is not a claim that there is no relationship in the sample data. Rather, it is a claim that weather and crowd behavior are independent, and that the pattern we got in the data was just chance variation.

The Alternative Hypothesis always contradicts the Null Hypothesis, in one way or another. In this topic, the Alternative claims:

H_a : There is a relationship between weather and crowd behavior.

Again, this is NOT a claim that weather and crowd behavior are related in the sample data: everyone agrees that there is a relationship in the sample. Rather, the Alternative is claiming that the relationship in the sample is so strong that it exists because of a *real* relationship between weather and crowd behavior, not because of chance variation in the 21 incidents under study.

Step Two. Compute a test statistic.

The test statistic is a number that depends on the data. The formula for the test statistic is chosen so that the bigger it is, the more the data differ from what the Null Hypothesis expects to see. This time, the test statistic is the chi-square statistic, and its value is about 4.07.

Step Three. Compute the P-value.

By definition, the P-value is always:

> the probability of getting a test statistic at least as extreme as the one you got, if H_0 were true.

We can approximate the P-value by simulating the study many times, under conditions where H_0 is true. Statisticians often have short-cut ways to calculate P-values, and you will learn about them as we go along.

This time, the P-value was a little over 5%, we think.

It is a very good idea to pause at this point and to construct a *practical intepretation* of the P-value that we have obtained. This interpretation is basically a restatement of the definition of the term "P-value", *with the specific context of the problem inserted in place of the abstract terminology.* Here is a fine practical interpretation of the P-value in this study:

> If there is no relationship between weather and crowd behavior, then there is a bit more than a 5% chance of getting a test statistic at least as big as the 4.07 value that we got in the actual study.

Notice that we used the numerical value of the P-value in our interpretation. This tells the audience just how likely or unlikely it is to get results of the sort we got, if the Null is true.

Step Four. Decide whether or not to reject H_0.

We always make this decision on the basis of our P-value, and on the "cut-off" value, which has been set at 0.05.

- If P < 0.05, we reject H_0, and we say that our results are *statistically significant*.
- If P ≥ 0.05, we do not reject H_0, and we do not say that our results are statistically significant.

This time, we don't reject H_0. (But we are suspicious anyway that H_0 might be wrong.)

Notice how important it is to be able to interpret your P-value. It's the interpretation of the P-value that really makes the case for whether we should rule out the the Null Hypothesis as unreasonable.

Step Five: Report a conclusion.

We always write a brief conclusion, stated in the context of the problem. This means that we don't use a lot of technical terminology such as "P-value" or "chi-square statistic": we just state our conclusion in such a way that anyone who can understand the Research Question can understand our conclusion.

In this example, we might conclude:

> The data do not quite provide strong evidence for a relationship between weather and crowd behavior.

As a rule, the conclusion should say how strong the evidence is against H_0. Equivalently, it could say how strong the evidence is for H_a (which is how we chose to state it just now).

Note: Interestingly, the tests of significance that we perform in this book often cannot provide positive evidence *for* a Null hypothesis. When you decide not to reject H_0, you should avoid wording your conclusion as "the data provided so-and-so much evidence for" the Null.

For additional practice, let's consider another example. This time our Research Question is:

> In the student population at Georgetown College, is there a relationship between one's sex and where one prefers to sit in a classroom?

To start our investigation, we perform descriptive statistics, so we can figure out if there is a relationship in the sample data.

Make the two-way table:

```r
SexSeat <- xtabs(~sex+seat,data=m111survey)
SexSeat
```

```
##         seat
## sex       1_front 2_middle 3_back
##   female      19       16      5
##   male         8       16      7
```

Compute row percents:

```r
rowPerc(SexSeat)
```

```
##         seat
## sex       1_front 2_middle 3_back  Total
##   female   47.50    40.00  12.50 100.00
##   male     25.81    51.61  22.58 100.00
```

As we saw in Chapter 2, it appears that in the sample the females are more likely than the males to prefer the front (47.5% vs. 25.81% for the guys).

Now we go for the inferential statistics: we will perform a test of significance.

Step One: The hypotheses are:

H_0: There is no relationship, in the GC population, between sex and seating preference.

H_a: There is a relationship, in the population, between sex and seating preference.

Step Two: Compute the test statistic.

Step Three: Compute the P-value.

For these two steps, let's use the app again:

```r
ChisqSimSlow(~sex+seat,
             data=m111survey,effects="random")
```

Note: When your data is a random sample from some larger population, we recommend that you set the *effects* argument to "random." (See GeekNotes for an explanation.)

The chi-square statistic was 3.73. After some work, we approximated the P-value as somewhere around 16%. This means:

> If **sex** and **seat** are unrelated in the GC population, then there is about a 16% chance of getting a test statistic of 3.73 or more, as we got in our study.

Step Four Make a decision about H_0.

Since P = $0.16 > 0.05$, we do not reject H_0.

Step Five: Write a conclusion.

The sample data did not provide strong evidence for a relationship between sex and seating preference in the Georgetown College population.

3.4.2 More on the Chi-Square Statistic

The *degrees of freedom* for a two-way table (*df* for short) is an important quantity in statistical theory. It is defined as:

$$df = (NumbRows - 1) \times (NumbCols - 1).$$

For the Sex and Seat table, the *df* is $(2-1) \times (3-1) = 2$.

Statistical theory says that, if H_0 is true and we have taken a fairly large sample, then in repeated re-sampling the chi-square statistic should be, on average, equal to the df. Also, the standard deviation of the re-sampled chi-square statistic should be about $\sqrt{2 \times df}$. In this case, the SD should be about $\sqrt{4} = 2$.

You can use the mean and the SD as an advance indicator to whether or not H_0 looks reasonable. If the test statistic is many SDs above the df, then things look pretty bad for the Null.

3.4.3 chisqtestGC()

Statisticians can often derive short-cut ways to compute P-values, and these methods are written up into R-functions. May we suggest the use of the function `chisqtestGC()`.

```
chisqtestGC(~sex+seat,data=m111survey)
```

```
## Pearson's Chi-squared test
##
## Observed Counts:
##         seat
## sex      1_front 2_middle 3_back
##   female      19       16      5
##   male         8       16      7
##
## Counts Expected by Null:
##         seat
## sex      1_front 2_middle 3_back
##   female   15.21    18.03   6.76
##   male     11.79    13.97   5.24
##
## Contributions to the chi-square statistic:
##         seat
## sex      1_front 2_middle 3_back
##   female    0.94     0.23   0.46
##   male      1.22     0.29   0.59
##
##
## Chi-Square Statistic = 3.734
## Degrees of Freedom of the table = 2
## P-Value = 0.1546
```

You get the test statistic and the P-value that you need to write up a test of significance. You get more information, too, such as a table of observed counts and a table of expected counts.

You should still write out all five steps of a test, even when you use a pre-made R-function. Here is how we would write up the test using such a function.

Step One: The hypotheses are:

H_0: There is no relationship, in the GC population, between sex and seating preference.

H_a: There is a relationship, in the population, between sex and seating preference.

Step Two: Compute the test statistic.

Just run the test:

```
chisqtestGC(~sex+seat,data=m111survey)
```

```
## Pearson's Chi-squared test
##
## Observed Counts:
##         seat
## sex       1_front 2_middle 3_back
##   female      19       16      5
##   male         8       16      7
##
## Counts Expected by Null:
##         seat
## sex       1_front 2_middle 3_back
##   female   15.21    18.03   6.76
##   male     11.79    13.97   5.24
##
## Contributions to the chi-square statistic:
##         seat
## sex       1_front 2_middle 3_back
##   female    0.94     0.23   0.46
##   male      1.22     0.29   0.59
##
##
## Chi-Square Statistic = 3.734
## Degrees of Freedom of the table = 2
## P-Value = 0.1546
```

The chi-square statistic is about 3.73. The degrees of freedom for this table is reported to be 2, so if the Null is right you would expect the chi square statistic to turn out to be around 2, give or take

$$\sqrt{2 \times df} = \sqrt{2 \times 2} = \sqrt{4} = 2$$

or so. The value of 3.73 that we got isn't even one SD above 2, so if the Null is right then our results aren't very surprising.

Step Three Report the P-value.

The test output showed the P-value as about 0.155. Sure enough, results like ours aren't terribly unlikely to occur, if the Null is right.

Step Four Make a decision about H_0.

Since P $= 0.155 > 0.05$, we do not reject H_0.

Step Five: Write a conclusion.

The sample data did not provide strong evidence for a relationship between sex and seating preference in the Georgetown College population.

Here is one more example. Learn about the General Social Survey:

```r
data(gss02)
View(gss02)
help(gss02)
```

The General Social Survey (GSS) is a nationwide poll that has been conducted since 1972 (semiannually since 1994). Most interviews are done face-to-face. The questions asked vary from year to year, and not every subject is asked the same set of question. The data frame `gss02` provides a selection of variables corresponding to questions asked of subjects in the year 2002.

Let say we are interested in the Research Question:

> *Is there any association between one's race and whether or not one owns a gun?*

This is a question about the relationship between two factor variables:

- **race**, with possible values:
  - African American
  - Hispanic
  - Other
  - White
- **owngun** (whether or not the subject owns a gun). The possible values are:
  - Yes
  - No

In this study we'll think of **race** as explanatory and **gunlaw** as response.

Here we go with the test of significance:

Step One: The hypotheses are:

H_0: There is no relationship, in the United States population, between race and gun ownership.

H_a: There is a relationship, in the U.S. population, between these two variables.

Step Two: Compute the test statistic.

We run the test:

```r
chisqtestGC(~race+owngun,data=gss02)
```

```
## Pearson's Chi-squared test
##
## Observed Counts:
##           owngun
## race         No Yes
##   AfrAm     106  16
##   Hispanic   20   3
##   Other      25   7
##   White     454 284
##
## Counts Expected by Null:
##           owngun
## race            No    Yes
##   AfrAm     80.67  41.33
##   Hispanic  15.21   7.79
##   Other     21.16  10.84
##   White    487.97 250.03
##
## Contributions to the chi-square statistic:
##           owngun
## race          No   Yes
##   AfrAm     7.96 15.53
##   Hispanic  1.51  2.95
##   Other     0.70  1.36
##   White     2.36  4.61
##
##
## Chi-Square Statistic = 36.9779
## Degrees of Freedom of the table = 3
## P-Value = 0
```

The chi-square statistic is about 36.98. The degrees of freedom for this table is reported to be 3, so if the Null is right you would expect the chi-square statistic to turn out to be around 3, give or take

$$\sqrt{2 \times df} = \sqrt{2 \times 3} = \sqrt{6} \sim 2.45$$

or so. The value of 36.98 that we got is many SDs above 3, so if the Null is right then our results are incredibly surprising.

Step Three Report the P-value.

The test output showed the P-value as about `4.651151e-08`. This is in scientific notation, and it equals

$$4.651151 \times 10^{-8} \sim 0.000000046,$$

or about 5 in 100 million. so if the null is right, there is only about a 5 in 100 million chance of getting a test statistic at least as big as the one we got in this study.

Step Four Make a decision about H_0.

Since $P = 4.65 \times 10^{-8} < 0.05$, we do reject H_0.

Step Five: Write a conclusion.

The sample data provided very strong evidence for a relationship between race and gun ownership in the United States population.

We might want to describe the relationship a bit more. In order to do this, we should look at some row percentages:

```
rowPerc(xtabs(~race+owngun,data=gss02))
```

```
##           owngun
## race          No    Yes  Total
##   AfrAm     86.89  13.11 100.00
##   Hispanic  86.96  13.04 100.00
##   Other     78.12  21.88 100.00
##   White     61.52  38.48 100.00
```

We see that white folks are more likely to own guns than people of other races are: for example, 38% of whites own a gun, whereas only 13% of African American own a gun.

This is not to say that one's race has a *causal* role in whether or not one owns a gun. We know that gun owners are predominantly rural, so if white people tend to be more rural than folks of other races are, that could make the difference. (In Chapter 6 we will look more deeply into questions about inferring causation from studies where an association is found between variables.)

3.4.4 Working With Summary Data

Sometimes you don't have the raw data available, but you still want to study the relationship between two factor variables. The descriptive and inferential procedures are the same as always, but you have to make your own table first.

Example: Suppose that in the ledge-jump study we had 42 incidents. In 18 of them the weather was cool and in the remaining 24 the weather was warm. In 4 of the 18 cool-weather incidents, the crowd was baiting. In 16 of the 24 warm-weather incidents, the crowd was baiting. We want to know whether weather and crowd behavior are related.

From the information we see:

- 4 cool-baiting incident
- 14 cool-polite incidents $(18 - 4 = 14)$
- 16 warm-baiting incidents
- 8 warm-polite incidents $(24 - 16 = 8)$

We want a table that looks like this:

```
##      baiting polite
## cool       4     14
## warm      16      8
```

To get this table into R, we proceed in steps. First, make lists of the cool and the warm counts:

```
cool <- c(4,14)
warm <- c(16,8)
```

Then "bind"" them together as two rows of a single table, using the `rbind()` function:

```
WeBe2 <- rbind(cool,warm)
WeBe2
```

```
##      [,1] [,2]
## cool    4   14
## warm   16    8
```

This is good, except that it does not have names for the columns, so we set them as follows:

```
colnames(WeBe2) <- c("baiting","polite")
WeBe2
```

```
##      baiting polite
## cool       4     14
## warm      16      8
```

That's better. To check for a relationship in the data, we compute row percents:

```
rowPerc(WeBe2)
```

```
##      baiting polite  Total
## cool   22.22  77.78 100.00
## warm   66.67  33.33 100.00
```

We see the same strong relationship as before. In this imaginary example, we just doubled all of the counts, so the percentages stay the same!

Now for the inferential statistics:

Step One: The hypotheses are:

H_0: There is no relationship between weather and crowd behavior.

H_a: There is a relationship, between weather and crowd behavior.

Step Two: Compute the test statistic. **Step Three**: Report the P-value.

Again we can combine these steps, by putting the table into `chisqtestGC()`:

```
chisqtestGC(WeBe2)
```

```
## Pearson's Chi-squared test with Yates' continuity correction
##
## Observed Counts:
##      baiting polite
## cool       4     14
## warm      16      8
##
## Counts Expected by Null:
##      baiting polite
## cool    8.57   9.43
## warm   11.43  12.57
##
## Contributions to the chi-square statistic:
##      baiting polite
## cool    2.44   2.22
## warm    1.83   1.66
##
##
## Chi-Square Statistic = 6.4611
## Degrees of Freedom of the table = 1
## P-Value = 0.011
```

The chi-square statistic is about 6.46, and the P-value is about 0.01.

Step Four: Make a decision about H_0.

Since $\mathrm{P} = 0.01 < 0.05$, we do reject H_0.

Step Five: Write a conclusion.

The sample data provided strong evidence for a relationship between weather and crowd behavior.

Note: The pattern in this imaginary study was exactly as strong as the pattern in the actual ledge-jump study, but it was based on twice as much data. Observe the effect on the P-value: it went down a lot from the original example! In general, the more data you have, the more *powerful* your test of significance will be, in the sense that it is more likely to reject H_0 when H_0 is false.

3.4.5 Simulating P-Values

Pre-made R-functions for tests are very convenient, but you have to be careful not to misuse them. When it comes to the **chisq.test**, statisticians say that you should only trust the P-value when the expected cell counts are all at least 5. (Some say that at least 80% of the expected cell counts should be at least 5, and that all of them should be at least 1.)

Consider what happens with the `ledgejump` data:

```
chisqtestGC(~weather+crowd.behavior,data=ledgejump)
```

```
## Pearson's Chi-squared test with Yates' continuity correction
##
## Observed Counts:
##        crowd.behavior
## weather baiting polite
##    cool       2      7
##    warm       8      4
##
## Counts Expected by Null:
##        crowd.behavior
## weather baiting polite
##    cool    4.29   4.71
##    warm    5.71   6.29
##
## Contributions to the chi-square statistic:
##        crowd.behavior
## weather baiting polite
##    cool    1.22   1.11
##    warm    0.91   0.83
##
##
## Chi-Square Statistic = 2.4858
## Degrees of Freedom of the table = 1
## P-Value = 0.1149
##
## Some expected cell counts are low:
##  the approximation of the P-value may be unreliable.
##  Consider using simulation.
```

The warning at the end was issued because a couple of the expected cell counts—the ones in the first row—were less than 5.

When you get a warning, it is best to back up and try simulation. You don't necessarily have to use the app `ChisqSimSlow()`, because R provides its own simulation routines, via the *simulate.p.value* argument. Another argument *B* specifies the number of re-samples to take.

```
chisqtestGC(WeBe,simulate.p.value=TRUE,B=2500)
```

```
## Pearson's chi-squared test with simulated p-value
##   (based on 2500 resamples)
##
## Observed Counts:
##        crowd.behavior
## weather baiting polite
##    cool       2      7
##    warm       8      4
##
## Counts Expected by Null:
##        crowd.behavior
## weather baiting polite
##    cool    4.29   4.71
##    warm    5.71   6.29
##
## Contributions to the chi-square statistic:
##        crowd.behavior
## weather baiting polite
##    cool    1.22   1.11
##    warm    0.91   0.83
##
##
## Chi-Square Statistic = 4.0727
## Degrees of Freedom of the table = 1
## P-Value = 0.0824
```

We recommend that you perform such “safety checks” during Step Two of any test of significance. `chisqtestGC()` comes in handy, here, because it shows the expected cell counts.

Note: The simulated P-value you got with **chisqtestGC()** was probably a bit higher than the one you got with `ChisqSimSlow()`. That’s because the two functions take slightly different approaches to re-sampling. (Interested persons should consult the GeekNotes for details.)

3.5 Simpson’s Paradox

Warning: What you are about to see appears to be impossible. But it really does happen.

Learn about the **deathpen** study:

```
data(deathpen)
View(deathpen)
help(deathpen)
```

The data frame is constructed from a study of 326 capital cases (cases where the defendant could receive the death penalty) in a court district in Florida, during the years 1976-1977. In all of the cases under study, the defendant had been convicted, and would either receive the death penalty or would not.

We are interested in the following Research Question:

Who is more likely to get the death penalty in capital cases: a black defendant or a white defendant?

First, some descriptive statistics to detect and describe relationships:

```
DefD <- xtabs(~defrace+death,data=deathpen)
DefD
```

```
##        death
## defrace  no yes
##   black 149  17
##   white 141  19
```

Of course to detect the relationship we need to consider row percents:

```
rowPerc(DefD)
```

```
##        death
## defrace     no    yes  Total
##   black  89.76  10.24 100.00
##   white  88.12  11.88 100.00
```

Surprisingly, it seems that white defendants are a little bit more likely to get the death penalty than black defendants are (11.88% compared to 10.24% for black defendants).

There is a third variable present in the study: **vicrace**, the race of the murdered victim. Let's break the data down into two groups, based on the two values of **vicrace**. This can be accomplished with R's `subset()` function:

```
deathpenWV <- subset(deathpen,vicrace=="white")
deathpenBV <- subset(deathpen,vicrace=="black")
```

Now let's study the relationship between **defrace** and **death** when the victim was white. First the two-way table:

```
DefDWV <- xtabs(~defrace+death,data=deathpenWV)
DefDWV
```

```
##        death
## defrace  no yes
##   black  52  11
##   white 132  19
```

Then row percents:

```
rowPerc(DefDWV)
```

```
##        death
## defrace     no    yes  Total
##   black  82.54  17.46 100.00
##   white  87.42  12.58 100.00
```

This is interesting: when the victim was white, a black murderer was considerably more likely to get the death penalty (17.46%, as compared to 12.58% for a white murderer).

Now let's study the relationship between **defrace** and **death** when the victim was black. Here's the two-way table:

```
DefDBV <- xtabs(~defrace+death,data=deathpenBV)
DefDBV
```

```
##        death
## defrace no yes
##   black 97   6
##   white  9   0
```

Now row percents:

```
rowPerc(DefDBV)
```

```
##        death
## defrace     no    yes  Total
##   black  94.17   5.83 100.00
##   white 100.00   0.00 100.00
```

When the victim was black, again a black murderer was more likely to get the death penalty (5.83% chance, as compared to 0% for a white murderer).

This puzzling situation is an example of *Simpson's Paradox.* Simpson's Paradox occurs when the direction of the relationship between two variables is one way when you look at the aggregate data, but turns out the opposite way when you break up the data into subgroups based on a third variable. This time, whites were more likely than blacks to get death in the aggregate data, but were less likely than blacks to get death in both of the subgroups.

Simpson's Paradox is mathematically possible—we just now saw an example—but it still seems unreal. Can we figure out WHY it has occurred, in this example?

In this example, our explanatory variable X is **defrace**, and the response variable Y is **death**. The third, lurking variable Z is **vicrace**. The key to understanding how Simpson's paradox occurs is as follows:

- Study the relationship between the explanatory X and the lurker Z;
- Study the relationship between Z and response Y.
- Synthesize the results of these two studies.

As for the relationship between **defrace** and **vicrace**, we have:

```
DefVic <- xtabs(~defrace+vicrace,data=deathpen)
DefVic
```

```
##        vicrace
## defrace black white
##   black   103    63
##   white     9   151
```

```
rowPerc(DefVic)
```

```
##        vicrace
## defrace  black  white  Total
##   black  62.05  37.95 100.00
##   white   5.62  94.38 100.00
```

We see that black folks are much more likely to kill black folks than white folks are.

As for the relationship between **vicrace** and **death**, we have:

```
VicD <- xtabs(~vicrace+death,data=deathpen)
VicD
```

```
##         death
## vicrace  no yes
##   black 106   6
##   white 184  30
```

```
rowPerc(VicD)
```

```
##         death
## vicrace      no    yes  Total
##   black  94.64   5.36 100.00
##   white  85.98  14.02 100.00
```

Interesting: when the victim is white, the defendant is about three times more likely to get the death penalty than when the victim is black (14.02% vs. 5.36%).

Synthesize the results and the mystery is solved: White defendants are indeed less likely to get the death penalty than black defendants are, both when the victim is white and when the victim is black, but the white defendants hamstring themselves by mostly killing white people. Killing a white person seems to have been a sure-fire way to incur the wrath of the Florida legal system, back in the day.

3.6 Thoughts on R

Some new R-function to learn:

- `colPerc()` (less commonly used than `rowPerc()`, if you plan to write the explanatory variable along rows)
- `chisqtestGC()`
- To make a two-way table from summary data:
  - `rbind()`
  - `rownames()`
  - `colnames()`
- `subset()` to make a subset of a data frame

Chapter 4

Two Numerical Variables

4.1 Outline

In the previous chapter, we investigated methods for describing relationships between two *factor* variables, using

- twoway tables and row percents for numerical desriptive statistics;
- barcharts for graphical descitpive statistics;
- the chi-square test, for inference.

In this chapter, we are interested in describing relationships between **two** *numerical* variables.

In chapter 2, we learned about methods to graphically and numerically summarize **one** numerical variable. The methods we used were:

Graphical

- Histogram
- Density Plot
- Stem Plot
- Box Plot (and the really cool Violin Plot)

Numerical

- Median
- Percentiles (quantiles)
- 5-number summary
- Mean
- SD
- IQR

We will now be learning to describe how *two numerical variables* relate to one another. Throughout this chapter, we're going to work with three different datasets: `m111survey`, `pennstate1`, and `ucdavis1`. The datasets `pennstate1` and `ucdavis1` are both surveys of students at their respective schools, very similar to what we've seen in `m111survey`. You can put these datasets into your Global Environment, take a quick look at them, and learn more about them with:

```
data(m111survey)
View(m111survey)
help(m111survey)

data(pennstate1)
View(pennstate1)
help(pennstate1)

data(ucdavis1)
View(ucdavis1)
help(ucdavis1)
```

4.2 Statistical Relationships

When we look at relationships between numerical variables, there are 2 main kinds of relationships that interest us.

- *Deterministic* relationships
- *Statistical* relationships

Deterministic relationships are the type you are used to seeing in algebra class. In this kind of relationship, the value of one variable can be *exactly* determined by the value of the other variable.

For example, consider the relationship between degrees Fahrenheit and degrees Celcius. If $y =°$ C and $x =°$ F, the deterministic relationship can be written

$$y = \frac{5}{9}(x - 32).$$

The degrees Celcius (y) is *exactly* determined by knowing the degrees Fahrenheit (x). The graph of this equation is a line. See Figure [Deterministic Relationship].

We can see from the graph that there is no variation in the pattern. Every temperature in Fahrenheit has exactly one corresponding temperature in Celcius. We might regard Fahrenheit and Celcius as having a perfect relationship.

Statistical relationships are the ones that we will study in this class. In this kind of relationship, there is variation from the average pattern. If we know the value of one variable, we can *estimate* the typical value of the other variable. However, this is only an estimation. There is no certainty!

For example, suppose we want to use the length of a person's right handspan (the measurement of one's outstretched right hand from thumb to pinky) to predict the person's height. There is certainly a *relationship* between the length of one's handspan and their height. People with large hands tend to be taller than people with small hands. However, there is not an equation (or a line) that will *exactly* tell us the height of a person with a certain handspan. Not every person with a handspan of 20 centimeters is exactly the same height. There is variation in their heights. What we've just described is a statistical relationship.

There are 3 tools that we will use to describe *statistical* relationships:

- Scatterplots
- Correlation
- Regression Equation

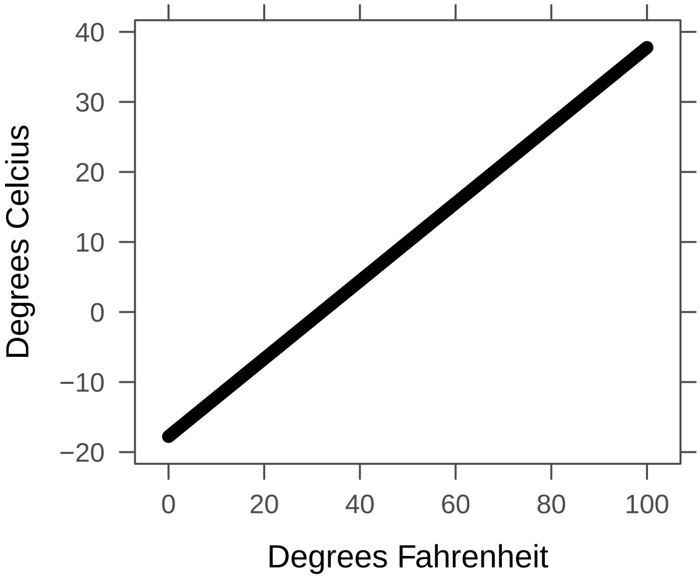


Figure 4.1: Deterministic Relationship. This graph shows the relationship between Fahrenheit and Celcius.

4.2.1 Scatterplots

As we just discussed, our intuition tells us that a person with a large right handspan tends to be tall. Let's investigate this idea with the following research question.

> **Research Question**: At Pennstate, how is a student's right handspan related to his/her height?

A *scatterplot* is how we will graphically display the relationship between two numerical variables. Scatterplots allow us to visually identify

- overall patterns,
- directions, and
- strength of association

between two numerical variables. The best way to get a feel for the relationship between two numerical variables is to take a look at the scatterplot.

4.2.1.1 Overall Patterns

Let's use the function `xyplot()` in `R` to create a scatterplot of the variables **RtSpan** and **Height** from the `pennstate1` dataset.

```
xyplot(Height~RtSpan,data=pennstate1,
       xlab="Right Handspan (cm)", ylab="Height (in)")
```

Each point, (x, y), that you see on the scatterplot represents an individual in the dataset - one of the 190 Penn State students in the survey. The x-coordinate of the point is that student's right handspan, in centimeters. The y-coordinate of the point is that student's height, in inches.

The 'formula-data' input syntax for `xyplot()` should be starting to become familiar to you. This is the same syntax that we used to produce the graphical outputs from chapter 2. `R` will plot the variable in front of

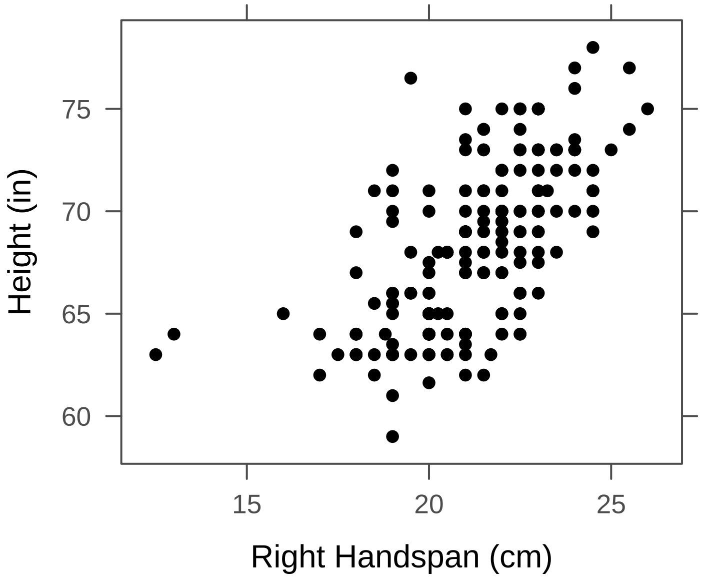


Figure 4.2: Hand/Height Scatterplot. Relationship Between Right Handspan and Height

the ~ along the vertical axis and the variable behind the ~ along the horizontal axis. Typically, we put the explanatory variable along the horizontal axis and the response variable along the vertical axis.

You can control how the points in the scatterplot appear using `pch` and `col` in the `xyplot()` function. For example, we can make the scatterplot have solid red points. See Figure[Red Points].

```
xyplot(Height~RtSpan,data=pennstate1,
       xlab="Right Handspan (cm)", ylab="Height (in)",
       col="red",pch=19)
```

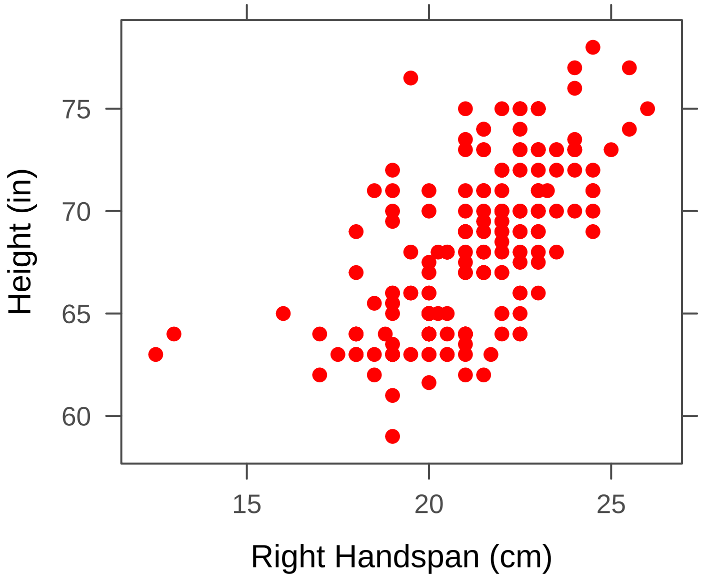


Figure 4.3: Red Points: Relationship Between Right Handspan and Height using solid red points

Note: If you are interested in making fancier scatterplots, there are different values of `pch` that produce various shapes for the points in the scatterplot. To learn more, consult GeekNotes. R will also provide a complete list of all available colors with:

```
colors()
```

When we look at the overall pattern in the scatterplot in Figure[Hand/Height Scatterplot], it appears that students with large right handspans also tend to be tall. One question that may arise is whether this observed relationship is simply the result of the student's sex. On average, males tend to have larger hands than females and also tend to be taller than females. As we did in chapter 2, we can look at parallel scatterplots in separate panels by "conditioning" on a category, such as **Sex**. See Figure[Parallel Hand/Height by Sex].

```
xyplot(Height~RtSpan|Sex,data=pennstate1,
       xlab="Right Handspan (cm)", ylab="Height (in)",pch=19)
```

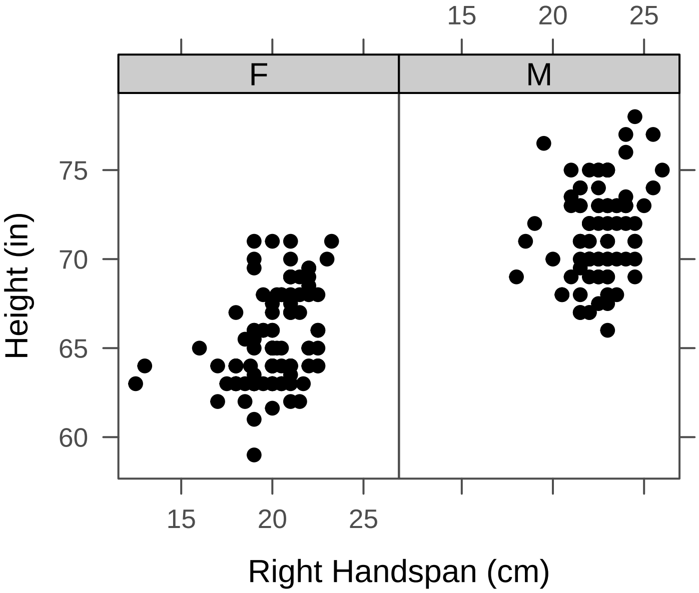


Figure 4.4: Parallel Hand/Height by Sex: Scatterplots showing the relationship between one's right handspan and height appear in separate panels.

We can see that the relationship we observed in the original scatterplot seems to hold separately for both males and females.

Parallel scatterplots can sometimes be hard to compare since they have separate x-axes. We can "overlay" these scatterplots, using one color for the points representing the males and another color for the points representing the females. Overlaying can be accomplished using the *groups* argument. See Figure[Overlayed Hand/Height by Sex].

```
xyplot(Height~RtSpan,groups=Sex,data=pennstate1,
       xlab="Right Handspan (cm)", ylab="Height (in)",
       pch=19,auto.key=TRUE)
```

Note: Given several numerical variables, R can produce a group of scatterplots, one for each pair of variables—all in one graph. Such a graph is called a *scatterplot matrix*. For more information, consult GeekNotes.

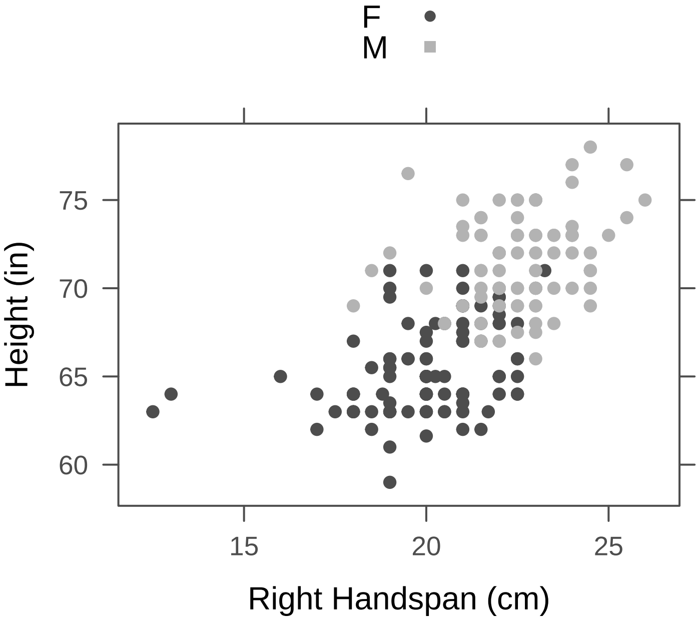


Figure 4.5: Overlayed Hand/Height by Sex: One scatterplot showing the relationship between right handspan and height colored by sex.

4.2.1.2 Direction

Viewing the scatterplot allows us to detect overall patterns. One way to describe the scatterplot is by giving a name to the *direction* of the observed pattern.

In the scatterplot Figure[Hand/Height Scatterplot], we could say that the variables `RtSpan` and `Height` are *positively associated* since students with larger handspans tend to be on the tall side and students with smaller handspans tend to be on the short side.

Positive Linear Association

: Two numerical variables have a *positive linear association* if high values of one variable tend to accompany high values of the other and low values of one variable tend to accompany low values of the other.

To give a visualization for positively associated variables, let's add a vertical line that marks the *mean* of the right handspans and a horizontal line that marks the *mean* of the heights to break the scatterplot into four "boxes". See Figure[Positive Association].

Notice that there are more points in the upper right box and lower left box than in the other two boxes. The upper right box includes points from individuals with higher than average right handspans **and** higher than average heights. The lower left box includes points from individuals with lower than average right handspans **and** lower than average heights. When most of the points in a scatterplot are located in these two boxes, the variables have a positive linear association.

Negative Linear Association Two variables have a *negative linear association* if high values of one variable tend to accompany low values of the other and low values of one variable tend to accompany high values of the other.

Let's take a look at an example of negatively associated variables from the `m111survey` dataset - `GPA` and `height`.

Notice that in this scatterplot, there are more points in the upper left box and the lower right box. (See Figure[Negative Association].) The upper left box includes points from individuals with lower than average

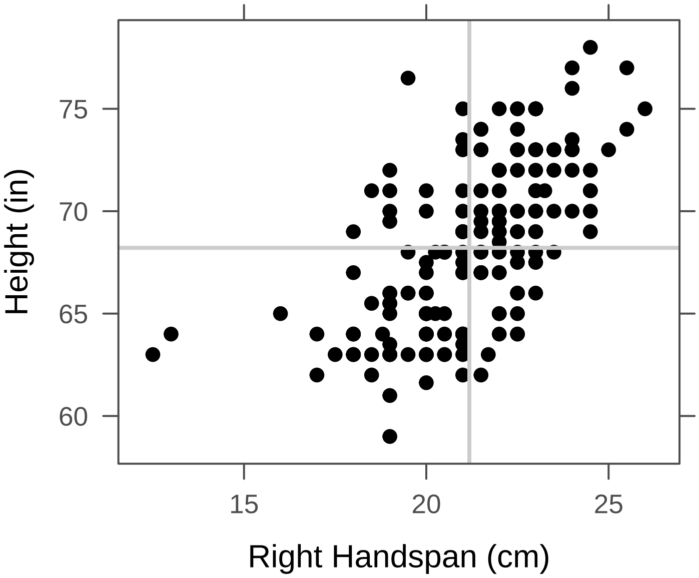


Figure 4.6: Positive Association: Most of the points are in the upper right and lower left box showing a positive association between the variables.

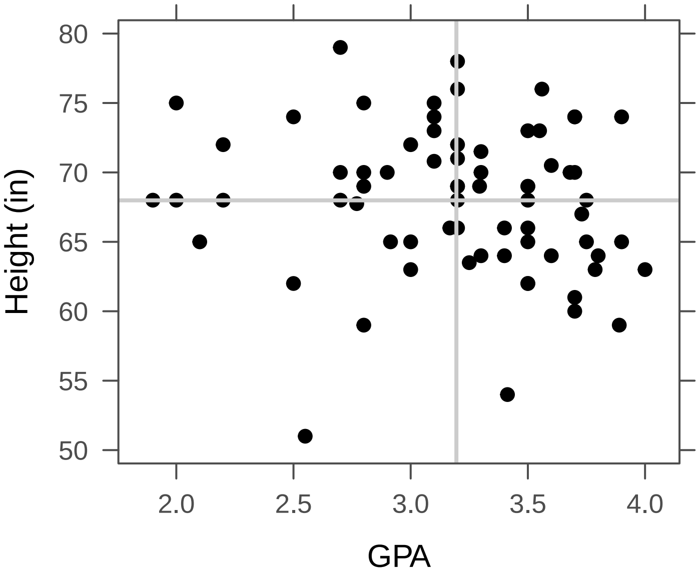


Figure 4.7: Negative Association: Most of the points are in the lower right and upper left box showing a negative association between the variables.

GPA's **and** higher than average heights. The lower right box includes points from individuals with higher than average GPA's **and** lower than average heights. When most of the points in a scatterplot are located in these two boxes, the variables have a *negative linear association.*

No Associaton Two variables have *no association* if there is no apparent relationship between the two variables.

Let's look at the *association* between an individual's height, `Height`, and the hours of sleep they got last night, `HrsSleep`, from the `pennstate1` dataset.

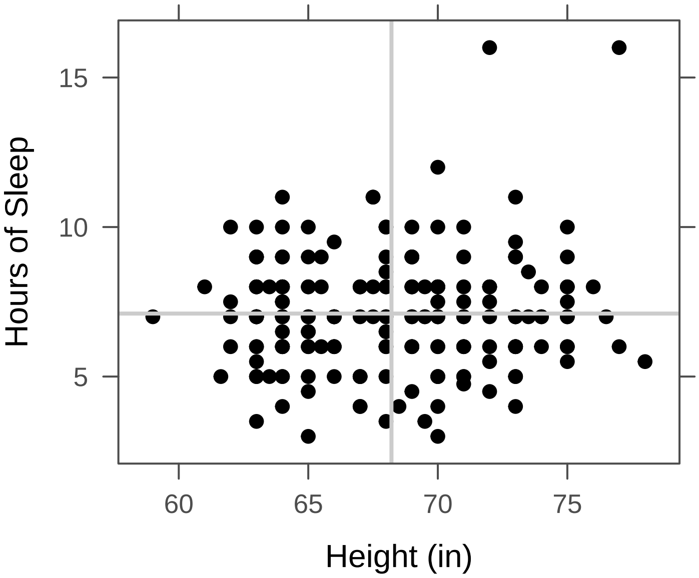


Figure 4.8: No Association: There is no apparent pattern in where the points lie.

In the scatterplot in Figure[No Association], it appears that all of the boxes have *about* the same number of points in them. When this is the case, the variables have *no linear association.*

So far we have been interested in variables that appear to have a *linear association*, and our goal has been to describe the linear association that we see in a scatterplot using the equation of straight line. Data with nonlinear association certainly exists! For example, *curvilinear* data follows the trend of a curve, rather than a line. You can see an example of this by looking at the `fuel` dataset. Read the `help` file on this data to understand what the variables are.

```
data(fuel)
View(fuel)
help(fuel)
```

The scatterplot clearly shows that this data would not be well represented by a line. (See Figure[Curvilinear].) As a vehicle's speed increases up to about 60 kph, the amount of fuel required to travel 100 kilometers decreases. However, as a vehicle's speed increases from about 60 kph, the fuel efficiency increases: this is one type of curvilinear relationship. We will offer a rudimentary discussion of nonlinear relationships later on in this Chapter.

For now, we'll stick with linear relationships.

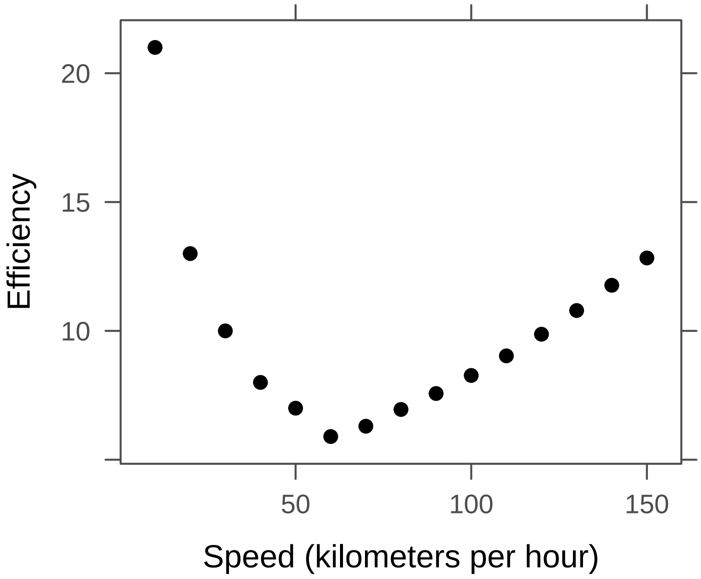
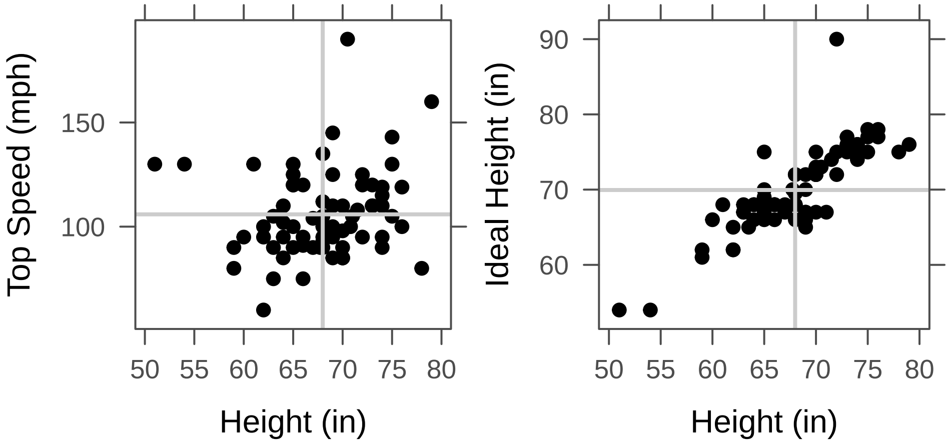


Figure 4.9: Curvilinear: Efficiency (liters of fuel required to travel 100 kilometers) versus Speed (kilometers per hour)

4.2.2 Correlation

4.2.2.1 Strength of Association

So far, we've investigated relationships between two numerical variables by looking at the scatterplot and making note of the observed pattern of the points, and their direction of association. However, as we can see in the following scatterplots, we should also consider the *strength* of association between two variables. See FigureStrength of Association.



These scatterplots both display 2 variables that are *positively* associated, but there is less "scatter" in the second plot. The variable **height** appears to have a stronger association with **ideal_ht** than it does with **fastest**. To distinguish between the amount of "scatter" in a plot, it is useful to assign a numerical value to the strength of association.

Correlation *Correlation* is the numerical measure of the direction and strength of the linear association between two numerical variables.

The formula for the *correlation coefficient*, r, is written:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where:

- n denotes the number of values in the list
- $\sum$ means summing
- x_i denotes the individual x values
- $\bar{x}$ denotes the average of the x's
- s_x denotes the SD of the x's
- y_i denotes the individual y values
- $\bar{y}$ denotes the average of the y's
- s_y denotes the SD of the y's

Since this calculation can be cumbersome, we will use R's built in function, `cor()`. The correlation coefficient for the variables **height** and **fastest** from the `m111survey` dataset plotted in the first scatterplot in FigureStrength of Association can be found by:

```
cor(fastest~height,data=m111survey,use="na.or.complete")
```

```
## [1] 0.1708742
```

Let's compare this to the correlation coefficient for the variables **height** and **ideal_ht** from the `m111survey` dataset plotted in the second scatterplot in FigureStrength of Association.

```
cor(ideal_ht~height,data=m111survey,use="na.or.complete")
```

```
## [1] 0.832047
```

What we've just observed holds true in general.

- *Positively associated* variables have a *positive* correlation coefficient, $r > 0$. Consult Geek Notes for a detailed explanation.

- The stronger the association is between variables, the larger the correlation coefficient, r, will be.

The same ideas hold true for *negatively* associated variables and variables with *no* association. Let's summarize all of the properties of the correlation coefficient, r.

Properties of r

- r always falls between 1 and -1.

- The sign of r indicates the *direction* of the relationship.

  - $r > 0$ indicates a *positive linear association*
  - $r < 0$ indicates a *negative linear association*

- The *magnitude* of r indicates the *strength* of the relationship. See Figure[Correlation Values].

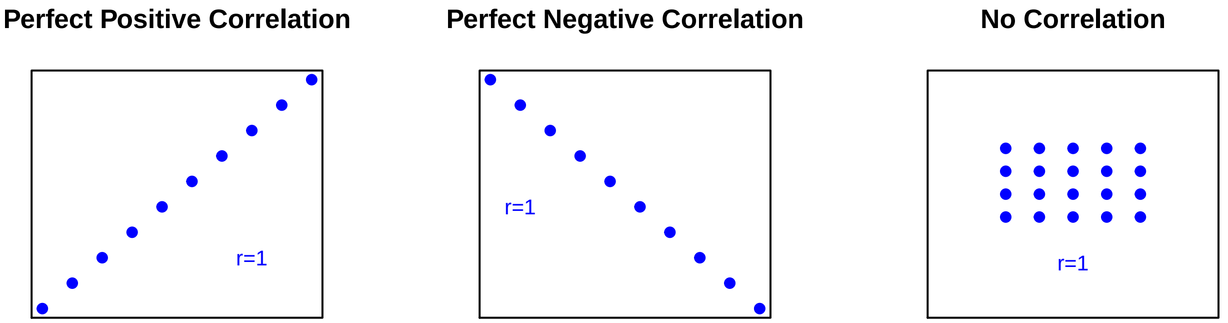


Figure 4.10: Correlation Values: The first scatterplot represents two variables that have a perfect positive linear relationship, $r = 1$. The second scatterplot represents two variables that have a perfect negative linear relationship, $r = -1$. The third scatterplot represents two variables that have no linear relationship, $r = 0$.

- $r = 1$ indicates a *perfect positive* linear relationship. All points fall exactly on a line sloping upward.
- $r = -1$ indicates a *perfect negative* linear relationship. All points fall exactly on a line sloping downward.
- $r = 0$ indicates *no* linear relationship.

You can investigate this with the following app, as well.

```
require(manipulate)
VaryCorrelation()
```

Let's look at an example.

> **Research Question**: At UCDavis, how is a student's mom's height (**momheight**) related to their dad's height (**dadheight**)?

Since we are interested in how a mother's height is related to a father's height, we will treat **dadheight** as the explanatory variable and **momheight** as the response variable. We will start by taking a look at the scatterplot. See Figure[Mom/Dad Height].

```
xyplot(momheight~dadheight,data=ucdavis1)
```

Since our cloud of points seems to be somewhat shaped upward to the right, it seems that students with tall dads tend to have tall moms as well and students with short dads also have short moms. This suggests that the correlation coefficient will be positive. However, since the cloud of points is not very tightly clustered, we might think that this association is not very strong. We might predict that the correlation coefficient, r, will be positive but closer to 0 than it is to 1. Let's find out.

```
cor(momheight~dadheight,data=ucdavis1,use="na.or.complete")
```

```
## [1] 0.2571501
```

4.2.3 Regression Equation

Regression analysis is used to numerically explain the linear relationship between two numerical variables using the equation of a line. It is much more specific than the visual analysis we've been making by looking at scatterplots.

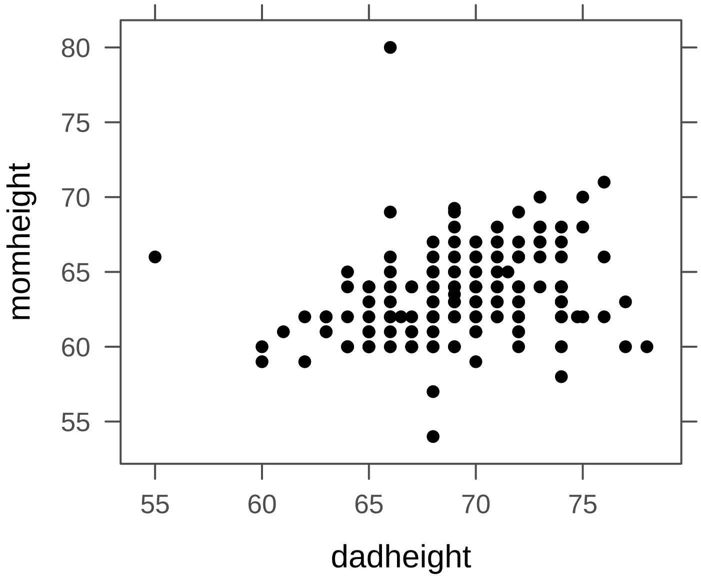


Figure 4.11: Mom/Dad Height

Recall that the equation of the line $y = \frac{5}{9}(x - 32)$ was used to describe the *deterministic* relationship between degrees Celcius and degrees Fahrenheit. This equation let us determine the *exact* temperature in Celcius by knowing the temperature in Fahrenheit.

The *regression equation* is the equation of a line that is used to *predict* the value for the response variable (y) from a known value of the explanatory variable (x) in a *statistical* relationship. It describes how, *on average*, the response variable is related to the explanatory variable.

In general, the equation of the regression line is

$$\hat{y} = a + bx,$$

where:

- a is the *y-intercept.* (The y-intercept is the point where the line crosses the vertical axis. It is the height of the line at $x = 0$.)
- b is the *slope.* (The slope is the inclination or orientation of the line. It is calculated by the ratio $\frac{\text{rise}}{\text{run}}$.)
- x is the known value of the explanatory variable.
- $\hat{y}$ is the *predicted* value of the response variable.

The regression line is the line that best approximates the data in the scatterplot. Returning to our original research question involving the variables **Height** and **RtSpan** from the `pennstate1` dataset, let's take a look at the regression line. See Figure[Hand/Height Regression Line].

Each point (x, y) on the scatterplot is an observation—known x and y-values. These points correspond to the measurements for an actual individual in the sample. Each point $(x, \hat{y})$ on the regression line is a known x-value and its predicted response, $\hat{y}$.

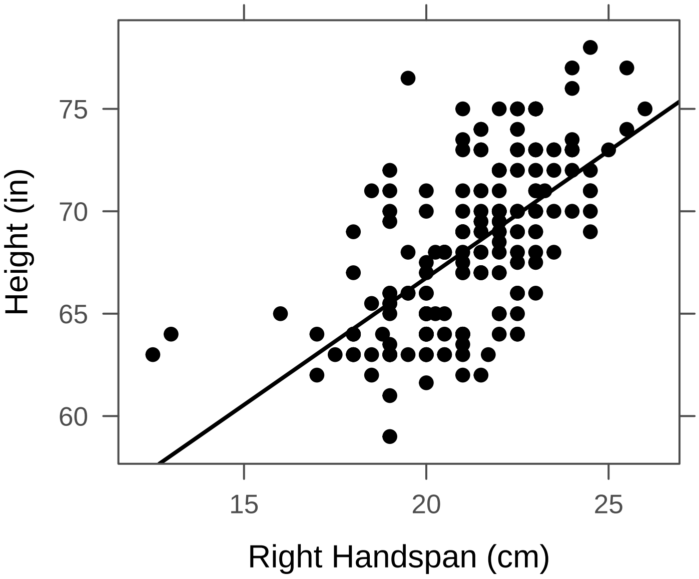


Figure 4.12: Hand/Height Regression Line: The regression line is plotted on the scatterplot showing the relationship between right handspan and height.

4.2.3.1 Residuals

You can see how the regression line in Figure[Hand/Height Regression Line] seems to do a good job describing the trend of the points in the scatterplot. In fact, it "best fits" the data. The regression (best fit) line is the line that is collectively the closest, in terms of vertical measurement, to all of the points on the scatterplot. These vertical measurements are called *residuals*. Residuals measure the size of the prediction errors.

Residuals For a given data point (x, y), the *residual* is the difference between the observed response and the response that is predicted by the regression line.

This can be written

$$y - \hat{y}.$$

On the graph, a residual is the vertical distance between a point and the regression line. Several residuals are plotted on the scatterplot shown in FigureResiduals.

```
## Predict Height is about 58.07,
## give or take 3.285 or so for chance variation.
```

```
## Predict Height is about 61.79,
## give or take 3.209 or so for chance variation.
```

```
## Predict Height is about 63.03,
## give or take 3.191 or so for chance variation.
```

A residual can be computed for every point on the scatterplot. The regression line $\hat{y} = a + bx$ is determined by choosing a (intercept) and b (slope) so that the sum of the (squared) residuals is minimized.

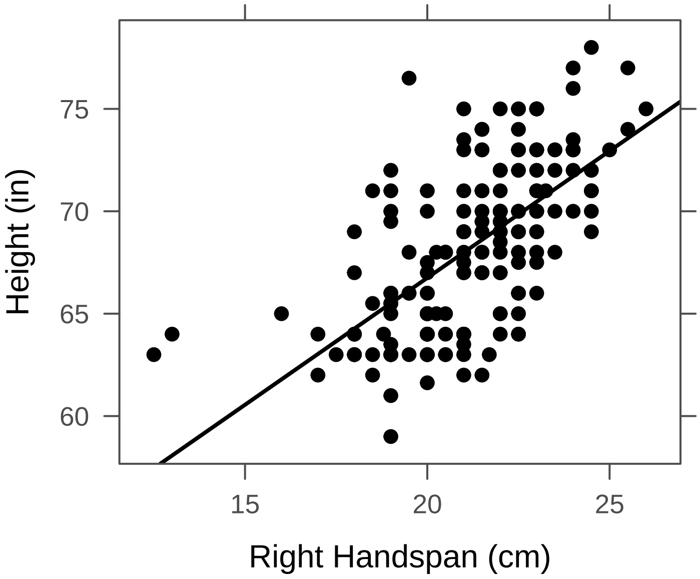


Figure 4.13: Residuals

$$\text{Sum of Squares } = \sum(\text{ residuals})^2 = \sum(y_i - \hat{y})^2$$

Investigate how to minimize the sum of squared residuals with the following app.

```
require(manipulate)
FindRegLine()
```

If you are interested in how the actual values of a and b are calculated, consult Geek Notes.

4.2.3.2 Predictions

Now that we know what the regression equation means and how it is found, let's use it to make some statements about the Pennstate scatterplot of heights and right handspans.

> **Research Question**: What is the predicted height of a Pennstate student with a right handspan measurement of 22 cm?

We have already seen the graph of the regression line in Figure[Hand/Height Regression Line]. Now, we would like to know the equation for this line. R has a built in function, `lmGC()`, to compute this equation. (The `lm` stands for "linear model".) This function also has the option to graph the regression line (set `graph=TRUE`).

```
lmGC(Height~RtSpan,data=pennstate1,graph=TRUE)
```

```
##
##  Linear Regression
##
## Correlation coefficient r =  0.6314
##
## Equation of Regression Line:
```

```
##
##   Height = 41.9593 + 1.2394 * RtSpan
##
## Residual Standard Error: s   = 3.1486
## R^2 (unadjusted):          R^2 = 0.3987
```

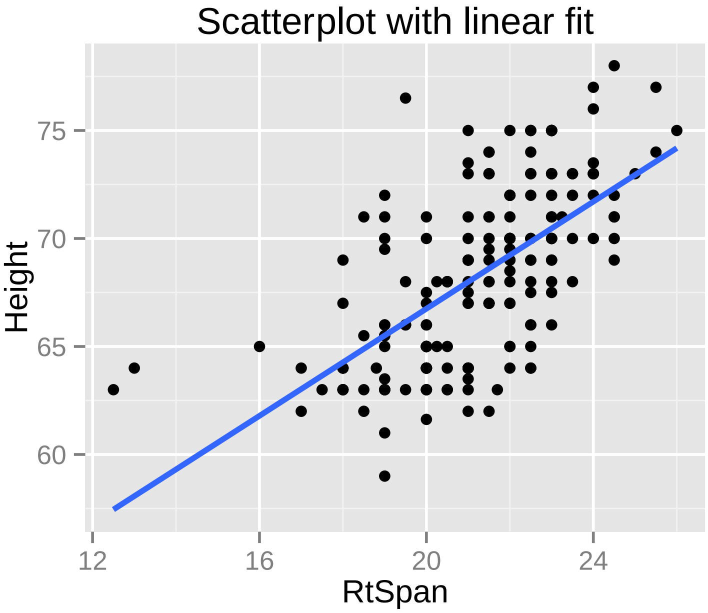


Figure 4.14: Height/Hand Regression: In the lmGC function, the parameter graph=TRUE will plot the regression line as well as giving the equation of the regresion line.

So the equation of our regression line is:

$$\hat{y} = 41.959349 + 1.239448x$$

We can now answer the research question two different ways. We can use R as a calculator and plug $x = 22$ into the equation above.

```
41.95935+1.239448*22
```

```
## [1] 69.22721
```

A Penn State student with a right handspan of 22 centimeters will have a predicted height of 69.22721 inches.

Another way to do this is using the `predict()` function. The `predict()` function requires two inputs:

- a linear model,
- a value of the explanatory variables, x.

In order to use the `predict()` function, you should store your linear model in a variable.

```
handheightmod <- lmGC(Height~RtSpan,data=pennstate1)
predict(handheightmod,22)
```

```
## Predict Height is about 69.23,
## give or take 3.158 or so for chance variation.
```

In addition to the prediction, R gives you a *prediction standard error* — a rough estimate of how much your prediction of the person's height is liable to differ from his or her actual height. In general, the regression line's prediction could easily differ from the actual y value (the height) by as much as a couple of prediction standard errors.

If you would like to have a better feel for where the actual value of y might lie, consider asking for a *prediction interval*, as follows:

```
predict(handheightmod,22,level=0.95)
```

```
## Predict Height is about 69.23,
## give or take 3.158 or so for chance variation.
##
## 95%-prediction interval:
##           lower.bound         upper.bound
##           62.997191           75.457220
```

You can be about 95%-confident that the actual height of a person with a handspan of 22 centimeters is somewhere between 63 and 74.5 inches — the stated bounds of the 95%-prediction interval above.

You can ask for prediction intervals at any level of confidence between 0% and 100%,simply by varying the value of the `level` parameter. For an 80%-prediction interval, run the command:

```
predict(handheightmod,22,level=0.80)
```

```
## Predict Height is about 69.23,
## give or take 3.158 or so for chance variation.
##
## 80%-prediction interval:
##           lower.bound         upper.bound
##           65.165568           73.288843
```

The prediction interval is narrower than before, but you pay a price for the extra precision: now you are only 80%-confident that the actual height lies somewhere inside of the interval.

4.2.3.3 Interpretation of Slope and Intercept

It's also important that we know how to interpret the *slope* and *intercept* of a regression line. Let's take a look at the `pushups` dataset.

```
data(pushups)
View(pushups)
help(pushups)
```

Now, we can view the equation of the regression line as well as the scatterplot with the regression line plotted. See Figure[Pushups].

```
##
##  Linear Regression
##
## Correlation coefficient r =  -0.4675
##
## Equation of Regression Line:
##
##   weight = 247.4886 + -0.8243 * pushups
##
## Residual Standard Error: s   = 39.5677
## R^2 (unadjusted):        R^2 = 0.2185
```

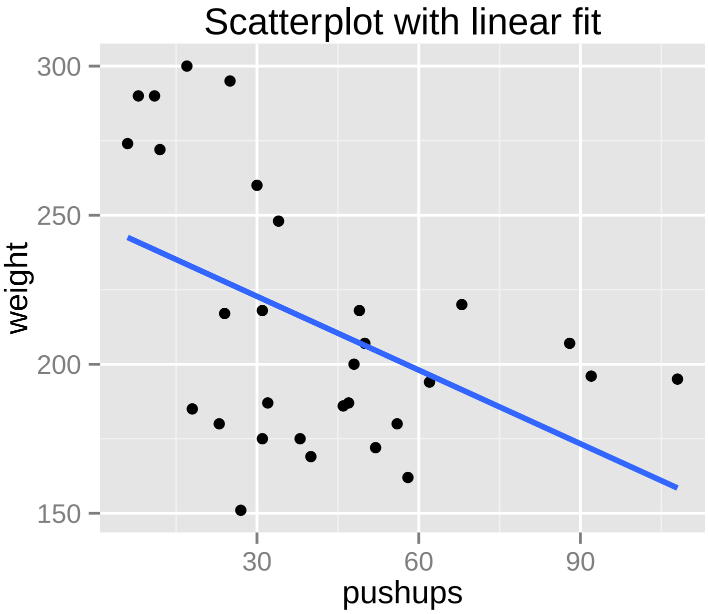


Figure 4.15: Pushups: Scatterplot and Regression line for Pushups versus Weight Relationship

- What does the intercept mean? 247.49 is the *predicted* weight of a GC football player who cannot do any pushups in 2 minutes.

- What does the slope mean? The *predicted* weight of a football player changes by -0.82 pounds as the max number of pushups in two minutes increases by 1. In other words, for every one pushup increase, the predicted weight of the football player decreases by 0.82 pounds. This tells us that there is a *negative* association between max number of pushups and weight!

Note: The interpretation of the intercept doesn't always make sense! Consider, for example, what would happen if we used **weight** as the explanatory variable and **pushups** as the response. See Figure[Slope/Intercept Interpretation].

```
##
##  Linear Regression
##
## Correlation coefficient r =  -0.4675
##
## Equation of Regression Line:
```

```
##
##   pushups = 97.6857 + -0.2651 * weight
##
## Residual Standard Error: s   = 22.4414
## R^2 (unadjusted):        R^2 = 0.2185
```

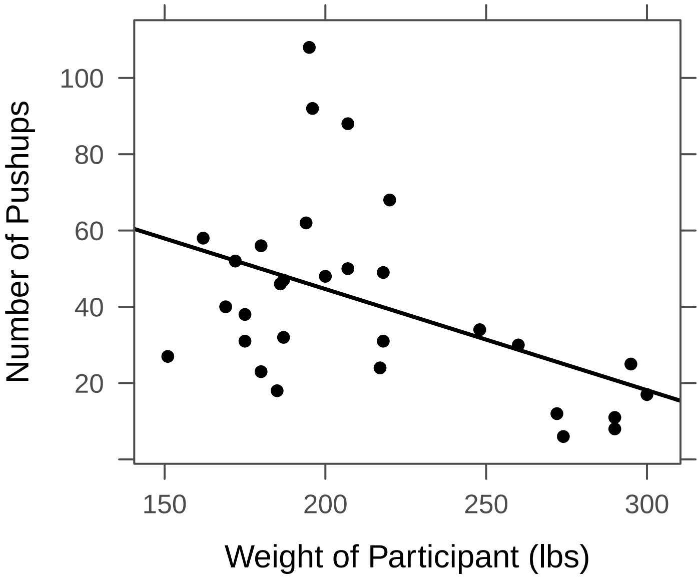


Figure 4.16: Slope/Intercept Interpretation: Slope and intercept interpretation for the regression line used to predict maximum number of pushups from weight of a football player.

- Intercept Interpretation: 247.49 is the *predicted* maximum number of pushups of a GC football player who weighs 0 pounds can do in 2 minutes. This doesn't make any logical sense!

- Slope Interpretation: The *predicted* maximum number of pushups that a football player can do in two minutes changes by -0.82 as the weight of the football player increases by 1 pound. In other words, for every one pound increase in weight, the predicted maximum number of pushups decreases by 0.82.

4.2.3.4 How Well does our Regression Line fit?

If all the data in a scatterplot lie *exactly on* the regression line, we say that our regression line perfectly explains the data. However, this is rarely the case. We usually have points that do not lie on the regression line. Anywhere that this happens, we have *variation* that is not explained by the regression line. If the data is not on a line, then a line will not be a *perfect* explanation of the data.

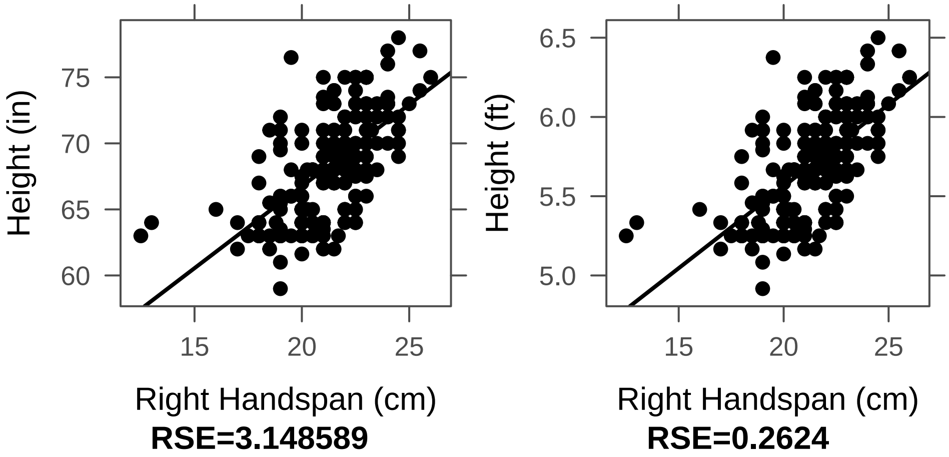
One way we can measure this variation is the *residual standard error* — sometimes called RSE, or even s, for short. This quantity does exactly what its name implies - it measures the *spread* of the residuals (those vertical distances between observations and the regression line). This value is in the output of the `lmGC` function.

However, we run into a problem when using the residual standard error to measure the variation in the plot. The residual standard error for the **Height** and **RtSpan** data from `pennstate1` is 3.148589.

```
lmGC(Height~RtSpan,data=pennstate1)
```

```
##
##  Linear Regression
##
## Correlation coefficient r =  0.6314
##
## Equation of Regression Line:
##
##   Height = 41.9593 + 1.2394 * RtSpan
##
## Residual Standard Error: s   = 3.1486
## R^2 (unadjusted):        R^2 = 0.3987
```

For this model, the unit of measure for height is inches. Consider what happens when we change this model by converting the unit of measure for height to feet. Take a look at the scatterplots and regression lines for these two models. See Figure[Different Units].



This change in unit of the response variable does not affect the way the scatterplot of regression line looks. The regression line did an equally good job of fitting the data, in both scatterplots. It does, however, directly affect the value of the residuals. The residuals in the first plot have a much larger value than the residuals in the second plot. This causes the *spread* of the residuals (the residual standard error) to be bigger in the first plot. Hence, residual standard error is not the best way of measuring the variation accounted for by the regression line.

Another measure of the "explained variation" in the scatterplot is the *squared correlation*, r^2. This also measures how well our regression line fits the data. However, it tells us the *proportion* of variation in the response variable that is explained by the explanatory variable. A change in unit (or scale) will not affect the value of r^2.

4.2.3.5 Properties of r^2

- r^2 always has a value between 0 and 1.
- $r^2 = 1$ implies perfect linear association between explanatory and response variables.
- $r^2 = 0$ implies no linear association.

Beware: A low r^2 value does not necessarily mean that there is *no* relationship between the explanatory variable and the response. It might mean that a linear model is not an appropriate model! So, a low r^2 value means one of two things for us:

- There is a linear relationship between the variables, but there's just alot of scatter in the data.

- You might have the wrong model. In other words, the data might not follow a linear pattern. For example, consider the relationship we saw in the curvilinear `fuel` data from before. See Figure[R-squared for Nonlinear Data].

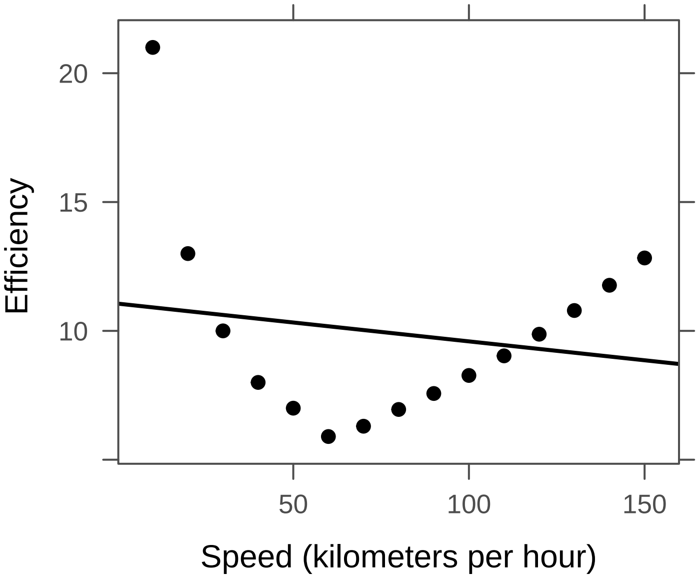


Figure 4.17: R-squared for Nonlinear Data

```
##
##  Linear Regression
##
## Correlation coefficient r =  -0.1716
##
## Equation of Regression Line:
##
##   efficiency = 11.0579 + -0.0147 * speed
##
## Residual Standard Error: s   = 3.9047
## R^2 (unadjusted):        R^2 = 0.0295
```

Here, we get a small r^2 value. However, there is certainly a relationship between **efficiency** and **speed** - just not a *linear* one! It is always important that you visually examine the data before drawing conclusions based on certain statistics!

4.3 Cautions

4.3.1 Extrapolation

The regression line we found for predicting **Height** from **RtSpan** using the `pennstate1` dataset, $\hat{y} = 41.959349 + 1.239448x$. can be used for *interpolation*, i.e. predicting a height for a person that was not in the original dataset, but is within the range of right handspans covered by the dataset. However, it is inappropriate to use this regression line to predict a height for someone with a hand span that is outside of the range covered by the dataset.

Extrapolation *Extrapolation* is using a regression line to predict $\hat{y}$-values for x-values outside the range of observed x-values.

Sultan Kosen holds the Guiness World Record for the tallest living male. (His right handspan measures 30.48 cm, which is considerably bigger than all of the hand-spans in our dataset.) If we extrapolate using the regression line from this dataset, then we would predict his height to be $\hat{y} = 41.959349 + 1.239448 \cdot 30.48 = 79.36$ inches ≈ 6' 7.5". But he was really 8' 3"! The prediction based on extrapolation was not very accurate.

4.3.2 Influential Observations

Influential Observations *Influential observations* are observations which have a large effect on correlation and regression:

- They have extreme x-values.
- They inflate or deflate correlation.
- They affect the slope of the regression line.

Investigate influential observations with the following app.

```
require(manipulate)
Points2Watch()
```

4.3.3 Association versus Causation

Two variables having a high positive (or high negative) correlation between two variables suggests that there is a linear association that exists between them. However, correlation does *not* imply causation. In other words, strong correlation does not imply that one variable *causes* the other.

Consider the variables **height** and **fastest** in the `m111survey` dataset. The scatterplot and correlation coefficient suggest that taller people tend to have driven at a higher maximum speed. See Figure[Height/Fastest Scatterplot].

```
cor(height~fastest,data=m111survey,use="na.or.complete")
```

```
## [1] 0.1708742
```

There a positive correlation between height and top speed. However, this does not necessarily mean that being tall *causes* you to drive faster. It's certainly not the case that tall people can't help but have a heavy foot! There may be a confounding variable that is (at least partially) responsible for the observed association. One possible confounder is **sex**. See Figure[Height/Fastest by Sex]. This suggests that males tend to drive faster than females. Since males also tend to be taller than females, this helps to explain the pattern we saw in the original scatterplot.

4.3.4 Simpson's Paradox

Recall: Simpson's Paradox occurs when the direction of the relationship between two variables is one way when you look at the aggregate data, but turns out the opposite way when you break up the data into subgroups based on a third variable. *You saw this back in Chapter 3 with two categorical variables!*

Simpson's Paradox can arise with two numerical variables as well!

Let's look at a new dataset, `sat`.

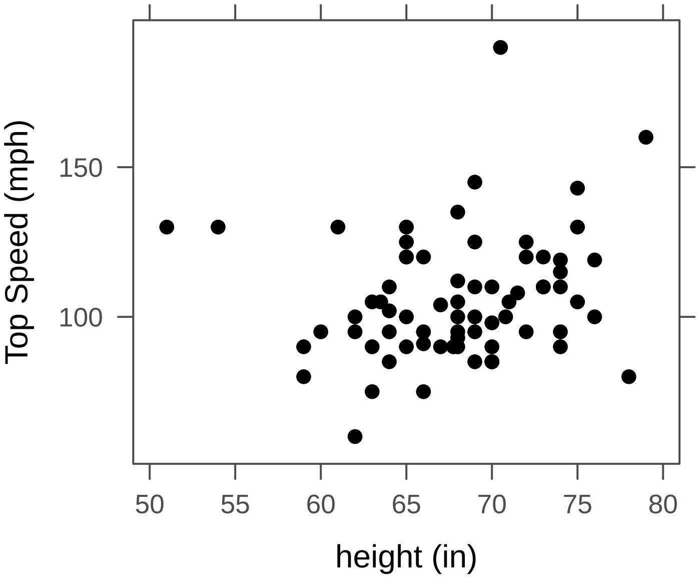


Figure 4.18: Height/Fastest Scatterplot

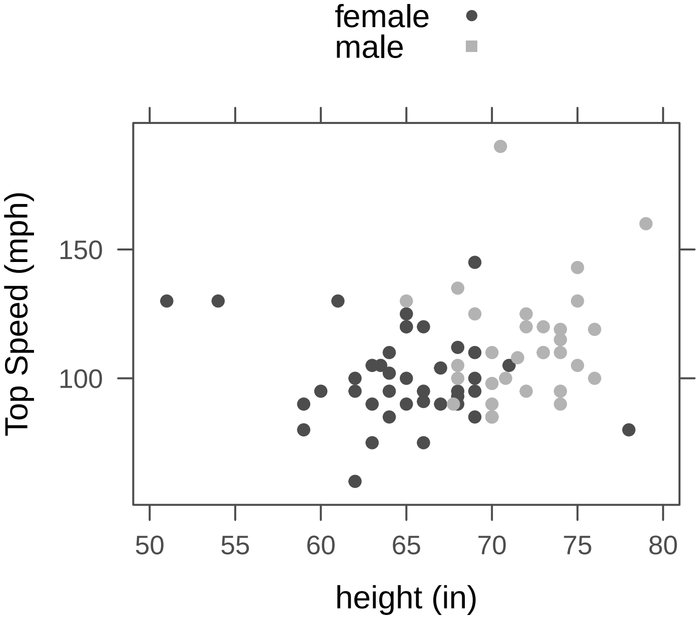


Figure 4.19: Height/Fastest by Sex: Grouping shows that sex may be a confounding variable that helps to explain why taller people tend to drive faster.

```r
data(sat)
View(sat)
help(sat)
```

We can describe the relationship between **salary** (mean annual teacher salary by state in $1000s) and **sat** (sum of mean Verbal and mean Math scores by state) using the correlation coefficient, scatterplot, and regression line. See Figure[SAT].

```
##
##  Linear Regression
##
## Correlation coefficient r =  -0.4399
##
## Equation of Regression Line:
##
##   sat = 1158.859 + -5.5396 * salary
##
## Residual Standard Error: s   = 67.8893
## R^2 (unadjusted):        R^2 = 0.1935
```

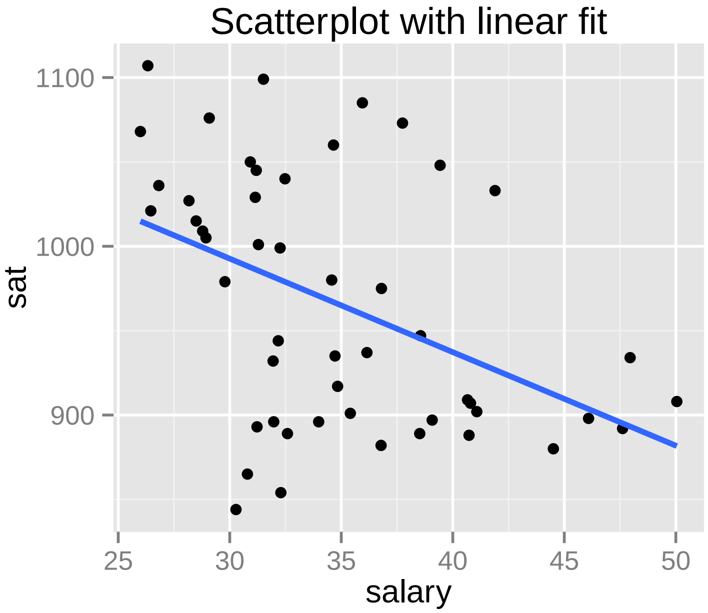


Figure 4.20: SAT: Association between Average Annual Teacher Salary and Average Cumulative SAT Score (by state)

Our regression analysis suggests that as teachers are paid higher salaries, the average SAT score drops. However, our intuition tells us that higher salaries would attract better teachers, who in turn would better prepare students for the SAT. What's going on here?

Use the following app to investigate what happens when we take into account a third variable, `frac`. This is the percentage of students in the state who take the SAT.

```r
require(manipulate)
DtrellScat(sat~salary|frac,data=sat)
```

We can see that within most subgroups, the slope of the regression line is positive, while the regression line for the overall data has a negative slope.

How can we explain this paradox?

First Observation: It turns out that states where education is considered a high priority pay their teachers higher salaries. But since education is a high priority in these states, a high proportion of students in these states want to go to college, so a high proportion of them take the SAT. Similarly, states where education isn't such an important part of the economy or the culture tend to pay their teachers less, and they also tend to have fewer students taking the SAT. So we've got a **positive** association between **frac** and **salary**.

Check this out in Figure[Frac/Salary Scatterplot].

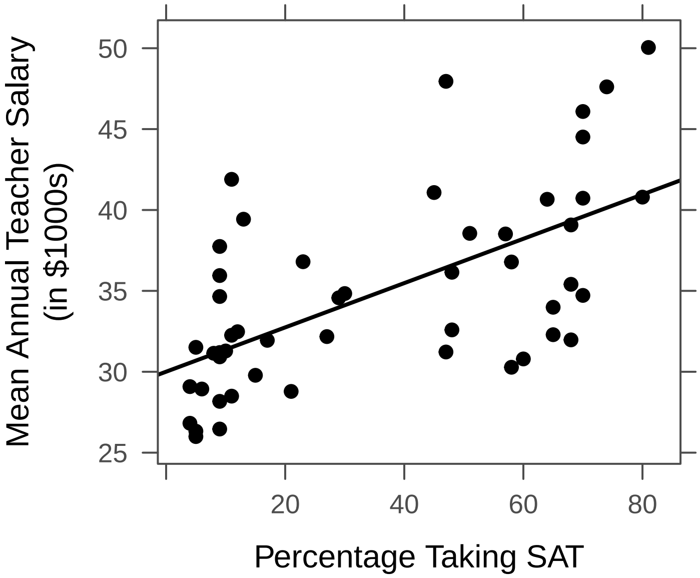


Figure 4.21: Frac/Salary Scatterplot: Percentage of students in the state that take the SAT vs. the average annual teacher salary.

Second Observation: When a high percentage of students in a state take the SAT, this pool of test-takers is bound to include both the very strong students and those who are not so strong, so the mean SAT score is liable to be low. On the other hand, in states where fewer students take the SAT, the students who opt to take the test are likely to be highly motivated, definitely college-bound, very studious individuals. This elite pool tends to boost the mean SAT score for the state. So we've got a **negative** association between **frac** and **sat**.

Again, check this out in Figure[Frac/SAT Scatterplot].

Finally, put it together: the positive association between **salary** and **frac** and the negative association between **frac** and **sat** results in the negative association between **salary** and **sat**.

To put it in a nontechnical nutshell: states where education is important tend to pay their teachers well, but they also "handicap" themselves by encouraging most students (including the weaker ones) to take tests like the SAT. The handicap is so pronounced that higher values of **salary** tend to go along with lower values of **sat**.

4.4 Curvilinear Fits

Frequently we have brought up the idea that the relationship between x and y — our two numerical variables — might not be a linear one. Let's look into this idea bit further.

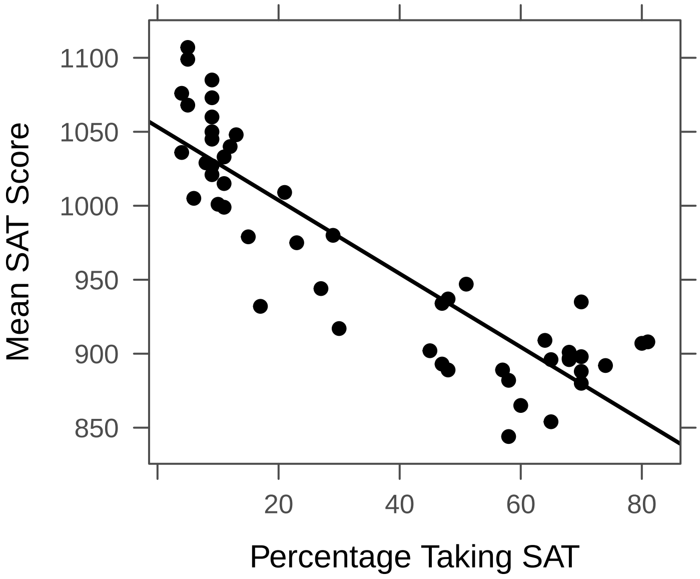


Figure 4.22: Frac/SAT Scatterplot: Percentage of students in the state that take the SAT vs. the average cumulative SAT score.

Consider the data frame `henderson` from the `tigertats` package:

```
data(henderson)
View(henderson)
help(henderson)
```

Ricky Henderson, who played in the Major Leagues from 1979 to 2003, is widely considered to be one of the finest leadoff hitters in baseball history. The `henderson` data frame gives some fo his most important offensive statistics, by season. Let's have a look at his slugging average (the average number of bases he got per at-bat) over the years (see Figure [Career Slugging]:

We have included a regression line, but as you can see the fit doesn't look right at all: the points on the scatter plot don't appear to follow a line. It's not just that there is a lot of "scatter" around the regression line; instead it seems that some other pattern – not a linear one — is at work in determining how Ricky's slugging average varied over the years.

One way to confirm our suspicion is to run a *check* on the linear fit, using the `check` parameter for `lmGC()`:

```
lmGC(SLG~Season,data=henderson,check=TRUE)
```

```
##
##  Linear Regression
##
## Correlation coefficient r =  -0.2719
##
## Equation of Regression Line:
##
##   SLG = 5.7685 + -0.0027 * Season
##
## Residual Standard Error: s   = 0.0661
## R^2 (unadjusted):        R^2 = 0.0739
```

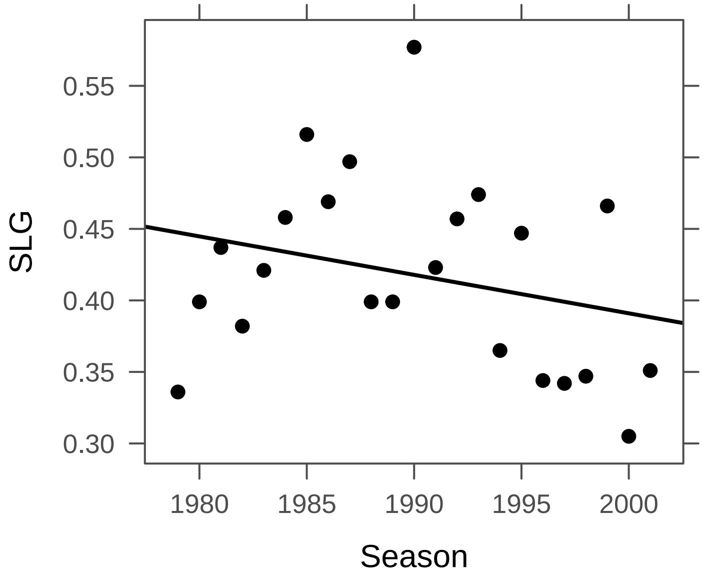
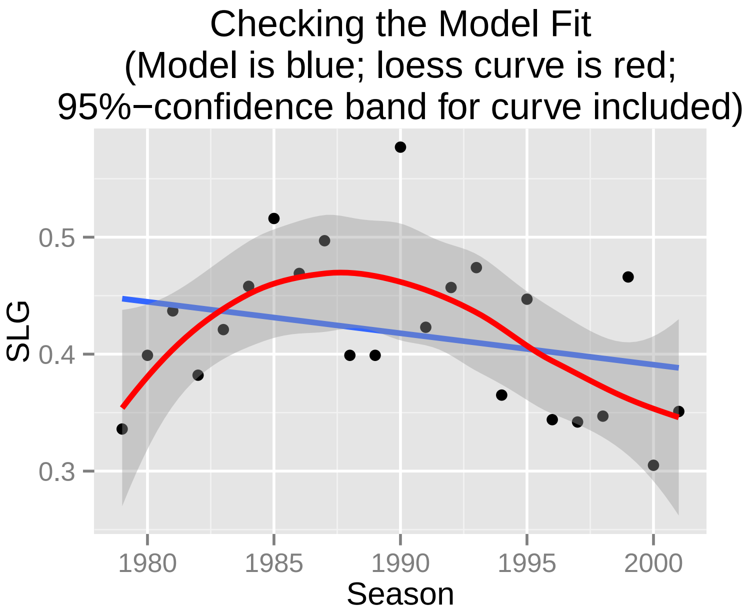


Figure 4.23: Slugging for Ricky Henderson, 1979-2001



In addition to the usual linear model information, the scatter plot contains two new features:

- a *loess* curve;
- an approximate 95%-confidence band around the loess curve.

The term "loess" is short for "local estimation": a loess curve is an attempt to use only the data itself to estimate the "real" deterministic part of the relationship between the x and y variables. In particular, it doesn't assume that the relationship is a linear one! hence it wobbles around, following the general pattern of the point son the scatter plot.

The confidence band is the slightly shaded area that surrounds the loess curve. A rough-and-ready way to interpret the interval is to say that we are pretty confident that the "real" relationship between **Season** and **SLG** — if it could somehow be graphed on the scatter plot – would lie somewhere within the band. The loess curve is simply our best "data-based" guess at that unknown relationship.

Note that the regression line wanders a bit outside of the band, and is often quite near the edge of it: this is a good indication that the "real" relationship between **Season** and **SLG** is not a linear one.

So we have "checked" the linear fit, and found it wanting. Let's use our common sense to consider what the "real" relationship might be like.

For many professional athletes, the first couple of years in the majors leagues are tough ones, as they go up against other top-of-the-line players. But after a while they adjust —if they don't their career stats come to an end quickly! — and a relatively long "peak" period of play ensues. Eventually, though, age and injuries catch up with the athlete, and his/her performance begins to decline.

Hence the relationship between season of play and some performance measure such as slugging average ought to be a curvilinear one, with a rise and then a fall.

Rise-and-fall curvilinear relationships can be modeled mathematically by means of quadratic equations, which have the general form:

$$y = ax^2 + bx + c,$$

since these equations graph as parabolas. Let's try to fit a second-degree curve (a parabola) to our data. This is accomplished using a new function, `polyfitGC()`. The function works like `lmGC()`, but has a new parameter `degree` to indicate the degree of the fitting curve:

- `degree=2` fits a quadratic (a one-humped rise/fall or fall/rise parabola) to the data;
- `degree=3` fits a cubic (as many as two "humps""");
- `degree=4` fits a cubic (as many as three "humps""");
- and so on to higher degrees.

Here's the R-code for our quadratic fit:

```
polyfitGC(SLG~Season,data=henderson,
          degree=2,graph=TRUE)
```

The output to the console does not give the equation of the parabola, but we do get the residual standard error and the R^2 value. (Note that there is no correlation: r doesn't make sense outside the context of linear fits.)

The graph looks promising: the parabola does appear to run through the points on the scatter plot a lot better than the line did. Also note that the R^2 value is a good bit higher than it was with the linear fit (0.3825 as compared to 0.0739).

We can check this new fit as follows:

```
polyfitGC(SLG~Season,data=henderson,
          degree=2,check=TRUE)
```

```
## Polynomial Regression, Degree = 2
##
## Residual Standard Error: s   = 0.0553
## R^2 (unadjusted):        R^2 = 0.3825
```

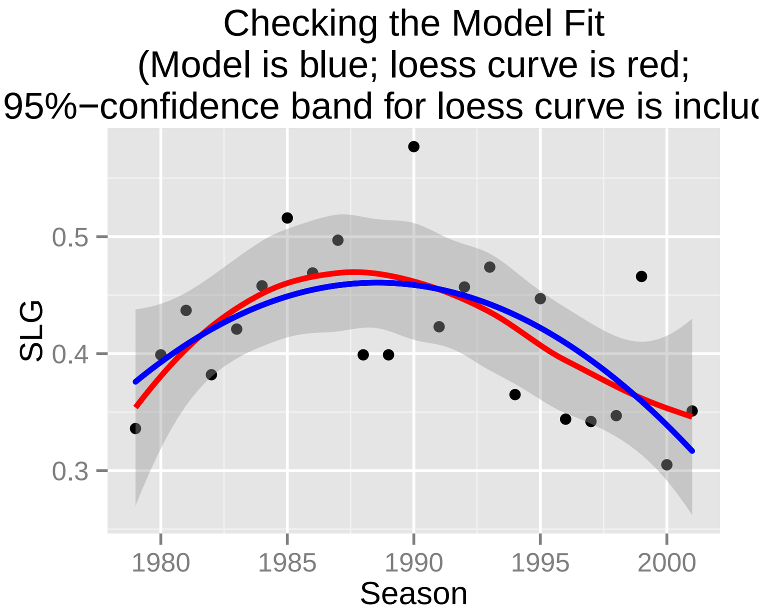


Figure 4.24: Checking Graph for the Quadratic Fit

Examining Figure [Checking Graph], we see that the parabola stays well within the band at all times: our quadratic fit looks quite promising as a way to predict y values from known x values!

The the data go only up through the 2001 season, but Henderson played 72 games with the Boston Red Sox in 2002, and 30 games in 2003 for the LA Dodgers. Let's use the parabola and the linear fit to predict his slugging averages for those two seasons.

First, the prediction based on the linear fit:

```
linModel <- lmGC(SLG~Season,data=henderson)
predict(linModel,x=2002)
predict(linModel,x=2003)
```

Now the prediction based on the quadratic fit:

```
quadModel <- polyfitGC(SLG~Season,degree=2,data=henderson)
predict(quadModel,x=2002)
predict(quadModel,x=2003)
```

Here are the predictions from each model, compared with Henderson's actual statistics:

Season	Linear Prediction	Quadratic Prediction	Actual
2002	0.386	0.293	0.352
2003	0.383	0.267	0.306

Table 4.1: Predicted/Actual Slugging Averages for R. Henderson

The regression line was better for 2002, but it falls too slowly: the parabola makes the better prediction for 2003. If we had to predict how Ricky would have done had he stayed on for the 2004 season, surely we would go with the prediction provided by the quadratic fit!

4.5 Thoughts on R

4.5.1 New R Functions

Know how to use these functions:

- `xyplot()`

- `lmGC()`
- `predict()`
- `polyfitGC()`

Chapter 5

Sampling and Surveys

5.1 Introduction

Recall from Chapter 3 that most research questions actually break down into 2 parts:

- **Descriptive Statistics**: What relationship can we observe between the variables, *in the sample*?
- **Inferential Statistics**: Supposing we see a relationship in the sample data, how much evidence is provided for a relationship *in the population*? Does the data provide lots of evidence for a relationship in the population, or could the relationship we see in the sample be due just to chance variation in the sampling process that gave us the data?

Both parts of answering research questions involve dealing with the sample. In order to make valid conclusions about any research question, we first need to make sure we are dealing with a *good* sample. This chapter will discuss various techniques for drawing samples, the strengths and weaknesses of these sampling techniques, and the uses and abuses of statistics.

5.2 Population versus Sample

In the past few of chapters, we have looked at both parts of research questions. An important distinction that we want to make sure has been made before we go any further is the distinction between a **sample** and a **population**.

Population A *population* is the set of all subjects of interest.

Sample A *sample* is the subset of the population for which we have data.

Let's consider these two definitions with a research question.

> **Research Question**: In the United States, what is the mean height of adult males (18 years +)?

The *population* that we are dealing with in this case is all U.S. adult males. One way to find an exact answer to this research question would be to survey the entire population. However, this is nearly impossible! It would be much quicker and easier to measure only a subset of the population, a *sample*.

However, if we want our sample to be an accurate reflection of the population, we can't just choose *any* sample that we wish. The way in which we collect our **sample** is very important and will be a topic of conversation in this chapter.

For the time being, let's suppose that we were able to choose an appropriate sample (and we'll talk more about how this is done later). Suppose that our sample of U.S. men is an accurate representation of the U.S. population of men. Then, we might discuss two different means: the *mean height of the sample* and the *mean height of the population*. These are both **descriptions**, as opposed to **inferences**. There are a couple of differences, however.

Mean Height of the Sample

- Statistic - describes the sample
- Can be known, but it changes depending on the sample
- Symbol - $\bar{x}$ (pronounced "x bar")

Mean Height of the Population

- Parameter - describes the population
- Usually unknown - but we wish we knew it!
- Symbol - μ (pronounced "mu")

Our goal is to use the information we've gathered from the *sample* to **infer**, or **predict**, something about the *population*. For our example, we want to predict the population mean, using our knowledge of the sample. The accuracy of our *sample mean* relies heavily upon how well our *sample* represents the *population* at large. If our sample does a poor job at representing the population, then any inferences that we make about the population are also going to be poor. Thus, it is very important to select a good sample!

Note: If we already knew everything about a population, it would be useless to gather a sample in order to *infer* something about the population. We would already have this information! Using statistics as an *inferential* tool means that you don't have information about the entire population to start with. If you are able to sample the entire population, this would be called a **census**.

It would be nice to see what a sample looks like in comparison to the population from which it is drawn. In `tigerstats` we have a dataset that represents an imaginary population, `imagpop`. Drawing samples from this "population" will help give an idea of the distinction between *sample* versus *population* and *statistic* versus *parameter*. Try out the following app, keeping in mind that information about the sample is displayed in *light blue* and information about the population is displayed in *red*.

```r
require(manipulate)
SimpleRandom()
```

5.3 Types of Samples

There are 2 main kinds of sampling:

- Random Sampling

- Non-Random Sampling

There are advantages and disadvantages of both.

5.3.1 Random Sampling

There are four different methods of random sampling that we will discuss in this chapter:

- **Simple Random Sampling (SRS)**
- **Systematic Sampling**
- **Stratified Sampling**
- **Cluster Sampling**

The simple random sample (SRS) is the type of sample that we will focus most of our attention on in this class. However, the other types have been included in the text to give you comparisons to the SRS and also to aid you in the future.

It will be helpful to work with an example as we describe each of these methods, so let's use the following set of 28 students from FakeSchool as our population from which we will sample.

```r
data(FakeSchool)
View(FakeSchool)
help(FakeSchool)
```

Keep in mind that we would not know information about an entire population in real life! **We are using this "population" for demonstration purposes only!**

Our goal is to describe how these different sampling techniques are implemented, the strengths and weaknesses of them, and to form a comparison between the techniques. We will try to answer the following question:

> Which random sampling method (*simple random sample*, *systematic sample*, *stratified sample*, or *cluster sample*) is the most appropriate for estimating the mean grade point average (GPA) for the students at FakeSchool?

We can easily compute the true mean GPA for the students at FakeSchool by averaging the values in the fourth column of the dataset. This will be the **population mean**. We will call it μ ("mu").

```r
mu <- mean(~GPA,data=FakeSchool)
mu
```

```
## [1] 2.766429
```

Again, the population parameter, μ, is not typically known. If it were known, there would be no reason to estimate it! However, the point of this example is to practice selecting different types of samples and to compare the performance of these different sampling techniques.

5.3.1.1 Simple Random Sample

Simple Random Sampling (SRS) In *simple random sampling*, for a given sample size n every set of n members of the population has the same chance to be the sample that is actually selected.

We often use the acronym SRS as an abbreviation for "simple random sampling".

Intuitively, let's think of simple random sampling as follows: we find a big box, and for each member of the population we put into the box a ticket that has the name of the individual written on it. All tickets are the same size and shape. Mix up the tickets thoroughly in the box. Then pull out a ticket at random, set it aside, pull out another ticket, set it aside, and so on until the desired number of tickets have been selected.

Let's select a *simple random sample* of 7 elements without replacement. We can accomplish this easily with the built in function `popsamp` in R. This function requires two pieces of information:

- the size of the sample
- the dataset from which to draw the sample

Remember that sampling *without replacement* means that once we draw an element from the population, we do not put it back so that it can be drawn again. We would not want to draw with replacement as this could possibly result with a sample containing the same person more than once. This would not be a good representation of the entire school. (By default, the `popsamp` function always samples without replacement. If you want to sample with replacement, you would need to add a third argument to the function: `replace=TRUE`. Typically, we will sample without replacement in this class.)

Since we may want to access this sample later, it's a good idea to store our sample in an object.

```
set.seed(314159)
srs <- popsamp(7,FakeSchool)
srs
```

```
##    Students Sex class  GPA Honors
## 6       Eva   F    Fr 1.80     No
## 18    Derek   M    Jr 3.10    Yes
## 7     Georg   M    Fr 1.40     No
## 11    Dylan   M    So 3.50    Yes
## 23      Bob   M    Sr 3.80    Yes
## 13     Eric   M    So 2.10     No
## 14  Gabriel   M    So 1.98     No
```

Let's calculate the mean GPA for the 7 sampled students. This will be the *sample mean*, $\bar{x}_{srs}$. We will use the subscript 'srs' to remind ourselves that this is the sample mean for the simple random sample.

```
xbar.srs <- mean(~GPA,data=srs)
xbar.srs
```

```
## [1] 2.525714
```

Strengths

- The selection of one element does not affect the selection of others.
- Each possible sample, of a given size, has an equal chance of being selected.
- Simple random samples tend to be good representations of the population.
- Requires little knowledge of the population.

Weaknesses

- If there are small subgroups within the population, a SRS may not give an accurate representation of that subgroup. In fact, it may not include it at all! This is especially true if the sample size is small.
- If the population is large and widely dispersed, it can be costly (both in time and money) to collect the data.

5.3.1.2 Systematic Sample

Systematic Sampling In a *systematic sample*, the members of the population are put in a row. Then 1 out of every k members are selected. The starting point is randomly chosen from the first k elements and then elements are sampled at the same location in each of the subsequent segments of size k.

To illustrate the idea, let's take a 1-in-4 systematic sample from our FakeSchool population.

We will start by randomly selecting our starting element.

```r
set.seed(49464)
start=sample(1:4,1)
start
```

```
## [1] 4
```

So, we will start with element 4, which is Daisy and choose every 4th element after that for our sample.

```
##    Students Sex class GPA Honors
## 4     Daisy   F    Fr 2.1     No
## 8    Andrea   F    So 4.0    Yes
## 12   Felipe   M    So 3.0     No
## 16 Brittany   F    Jr 3.9     No
## 20   Eliott   M    Jr 1.9     No
## 24     Carl   M    Sr 3.1     No
## 28    Grace   F    Sr 1.4     No
```

The mean GPA of the systematic sample, the *sample mean*, $\bar{x}_{sys}$, is 2.7714286.

Strengths

- Assures an even, random sampling of the population.
- When the population is an *ordered* list, a systematic sample gives a better representation of the population than a SRS.
- Can be used in situations where a SRS is difficult or impossible. It is especially useful when the population that you are studying is arranged in time.

For example, suppose you are interested in the average amount of money that people spend at the grocery store on a Wednesday evening. A *systematic sample* could be used by selecting every 10th person that walks into the store.

Weaknesses

- Not every combination has an equal chance of being selected. Many combinations will never be selected using a systematic sample!
- Beware of *periodicity* in the population! If, after ordering, the selections match some pattern in the list (skip interval), the sample may not be representative of the population.

The list of the FakeSchool students is ordered according to the student's year in school (freshmen, sophomore, junior, senior). Taking a systematic sample ensures that we have a person from each class represented in our sample. However, there is an underlying pattern, or periodicity, in the data. The students are also listed according to their GPA. For instance, Alice is ranked first in the freshmen class and George is ranked last in the freshmen class.

Consider what would have happened if we had used a *systematic sample* of 4 students to estimate the average GPA of the students at the school.

```
##    Students Sex class  GPA Honors
## 1     Alice   F    Fr 3.80    Yes
## 8    Andrea   F    So 4.00    Yes
## 15     Adam   M    Jr 3.98    Yes
## 22   Angela   F    Sr 4.00    Yes
```

Notice that even thought the *systematic sample* ensured that we got one person from each class, we also ended up getting students of the same class rank due to the underlying pattern. Our estimate for the average GPA is not going to truly reflect the population of the school! It may be **biased** since the GPA pattern coincided with the skip interval.

5.3.1.3 Stratified Sample

Stratified Sampling In a *stratified sample*, the population must first be separated into homogeneous groups, or *strata*. Each element only belongs to one stratum and the stratum consist of elements that are alike in some way. A simple random sample is then drawn from each stratum, which is combined to make the stratified sample.

Let's take a stratified sample of 7 elements from FakeSchool using the following strata: Honors, Not Honors. First, let's determine how many elements belong to each strata:

```
## Honors
##  No Yes
##  16  12
```

So there are 12 Honors students at FakeSchool and 16 non-Honors students at FakeSchool.

There are various ways to determine how many students to include from each stratum. For example, you could choose to select the same number of students from each stratum. Another strategy is to use a *proportionate stratified sample*. In a *proportionate stratified sample*, the number of students selected from each stratum is proportional to the representation of the strata in the population. For example, $\frac{12}{28}$ X 100% = 42.8571429% of the population are Honors students. This means that there should be 0.4285714 X 7 = 3 Honors students in the sample. So there should be 7-3=4 non-Honors students in the sample.

Let's go through the coding to draw these samples. Check out the how we use the `subset` function to pull out the Honors students from the rest of the populations:

```
set.seed(1837)
honors=subset(FakeSchool,Honors=="Yes")
honors
```

```
##    Students Sex class  GPA Honors
## 1     Alice   F    Fr 3.80    Yes
## 2      Brad   M    Fr 2.60    Yes
## 8    Andrea   F    So 4.00    Yes
## 9     Betsy   F    So 4.00    Yes
## 10    Chris   M    So 4.00    Yes
## 11    Dylan   M    So 3.50    Yes
## 15     Adam   M    Jr 3.98    Yes
## 17   Cassie   F    Jr 3.75    Yes
## 18    Derek   M    Jr 3.10    Yes
## 19    Faith   F    Jr 2.50    Yes
## 22   Angela   F    Sr 4.00    Yes
## 23      Bob   M    Sr 3.80    Yes
```

Next, we take a SRS of size 3 from the Honors students:

```
honors.samp=popsamp(3,honors)
honors.samp
```

```
##    Students Sex class GPA Honors
## 9     Betsy   F    So 4.0    Yes
## 11    Dylan   M    So 3.5    Yes
## 8    Andrea   F    So 4.0    Yes
```

The same method will work for non-Honors students.

```
set.seed(17365)
nonhonors=subset(FakeSchool,Honors=="No")
nonhonors.samp=popsamp(4,nonhonors)
nonhonors.samp
```

```
##    Students Sex class  GPA Honors
## 25    Diana   F    Sr 2.90     No
## 13     Eric   M    So 2.10     No
## 14  Gabriel   M    So 1.98     No
## 28    Grace   F    Sr 1.40     No
```

We can put this together to create our stratified sample.

```
##    Students Sex class  GPA Honors
## 9     Betsy   F    So 4.00    Yes
## 11    Dylan   M    So 3.50    Yes
## 8    Andrea   F    So 4.00    Yes
## 25    Diana   F    Sr 2.90     No
## 13     Eric   M    So 2.10     No
## 14  Gabriel   M    So 1.98     No
## 28    Grace   F    Sr 1.40     No
```

The sample mean for the stratified sample, $\bar{x}_{strat}$, is 2.84.

Strengths

- Representative of the population, because elements from all strata are included in the sample.
- Ensures that specific groups are represented, sometimes even proportionally, in the sample.

- Since each stratified sample will be distributed similarly, the amount of variability between samples is decreased.
- Allows comparisons to be made between strata, if necessary. For example, a stratified sample allows you to easily compare the mean GPA of Honors students to the mean GPA of non-Honors students.

Weaknesses

- Requires prior knowledge of the population. You have to know something about the population to be able to split into strata!

5.3.1.4 Cluster Sample

Cluster Sampling *Cluster sampling* is a sampling method used when natural groups are evident in the population. The clusters should all be similar each other: each cluster should be a small scale representation of the population. To take a **cluster sample**, a random sample of the clusters is chosen. The elements of the randomly chosen clusters make up the sample.

Note: There are a couple of differences between stratified and cluster sampling.

- In a stratified sample, the differences *between* stratum are high while the differences *within* strata are low. In a cluster sample, the differences *between* clusters are low while the differences *within* clusters are high.
- In a stratified sample, a simple random sample is chosen from *each* stratum. So, all of the stratum are represented, but not all of the elements in each stratum are in the sample . In a cluster sample, a simple random sample of clusters is chosen. So, not all of the clusters are represented, but all elements from the chosen clusters are in the sample.

Let's take a cluster sample using the grade level (freshmen, sophomore, junior, senior) of FakeSchool as the clusters. Let's take a random sample of 2 of them.

```
##    Students Sex class  GPA Honors
## 15     Adam   M    Jr 3.98    Yes
## 16 Brittany   F    Jr 3.90     No
## 17   Cassie   F    Jr 3.75    Yes
## 18    Derek   M    Jr 3.10    Yes
## 19    Faith   F    Jr 2.50    Yes
## 20   Eliott   M    Jr 1.90     No
## 21    Garth   M    Jr 1.10     No
## 22   Angela   F    Sr 4.00    Yes
## 23      Bob   M    Sr 3.80    Yes
## 24     Carl   M    Sr 3.10     No
## 25    Diana   F    Sr 2.90     No
## 26    Frank   M    Sr 2.00     No
## 27       Ed   M    Sr 1.50     No
## 28    Grace   F    Sr 1.40     No
```

The sample mean for the clustered sample, $\bar{x}_{clust}$, is 2.7807143.

Strengths

- Makes it possible to sample if there is no list of the entire population, but there is a list of subpopulations. For example, there is not a list of **all** church members in the United States. However, there is a list of churches that you could sample and then acquire the members list from each of the selected churches.

Weaknesses

- Not always representative of the population. Elements within clusters tend to be similar to one another based on some characteristic(s). This can lead to over-representation or under-representation of those characteristics in the sample.

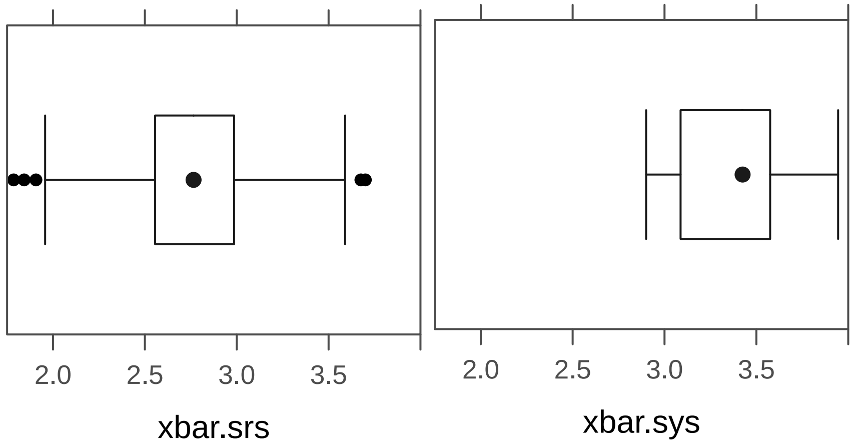
5.3.2 Comparison of Sampling Methods

Now that you have an idea about how to take each of these kinds of samples, let's compare them by doing repeated samples. There is no general rule for determining which sampling method is best. The choice of sampling method depends on the data that is being analyzed and will require the statistician's judgment.

We will compare the *simple random sample* and the *systematic sample* by determining which sample produces the **least variable mean GPA estimate** after repeated sampling.

Putting that another way: Let's start by taking 1000 simple random samples and 1000 systematic samples. We will compute $\bar{x}_{srs}$ and $\bar{x}_{sys}$ for each of the samples. Then, these sample means will be compared using some graphical and numerical summaries (specifically standard deviation) that you learned about in Chapter 2.

We're going to take the SRS's and systematic samples just like did before. The only difference is that now we'll be taking 1000 of them instead of just 1. Since we only care about the sample mean for each sample, we'll create a boxplot of the $\bar{x}_{srs}$'s and a boxplot of the $\bar{x}_{sys}$'s. These two boxplots allow us to compare the amount of variation, or spread, in the estimates for the mean GPA generated from the two different sampling methods (SRS and systematic sampling). See Figure[Boxplots].



To support this visualization of the variability of the mean estimate for GPA, let's also look at `favstats`. For the 1000 simple random samples, the numerical summaries of the sample means is:

```
##       min       Q1 median       Q3 max     mean        sd    n missing
##  1.785714 2.556429  2.765 2.985714 3.7 2.768701 0.3173966 1000       0
```

For the 1000 systematic samples, the numerical summaries of the sample means is:

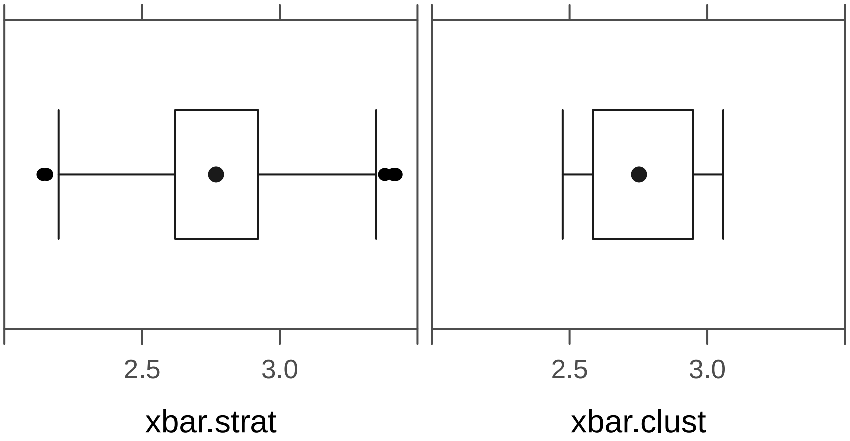
```
##  min      Q1 median    Q3   max    mean       sd    n missing
##  2.9 3.18125  3.425 3.575 3.945 3.42042 0.381641 1000       0
```

Recall that the true average GPA for the population of students at FakeSchool was 2.7664286. Notice that the average value for the sample means from the 1000 simple random samples is 2.7687014. This is pretty close to the population parameter. (We will talk about what "pretty close" means in later chapters.) Compare this to the average value for the sample means from the 1000 systematic samples: 3.42042. On average, the SRS does a better job of producing an estimate for the mean GPA than the systematic sample.

Additionally, there is less variability in the 1000 $\bar{x}_{srm}$'s (0.3173966) than in the 1000 $\bar{x}_{sys}$'s (0.381641).

If we could only pick one of these types of samples to estimate the mean GPA, it appears the a SRS is a better choice than a systematic sample.

Let's do a similar analysis to compare the two sampling methods, *stratified sampling* or *cluster sampling.* We will compare the *stratified sample* and the *cluster sample* by determining which sample produces the **least variable mean GPA estimate** after repeated sampling.



To support this visualization of the variability of the mean estimate for GPA, let's also look at `favstats`. For the 1000 stratified samples, the numerical summaries of the sample means is:

```
##   min       Q1   median       Q3      max     mean        sd    n missing
##  2.14 2.620714 2.768571 2.921429 3.422857 2.772663 0.2215186 1000       0
```

For the 1000 cluster samples, the numerical summaries of the sample means is:

```
##    min       Q1   median       Q3      max     mean        sd    n missing
##  2.475 2.584286 2.752143 2.948571 3.057857 2.762486 0.1965341 1000       0
```

Both of these sampling methods produce an average of the sample means that is pretty close to the true mean GPA for the population. However, the sample means from the clustered samples have less variability. (This can be seen by comparing the standard deviations.) In other words, the 1000 cluster samples are closer, on average, to the true mean than the 1000 stratified samples.

If we could only pick one of these types of samples to estimate the mean GPA, it appears the cluster sample is a better choice than a stratified sample.

5.4 Bias in Surveys

Bias We say that a sampling method is *biased* if it exhibits a systematic tendency to result in samples that do not reflect the population, in some important respect.

You can think of a survey as occurring in three stages:

1. Select subjects to invite into your sample. this is the *sampling* stage.
2. Get them to accept your invitation. This is the stage where you contact the subjects you have sampled, and ask them to participate in your survey.
3. Obtain their responses to your questions.

At each of these stages, some bias can creep in!

5.4.1 Selection Bias

Selection bias is the type of bias that can occur in the first stage, in which you are selecting the subjects who will be your sample.

Selection Bias We say that a sampling method exhibits *selection bias* if its mechanism for selecting the sample has a systematic tendency to over-represent or under-represent a particular subset of the population.

One sampling method that can result in selection bias is convenience sampling.

Convenience Sampling *Convenience sampling* is the practice of sampling subjects that the researcher can reach easily. This may result in certain subgroups of the population being underrepresented or completely left out.

Example: A math professor wants to know what percentage of young adults, ages 18-22, consider education a top priority. She gathers a sample by surveying all of her advisees.

This method of sampling is quick and easy. However, only including students that are enrolled in college leaves out a large part of the population - those young adults that did not go to college or enrolled in a different type of higher education. Only including college students in this study might make it appear that a high percentage of young adults consider education a top priority. The subjects in the study surely consider it a priority since they are seeking a college degree.

Another form of selection bias occurs when you attempt to sample everyone in the population, but you leave it up to each member of the population to find out about your survey and to take part in it. This is called "volunteer" sampling.

Volunteer Sampling A *volunteer sample* is a sample of only those subjects that have volunteered to be part of a study. There may be common characteristics about the people that volunteer to be part of a particular survey that creates bias.

Example: A radio station wishes to examine the proportion of its listeners which candidate they voted for in the last presidential election. They conduct a poll by asking listeners to call the station.

Conducting a survey in this manner is also quick and easy, but there are groups in the population that are *underrepresented* or *not represented at all*! Only those listeners who want to disclose this information will be part of the survey. Those volunteers may have something else in common that will bias the results: for example, they may have stronger opinions on the question at hand than do other folks who did not choose to go out of their way to phone in their thoughts.

One great advantage of the simple random sampling and proportionate stratified sampling – two of the methods we discussed earlier – is that they are not subject to selection bias.

5.4.2 Nonresopnse Bias

Even if you succeed in selecting a sample of subjects in an unbiased way, you still face the task of acquiring their consent to be in your survey. Some people may refuse to take part, or perhaps you will be unable to contact them all. In that event, they won't respond to the survey, and this could lead to bias.

Nonresponse Bias We say that a sampling method exhibits *nonresponse bias* if there is a systematic tendency for the people who elect to take part in the survey to differ from the population in some important way.

Example: The faculty at Georgetown College wanted to know what proportion of students thought that Foundations should be required for all freshmen. A simple random sample of 200 students was selected from a list obtained from the registrar. A survey form was sent by email to those students. After analyzing the results from the 20 people that reply, the faculty report that 90% of the students oppose the requirement for Foundations.

- What was the population?

  Answer: The population of interest is the entire student body at Georgetown College.

- What was the intended sample size?

  Answer: The intended sample size was 200.

- What was the sample size actually observed?

  Answer: The sample size that was actually observed was the number of students that responded to the survey, 20.

- What was the percentage of nonresponse?

  Answer: Since 20 of the 200 students selected for the survey actually respond, 180 did not respond. The percentage of nonresponse was 90%. It can be found by:

```
(180/200)*100
```

```
## [1] 90
```

- Why might this cause the results to be *biased*?

  Answer: If all of the 200 randomly selected students had responded to the survey, we would have had a true SRS. However, nonresponse bias has occurred in this study because 90% of the sampled subjects either *were not reached*, *refused to participate*, or *failed to answer the question.* A couple of possible explanations for the nonresponding students might be that they do not check their email or they simply did not have a strong enough opinion on the topic to feel the need to take the time to respond. (There may be other legitimate reasons.) So, it could be that the students that responded had very strong feelings about the Foundations requirement. If these are the only answers that are acquired, the results may be heavily biased in the direction of the opinion of the respondents. However, this does not mean that all students feel this way.

5.4.3 Response Bias

The wording and presentation of the questions can significantly influence the results of a survey. The main type of bias that can result from a poorly-worded survey is *response bias*.

Response Bias We say that a sampling method exhibits *response bias* if the way the questions are asked or framed tends to influence the subjects' responses in a particular way.

Many things can subject a survey response bias. Here are a few:

- Deliberate Wording Bias
- Unintentional Wording Bias
- Desire of the Respondents to Please
- Asking the Uninformed
- Unnecessary Complexity
- Ordering of Questions
- Confidentiality Concerns

> **Deliberate Response Bias -** If a survey is being conducted to support a certain cause, questions are sometimes deliberately worded in a biased manner. The wording of a question should not indicate a desired answer.

Example: Consider the following research question: "Seeing as Dr. Robinson and Dr. White are the greatest professors you have ever had, is it worth even offering the peer tutoring sessions for MAT111?"

This question is prefaced in a way that encourages a desired response from the subjects in the study.

> **Unintentional Response Bias -** Some questions are worded in such a way that the meaning is misinterpreted by the respondents.

Example: Consider the following research question: "Do you use drugs?" The word *drugs* can cause unintentional confusion for the respondent. The intended definition of drugs is not made clear in the wording of the question. Does the researcher mean illegal drugs, prescription drugs, over the counter drugs, or possibly even caffeine?

> **Desire of Respondents to Please -** People may respond differently depending on how they are being asked - face-to-face, over the telephone, on paper, on the internet.

For example, a person may tend to be more honest when answering questions on paper or over the internet. When speaking directly to the researcher, the respondent may feel the need to answer the question how they perceive the researcher wants.

> **Asking the Uninformed -** If a question is about a topic that the respondent does not know anything about, they often do not like to admit it. Respondents may tend to give an answer, even though they do not understand the question.

> **Unnecessary complexity -** Questions should be kept simple. Try to only ask one question at a time.

Example: Consider the following survey question: “Most semesters are 15 weeks long; while most quarters are 10 weeks long. Most schools on a quarter system get 2 days for Thanksgiving, one for Veteran’s Day, and one for Columbus Day. Most semester schools get Labor Day off and some take more than 2 days at Thanksgiving. However, semester schools typically start several weeks earlier in the fall and generally attend school farther into December. Considering the above, which system would you prefer?”

This question has too much information in it. By the time you get done reading it, you may have forgotten what the question is even referring to. It is unnecessarily complex!

> **Ordering of Questions** - If one question requires respondents to think about something that they may not have otherwise considered, then the order in which questions are presented can change the results.

Example: Suppose a researcher wants to know how many hours a day people spend on the Internet. Consider the following sequence of questions:

- “Do you own a smartphone?”"
- “How many hours a day do you spend on the Internet?”

Placing the question about the smartphone before the question about time spent on the Internet causes the respondent to take into consideration that they are often on the Internet when they are using their phone. Putting the questions in this order may change the answers received for the second question.

> **Confidentiality Concerns** - Some personal questions will be answered differently depending on how confident the respondent is that their identity will be concealed.

5.5 Thoughts on R

Know how to use this function:

- `popsamp`

Chapter 6

Design of Studies

6.1 Observational Studies and Experiments

The `attitudes` data frame contains the results of a large-scale survey conducted at Georgetown College in the fall semester of 2001:

```
data(attitudes)
View(attitudes)
help(attitudes)
```

This dataset prompts many research questions concerning the relationship between two variables. For example, the numerical variable **sentence** gives the sentence recommended for a hypothetical defendant who has been convicted for involuntary manslaughter in a drunk-driving incident. Another variable—the factor **major**—records the type of major that the survey participant intends to pursue. A Research Question on the relationship between these two variables might take the form:

> *Which type of major tends to be hardest on crime?*

Let's quickly explore this question. We can take a numerical approach:

```
favstats(sentence~major,data=attitudes)
```

```
##       .group min    Q1 median    Q3 max     mean       sd   n missing
## 1 humanities   5 11.25     20 38.75  50 25.26316 15.96565  38       0
## 2   math.sci   5 20.00     30 50.00  50 30.49351 15.64964  77       0
## 3   pre.prof   3 14.00     25 32.50  50 25.02336 15.05342 107       1
## 4 social.sci   4 15.00     25 40.00  50 25.48889 14.66694  45       0
```

Comparing the means of the four major groups, it seems that math and science majors recommend sentences that are, on average, about five years longer than sentences recommended by the other three major groups.

We can also take a graphical approach. Figure [Sentence by Major] also indicates that math and science majors are harder on crime. (Note, for instance, that their median is about five years above the medians for the other three groups.)

The data for this research question – the variables **major** and **sentence** – were gathered by means of an *observational study*:

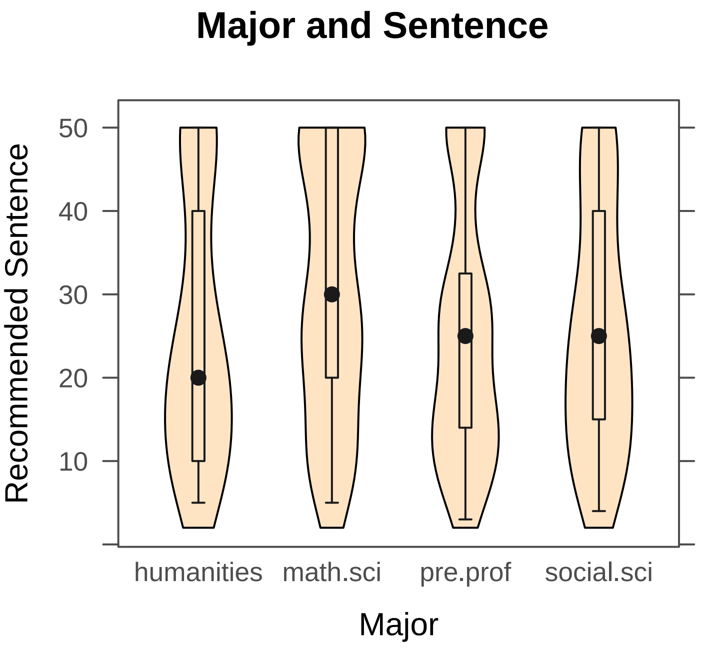


Figure 6.1: Sentence by Major. The survey respondents are broken into four groups, accordng to the type of majors they intend to pursue.

Observational Study

: In an *observational study* researchers simply observe or question the subjects. In particular, they measure the values of the explanatory variable X and measure the values of the response variable Y, for each subject.

For each student in the survey, we measured the value of the variable **major** for each student by recording his/her choice, and observed the value of **sentence** by recording the sentence he/she recommended. Thus, the Research Question at hand was addressed by means of an observational study.

But not all of the variables in the `attitudes` data were simply observed. From the Help file on `attitudes`, we learn that not all subjects received the same survey form. In particular, in the question for the drunk-driving incident, some of the subjects got a form where the question was stated as follows:

> You are on a jury for a manslaughter case in Lewistown, PA. The defendant has been found guilty, and in Pennsylvania it is part of the job of the jury to recommend a sentence to the judge. The facts of the case are as follows. The defendant, Tyrone Marcus Watson, a 35-year old native of Lewistown, was driving under the influence of alcohol on the evening of Tuesday July 17, 2001. At approximately 11:00 PM Watson drove through a red light, striking a pedestrian, Betsy Brockenheimer, a 20-year old resident of Lewistown. Brockenheimer was taken unconscious to the hospital and died of her injuries about one hour later. Watson did not flee the scene, nor did he resist arrest.
>
> The prior police record for Mr. Watson is as follows: two minor traffic violations, and one previous arrest, five years ago, for DUI. No one was hurt in that incident.
>
> Watson has now been convicted of DUI and manslaughter. The minimum jail term for this combination of offenses is two years; the maximum term is fifty years. In the blank below, write a number from 2 to 50 as your recommended length of sentence for Tyrone Marcus Watson.

Other subjects received a survey form in which the details of the incident were exactly the same as in the question above, except that the defendant's name was given as "William Shane Winchester." In other forms, the name of the victim varied: “Latisha Dawes”" instead of “Betsy Brockenheimer.” All told, there were four variants of the drunk-driving question:

Defendant’s Suggested Race	**Victim’s Suggested Race**
Black	Black
Black	White
White	Black
White	White

Table 6.1: Form-Variation in the Drunk-Driving Question

The purpose of varying the names—and with the names, the suggested race of the person—was to investigate Research Questions like the following:

> **Research Question 1**: *Does the suggested race of the defendant have an effect on the length of sentence a subject would recommend?*

> **Research Question 2**: *Does the suggested race of the victim have an effect on the length of sentence a subject would recommend for the defendant?*

The data we collected for these Research Questions does **not** constitute an observational study: the subjects did not choose which type of form to fill out; instead the researchers (randomly) assigned forms to subjects. This makes the study an *experiment.*

Experiment In an *experiment*, researchers manipulate something and observe the effects of the manipulation on a response variable.

Most commonly, the manipulation consists in assigning the values of an explanatory variable X to the subjects.

In Research Question 1 above, the explanatory variable is **defrace**, the race of the defendant that is suggested by the name given to him in the form. (The response variable is **sentence**.) Since researchers assigned the values of **defrace** to each subject, Research Question 1 is being addressed through an experiment.

In Research Question 2 above, the explanatory variable is **vicrace**, the race of the victim that is suggested by the name given to her in the form. (The response variable is **sentence**.) Since researchers assigned the values of **vicrace** to each subject, Research Question 2 also is being addressed through an experiment.

For further practice in distinguishing between experiments and observational studies, study the following examples:

> **Example**: Researchers test whether zinc lozenges help people with colds recover more quickly than if they took no lozenges at all. They gather 40 people who are suffering from a cold, and randomly choose 20 of them to take zinc lozenges. The remaining 20 also take a lozenge, but it the lozenge is made of an inert substance that is known to have no effect on the body.

In this situation, the explanatory variable is whether or not one takes a zinc lozenge, and the response variable is the time it takes to recover from the cold. Since researchers randomly picked who would get a zinc lozenge, the values of the explanatory variable were assigned by the researchers. This makes the study as experiment.

Example: Researchers gather a number of people. They find that, of those who wear bifocals, 5% have colon cancer. Of those who do not wear bifocals, only 1% have colon cancer.

In the second situation, the Research Question of interest appears to concern the relationship between whether or not one wears bifocals and whether or not one gets colon cancer. It's not clear which variable—if either–is supposed to be the explanatory, but since neither variable had its values assigned to subjects (researchers did not make anyone wear bifocals, or give colon cancer to anyone!) the study is clearly just an observational study.

6.2 Why Perform Experiments?

Because in an experiment researchers assign values of the explanatory variable X to subjects, experiments can be difficult to perform. For example, in the lozenge example above, most people would prefer to choose for themselves what medicine to take for a cold. In an experiment, someone else assigns the treatment to you. It's difficult to recruit subjects for such a study!

Why, then, do researchers go to the trouble to do experiments? Why not just perform observational studies all of the time?

The answer has to do with the distinction, introduced in Chapter Four, between association and causation. We know already that two variables can be associated without there being a *causal* relationship between them. But there are many situations in which we really do want to know if an observed association is due to a causal relationship. Here are some examples:

1. We may find that smokers tend to develop lung cancer at a higher rate than non-smokers; we would like to know whether smoking actually *causes* lung cancer.

2. We may find that men drive faster than women, but we wonder whether a person's sex is a causal factor in how fast he or she drives.

3. Bifocal wearers have colon cancer at a higher rate than do those who do not wear bifocals, but does this mean that wearing bifocals increases your risk of colon cancer?

Consider the last example above: we don't think that wearing bifocals causes lung cancer, because we know that folks who wear bifocals tend to be older than those who don't, and we know that the risk of colon cancer increases with age. Age is considered a *confounding variable* that explains the association between bifocal-wearing and colon cancer.

Confounding Variable In a study with an explanatory variable X and a response variable Y, the variable Z is called a *confounding* variable if it meets the following three conditions:

1. It is a third variable (different from X and different from Y);
2. It is associated with X, but not caused by X.
3. It is a causal factor in Y (is at least part of the cause of Y)

Thus, **age** (Z) is indeed a confounding variable in a study of the relationship between **whether or not one wears bifiocals** (X) and **whether or not one has colon cancer** (Y). We know this because:

- **Age** is a third variable;
- **Age** causes bifocal wearing–older people have worse eysesight, and so choose to wear corrective lenses, including bifocals—and so **age** is associated with bifocal wearing.

- **Age** helps to cause colon cancer.

> **Note**: In Condition 2 in the definition of confounding variables we included the requirement that Z must not be caused by X. The reason for doing this is that if Z were caused by X, then, since Z helps to cause Y, X would help to cause Y indirectly after all. (We consider indirect causes to be legitimate causes; in fact, pretty much anything that we think of as a "cause" does its causal work indirectly.)

A difficulty with observational studies is that because subjects come to you with the value of their X variable already in place, observational studies are pervasively subject to the possibility of confounding variables. In the bifocal and colon cancer study, people choose whether or not to wear bifocals, and that choice is caused somewhat by their age. Hence the two groups being compared in the study—the bifocal-wearers and the non-bifocal wearers—differ with respect to age. Therefore we cannot attribute an observed difference in their colon cancer rates to the wearing of bifocals: the difference is probably due to age (or to other variables associated with age).

For another example, think back to the **sat** data, with which we attempted to study the relationship between teacher salary (X) and SAT score (Y) for states. Although we observed an association (states that paid their teachers more tended to have lower SAT scores), we could not conclude that paying teachers more *causes* their students to perform worse: the variable **frac** was a confounder. After all,

- **frac** is a third variable (different from **salary** and from **sat**);
- states where **frac** is high are states that tend to place value on education, so they are liable to pay their teachers well, so **frac** and **salary** are related;
- It is likely that states with a high percentage of students taking the SAT have lower scores due to the presence in the testing pool of non-elite students, so **frac** helps to cause **sat**.

Observational studies are not rendered worthless by the mere possibility of confounding variables, but they do make things more difficult for researchers, in at least two respects:

- You have to identify possible confounders BEFORE you collect your data, so you can record the value of the possible confounder for each subject in the study.
- You have to figure out how to *correct* for the confounders.

The second requirement can be pretty tough to satisfy. One possible way to proceed is to break your data up into little groups, such that the values of the confounders are the same, or nearly the same, for the subjects in each group. That's what we did with the **sat** data: recall the app:

```
require(manipulate)
DtrellScat(sat~salary|frac,data=sat)
```

The problem with breaking up the data into small subsets and studying them separately is that any relationship you see between X and Y within a subset is based on a very small sample—too small by itself to provide much evidence for a relationship in the population. Even if most of the subsets indicate the same relationship between X and Y, the fact that they were studied separately prevents us from synthesizing the results to infer a relationship between X and Y in the population at large.

Hence we would prefer to have no confounding variables in our study at all. In other words:

> **An Ideal**: In an ideal world the different groups in the study would be exactly alike with respect to every variable that might help to cause the response variable Y, with the sole exception that the groups differ in their X-values.

For example, in the bifocal study, you would prefer to compare two groups of people that are the same with respect to:

- age,
- diet,
- exercise patterns,
- genetic history,
- and anything else that might affect the risk of colon cancer!

The only way the groups would differ is that one group wears bifocals, and the other group—sometimes called the *control* group—does not. If two such groups were to exhibit a statistically significant difference in their colon cancer rates, then that difference could be ascribed to the bifocals, and not to anything else: there would be no confounding factors to worry about.

6.3 Experimental Designs

The ideal laid out in the preceding section cannot be attained in practice, but it can be approximated fairly well, provided that you *design* your experiment well.

It may surprise you to learn that good experimental designs make use of brute chance, at some point, in order to assign subjects to treatment groups. We will discuss some of these ways in this Section.

6.3.1 Completely Randomized Designs

In a *completely randomized* experiment, subjects are assigned to their treatment groups by randomization alone.

What is randomization, exactly, and why is it so effective in making groups similar? To see learn more about this consider again the following imaginary population:

```
data(imagpop)
View(imagpop)
help(imagpop)
```

Let's say that the first 200 people in this population agree to be part of an experiment to see whether taking aspirin reduces the risk of heart disease.

```
AspHeartSubs <- imagpop[1:200,]
```

The experimenters plan to choose 100 of the subjects at random, and assign them to take an aspirin each morning for ten years. The remaining 100 subjects—the control group—are given a pill each morning that looks and tastes like aspirin, but has no effect on the body. (Such a substance is called a *placebo*, and is used to keep the subjects unaware of which group they are in. If subjects knew their group, this knowledge might determine lifestyle choices that in turn affect the risk of heart disease. When subjects in an experiment don't know their group, the experiment is said to be *single-blind.*)

Together, the two groups are called the *treatment groups*, because they are treated differently: one groups gets aspirin, the other does not.

The X variable is whether or not the subject takes aspirin, and the Y variable is some measure of the heart-health of the subject. The values of the X variable were assigned randomly to the subjects, so this experiment has what we call a *completely randomized* design.

The R-function `RandomExp()` carries out the randomization:

```
Assignment <- RandomExp(AspHeartSubs,
                  sizes=c(100,100),
                  groups=c("placebo","aspirin"))
```

You should take a look at the result:

```
View(Assignment)
```

As you can see, 100 of the subjects are assigned to the Placebo group, and the other 100 are assigned to the Aspirin group. The assignment is done at random: each subject was equally likely to be assigned to either group.

Recall that we would like the treatment groups to be similar in every way, except that one group takes aspirin and the other does not. In order to verify that they are reasonably similar, we can check to see whether the variable **treat.grp** is related to any of the other variables in the `Assignment` data frame.

Do the treatment groups differ much with respect to **sex** (i.e., is **treat.grp** related to **sex**)? Let's see:

```
SexGrp <- xtabs(~treat.grp+sex,data=Assignment)
rowPerc(SexGrp)
```

```
##          sex
## treat.grp female male Total
##   placebo     51   49   100
##   aspirin     56   44   100
```

The distribution of **sex** in the two treatment groups appears to be quite similar.

Do the groups differ much with respect to income? (Is **treat.grp** related to **income**?) Let's see:

```
favstats(income~treat.grp,data=Assignment)
```

```
##    .group  min    Q1 median    Q3    max  mean       sd   n missing
## 1 placebo 1200 18325  33600 50000 154700 39215 26995.24 100       0
## 2 aspirin 1500 16675  33550 55650 165700 39486 29352.24 100       0
```

The means and medians of the two groups appear to be quite similar.

For further practice, pick another variable in the `Assignment` data frame, and see if the treatment groups differ much with respect to that variable.

The randomization procedure is not perfect: if you repeat the procedure many times:

```
Assignment <- RandomExp(AspHeartSubs,
                sizes=c(100,100),
                groups=c("placebo","aspirin"))
```

From time to time you will discover that for one or two of the variables the treatment groups are somewhat different. But on the whole it appears likely that for the most part the treatment groups will be pretty similar.

When R was doing the randomization to create the **treat.grp** variable, it worked "blindly": it did not take into account the sex, income, etc. of any of the subjects. This means that when researchers adopt a completely randomized design, they don't have to keep track of *any* characteristics of their subjects – not even those characteristics that might be confounding factors. Blind chance by itself is likely to produce similar treatment groups.

Statistical theory tells us that, the larger the number of subjects, the more likely it is that the treatment groups will be similar, and the more similar they are likely to be! That's one reason why researchers like to get as many subjects as possible into each treatment group. This principle is called *replication*:

Replication An experiment is said to involve *replication* if each the treatment group contains more than one subject. (The more subjects there are in each group, the more replication the experiment has.)

6.3.2 Randomized Block Designs

6.3.2.1 Dealing with a Small Number of Subjects

So far we have looked at experiments with a completely randomized design. This design works very well when you have a lot of subjects, because then you can get a lot of subjects into each treatment group. Blind chance then guarantees—more or less—that the treatment groups will be similar. When you don't have a lot of subjects, however, things might not go so well.

Suppose, for example, that you want to do an experiment to compare two weight-lifting programs (Program A and Program B), to see which one is more effective. You have 16 subjects in your group:

```r
data(SmallExp)
View(SmallExp)
help(SmallExp)
```

Note that the sex and athletic status (athlete or not) of each subject has been recorded.

Your plan is to break the subjects into two groups of size 8 each. Group A will train with Program A for ten weeks, and Group B will train with Program B. At the end of the ten-week period, you will measure the increase in strength for each subject.

Let's say you decide on a completely randomized design:

```r
RandomExp(SmallExp,sizes=c(8,8),
          groups=c("Program.A","Program.B"))
```

Perform the randomization several times by re-running the above code. Are you *always* happy with the results, or do the treatment groups often look rather different with respect to sex and/or with respect to athletic status?

When the number of subjects is small, randomization can easily produce dis-similar treatment groups. If the treatment groups are dis-similar with respect to possible confounders (such as **Sex** and **athlete** in this study) then we have a problem.

6.3.2.2 Blocking

Accordingly, when an experiment is expected to have only a small number of subjects, researchers will often adopt a *randomized block design*. This means that they carry out the following procedure:

1. For each subject, record the values of variables that you consider to be potential confounders. In the current study, **sex** and **athlete** have been recorded.

2. Break the subjects into *blocks* based on combinations of values of the potential confounders. In this study the blocks would be:
    - the four females who are athletes
    - the four females who are not athletes
    - the four males who are athletes
    - the four males who are not athletes
3. Within each block randomly assign subjects to treatment groups. In this study, for example, you would randomly pick two of the female athletes to be in Group A, with the other two going to Group B, and do the same for the other three blocks.

Randomized blocking is accomplished with `RandomExp()` as follows:

```
RandomExp(SmallExp,sizes=c(8,8),
          groups=c("Program.A","Program.B"),
          block=c("sex","athlete"))
```

Re-run the randomization several times, and observe that:

- In terms of the individuals, the treatment groups differ from one run to another.
- But they are the same every time, as far as the values of **sex** and **athlete** are concerned.

The basic idea of blocking is to ensure that—at least with respect to a few variables whose values you can record in advance—the two groups will be similar. Naturally, the variables for which you would choose to block should be variables that are strongly associated with the response variable. If you block for them, then the treatment groups are the same with respect to these variables and you will have reduced the number of confounding variables you need to worry about.

As an example of a randomized block design, consider the `saltmarsh` data frame:

```
data(saltmarsh)
View(saltmarsh)
help(saltmarsh)
```

The data frame contains the results of an experiment to study the effects of salinity levels in soil on the growth of plants. Researchers had access to four different fields of equal size at an agricultural research station. They divided each field into six plots of equal size and treated each plot with one of six concentrations of salt, ranging from 10 to 35 parts per million. They allowed plants to grow on the fields for a period of time, and then measured the total biomass of each plot.

In this experiment the subjects are the plots—24 of them, all told, six in each field. The explanatory variable is the salinity level **salt**, and the response variable is **biomass**. Since the four fields probably contained different types of soil and were subject to differing environmental conditions, the field to which a plot belongs could have a considerable effect on plant growth: hence it was considered important to block for "field", and so each field was subdivided into six equal plots. Within each field, the six levels of salinity were assigned randomly to the six plots in that field. In the data frame, the field-variable is recorded as **block**.

You can check directly that blocking was indeed performed:

```
xtabs(~salt+block,data=saltmarsh)
```

```
##     block
## salt Field1 Field2 Field3 Field4
##   10     1     1     1     1
##   15     1     1     1     1
##   20     1     1     1     1
##   25     1     1     1     1
##   30     1     1     1     1
##   35     1     1     1     1
```

Sure enough: the six treatments groups (the salt levels) are distributed identically with respect to **block**.

In order to address the Research Question, we note that the explanatory and response variables are both numerical, so a scatterplot could be an appropriate graphical tool. Figure [Biomass and Salinity] shows the scatterplot, and indeed it does appear that there is a negative relationship between salinity-level and biomass.

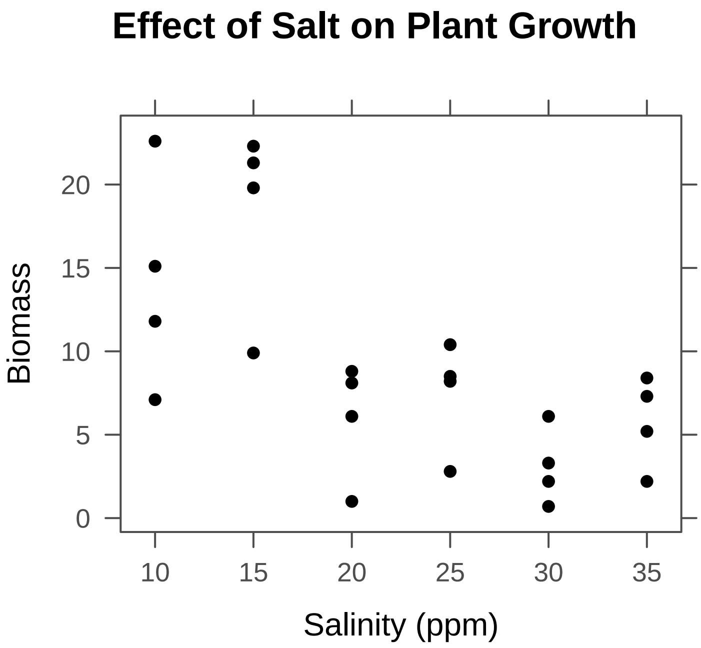


Figure 6.2: Biomass and Salinity. The higher the concentration of salt, the lower the biomass in the plot.

6.3.3 Matched Pair Designs

A *matched pair* design is really an extreme form of blocking.

Suppose that you want to form two treatment groups: Group A and Group B. You pair up your subjects so that the members of each pair are as similar as possible to one another. Each pair is a considered to be a “block.” Then for each pair, one member is assigned randomly to Group A, and the other is assigned to Group B.

Here is a well-known example of a matched pair design. Suppose you want to see which type of shoe sole, type A or type B, wears out more quickly, and say that you have 20 people who agree to participate in your experiment. Each person has a pair of feet, and these forty feet constitute the “subjects” in the experiment. The two members of any given pair of feet are pretty similar with respect how much wear they will put on a shoe—after all, both members of the pair belong to the same person, and hence are engaged in pretty nearly the same set of daily activities—so you randomly assign 10 of the subjects to wear a type-A sole on their left foot and a type-B sole on their right foot. The other ten will do the reverse. In this way, the members of each pair of feet has been assigned randomly to one of the treatment groups! (Randomization is especially important here, because most people tend to favor one foot over the other.)

6.3.4 Repeated-Measure Designs

For the purposes of this course, an experiment is said to employ a *repeated-measure* design if the researchers make two or more similar measurements on each individual in the study, with a view to studying the differences between these measures.

As an example, consider the data frame `labels`:

```
data(labels)
View(labels)
help(labels)
```

The `labels` data frame contains the results of an experiment conducted at Georgetown College by two students. The students wanted to investigate whether one's belief about the price of a brand item would, in and of itself, affect one's perception of its quality. Thirty fellow students served as subjects. One at a time, each subject was brought into a room where two jars of peanut butter were placed on a table. One jar had the well-known "Jiff" label, and the other had the Great Value label. (Great Value is the Wal-Mart brand, and is considered "cheaper.") The subject tasted each peanut butter and rated it on a scale of 1 to 10. Unknown to the subjects, both jars contained the exactly the same peanut butter: the Great Value brand.

In this experiment, the two similar measurements are the rating given to the jar of peanut butter that is labeled Jiff, and the rating given to the jar labeled Great Value. What makes this repeated-measures study an experiment is not that researchers assigned the values of some explanatory variable to subjects, but rather that they manipulated the situation so the subjects believed that the experiment was a taste test to compare two different types of peanut butter, when it fact it was designed to compare their reactions to different labels on the SAME peanut butter.

(**Note**: Some people like to say that every experiment involves assigning the values of the X variable to subjects. Such persons would maintain that in this experiment, the X variable is the kind of label (Jiff or Great Value), and that the experimenters were able to assign BOTH values of X to each subject. When you think about it this way, the two treatment groups are quite similar indeed, because they are composed of exactly the same set of subjects!)

When measures are repeated, there is always the possibility that the second measurement might depend in some way on the first. In this case, the rating given to the peanut butter in the second jar tasted might be affected by the fact that the subject has recently tasted some peanut butter (from the first jar). In order to wipe out the effects of any such dependence, it is good practice to randomly vary the order in which the measurements are made. In this experiment, for example, you might determine in advance for each subject—by coin toss perhaps—whether he or she is to taste from the Jiff-labeled jar first or from the Great Value-labeled jar first. (It is also good practice to record the order of tasting, though unfortunately the researchers did not do this.)

We would like to see whether subjects tended to rate the jar labeled with the more expensive brand as more highly, on the whole, so we are interested in the difference of the repeated measures

```
diff <- labels$jiffrating - labels$greatvaluerating
```

We can look at the difference in a variety of ways. Numerically:

```
favstats(~diff)
```

```
##  min Q1 median Q3 max     mean       sd  n missing
##   -5  1    2.5  4   8 2.366667 2.809876 30       0
```

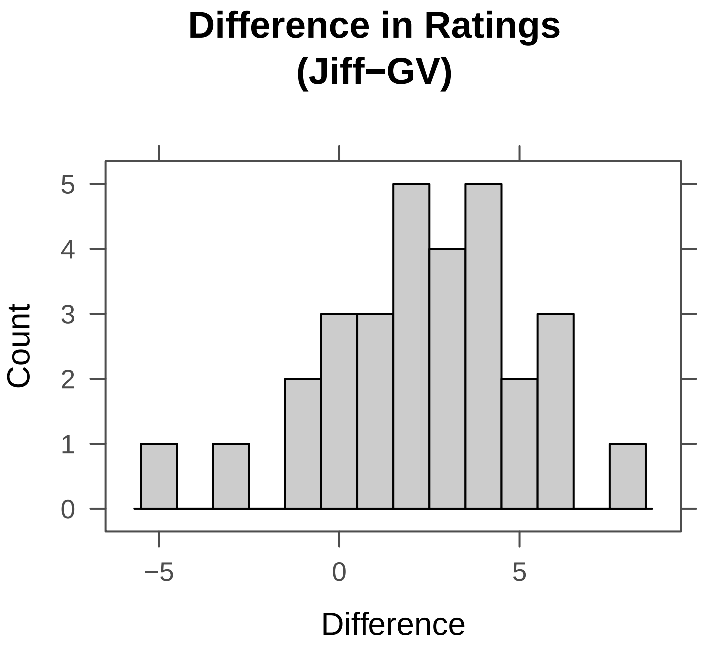


Figure 6.3: Ratings Difference. Most of the differences are positive, indicating that subjects usually rated the Jiff-labeled peanut butter more highly.

We see that the subjects rated the Jiff-labeled peanut butter an average of 2.37 points higher than they rated the peanut butter from the Great-Value-labeled jar. A graphical approach is shown in Figure [Ratings Difference].

A repeated-measures design appears to be the gold-standard among experiments, since your treatment groups are – in some sense – the same group. But is it always feasible to perform a repeated-measures experiment?

Consider, for example, the following (thankfully hypothetical) Knife-or-Gun study:

```
data(knifeorgunblock)
View(knifeorgunblock)
help(knifeorgunblock))
```

The Research Question was:

> *What will make you yell louder: being killed with a knife or being killed with a gun?*

Would it have been feasible to perform this experiment with a repeated measures design? Probably not: once a subject is slain by one method, be it Knife or Gun, he/she probably will make no noise at all whilst being subjected to the subsequent method of slaying! As a result, we would have only one useful measurement on each subject.

On a more serious note:

> Consider the `attitudes` study, in which a completely randomized design was employed. Suppose we had tried to run the study as a repeated measures experiment, giving each subject all of the different kinds of form, perhaps in some random order determined by a coin toss: what would have happened? Would the repeated-measures design have been any improvement over a completely randomized design?

6.4 Considerations About Experiments

6.4.1 Reproducibility

When you do any sort of scientific work, you want your results to be *reproducible* as much as possible, so that others may see exactly how you obtained these results, and thereby check them. Accordingly, when you are actually doing an experiment, even your randomization needs to be done in such a way that another person could reproduce it entirely. In R this is accomplished by the `set.seed()` function.

Consider the following small block of code:

```
set.seed(314159)
RandomExp(SmallExp,sizes=c(8,8),
          groups=c("Program.A","Program.B"),
          blocks=c("sex","athlete"))
```

You will recognize the call to `RandomExp()` as the same call you performed earlier in the chapter, in order to randomize subjects into treatment groups after blocking for **sex** and **athlete**. This time, however, the call to `RandomExp()` is preceded by a call to `set.seed()`. This call "starts out" the randomization from a fixed point in R—called a *seed*—so that no matter how many times you run the above code chunk you will get the same results each time.

The seed is the integer that is supplied to the `set.seed()` function. This time, the seed was 314159: the first few digits of the number π. In general, you should choose a seed that stands for something familiar, so that others who examine your code can see that you did not just keep on trying out different seeds until you got the sort of results you wanted.

6.4.2 Subjects and the Population

So far we have highlighted a major advantage of experiments over observational studies, namely that in an experiment you have the opportunity to reduce or eliminate the effect of confounding factors by the appropriate use of such techniques as randomization, blocking, matched pairs, and so on.

However, researchers wishing to perform experiments on human subjects face a problem that does not crop up—at least not in such an extreme form—in the course of conducting an observational study. The problem is one of *consent*.

You can't just *compel* people do one thing or another. You can't grab people at random off the street and assign a value of an explanatory variable to them:

> "Hey, you, there! Yes, you! Come over here! You are hereby inducted into my experiment. Um, wait a minute while I flip this coin. OK, it's settled: you will ______ [insert `smoke` if toss was heads, `not smoke` if toss was Tails] for the next twenty years! Got it?"

Since people must consent to be in an experiment, the set of people who are willing to be in the experiment may be quite different from the population at large. It is well known, for instance, that people who volunteer for medical experiments tend to be more educated and to have higher incomes than those who refuse to participate in such experiments. Anyone who would consent to be part of the Knife or Gun experiment (see the previous section) would have to be stark raving mad—surely that makes them different from the general population.

In general:

> If the subjects differ from the general population in ways that are associated with how they respond to the different treatments available in the experiment, then the results of the experiment can at most be said to apply to the subjects themselves.

In such a case we must exercise a great deal of caution when applying our results to some larger population.

6.4.3 Statistical Significance

Even when the subjects are quite different from the population, and we are reduced to wondering whether the the treatments made a difference *for those subjects only*, statistical significance is still an issue. Consider once again the Knife or Gun experiment.

Who did yell louder, in the experiment: the subjects who were killed by knife, or the subjects who were killed by gun? Let's see:

```
favstats(volume~means,data=knifeorgunblock)
```

```
##   .group  min     Q1 median     Q3  max  mean        sd  n missing
## 1    gun 40.3 45.925  53.25 59.925 64.5 53.00  8.761152 10       0
## 2  knife 61.2 65.200  72.00 77.125 90.9 73.13 10.071528 10       0
```

A nice graphical tool for small data sets is a strip plot, which we combine here with a violin plot (see Figure [Knife or Gun?]):

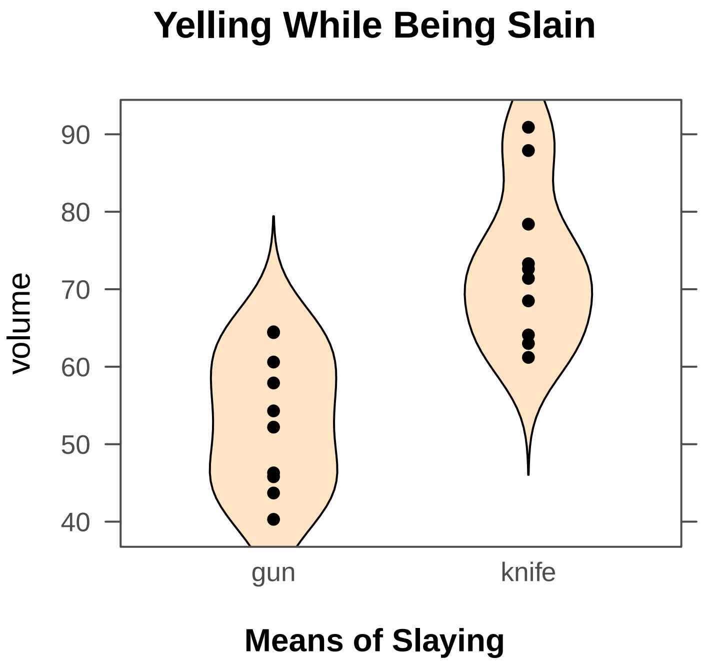


Figure 6.4: Knife or Gun? Apparently those slain by Knife yell louder!

Hmm, it appears the subjects who were slain by knife yelled louder: the volume-points, and their violin, is somewhat higher than the points and violin for those slain by a gun.

But are the results statistically significant? By this question we do NOT mean:

> "Could the difference reasonably be ascribed to chance variation in the process of sampling subjects from the population?"

These subjects aren't a random sample from the population at all. Maybe they were the ONLY 20 people in the world who would agree to be part of the experiment, so there was no chance at all involved in their selection.

Instead, the question of statistical significance addresses the randomness that *was* at play in the experiment: namely, the random assignment of subjects to Knife or Gun group.

So we are REALLY asking:

> “Is the observed difference in dying screams due to the means of slaying, or is is reasonable to believe that in the random process of assigning subjects to groups, a lot of naturally loud yellers just happened to get into the Knife group?”

In any experiment where randomization was employed, the question of statistical significance is meaningful, and it can always be thought of as asking:

> “Do the results indicate that X causes Y, or is reasonable to ascribe the observed difference between the treatment groups to chance variation in the assignment of the subjects to their groups?”

See the GeekNotes for some preliminary assessment of statistical significance in some of the experiments discussed in this chapter.

6.5 Terminology for Experiments

Here follows a list of technical terms, associated with experiments and observational studies, that we would like you to know. If a term has already been introduced, then it is simply listed below with its definition. If it has not been discussed yet, then we offer a brief explanation and/or examples.

Observational Study In an *observational study* researchers simply observe or question the subjects. In particular, they measure the values of the explanatory variable X and measure the values of the response variable Y, for each subject.

Experiment In an *experiment* researchers manipulate something and observe the effects of the manipulation on a response variable.

Most commonly, the manipulation consists in assigning the values of an explanatory variable of X to the subjects.

Confounding Variable In a study with an explanatory variable X and a response variable Y, the variable Z is called a *confounding* variable if it meets the following three conditions:

1. It is a third variable (different from X and different from Y);
2. It is associated with X, but not caused by X.
3. It is a causal factor in Y (is at least part of the cause of Y)

Experimental Units The *experimental units* (also called the *individuals*) are the subjects in an experiment.

Subjects do not have to be human beings, or animals. Recall, for example, the matched-pair shoe experiment in which the subjects were the 40 feet of the twenty human participants!

Treatment Groups The *treatment groups* of an experiment are the groups into which the subjects are divided.

Control Group If one group in an experiment is not treated in any special way, or is present for comparative purposes, then this group is often called the *control group.*

Single-Blind Experiment If the subjects in an experiment do not know which treatment group they are in, then the experiment is said to be *sinble-blinded*.

If the people who measure the response variable do not know which groups the subjects are in, then the experiment is also-said to be *single-blinded*.

Placebo A *placebo* is an inert substance given to subjects in the control group. It resembles the substances given to subjects in the other treatment groups, and thus allows the experiment to be single-blinded.

Double-Blinded Experiment If neither the subjects nor the people responsible for measuring the response variable know the group assignments of the subjects, then the experiment is said to be *double-blinded*.

Blinding is useful as a means of reducing bias. If subjects know their treatment group, then this knowledge can affect their behavior in ways that have a bearing on the response variable. For example, in a study on the effectiveness of a vaccine against a particular diseases, subjects who know they are receiving the vaccine may behave more recklessly than subjects who know that they are in a control group (not receiving any vaccine). These differences in behavior could bias the study against the vaccine.

The people who are responsible for measuring the response variable could “skew” the measurements in one direction or another, if they have some stake in the outcome of the experiment. Keeping these researchers in the dark as to the group assignment of subjects with whom they work can reduce bias that arises from a desire to have the experiment “succeed”.

Replication An experiment is said to involve *replication* if each the treatment group contains more than one subject. (The more subjects there are in each group, the more replication the experiment has.)

Double-Dummy Design A *double-dummy* design is a procedure to blind an experiment, even when the treatments don’t resemble each other at all.

A classic example of a double-dummy design is an experiment to compare two methods of delivering nicotine to the bodies of people who desire two quit smoking. One method involves wearing a patch that delivers nicotine through the skin, and the other involves chewing gum that contains nicotine. In a double-dummy experiment:

- Members of the Patch Group wear a real nicotine patch, and chew a gum that looks and tastes like real nicotine gum but which actually contains no nicotine;
- Members of the Gum Group chew real nicotine gum, and wear a patch that looks the same as a nicotine patch but which delivers no nicotine.

In neither group would subjects be able to tell which group – Patch or Gum – they are in.

6.6 Thoughts on R

6.6.1 New R-functions

Just one new R-function to learn: `RandomExp()`, for completely randomized designs and for blocking. Here are some examples of its use. Suppose you have a data frame called `MyData`, that contains the list of subjects, and possibly some other variables of interest for which you might like to block.

For a completely randomized design into two groups with groups sizes n_1 and n_2, with group names “Group.A” and “Group.B”, use:

```
RandomExp(MyData,sizes=c(n1,n2),
          groups=c("Group.A","Group.B")
```

For three groups (“Group.1”, “Group.2” and “Group.3”) with sizes n_1, n_2 and n_3, blocking with respect to the variable **Var**, use:

```
RandomExp(MyData,sizes=c(n1,n2,n3),
    groups=c("Group.1","Group.2","Group.3"),
    block="Var")
```

To block for two variables **Var1** and **Var2** at once, use:

```
RandomExp(MyData,sizes=c(n1,n2,n3),
    groups=c("Group.1","Group.2","Group.3"),
    block=c("Var1","Var2"))
```

Make sure that the group sizes sum to the number of rows in the data frame, and that the blocking variables are factors.

6.6.2 “Book-Chapter” Notation

Sometimes you have to manipulate variables in a data frame directly, outside of the context of a function that takes formula-data input. In such a case, you can use what we call “book-and-chapter” notation to help R locate the variables of interest.

If variable **Var** is in data frame `MyData`, then R can locate it if you use a `$`-sign and write it as:

```
    Mydata$Var
```

For example, to compute the difference of two variables in `MyData`, you could use:

```
diff <- MyData$Var1 - MyData$Var2
```

The data frame `MyData` is like a book, and the `$`-sign instructs R to look inside the “book” `MyData` at “chapters” **Var1** and **Var2**.

Chapter 7

Basic Probability

7.1 Introduction

Statistical inference is the process of forming judgments about a population based on information gathered from a sample of that population. Our goal in this chapter is to describe populations and samples using the language of probability.

7.2 Probability

By **probability**, we mean some number between 0 and 1 that describes the *likelihood*, or *chance*, that an event occurs. Probability is a way to quantify *uncertainty.*

There are several different interpretations of the word **probability** that come from understanding how the numbers are generated.

7.2.1 Subjective Probability

Subjective Probabilbities Probabilities that are assigned or postulated based on a personal belief that an outcome will occur are called *subjective probabilities.*

Example: A surgeon, who is performing a surgery for the very first time, tells his patient that he feels that the probability that it will be successful is 0.99. This probability is *subjective* because it is based entirely on the surgeon's belief.

In this class, we will not be very interested in subjective probabilities since they are not supported by data.

7.2.2 Theoretical Probability

To discuss theoretical probabilities, let's first state some definitions.

Sample Space The *sample space* is the set of all possible outcomes of ane experiment.

Event An *event* is a subset of the sample space. In other words, an event is a collection of outcomes.

Suppose that all outcomes in a sample space are equally likely - i.e. they have the same chance of occurring. Then, the probability of an event is the number of outcomes in the event divided by the total number of outcomes in the sample space. In symbols, this is

$$\text{P(event)} = \frac{\text{number of outcomes in the event}}{\text{total number of outcomes in the sample space}}.$$

Example: Consider tossing a fair coin. There are two possible outcomes - tossing a head or tossing a tail. The sample space is the set of all possible outcomes, so the sample space is {H, T}. Since the coin is *fair*, the outcomes are equally likely. The probability of the event **toss a head** is $\frac{1}{2} = 0.5$. In symbols, we can write this as P(toss a head) = 0.5 or, for short, P(H) = 0.5.

Example: Consider tossing two fair coins. The sample space is {HH, HT, TH, TT}. Here, HH represents both coins landing heads up and TH represents the first coin landing tails and second coin landing H. Since the coin is fair, each of the four outcomes is equally likely. We can compute various probabilities.

- What is P(HH)?

  Answer: P(HH) $= \frac{1}{4} = 0.25$. This is the probability of tossing 2 heads in two tosses of a fair coin. One of the four events in the sample space corresponds to 2 heads being tossed - HH.

- What is the probability of tossing exactly one head?

  Answer: P(toss exactly one head) = P(HT or TH) $= \frac{2}{4} = 0.5$. This is the probability of tossing exactly 1 heads in two tosses of a fair coin. Two of the four events in the sample space corresponds to exactly 1 head being tossed - HT, TH.

- What is the probability of getting a head on the first toss?

  Answer: P(toss a head on Toss 1) = P(HT or HH) $= \frac{2}{4} = 0.5$. This is the probability of tossing a head on the first toss in two tosses of a fair coin. Two of the four events in the sample space corresponds to this - HT, HH.

7.2.3 Long-run Frequency Probability

The final type of probability we should discuss is really an *approximation* to the theoretical probability.

Long-Run Frequency Probability A *long-run frequency probability* comes from knowing the proportion of times that the event occurs when the experiment is performed over and over again. This is an approximation to a theoretical probability.

Example: Suppose you toss a fair coin 1000 times and it comes up as a tail 502 times. Then the *long run frequency* is $\frac{502}{1000} = 0.502$. We know that the theoretial (true) probability of a fair coin landing heads is 0.50. The long run frequency is an approximation this. The more times the coin is tossed, the better the long run frequency does at approximating the true probability of landing tails.

7.3 What is a Random Variable?

Here is a very important definition:

Random Variable A *random variable* is a variable whose value is the outcome of a chance experiment.

Calling it a *variable* may be somewhat confusing! It is actually a function on the sample space of an experiment. It can take on a set of possible different values. We make a distinction between whether or not we know, or have observed, the value of the random variable. Once observed, the random variable is known. Prior to being observed, it is full of potential - it can take on any value in the set of possible values. However, some of the possible values may be more likely than others. There is an associated *probability* that the random variable is equal to some value (or is in some range of values).

7.3.1 Notation

Letters near the end of the alphabet are typically used to symbolize a random variable. If the random variable has not yet been observed, we use uppercase letters, such as X, Y, and Z. If the value of the random variable is known, we use lowercase letters, such as x, y, and z, to refer to the random variable.

Just as numeric data is either discrete or continuous, random variables are classified as either **discrete** or **continuous**. This is determined by what kinds of numbers are in the set of possible values that the random variable can assume.

7.4 Discrete Random Variables

Discrete random Variable A *discrete random variable* is a random variable whose possible values come from a set of discrete numbers.

Example: Suppose that we toss a fair coin two times.

- The *sample space* is {HH, HT, TH, TT}.
- A possible random variable associated with this experiment is $X =$ number of heads tossed.
- The set of possible values that X can assume is {0, 1, 2}. Since this is a set of discrete values, we classify X as a **discrete random variable**.
- Recall that there is an associated **probability** that the random variable is equal to some value. For this example, it is more likely that $X = 1$ than $X = 0$ or $X = 2$. This can be seen by looking at the probabilities:
  - P($X = 0$) $= 1/4 = 0.25$. Keep in mind that X is the random variable that represents the "number of heads tossed in two tosses of a fair coin". This is just another way to represent the P(TT).
  - P($X = 1$) $= 2/4 = 0.5$. This is just another way to represent P(HT or TH).
  - P($X = 2$) $= 1/4 = 0.25$. This is just another way to represent the P(HH).

7.4.1 Probability Distribution Functions for Discrete Random Variables

It is nice to display the probabilities associated with a random variable in a table or graph.

Probability Distribution Function (pdf) The *probability distribution function* (pdf, for short) for a discrete random variable X is a function that assigns probabilities to the possible values of X. It may be viewed in the form of a table or a histogram.

For the two-coin toss example, the pdf for the random variable $X =$ number of heads tossed looks like:

x	0	1	2
$P(X = x)$	0.25	0.50	0.25

Table 7.1: Probability distribution function for two-coin toss

From the pdf, we can easily see that the probabilities of all possible values of a discrete random variable add up to 1.

$$\sum \mathrm{P}(X = x) = 1$$

We can use a histogram (See Figure[Two Coin Toss pdf]) to visualize the pdf where

- the possible outcomes for the random variable X are placed on the horizontal axis,
- the probabilities for the possible outcomes are placed on the vertical axis, and
- a bar is centered on each possible value and its height is equal to the probability of that value.

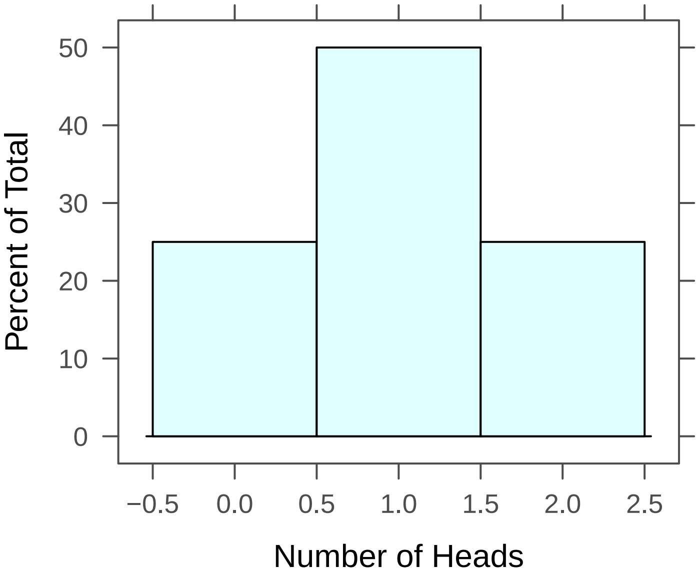


Figure 7.1: Two Coin Toss pdf: The probability distribution function, in histogram form, for a two coin toss.

We can use the pdf and the histogram of the two coin toss example to calculate probabilities.

Example: $P(X = 1) = 0.5$. This is the probability of tossing *exactly* 1 head in two tosses of a fair coin. This corresponds to the area of the pink rectangle in the histogram shown below. See Figure[Exactly 1 Head].

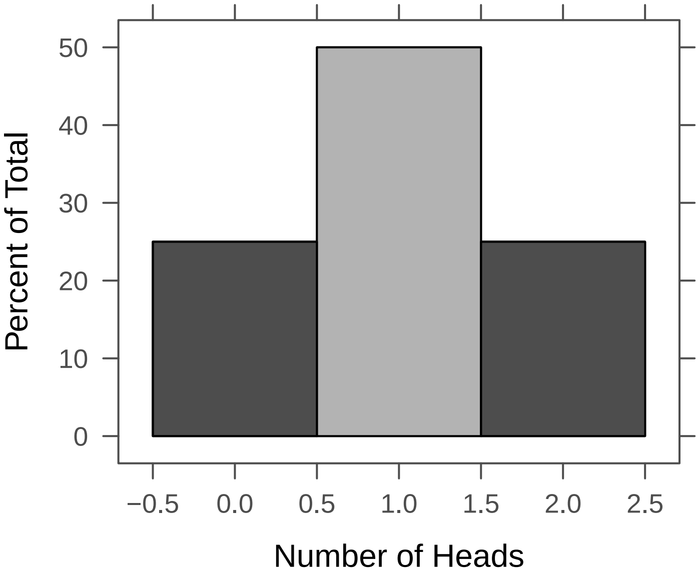


Figure 7.2: Exactly 1 Head: The shaded region of the pdf represents the probability of tossing exactly 1 head in two tosses of a fair coin.

Example: $P(X \geq 1) = 0.5 + 0.25 = 0.75$. This is the probability of tossing *at least* 1 head in two tosses of a fair coin. *At least* 1 head means that you toss 1 head **OR** 2 heads. In two tosses of a fair coin, the chance that you will toss 1 head is 0.5 and the chance that you will toss 2 heads is 0.25. So the likelihood that you will toss 1 **OR** 2 heads is 0.5 + 0.25 = 0.75. This corresponds to summing the area of the pink rectangles in the histogram shown below. See Figure[At Least 1 Head].

Example: $P(X \leq 1) = 0.25 + 0.5 = 0.75$. This is the probability of tossing *at most* 1 head in two tosses of a fair coin. *At most* one head means that we toss 0 heads **OR** 1 head. In two tosses of a fair coin, the chance that you will toss 0 heads is 0.25 and the chance that you will toss 1 head is 0.5. So the likelihood that you will toss 0 **OR** 1 head is 0.25 + 0.5 = 0.75. This is called a **cumulative probability** - it's the probability that a random variable is less than or equal to some value. This corresponds to summing the area of the pink rectangles in the histogram shown below. See Figure[At Most 1 Head].

7.4.2 Expectation of Discrete Random Variables

So far, we have talked about the possible values that a random variable might be. We have also talked about the probability, or chance, that the random variable equals one of those values. Some of the possible values for a random variable are *more likely* than others. This prompts the question: In the long run, what value do we **expect** the random variable to be?

Let's return to our two coin toss and recap what we know.

- If our random variable X = number of heads tossed in two tosses of a fair coin, then the *possible* values for X are $\{0, 1, 2\}$.
- We've also found that X is more likely to be 1 than it is to be 0 or 2.
- So, if we repeated the two coin toss over and over again, we would *expect* to see $X = 1$ more often than $X = 0$ or $X = 2$.

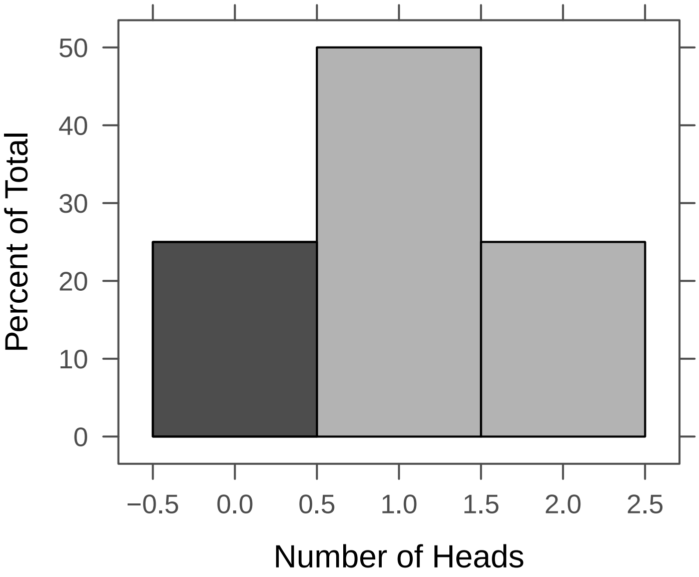


Figure 7.3: At Least 1 Head: The shaded region of the pdf represents the probability of tossing at least 1 head in two tosses of a fair coin.

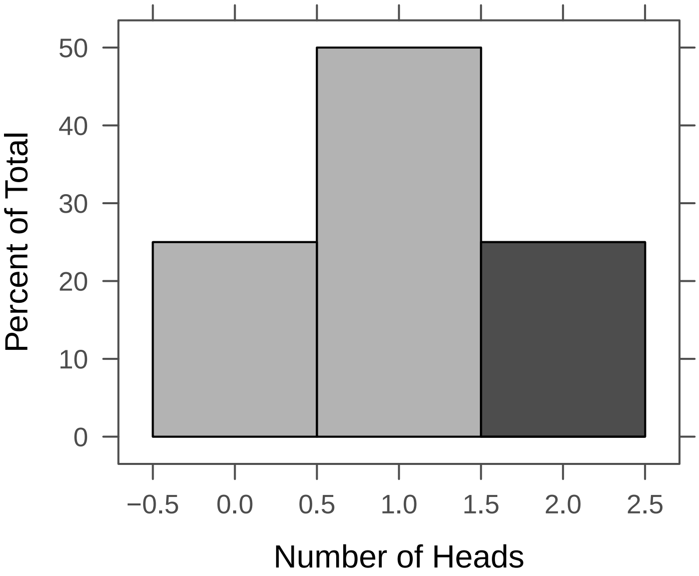


Figure 7.4: At Most 1 Head: The shaded region of the pdf represents the probability of tossing at most 1 head in two tosses of a fair coin.

Expected Value The *expected value* of a random variable X is the value of X that we would expect to see if we repeated our experiment many times.

Expected value is like an *average*, of sorts. In fact, it's a *weighted* average. Each possible value of X is weighted by the likelihood that X is that value. The expected value of X, $\mathrm{E}(X)$, can be calculated by the formula.

$$\mathrm{E}(X) = \sum x \cdot \mathrm{P}(X = x)$$

Step by step, this is:

- Multiply every possible value for X by it's corresponding probability.
- Sum these products.

For our two coin toss example, the expected value of X can be found with the following R code.

```
x<-c(0,1,2) #Possible values for X.
prob.x<-c(0.25,0.5,0.25) #Corresponding probabilities
EV.X<-sum(x*prob.x) #Expected Value of X.
EV.X
```

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

It turns out that $\mathrm{E}(X) = 1$, exactly as we had thought!

7.4.3 Standard Deviation of Discrete Random Variables

Another measure that we are interested in is the *standard deviation* for random variables.

Standard Deviation The *standard deviation* is a measurement of how much the random variable random variable can be expected to differ from its expected value.

In other words, if we repeat an experiment over and over, the standard deviation of a random variable is the average distance that it will fall from its expected value. Of course, this average is *weighted* by the probabilities. The standard deviation of X, $\mathrm{SD}(X)$, can be calculated by the formula.

$$\mathrm{SD}(X) = \sqrt{\sum \left(x - \mathrm{E}(X)\right)^2 \cdot \mathrm{P}(X = x)}$$

Step by step, this is:

- Calculate the difference between each of the possible values for X and the expected value for X.

- Square these differences.

- Multiply the squared differences by the probability that X equals that value.
- Sum these products.
- Take the square root.

For our two coin toss example, the standard deviation of X can be found with the following R code.

```
x<-c(0,1,2) #Possible values for X.
prob.x<-c(0.25,0.5,0.25) #Corresponding probabilities.
EV.X<-sum(x*prob.x) #Expected value of X.
SD.X<-sqrt(sum((x-EV.X)^2*prob.x)) #Standard deviation of X.
SD.X
```

```
## [1] 0.7071068
```

So, if we toss a fair coin two times, the number of heads that we would expect to toss is 1. On average, the number of heads will differ from 1 by 0.7071068.

7.4.4 Using EV and SD to Evaluate a Game

The **expected value** and **standard deviation** work together to describe how a random variable is *likely* to turn out. This comes in handy for situations like the following example.

Example: Suppose that you decide to invest $100 in a scheme that will hopefully make you some money. You have 2 investment options, Plan 1 and Plan 2. Let the random variable X = net gain from Plan 1 and the random variable Y = net gain from Plan 2. The probability distribution functions for the plans are shown below.

Plan 1:

x	$5,000	$1,000	$0
$P(X = x)$	0.001	0.005	0.994

Table 7.2: Probability Distribution Function for Investment Plan 1

Plan 2:

y	$20	$10	$4
$P(Y = y)$	0.30	0.20	0.50

Table 7.3: Probability Distribution Function for Investment Plan 2

Which plan do you choose? Plan 1 is the riskier plan. You have a chance to get rich off of your investment, but it's also very likely that will gain nothing! On the other hand, Plan 2 is a safe plan. You won't get rich, but you are *guaranteed* to gain some money.

Putting personal desires aside, let's think about this mathematically by calculating the expected value of each plan.

Coding Tip: Since we want to calculate the expected value for 2 different random variables, it's a good idea to name these variables something slightly different. This helps to keep track of which random variable we're talking about.

```
#EV Plan 1
x1 = c(5000, 1000, 0)
prob.x1 = c(0.001, 0.005, 0.994)
```

```r
EV.X1 = sum(x1 * prob.x1)
EV.X1
```

```
## [1] 10
```

```r
#EV Plan 2
x2 = c(20, 10, 4)
prob.x2 = c(0.30, 0.20, 0.50)
EV.X2 = sum(x2 * prob.x2)
EV.X2
```

```
## [1] 10
```

The expected amount of money you gain from investing in Plan 1 is $10. So, if you could invest your $100 in Plan 1 over and over again, on average you would walk away with a net gain of $10.

The expected gain from investing in Plan 2 is also $10. So, if you could invest your $100 in Plan 2 over and over again, on average you would also have a net gain of $10.

Based on expected values, it seems that the plans are the same! Before you get too excited and invest all of your money in the plan with the highest possible reward, let's compare the standard deviations. This will give us an idea about how much the net gain for each plan will differ, on average, from the expected value.

```r
#SD Plan 1
x1 = c(5000, 1000, 0)
prob.x1 = c(0.001, 0.005, 0.994)
EV.X1 = sum(x1 * prob.x1)
SD.X1=sqrt(sum((x1-EV.X1)^2*prob.x1))
SD.X1
```

```
## [1] 172.9162
```

```r
#SD Plan 2
x2 = c(20, 10, 4)
prob.x2 = c(0.30, 0.20, 0.50)
EV.X2 = sum(x2 * prob.x2)
SD.X2=sqrt(sum((x2-EV.X2)^2*prob.x2))
SD.X2
```

```
## [1] 6.928203
```

Now, we have a better idea of how X and Y are likely to turn out!

- The net gain for Plan 1, X, has an expected value of $10 with a standard deviation of $172.92.
- The net gain for Plan 2, Y, has an expected value of $10 with a standard deviation of $6.93.

Knowing all of this, which plan would you choose?

7.4.5 Independence

Independence Two events are considered *independent* if the outcome of one event does not affect the outcome of the other.

Suppose that we draw two cards from a standard 52-card deck. We want to know the probability of drawing a Jack.

If we draw the cards *with replacement*, then we draw a card from the deck, record the result, and replace the card before we draw again. In other words, we return the deck to it's original state before drawing the second card. This ensures that the probability of drawing a Jack on the first try is the same as the probability of drawing a Jack on the second try.

P(first draw is a Jack) $= \frac{4}{52}$ and P(second draw is a Jack) $= \frac{4}{52}$. The outcome of our first try did not affect the outcome of our second try. Drawing with replacement makes our two draws *independent*.

On the other hand, if we draw the cards *without replacement*, then we do not return the first card to the deck before drawing the second card. The first draw is as before, so P(first draw is a Jack) $= \frac{4}{52}$. However, now the state of the deck has been changed, so the chance of drawing a Jack on the second try *depends* on what happened on the first try.

If our first draw was a Jack, then the deck is left with only 3 Jacks out of 51 total cards. So, P(second draw is a Jack given that the first draw was a Jack) $= \frac{3}{51}$.

If our first draw was not a Jack, then the deck is left with 4 Jacks out of 51 total cards. So, P(second draw is a Jack given that the first draw was not a Jack) $= \frac{4}{51}$. Drawing without replacement makes our two draws *dependent*.

7.4.6 A Special Discrete Random Variable: Binomial Random Variable

The coin toss example that we have been looking at is actually an example of a special type of discrete random variable called a *binomial random variable*.

Binomial Random Variable A *binomial random variable* is a random variable that counts how often a particular event occurs in a specified number of tries.

To be a *binomial random variable*, a random variable must meet all of the following conditions:

1. There are a specified number of tries. This is sometimes referred to as "size".
2. On each try, the event of interest either occurs or it does not. In other words, we have a *success* or a *failure* on each try.
3. The probability of success is the same on each try. (We will denote the probability of success as p. In each trial, you will either succeed or you will fail, so the probability of failure will be the *complement* of a success: $1 - p$.)
4. The tries are independent of one another.

Let's revisit the two coin toss example using the language of a binomial random variable. Suppose we toss a fair coin twice and let the random variable X = number of heads tossed. This is a binomial random variable because it fits the 4 conditions:

- Since we are tossing the coin twice, the specified number of trials, or size, is $n = 2$.
- Each toss is either heads (*success*) or tails (*failure*). Since it is a fair coin, the probability of success is $p = 0.5$ and the probability of failure is $1 - p = 1 - 0.5 = 0.5$.
- The chance of tossing a head is the same on each toss of the coin.

- The trials are **independent**. What you tossed on the first toss does not affect what happens on the second toss.

Let's try another example.

Example: Suppose we toss a fair coin 5 times. Let the random variable $X =$ number of tails tossed. This is a binomial random variable:

- Since we are tossing the coin five time, the specified number of trials, or size, is $n = 5$.
- Each toss is either heads (*success*) or tails (*failure*). Since it is a fair coin, the probability of success is $p = 0.5$ and the probability of failure is $1 - p = 1 - 0.5 = 0.5$.
- The chance of tossing a head is the same on each toss of the coin.
- The trials are **independent**.

Listing out the sample space, and calculating probabilities from it, for a five-coin toss would be rather tedious. There is a function in `tigerstats` that calculates these probabilities (and produces graphs of the pdf) for us. Let's calculate a few probabilities using this function `pbinomGC`.

7.4.6.1 P$(X > 2)$

We will start by calculating P$(X > 2)$. To use `pbinomGC` we need to supply several inputs:

- The *bound* - for our example, the bound is 2
- The *region* - this is given by the sign in the probability. For our example, we are dealing with >, so our region is `"above"`. Other options for the region are `"above"`, `"below"`, and `"outside"`.
- The *size* - for our example, the number of trials is $n = 5$.
- The *prob*ability of success - for our example, $p = 0.5$.
- Whether or not you want to display a graph of the probability. The default of this function is to not produce a graph. If you would like to see one, you should include `graph=TRUE`.

Let's compute P$(X > 2)$ and view a graph of the probability. See Figure[Binomial Greater Than].

```
pbinomGC(2,region="above",size=5,prob=0.5,graph=TRUE)
```

```
## [1] 0.5
```

To just find the probability without producing the graph:

```
pbinomGC(2,region="above",size=5,prob=0.5)
```

```
## [1] 0.5
```

So, in five tosses of a fair coin, the probability of tossing more than 2 heads is P$(X > 2) = 0.5$.

7.4.6.2 P$(X \geq 2)$

Now, suppose you want to know $P(X \geq 2)$. Since the possible values for X are $\{0,1,2, 3, 4, 5\}$, $X \geq 2$ is the same as $X > 1$. The following code will calculate $P(X > 1)$. Look at the graph in Figure[Binomial Greater Than or Equal].

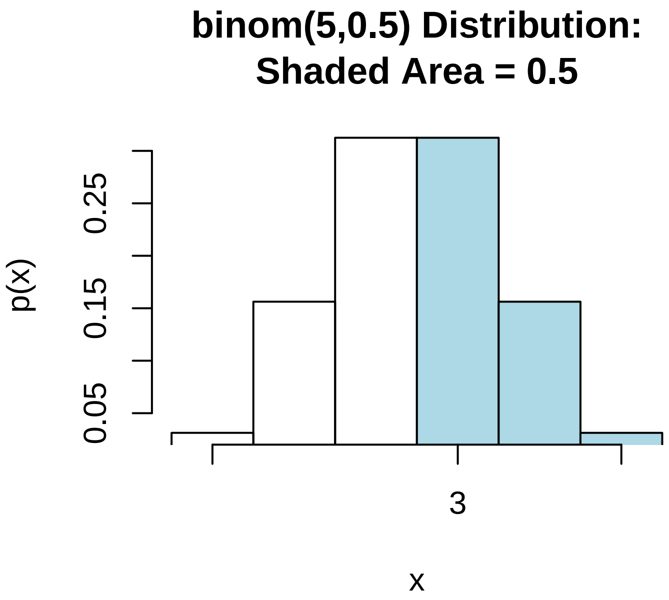


Figure 7.5: Binomial Greater Than: Shaded region represents the probability that more than 2 heads are tossed in 5 tosses of a fair coin.

```
pbinomGC(1, region="above",size=5,prob=0.5, graph=TRUE)
```

```
## [1] 0.8125
```

Thus, the probability of tossing at least 2 heads is $\mathrm{P}(X \geq 2) = 0.8125$

7.4.6.3 $\mathbf{P}(X \leq 3)$

Now, let's look at finding the probability that there are at most than 3 heads in five tosses of a fair coin. See Figure[Binomial Less Than or Equal].

```
pbinomGC(3,region="below",size=5,prob=0.5,graph=TRUE)
```

```
## [1] 0.8125
```

Thus, $\mathrm{P}(X \leq 3) = 0.8125$.

7.4.6.4 $\mathbf{P}(X < 3)$

Now, let's look at finding the probability that there are less than 3 heads in five tosses of a fair coin. Note that for a binomial random variable, $X < 3$ is the same as $X \leq 2$. See Figure[Binomial Less Than or Equal].

```
pbinomGC(2,region="below",size=5,prob=0.5,graph=TRUE)
```

```
## [1] 0.5
```

Thus, $\mathrm{P}(X \leq 3) = 0.8125$.

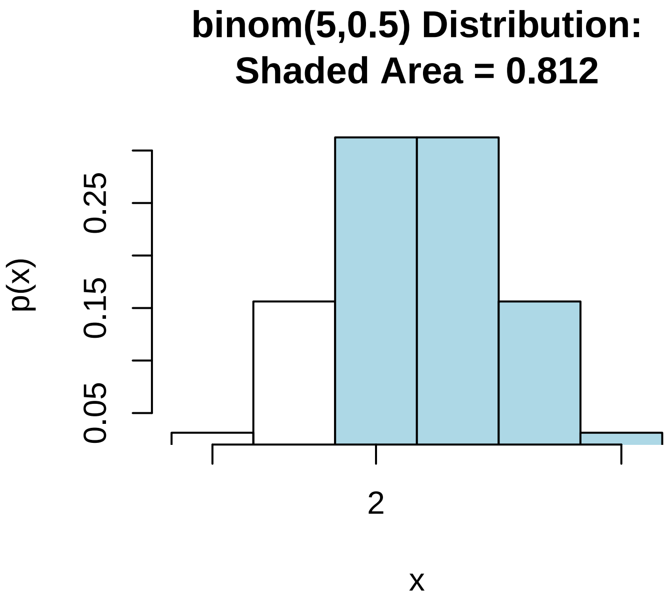


Figure 7.6: Binomial Greater Than or Equal: Shaded region represents the probability that at least 2 heads are tossed in 5 tosses of a fair coin.

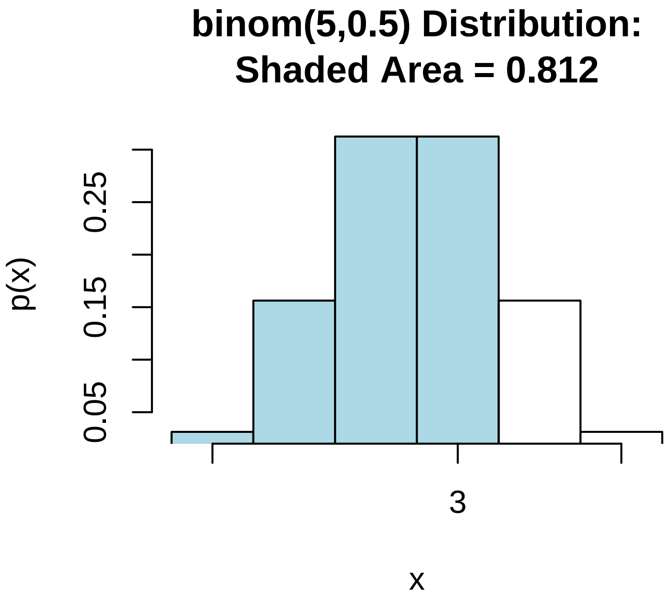


Figure 7.7: Binomial Less Than or Equal: Shaded region represents the probability that at most 3 heads are tossed in 5 tosses of a fair coin.

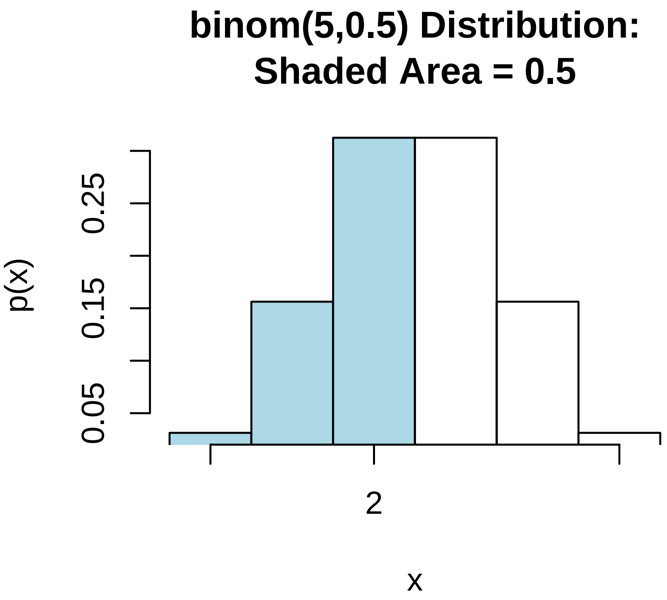


Figure 7.8: Binomial Less Than: Shaded region represents the probability that there are less than 3 heads are tossed in 5 tosses of a fair coin.

7.4.6.5 P$(2 \leq X \leq 4)$

Say we are interested in finding the probability of tossing at least 2 but not more than 4 heads in five tosses of a fair coin, P$(2 \leq X \leq 4)$. Put another way, this is the probability of tossing 2, 3, or 4 heads in five tosses of a fair coin. See Figure [Binomial Between].

```
pbinomGC(c(2,4),region="between", size=5,prob=0.5,graph=TRUE)
```

```
## [1] 0.78125
```

Thus, P$(1 \leq X \leq 4) = 0.78125$.

7.4.6.6 P$(X = 2)$

Finally, suppose we want to find the probability of tossing exactly 2 heads in five tosses of a fair coin. See Figure[Binomial Equal].

```
pbinomGC(c(2,2),region="between", size=5,prob=0.5,graph=TRUE)
```

```
## [1] 0.3125
```

So, P$(X = 2) = 0.3125$.

7.4.6.7 Expected Value and Standard Deviation for a Binomial Random Variable

Although expected value and standard deviation can be calculated for binomial random variables the same way as we did before, there are nice formulas that make the calculation easier!

For a binomial random variable, X, based on n independent trials with probability of success p,

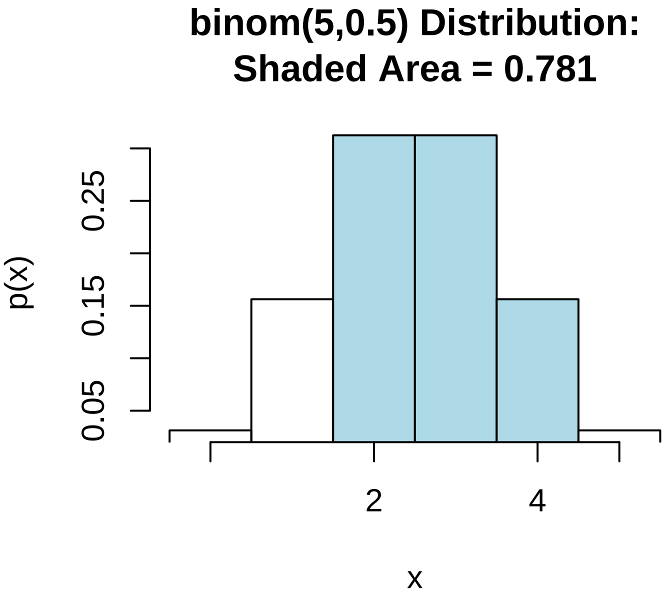


Figure 7.9: Binomial Between: Shaded region represents the probability that at least 2 but at most 4 heads are tossed in 5 tosses of a fair coin.

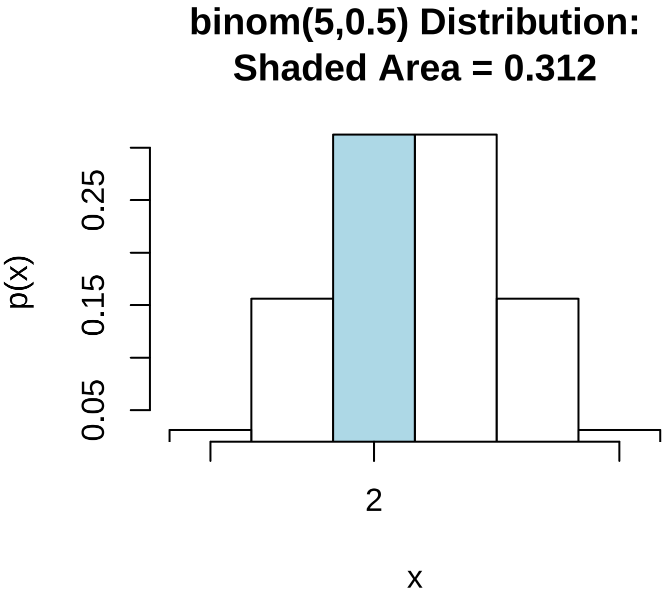


Figure 7.10: Binomial Equal: Shaded region represents the probability that exactly 2 heads are tossed in 5 tosses of a fair coin.

- the expected value is $EV(X) = n \cdot p$
- the standard deviation is $SD(X) = \sqrt{n \cdot p \cdot (1-p)}$.

Example: Let's compute the expected value for the random variable $X =$ number of heads tossed in the five coin toss example.

The expected value, $EV(X) = 5 \cdot 0.5 = 2.5$. Using R,

```
5*0.5
```

```
## [1] 2.5
```

The standard deviation, $SD(X) = \sqrt{5 \cdot 0.5 \cdot (1-0.5)} = 1.118034$. Using R,

```
sqrt(5*0.5*(1-0.5))
```

```
## [1] 1.118034
```

7.5 Continuous Random Variables

Continuous Random Variables A *continuous random variable* is a random variable whose possible values come from a range of real numbers, with no smallest difference between values.

Example: If you let X be the height in inches of a randomly selected person, then X is a continuous random variables. That's because there is no smallest possible difference between thow heights: two people could be an differ by one inch, 0.1 inches, 0.001 inches, and so on.

Non-Example: If you let X be the number of shoes a randomly-selected person owns, then X is not a continuous random variable. After all, there is a smallest difference between two values of X: one person can have two shoes and another could have three, but nobody can have any value in between, such as 2.3 shoes!

Note: For a *discrete* random variable, X, we could find the following types of probabilities:

- $P(X = x)$
- $P(X \leq x)$
- $P(X < x)$
- $P(X \geq x)$
- $P(X > x)$
- $P(a \leq X \leq x)$

For a *continuous* random variable, X, we can only find the following types of probabilities:

- $P(X \leq x)$
- $P(X < x)$
- $P(X \geq x)$
- $P(X > x)$
- $P(a \leq X \leq x)$

In other words, we were able to find the probability that a discrete random variable took on an *exact* value. We can only find the probablity that a continuous random variable falls in some *range* of values. Since we cannot find the probability that a continuous random variable equals an *exact* value, the following probabilities are the same for continuous random variables:

- $P(X \leq x) = P(X < x)$
- $P(X \geq x) = P(X > x)$
- $P(a \leq X \leq x) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$

7.5.1 Probability Density Functions for Continuous Random Variables

For *discrete* random variables, we used the **probability distribution function (pdf)** to find probabilities. The **pdf** for a discrete random variable was a table or a histogram.

For *continuous* random variables, we will use the **probability density function (pdf)** to find probabilities. The **pdf** for a continous random variable is a smooth curve.

The best way to get an idea of how this works is to examine an example of a continous random variable.

7.5.2 A Special Continuous Random Variable: Normal Random Variable

The only special type of continuous random variable that we will be looking at in this class is a **normal random variable**. There are many other continuous random variables, but normal random variables are the most commonly used continuous random variable.

A **normal random variable**, X

- is said to have a **normal distribution**,
- is completely characterized by it's mean, $EV(X) = \mu$, and it's standard deviation, $SD(X) = \sigma$, (These are the symbols that were introduced in Chapter 5.)
- has a probability density function (pdf) that is bell-shaped, or symmetric. The pdf is called a **normal curve**. An example of this curve is shown in Figure[Normal Curve].

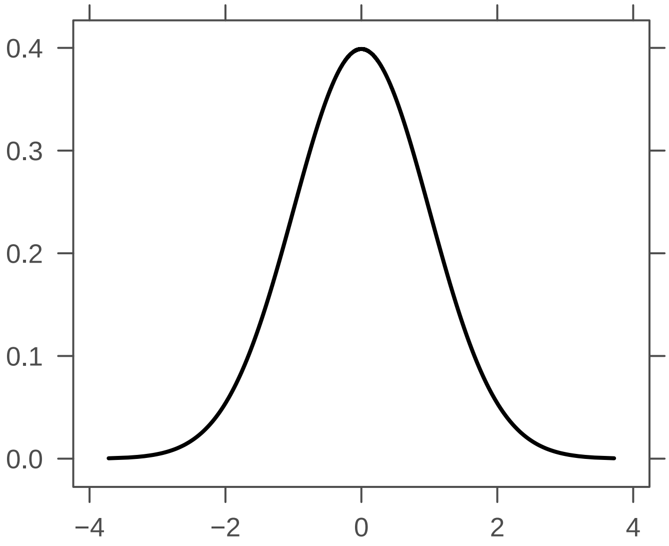


Figure 7.11: Normal Curve

Here is an important special type of normal random variable:

Standard Normal Random Variable A normal random variables with mean, $\mu = 0$, and standard deviation, $\sigma = 1$ is called a *standard normal random variable.*

The following is a list of features of a normal curve.

- Centered at the mean, μ. Since it is bell-shaped, the curve is symmetric about the mean.
- $P(X \leq \mu) = 0.5$. Likewise, $P(X \geq \mu) = 0.5$.
- Normal Random Variables follow the **68-95 Rule** (also called the **Empirical Rule**)
  - The probability that a random variable is within one standard deviation of the mean is about 68%. This can be written:
  $$P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$$
  - The probability that a random variable is within two standard deviations of the mean is about 95%. This can be written:
  $$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$$
  - The probability that a random variable is within three standard deviations of the mean is about 99.7%. This can be written:
  $$P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997$$

Example: Suppose that the distribution of the heights of college males follow a normal distribution with mean $\mu = 72$ inches and standard deviation $\sigma = 3.1$ inches. Let the random variable $X =$ heights of college males. We can approximate various probabilities using the 68-95 Rule.

- About 95% of males are between what two heights?
  P(____________ $< X <$ ____________) $\approx$ 0.95

  Answer: We can determine this using the second statement of the 68-95 Rule. The two heights we are looking for are:

```
72-2*3.1
```

```
## [1] 65.8
```

```
72+2*3.1
```

```
## [1] 78.2
```

So, P(65.8 $< X <$ 78.2) $\approx$ 0.95. See Figure[68-95 Rule Between 65.8 and 78.2].

- About what percentage of males are less than 65.8 inches tall?
  P($X <$ 65.8) $\approx$ ______________

  Answer: We know that 65.8 is two standard deviations below the mean. We also know that about 95% of males are between 65.8 and 78.2 inches tall. This means that about 5% of males are either shorter than 65.8 inches or taller than 78.2 inches. Since a normal curve is symmetric, then 2.5% of males are shorter than 65.8 inches and 2.5% of males are taller than 78.2 inches.

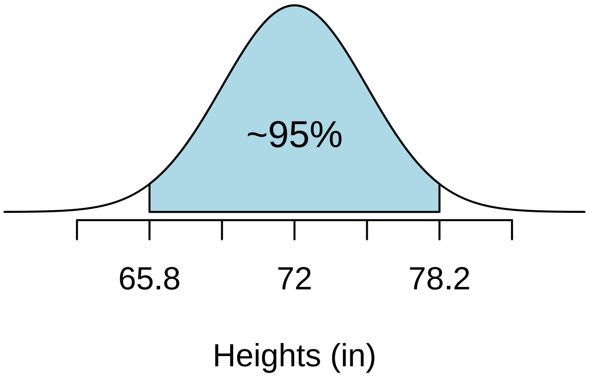


Figure 7.12: 68-95 Rule Between 65.8 and 78.2: The shaded part of this graph is the percentage of college males that are between 65.8 inches and 78.2 inches tall.

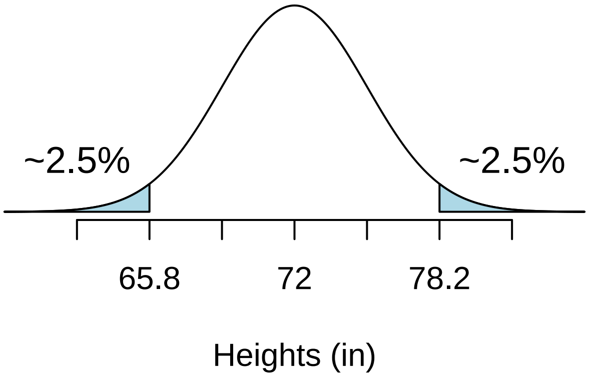


Figure 7.13: 68-95 Rule Below 65.8: The shaded part of this graph is the percentage of college males that are shorter than 65.8 inches.

So, P($X < 65.8$) ≈ 2.5%. See Figure[68-95 Rule Below 65.8].

- About what percentage of males are more than 65.8 inches tall? P($X > 65.8$) ≈ ____________________

  Answer: Since about 2.5% of males are shorter than 65.8 inches, then 100%-2.5%=97.5% of males are taller than 65.8 inches.

  So, P($X > 65.8$) ≈ 97.5%. See Figure [68-95 Rule Above 65.8].

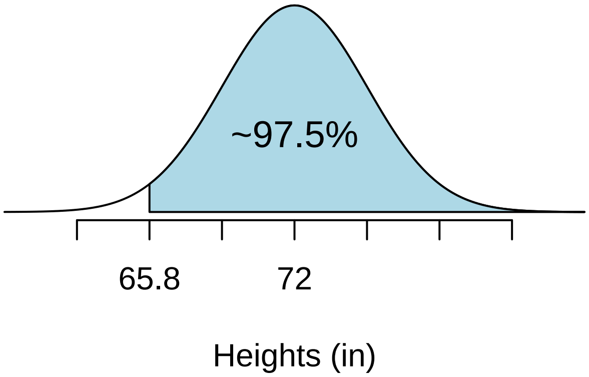


Figure 7.14: 68-95 Rule Above 65.8: The shaded part of this graph is the percentage of college males that are taller than 65.8 inches.

We can see the 68-95 Rule in action using the following app. You may find this app useful for various problems throughout the semester. All you have to do is change the mean and standard deviation to match the problem you are working on.

```
require(manipulate)
EmpRuleGC(mean=72,sd=3.1, xlab="Heights (inches)")
```

There is a function in R, similar to the one we used for binomial probability, that we can us to calculate probabilities other than those that are apparent from the 68-95 Rule. The `pnormGC` function will do this for you.

P($X > 70.9$) can be found using the following code. See Figure[Normal Greater Than].

```
pnormGC(70.9,region="above", mean=72,sd=3.1, graph=TRUE)
```

```
## [1] 0.6386448
```

Thus, P($X > 70.9$) = 0.6386448.

P($X < 69.4$ or $X > 79.1$) can be found using the following code. See Figure[Normal Outside].

```
pnormGC(c(69.4,79.1),region="outside", mean=72,sd=3.1,graph=TRUE)
```

```
## [1] 0.2118174
```

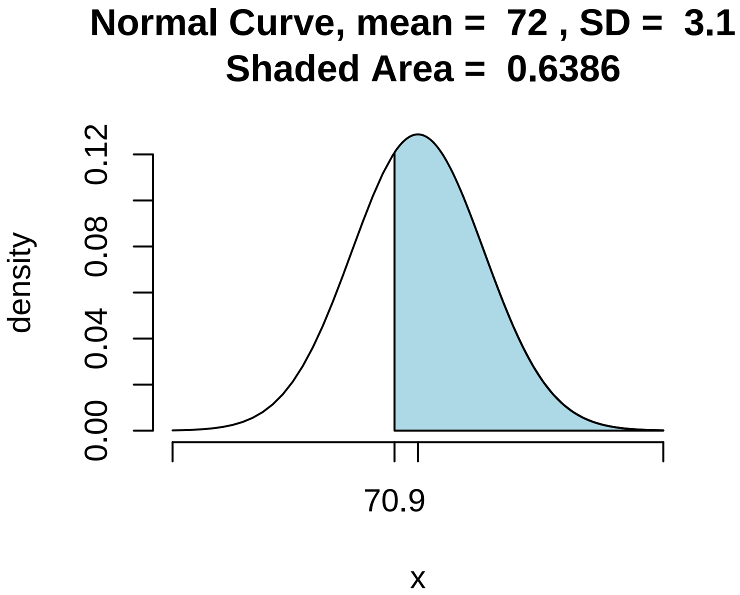


Figure 7.15: Normal Greater Than: The area of the shaded region is the percentage of males that are taller than 70.9 inches.

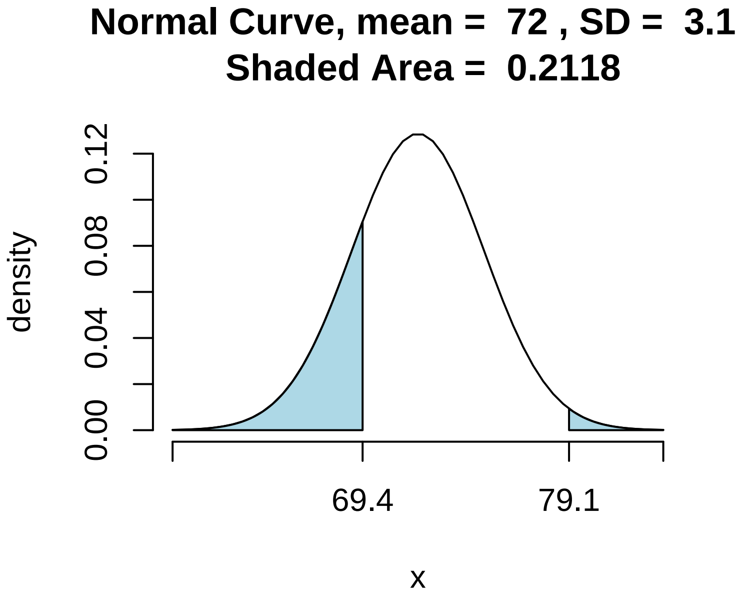


Figure 7.16: Normal Outside: The area of the shaded region is the percentage of males that are shorter than 69.4 inches or taller than 79.1 inches.

Let's switch this up a little. Suppose we want to know the height of a male that is taller than 80% of college men. Now we know the probability (or quantile) and we would like to know the x that goes with it. Here, x is called a **percentile ranking**. We are looking for $P(X \leq x) = 0.80$.

This can be found using the `qnorm` function. This function requires three inputs - quantile, mean, and standard deviation. It returns the percentile ranking.

```
qnorm(0.80,mean=72,sd=3.1)
```

```
## [1] 74.60903
```

So, $x = 74.6$. In other words, a male that is 74.6 inches tall is taller than about 80% of college men: P($X <\leq 74.6$) = 80%. We can use this number to look at the graph. See Figure[Quantile].

```
pnormGC(74.60903,mean=72,sd=3.1,region="below",graph=TRUE)
```

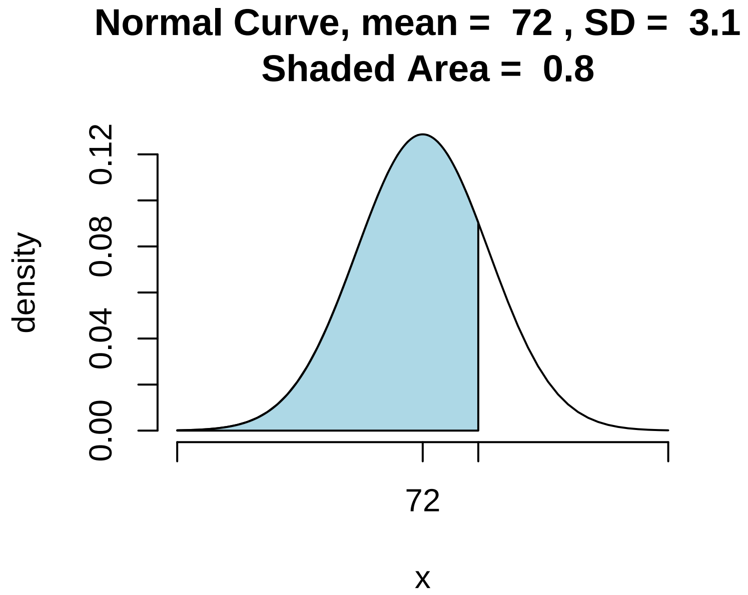


Figure 7.17: Quantile: This graph represents the 80th quantile.

```
## [1] 0.8000004
```

7.6 Approximating Binomial Probabilities

Recall that a **binomial random variable** is defined by the number of trials, n, and the probability of success, p. If the number of trials, n, is large enough, a binomial random variable can be well approximated by a **normal distribution** on *two conditions*:

- There are at least 10 successes, $n \cdot p \geq 10$.
- There are at least 10 failures, $n \cdot (1 - p) \geq 10$.

This can be seen by looking at the graphs of a binomial random variable. You can see that as n increases, the binomial distribution begins to look more and more like the normal curve using the following app.

```
require(manipulate)
BinomNorm()
```

You can also use the following app to see that if either the *expected number of successes* is too small ($n \cdot p \leq 10$) **or** the *expected number of failures* is too small ($n \cdot (1-p) \leq 10$), the normal curve does not do a very good job at approximating the binomial distribution.

```
require(manipulate)
BinomSkew()
```

7.7 Thoughts on R

7.7.1 New R Functions

Know how to use these functions:

- pbinomGC
- pnormGC
- qnorm

Chapter 8

Probability in Sampling

8.1 The Population and the Sample

8.1.1 Parameters

We begin by recalling our imaginary population, consisting of 10,000 individuals:

```
data(imagpop)
View(imagpop)
help(imagpop)
```

Let's examine the *population distribution* of the variable **height** (see Figure [Imagpop Height]): the distribution appears to be roughly normal.

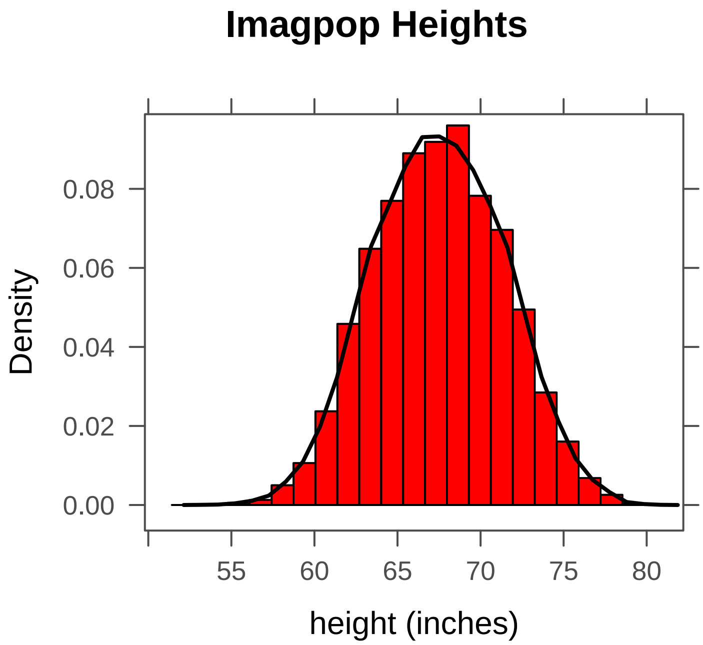


Figure 8.1: Imagpop Height.

From the population data, we can compute a wide variety of numbers. Numbers that are computed using the entire population are called *parameters*.

Parameter A *parameter* is a number associated with a population.

Here are a few parameters for `imagpop`:

```
favstats(~height,data=imagpop)
```

```
##   min   Q1 median   Q3  max     mean       sd     n missing
##  53.8 64.7   67.5 70.3 80.2 67.53012 3.907014 10000       0
```

When we apply `favstats()` to a population, everything it returns is a parameter. We have such things as:

- the population mean, written μ or just "mu". Its numerical value is 67.53012.
- the population median, usually just written as m. Its numerical value is 67.5.
- the population standard deviation, written σ, or just "sigma". Its numerical value is 3.9070139.

Let's look at another variable from `imagpop`, namely the categorical variable `sex`:

```
xtabs(~sex,data=imagpop)
```

```
## sex
## female   male
##   4968   5032
```

We see that, of the 10,000 members of the population, 5032—or 50.32%—are male. The 50.32% figure is the *population percentage* of males; we also say that the *population proportion* of males is 0.5032. Both of these quantities are parameters, since they are numbers associated with the population.

If you are lucky enough to have information on all the entire population, as we do with `imagpop`, then you can see a lot of parameters at once by calling the `summary()` function:

```
summary(imagpop)
```

We show below only the results for the first four variables in `imagpop`:

```
##      sex          math          income           cappun
##  female:4968   no :9537   Min.   :   200   favor :2976
##  male  :5032   yes: 463   1st Qu.: 19300   oppose:7024
##                           Median : 33600
##                           Mean   : 40317
##                           3rd Qu.: 54100
##                           Max.   :262200
```

8.1.2 Statistics

We have learned that a *parameter* is a number associated with a population. Usually we would like very much to know the numerical value of one or more parameters, but for practical reasons we are unable to examine every member of the population in order to compute these parameters. Usually the best we can do is to take a random sample from the population, and compute numbers based on that sample. Such numbers are called *statistics*.

Statistic A *statistic* is a number that we can compute from sample data.

Usually, we are interested in computing statistics that might serve to *estimate* some parameter.

Let's do an example. Suppose that we only have time to take a sample of 10 members from `imagpop`. The function `popsamp()` will take a simple random sample from any given data frame, of any given size:

```
popsamp(n=10,pop=imagpop)
```

```
##         sex math income cappun height idealheight diff kkardashtemp
## 7210 female   no  39700 oppose   65.6          67  1.4            2
## 8757 female   no  81000  favor   64.3          66  1.7            4
## 7609   male   no  35100  favor   70.9          74  3.1           78
## 8859   male   no  35900 oppose   70.4          73  2.6           80
## 4563   male   no  16100 oppose   73.5          76  2.5           80
## 1663 female   no  66600 oppose   69.0          71  2.0           10
## 3250   male   no  37900 oppose   63.1          65  1.9           99
## 5089   male   no  77300 oppose   68.4          71  2.6           93
## 7272   male   no  62600  favor   70.4          73  2.6           96
## 9889 female   no  59300 oppose   59.0          61  2.0            3
```

If you run the function several times, you will most likely get a different sample each time.

Now let's compute some statistics from a sample: this time, let's sample n = 100 members from the population by simple random sampling. We'll use the summary function to compute a lot of statistics at once (but we'll show the output just for the first few variables:

```
summary(popsamp(n=100,pop=imagpop))
```

```
##      sex      math        income          cappun
##  female:49   no :91   Min.   :   400   favor :25
##  male  :51   yes: 9   1st Qu.: 14625   oppose:75
##                       Median : 28350
##                       Mean   : 36676
##                       3rd Qu.: 48325
##                       Max.   :207000
```

If you try it yourself a few times, you will see that that the statistics vary, from sample to sample. This makes, sense, because the sample itself is random. In fact, a statistic is always a *random variable*!

You can practice more with the idea of a statistic as a random variable by playing again with the app `SimpleRandom()`, introduced in Chapter 5:

```
require(manipulate)
SimpleRandom()
```

For any variable you selected, you see that the statistics hop around from sample to sample, but on the whole each statistic more or less resembles the parameter in the table one row above it. In fact, that's why we compute statistics:

> **What Statistics Are For**: We use statistics to estimate parameters.

Used in this way, a statistic is called a *point estimate* for the parameter it is used to estimate.

Point Estimator A *point estimator* for a population parameter is a statistic that is used to estimate the parameter.

Even though we know it is highly unlikely to be exactly equal to the parameter, a well-chosen point estimator constitutes our single "best guess" as to the value of the parameter in question.

To summarize:

- A parameter is a number associated with a population.
  - We regard it as fixed (at least at some fixed time.)
  - However, it is usually unknown.
- A statistic is a number we can compute from a sample.
  - We can know its value, because we can compute it from our sample.
  - But it varies from sample to sample.
- We use statistics to estimate parameters.

8.2 From Questions to Parameters

Often—not always, but quite often—a Research Question can be construed as a question about the value of one or more population parameters. If you can learn how to turn Research Questions into questions about the value of parameters, then you gain access to an array of powerful statistical techniques.

We will learn how to turn Research Questions into questions about parameters by trying out lots of examples. To start, we'll keep these examples focused on `imagpop`.

Consider the following

> **Research Question**: We would like to know the center and spread of height in the population.

From the statement of the Research Question, we see that we are interested in:

- μ = mean height of all people in `imagpop`, and
- σ = the standard deviation of all people in `imagpop`.

One could also construe this Research Question as a question about population median, and population IQR. As a rule, though, we will go with mean and SD as a first preference.

We next consider a sequence of Research Questions that connect to a standard set of parameters. These parameters crop up so frequently that each will be accorded its own separate sub-section in this Chapter.

8.2.1 One Mean

> **Research Question**: We would like to know the mean rating that people in `imagpop` give to the celebrity Kim Kardashian.

In this example, the parameter of interest is pretty obvious. We are interested in knowing:

> μ = mean Kim Kardashian temperature rating for `imagpop`.

8.2.2 One Proportion

> **Research Question**: We would like to know the proportion of people in the population who majored in math.

Again the parameter of interest is fairly obvious. We are interested in knowing:

> $p =$ the proportion of all persons in `imagpop` who majored in math.

8.2.3 Difference of Two Means

> **Research Question**: We wonder how much taller guys are, on average, than gals.

This time, we break `imagpop` into two separate populations: all of the females, and all of the males. The parameters we are interested in are:

> $\mu_1 =$ mean height of all males in `imagpop`

and

> $\mu_2 =$ mean height of all females in `imagpop`.

Wanting to know how much taller the guys are, on average, is the same thing as wanting to know the *difference of means* $\mu_1 - \mu_2$. This becomes our parameter of interest.

Here is another example. Suppose we are interested in the following

> **Research Question**: Do math majors make more money, in average, than non-math majors do?

Again, we break `imagpop` into two populations: all of the math majors, and all of the non-math majors. The parameters we are interested in are:

> $\mu_1 =$ mean annual income for all math majors in `imagpop`

and

> $\mu_2 =$ mean annual income for all non-math majors in `imagpop`.

Wanting to know whether math majors make more money is the same as wanting to know whether the difference of means, $\mu_1 - \mu_2$, is positive.

Watch out, though, for the following type of question:

> **Research Question**: It is known that the mean height for the population of Australia is 67 inches. Do people in `imagpop` have the same height, on average?

You might think that this is another question about the difference of two means. However, we *already know* the mean height of all Australians, so really we are interested in just one population mean:

> $\mu =$ mean height of all people in `imagpop`.

We just want to know whether or not $\mu = 67$.

In order to be in a "difference of two means" situation, we should be dealing with two populations, and we should not know the mean of either population. Also, our plan should be to take two independent samples, one from each population, and to estimate each of the means from these samples.

(**Note**: "Independent" means that the samples don't have anything to do with each other: knowing who is in one sample tells you nothing about who is in the other sample.)

8.2.4 Difference of Two Proportions

Suppose we are interested in the following

> **Research Question**: Who is more likely to favor capital punishment: a guy or a gal?

This one is tricky for many students; the key is to interpret "likelihood" as probability. Probabilities are between 0 and 1, so they are proportions. Hence the question is really asking:

> Which is bigger: the proportion of all males who favor capital punishment, or the proportion of all females who favor capital capital punishment?

When you ask the question this way, you see that it is a question about two populations:

- the population of all males, and
- the population of all females.

Within each population, some proportion of people favor capital punishment, and we want to know which of the two proportions is larger. So are parameters of interest are:

> p_1 = the proportion of all males in `imagpop` who favor capital punishment;

and

> p_2 = the proportion of all females in `imagpop` who favor capital punishment.

We are interested in the difference of these two proportions: $p_1 - p_2$.

8.2.5 Mean of Differences

For our next research question, we will move back to the familiar `m111survey` data. The population from which this sample was drawn is, of course, the population of all GC students.

> **Research Quetion**: Do people at Georgetown want to be taller than they actually are?

This is another tricky one. Many people say: "there are two populations means at issue here: the mean actual height of the population, and the mean ideal height of the population. We want to know if the second mean is larger than the first. Hence we are interested in the difference of two means."

This isn't quite right. Recall that in order to be dealing with the difference of two means, we need to be in the situation of taking two independent samples from two populations. But this time we have one sample, from one population: it's just that we did a repeated-measures study by asking each individual two questions—actual height and ideal height.

Remember that in a repeated-measures study, you focus on the *difference* between the two measurements. Hence we are actually interested in the following parameter:

> μ_d = the mean of the difference between ideal height and actual height, for all students at Georgetown College.

For another example, let's look back at the labels-and-perception experiment:

```
data(labels)
View(labels)
help(labels)
```

In connection with the `labels` data, we are probably interested in the following

> **Research Question**: On average, which label results in the higher rating for the peanut butter: Jiff or Great Value?

This was a repeated measures design, and we are interested in the difference between the two ratings. So we are interested in:

> μ_d = mean difference in ratings (Jiff rating - GC rating) for all Georgetown College students.

Notice that we were careful to make the parameter refer to the entire population from which the sample of 30 students was drawn. **A parameter is a numerical feature of a population!**

In general, research questions associated with matched-pair and repeated-measure studies are likely to turn into questions about a mean of differences.

> Practice: For each of the following Research Questions, define the parameter(s) of interest, and then restate the Research Question in terms of that (or those) parameters.
>
> 1. (For the `m111survey` data.) Research Question: UK students are known to have a mean GPA of 3.0. Is the mean GPA of all Georgetown College students higher than the mean at UK?
> 2. (For the `m111survey` data.) Research Question: Who drives faster on average: Georgetown College males or GC females?
> 3. (For the `m111survey` data.) Research Question: Do a majority of GC males believe in love at first sight?
> 4. (For the `m111survey` data.) Research Question: Who is more likely to believe in love at first sight: A GC male or a GC female?
> 5. (For the `m111survey` data.) Research Question: Who wants to increase their height more, on average: GC males or GC females? (Think carefully about this one!)

8.2.6 The "Basic Five" Parameters

At the elementary level there are five types of Research Question that come up so frequently that the parameters associated with them are called the Basic Five:

1. One Mean (μ)

2. One Proportion (p)
3. Difference of Two Means ($\mu_1 - \mu_2$)
4. Difference of Two Proportions ($p_1 - p_2$)
5. Mean of Differences (μ_d)

We have already seen one or more examples of each of the Basic Five, and we'll see many more examples in the future.

8.3 Parameters in Experiments

Most of the examples of Research Questions that we considered in the previous section were based on observational studies, in which the data was considered to be a sample from some larger population. Because of the issues involved in obtaining consent for inclusion in an experiment, we may or may not be able to consider the subjects in an experiment as a random sample from some larger population, and even when we can, we often speak about that population a bit differently than we do for observational studies.

Let's consider a couple of examples.

8.3.1 Anchoring in m111survey

To begin with, look at `m111surveyfa12`:

```
data(m111surveyfa12)
View(m111surveyfa12)
help(m111surveyfa12)
```

Most of the Research Questions associated with this data frame may be considered as arising from an observational study. For example, the Research Question: "who drives faster, on average a male or female?" is based on an observational study, since the explanatory variable here is **sex**, and the values of **sex** cannot be assigned to subjects by researchers.

On the other hand, consider the question about the population of Canada. When subjects filled out their survey forms, some of the subjects were looking at forms where the question about Canada was stated as follows:

> "The population of Australia is about 23 million. What do you think is the population of Canada? (Give your answer in millions.)"

The rest of the subjects were looking at forms where the question about Canada was stated as follows:

> "The population of the United States is about 312 million. What do you think is the population of Canada? (Give your answer in millions.)"

The country whose population is given is called an *anchor*. Behavioral psychologists tell us that anchors can affect the way we think about that question, even when the anchor has no logical bearing on the question itself.

If everybody processes information in a completely rational way, then one's answer about Canada would be the same, no matter whether one is told first about Australia or about the United States. The "Canada" question was not asked because we were interested in how well Georgetown College students know geography trivia: we were actually interested in the question of whether Georgetown College students process information rationally! In other words, the Research Question was:

> **Research Question**: Who gives a higher estimate, on average, for the population of Canada: a person who was first told the population of the United States, or a person who was first told the population of Australia?

In this question, the explanatory variable is **anchor** (the type of form) and the response variable is **canada**. Since researchers were able to assign forms to subjects (for the most part they attempted to do so randomly), they were performing an experiment.

Here is how we translate the Research Question into a question about parameters. The parameters of interest are:

> μ_1 = the mean estimate of the population of Canada given by all GC students, if all of them could have been given a form in which they are first told the population of Australia.

> μ_2 = the mean estimate of the population of Canada given by all GC students, if all of them could have been given a form in which they are first told the population of the United States.

The Research Question turns into a question about whether the difference of means, $\mu_1 - \mu_2$, is zero or not. A difference of zero would indicate that, on average, the anchor made no difference in the response.

We have met several examples before of the difference of two means. The previous examples, though, were based on observational studies, and in that case there were two populations, and the plan was to take two independent simple random samples separately from the two populations.

In this experiment, there is just one population: all Georgetown College students. We simply imagine that the population is treated in two different ways:

- everyone in the population answers the Canada question after learning about Australia. The mean of their answers is μ_1.
- everyone in the population answers the Canada question after learning about the United States. The mean of their answers is μ_2.

By taking one sample from the population—the MAT 111 students who took the survey—and breaking them into two treatment groups (students with the Australia form and students with the U.S. form) we obtained two samples: one from each imaginary population. The samples aren't really independent (if George is given an Australia form then you know for sure that he will not be picked to get a U.S. form), but that's how it is with experiments.

8.3.2 Knife or Gun (Again)

Let's look at another example. Recall the Knife or Gun study:

```r
data(knifeorgunblock)
View(knifeorgunblock)
help(knifeorgunblock)
```

Remember that we were interested in the

> **Research Question**: What makes a person yell louder: being killed with a knife or being killed with a gun?

For this experiment, we know in advance that the results probably don't apply to some larger population, because anyone who would agree to be part of an experiment in which they get killed is liable to be extremely different from the general population—perhaps in ways associated with how they respond when being attacked with knives and guns! This fact affects how we define the parameters of interest. In this case, the parameters are defined as follows:

> $\mu_1 =$ the mean volume of yells for all 20 subjects, if they could have all been killed with a knife.

> $\mu_2 =$ the mean volume of yells for all 20 subjects, if they could have all been killed with a gun.

This time the so-called population is just the set of subjects themselves, not some larger group out of which the subjects are a sample. Just as in the previous example, though, we do imagine that this set of subjects is treated in two different ways, so we still have two "imaginary" populations.

You should define your parameters in this more restrictive way whenever the subjects in your experiments cannot be considered as representative of some larger population.

8.3.3 More Anchoring

In both of the examples above, the experiment had two treatment groups and the response variable was numerical. This resulted in a *difference of two means* situation. If your experiment has two groups and the response variable is categorical with two values (e.g., "yes" or "no"), then you will end up being interested in the difference of two proportions.

As an example, consider the attitudes experiment:

```r
data(attitudes)
View(attitudes)
help(attitudes)
```

The question about spending habits involved an experiment: some of the subjects were looking at a form in $20 was lost simply by the money falling out of a purse or wallet somehow. Other subjects faced a scenario in which they had lost $20 by purchasing a ticket and then losing that ticket. Either way, the subject is down by $20, so if everyone makes financial decisions on a purely rational basis, then one's decision about whether to attend the rock concert anyway should be unaffected by what form one is looking at: a person who would elect to attend the concert after having lost money would also elect to attend after having lost a ticket, and vice versa.

Obviously we are interested in the same sort of behavioral psychology question as in the earlier anchor experiment on guessing populations:

> **Research Question**: Who is more likely to elect to attend the rock concert: a person who has lost $20 in cash, or a person who has lost a $20 ticket?

The response variable is the categorical variable **conc.dec**, which has two values ("buy" and "not.buy"), so the parameters of interest must be proportions:

> $p_1 =$ the proportion of all GC students who would elect to attend the rock concert, if all of them could be given a survey form describing a scenario where they have lost $20 in cash.

> $p_2 =$ the proportion of all GC students who would elect to attend the rock concert, if all of them could be given a survey form describing a scenario where they have lost a $20 ticket.

We are interested in the difference of two proportions $p_1 - p_2$. If this difference is 0, then the way loses one's money has no effect on the likelihood of whether one will buy a ticket anyway.

> Practice. Consider the **attitudes** dataset, and the Research Question: *On average, does the suggested race of the defendant affect the length of sentence that would be recommended by a Georgetown College student?* Define appropriate parameters and translate the Research Question into a question about these parameters. Which one of the Basic Five is represented here?

8.4 EV and SD of Estimators

As we said earlier, people like to use a statistic – a random variable that is computed from a sample – to estimate a parameter. Let's talk about the statistics that are used to estimate the Basic Five parameters.

8.4.1 Estimating One Mean

To estimate one mean, μ, we take a simple random sample and compute the *sample mean*, which is written $\bar{x}$. As we know from previous chapters,

$$\bar{x} = \frac{\sum x_i}{n},$$

which is to say that it is the sum of the values in the sample, divided by the sample size.

This makes good sense. After all, the population mean μ is the sum of all of the values in the population, divided by the number of individuals in the population, and a random sample is (hopefully) a fair representation of the population. So it seems we should estimate the mean of the population by calculating the mean of the sample.

You might wonder how good a job $\bar{x}$ does, as an estimator of μ. Statisticians have studied this question, and the short answer is: "It does a very good job indeed!" A slightly more detailed answer starts by pointing out that $\bar{x}$ is a random variable—it is, after all, a number that depends on chance—so it has an expected value (EV for short) and a standard deviation (SD).

Recall from previous chapters that the EV of a random variable is what you expect to get, on average, in many "tries" of the random variable. Statisticians have proven that

$$EV(\bar{x}) = \mu.$$

Recall also that the SD of a random variable is about how much the random variable is liable to differ from it's EV. Statisticians have proven that the SD of $\bar{x}$ is:

$$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}},$$

where σ is the SD of the population, and n is the size of the sample.

Always keep in mind how EV and SD work together to describe how a random variable is liable to turn out. Now that we know the formulas for EV and SD of $\bar{x}$, we can say:

> When you take a simple random sample of size n from a population, the sample mean $\bar{x}$ is liable to be about μ, give or take $\frac{\sigma}{\sqrt{n}}$ or so.

8.4.2 Estimating the Difference of Two Means

When you want to estimate the difference $\mu_1 - \mu_2$ between the means of two populations, you could take a simple random sample from each population and compute:

- $\bar{x}_1$, the mean of the sample from the first population, as well as
- $\bar{x}_2$, the mean of the sample from the second population.

You could then subtract these two sample means. The difference, $\bar{x}_1 - \bar{x}_2$, would be an estimator for $\mu_1 - \mu_2$.

Statisticians have shown that the EV of $\bar{x}_1 - \bar{x}_2$ is:

$$EV(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2.$$

The SD of $\bar{x}_1 - \bar{x}_2$ is:

$$SD(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

This is a fairly complicated-looking quantity. Let's work with a numerical example:

> **Example:** Suppose that the population of all GC males has a mean height of 71 inches, with a standard deviation of 3 inches. Suppose also that the population of all GC females has a mean height of 68 inches, with a standard deviation of 2.5 inches. You plan to take a simple random sample of 25 males and an independent simple random sample of 36 females. You plan to estimate $\mu_1 - \mu_2$, the mean GC male height minus the mean GC female height, by $\bar{x}_1 - \bar{x}_2$, the difference of your sample means.
>
> (1). About what do you expect $\bar{x}_1 - \bar{x}_2$ to work out to be?
>
> (2). Give or take about how much?

As for the first question, we expect $\bar{x}_1 - \bar{x}_2$ to work out to about its EV, which is:

$$\mu_1 - \mu_2 = 71 - 68 = 3$$

inches.

For the second question we need to compute the SD, and we might as well have R do the work for us. From the given information, we seen that:

- $\sigma_1 = 3$
- $\sigma_2 = 2.5$
- $n_1 = 25$
- $n_2 = 36$

We then plug these values into the formula for $SD(\bar{x}_1 - \bar{x}_2)$, using R as a calculator to do the arithmetic for us:

```r
sqrt(3^2/25+2.5^2/36)
```

```
## [1] 0.7304869
```

So we expect $\bar{x}_1 - \bar{x}_2$ to work out to about 3 inches, give or take 0.73 inches or so.

8.4.3 Estimating One Proportion

When you need to estimate a population proportion p, you could take a simple random sample (let's abbreviate this as SRS, shall we?) and count the number X of individuals in the sample who possess the characteristic of interest. You could then divide X by n, the size of the sample, to get a proportion. This is called – naturally enough – the *sample proportion*, and it is written symbolically as $\hat{p}$. So the sample proportion is:

$$\hat{p} = \frac{X}{n},$$

and it is used to estimate the parameter p. Roughly, you can think of the sample proportion as: "the number of yesses, divided by the number of people you asked." The population proportion can be thought of as: "the number of yesses in the population, divided by the population size."

The EV of $\hat{p}$ is just p, the population proportion. The SD of $\hat{p}$ is:

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}},$$

where n is again the sample size.

Here is an example of the use of these formulas.

> **Example:** 40% of the individuals in a certain population think that marijuana use should be legal. A social scientist, who does not know this fact, would like to estimate the percentage of folks who think marijuana should be legal. She plans to take a SRS of 400 individuals, and compute the proportion of the sample who think marijuana should be legal. Fill in the blanks: "Her estimate is liable to be around __________ %, give or take __________ % or so."

Answer: The EV of $\hat{p}$ is p, which in this case is 0.40. Hence the first blank should be filled in with 40 (percent). The second blank should contain the SD of $\hat{p}$, multiplied by 100 to make it a percentage. Again we can just use R as a calculator:

```r
sqrt(.4*(1-.4)/400)*100
```

```
## [1] 2.44949
```

So we should fill in the second blank with 2.45 (percent).

8.4.4 Estimating the Difference of Two Proportions

When you need to estimate the difference between two population proportions p_1 and p_2, you can:

- take a SRS of size n_1 from the first population;
- compute the sample proportion $\hat{p}_1$, the number of yesses in this sample divided by n_1;
- take an independent SRS of size n_2 from the second population;
- compute the sample proportion $\hat{p}_2$, the number of yesses in this sample divided by n_2;
- subtract: $\hat{p}_1 - \hat{p}_2$. This difference is your estimator for $p_1 - p_2$.

The EV of $\hat{p}_1 - \hat{p}_2$ is:

$$EV(\hat{p}_1 - \hat{p}_2) = p_1 - p_2.$$

The SD of $\hat{p}_1 - \hat{p}_2$ is:

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

8.4.5 Estimating the Mean of Differences

You estimate μ_d for a population by taking a SRS of size n, say, and computing the $\bar{d}$, the sample mean of the differences. The EV of $\bar{d}$ is μ_d, and the SD of $\bar{d}$ is:

$$SD(\bar{d}) = \frac{\sigma_d}{\sqrt{n}},$$

where σ_d is the standard deviation of all of the differences in the entire population.

8.4.6 Properties of EV and SD of Estimators

8.4.6.1 Unbiased Estimators

You have probably observed that, for each of the Basic Five, the EV of the estimator is exactly the parameter you are trying to estimate! When the EV of an estimator equals the target parameter, then the estimator is said to be *unbiased.* The estimators for the Basic Five are unbiased, but not all estimators are unbiased.

8.4.6.2 Sample Size and SD

Probably you have also noticed that for the SD of each of the Basic Five estimators, there are samples sizes in the denominators:

- n for one mean, one proportion, and the mean of differences;
- n_1 and n_2 for the difference of two means, and for the difference of two proportions.

Now when the denominator of a fraction gets bigger, the fraction gets smaller, for example:

- $1/10 = 0.1$,
- $1/100 = 0.01$,

- 1/1000 = 0.001, and so on.

This observation leads to the following important point:

> *For means and proportions, the larger the sample size the smaller the SD of the estimator will be.*

In other words:

> *The larger the sample size, the less chance variation there is liable to be.*

This makes sense intuitively, as well. You might want to revisit the `SimpleRandom()` app, and this time vary the sample size.

```
require(manipulate)
SimpleRandom()
```

You will notice that at large sample sizes the estimators tend to approximate their target parameters quite closely. This is due to the fact that at large sample sizes their SDs are small.

Important Note!! The formulas for SD of the Basic Five Estimators are all based on taking a *simple random sample* from the population. For other probability sampling methods—such as stratified sampling, cluster sampling, and systematic sampling—the formulas are different and more complicated. Also, even for simple random sampling the formulas we have given are only approximately right. See the GeekNotes for the full story!

8.5 Estimators: Shape of the Distribution

For each of the Basic Five estimators, we can now say two things about its distribution:

- We can specify the center (it's the EV);
- We can specify the spread (it's the SD).

When we describe a distribution, we also want to describe its *shape*. What are the shapes of these estimators? The following sequence of apps may help you to formulate a partial answer to this question.

The first app deals with sampling from one population, in order to estimate one proportion. For a fixed sample size, take samples one by one, and watch the density plot of the sample proportions take shape.

```
require(manipulate)
PropSampler(~cappun,data=imagpop)
```

Try the app first for a small sample size. Ask yourself these questions:

- What is the shape, roughly:
  - Unimodal or bimodal?
  - Symmetric or skewed?

Try a much larger sample size, and ask yourself the same question.

The second app deals with sampling from one population, in order to estimate one mean:

```
MeanSampler(~income,data=imagpop)
```

Again for a fixed sample size take samples one by one, and watch the density plot of the sample take shape. Try at least three samples sizes, including $n = 1$ and $n = 30$. At each sample sizes, ask yourself the same questions about shape as before:

- What is the shape, roughly:
  - Unimodal or bimodal?
  - Symmetric or skewed?

The third app moves a bit more quickly. For any numerical variable you choose, and for any sample size you choose, the computer will draw 1000 simple random samples, compute the mean of each sample, and make a histogram of the results.

```
SampDistMean(imagpop)
```

Try the same sample size several times. Then change the sample size, and try it a few more times. Keep changing sample sizes. Then do the whole thing over, for a new numerical variable. Is there a pattern to what you have observed?

The fourth app is similar, except that it deals with estimating the difference of two means:

```
require(manipulate)
SampDist2Means(imagpop)
```

Again, play with sample sizes and with the various combinations of variables, and think a bit about what you noticed.

Finally, an app that that explores estimation of the difference between two proportions:

```
require(manipulate)
SampDist2Props(~sex+cappun,data=imagpop)
```

Try it as before, but also vary the combination of variables. (For example, try `~sex+math`.) Again, think about what you notice.

> **Before you go any further, make sure you have thought carefully about any patterns you might have observed.**

One of the things that you may have noticed is that no matter how the underlying population is distributed, the distribution of each of the Basic Five estimators looks more and more "bell-shaped" as the sample size increases. This phenomena comes from the famous:

> **Central Limit Theorem**: For any population with a finite mean μ and finite standard deviation σ, the distribution of the sample mean $\bar{x}$ gets closer and closer to
>
> $$norm(\mu, \frac{\sigma}{\sqrt{n}})$$
>
> as the sample size n gets larger and larger.

Although the Central Limit Theorem applies directly to the sample mean only, it can be used to show that each of the other Basic Five Estimators looks normally distributed, at "large enough" sample sizes. A sample proportion, for example is like a mean, because it involves dividing by the sample size n. Hence $\hat{p}$ is normal-looking, if n is big enough. The difference of two sample means will also look normal, because the difference of two independent normal random variables is also normal. The same idea helps us to see that the difference of two sample proportions will also be normal, when both samples sizes are large.

How large does n have to be to be "large enough"? For most populations you encounter:

- $n \geq 30$ is big enough for $\bar{x}$ to look normal;
- $n_1 \geq 30$ and $n_2 \geq 30$ is enough for $\bar{x}_1 - \bar{x}_2$ to look normal;
- $n \geq 30$ is big enough for $\bar{d}$ to look normal;
- we saw back in Chapter 7 that when $np \geq 10$ and $n(1-p) \geq 10$ then $\hat{p}$ looks normal;
- for $\hat{p}_1 - \hat{p}_2$ to look normal then we only need:
    - $np_1 \geq 10$,
    - $n(1-p_1) \geq 10$,
    - $np_2 \geq 10$, and
    - $n(1-p_2) \geq 10$.

The more skewed the population is, the larger the sample size needs to be before the distribution of $\bar{x}$ looks normal—you may have noticed this when you were playing with the apps. Nevertheless, we don't often run across a population that is so skewy that the sample size $n = 30$ is still "too small". Also, if the underlying population is pretty close to bell-shaped, then the distribution of $\bar{x}$ will be approximately normal, even when the sample size n is quite small. (You may have noticed that in the apps, too!)

8.6 Probability for Estimators

Imagine that you are a very powerful being: you can find out everything about the present world that you would like to know, all in a flash. But you are not all-knowing: you cannot know the future. Hence you are not God—-maybe you are more like Zeus or Athena, or one of the other deities from Mount Olympus.

Since you can find out anything you want about the present world, you can know everything about a population. In particular you can know any population parameter that you like.

Now imagine that you are looking down from Mount Olympus, watching a poor statistician—a mere mortal—about to take a random sample from some population about which you know everything. Since you can't know the future, you don't know what his or her sample will contain, so you don't know what the values of any of the estimators will be. But since you know the population, you know the *distribution* of the estimators: you known mean, the SD and the shape of each estimator.

Armed with this knowledge, you can answer all sorts of *probability* questions about an estimator: that is, you can determine how likely it is that the estimator will fall within any given range. Let's try a few examples of this.

> **Example (1):** A statistician is about to take a SRS of size $n = 25$ from `imagpop` and compute $\bar{x}$, the sample mean of the heights of the 25 selected individuals. What is the probability that the sample mean will exceed 68.3 inches? In other words, what is:

$$P(\bar{x} > 68.3)?$$

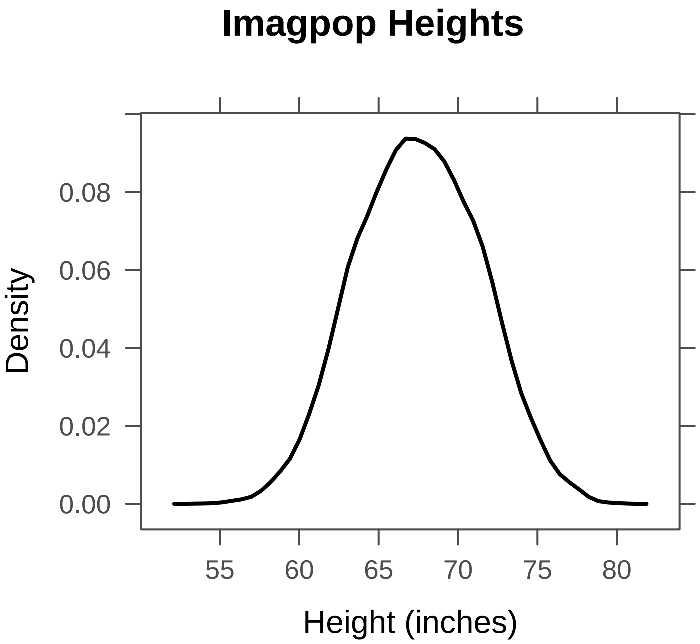
Solution: First of all, we use our god-like powers to find the mean μ and the standard deviation σ of the heights in the population:

```
favstats(~height,data=imagpop)[6:7]
```

```
##      mean       sd
##  67.53012 3.907014
```

Now the distribution of $\bar{x}$ is probably very close to normal. That's because the underlying population of heights is already pretty bell-shaped. We know this because we can use our god-like powers to draw a density plot of the entire population:

```
densityplot(~height,data=imagpop,
            xlab="Height (inches)",
            main="Imagpop Heights",
            plot.points=FALSE)
```



The results appear in Figure [Imagpop Heights]. Indeed, the distribution of **height** is quite bell-shaped. Therefore, even though the intended sample size ($n = 25$) is a bit lower than the suggested "safe" level of 30, $\bar{x}$ should be quite normal, with mean 67.53 inches and SD equal to

$$\frac{\sigma}{\sqrt{n}} = \frac{3.907}{\sqrt{25}} \approx 0.78.$$

In symbols, we might say:

$$\bar{x} \sim norm(67.53, 0.78).$$

So in order to find the probability we can just ask R to tell us the area under the appropriate normal curve after 68.3:

```
pnormGC(68.3,region="above",
        mean=67.53,sd=0.78)
```

```
## [1] 0.1617773
```

So there is about a 16.2% chance that $\bar{x}$ will exceed 68.3 inches.

> **Example (2):** A statistician plans to take a SRS of 30 males from the population of all males in `imagpop`, and an independent SRS of 40 females from the population of all women in `imagpop`. She will then compute $\bar{x}_1 - \bar{x}_2$, the sample mean height of the males minus the sample mean height of the females. Approximately what is the chance that the difference of sample means will be between 4 and 6 inches?

Solution: This time the samples sizes are $n_1 = 30$ and $n_2 = 40$. Both sample sizes are fairly large, so we can use the Central Limit Theorem to conclude that $\bar{x}_1 - \bar{x}_2$ is approximately normal.

Next, we need to compute the EV and SD of $\bar{x}_1 - \bar{x}_2$. For this we will need means and standard deviations of of the heights for:

- all males in `imagpop`, and
- all females in `imagpop`.

Hence we ask for:

```
favstats(height~sex,data=imagpop)[6:7]
```

```
##    max     mean
## 1 76.2 64.99624
## 2 80.2 70.03178
```

So the EV of $\bar{x}_1 - \bar{x}_2$ is

$$70.03 - 65 = 5.03$$

inches, and the SD of $\bar{x}_1 - \bar{x}_2$ is

```
sqrt(3.013^2^2/30+2.962^2/40)
```

```
## [1] 1.722336
```

So $\bar{x}_1 - \bar{x}_2$ is approximately

$$norm(5.03, 1.72).$$

Now we can get the desired probability:

```
pnormGC(c(4,6),region="between"
        ,mean=5.03,sd=1.72)
```

```
## [1] 0.4389664
```

So there is about a 43.9% chance that $\bar{x}_1 - \bar{x}_2$ will turn out to be between 4 and 6 inches. That is, there is about a 43.9% chance that the sample guys will be, on average, between 4 and 6 inches taller than the gals, on average.

> **Example (3):** A certain roughly normal population has a mean height of 65 inches, with a standard deviation of 3 inches.

> (1). You plan to select one individual at random from the population. Fill in the blanks, and explain: *There is about a 68% chance that the selected individual will be between ______ and _____ inches tall.*
>
> (2). You plan to select 16 individuals from the population by simple random sampling. Fill in the blanks, and explain: *There is about 95% chance that the mean height of the selected individuals will be between ______ and ______.*

Answer to 1: The number 68 should remind you of the "68" part of the 68-95 Rule for Random Variables:

> *When the distribution of a random variable is bell-shaped, there is about a 68% chance that the random variable will land within one SD of it's EV.*

The EV of the height of a single randomly-selected person is just μ, the population mean, so the EV of this random adult is 65 inches. The SD for the height of the person is the SD for the population, and that's 3. (You could also think of the height of one random person as the sample mean for a sample of size 1. Then the SD of this sample mean is $\sigma/\sqrt(1) = 3/1 = 3$.)

Therefore, by the 68-95 Rule for Random Variables, there is about a 68% chance that this randomly selected person is between $65 - 3 = 62$ and $65 + 3 = 68$ inches. So we fill in the blanks with 62 and 68.

Answer to 2: When you sample $n = 16$ individuals, the EV and SD of $\bar{x}$ are:

$$EV(\bar{x}) = \mu = 65, SD(\bar{x}) = 3/\sqrt{16} = 0.75.$$

By the "95" part of the 68-95 Rule, the approximately normal random variable $\bar{x}$ has about a 95% chance of landing within two SDs of its EV, so we should fill in the blanks with $65-2(0.75) = 63.5$ and $65+2(0.75) = 66.5$.

> **Example (4):** Suppose that forty-five percent of all registered voters approve of the job that the President of the U.S. is doing. Approximately what is the probability that, in a random sample of 1600 voters, between 43% and 47% will approve of the job that the President is doing?

Solution 1: We could turn this into a problem from Chapter 7. The *number* X of voters in the sample who approve is a binomial random variable, with the number of trials n is set at 1600 and with a probability of success on each trial equal to 0.45. Also:

- 43% of 1600 = 0.43 x 1600 = 688;
- 47% of 1600 = 0.47 x 1600 = 752.

So we are really asking for: $P(688 \leq X \leq 752)$. We can get this with R:

```r
pbinomGC(c(688,752),region="between",
         size=1600,prob=0.45)
```

```
## [1] 0.8976011
```

So there is about an 89.7% chance that the sample proportion $\hat{p}$ will be between 0.43 and 0.47.

Solution 2: This time we will use the normal approximation for $\hat{p}$. It is certainly safe to do so, because for us:

- $np = 1600(0.45) = 720 \geq 10$, and

- $n(1-p) = 100(1-0.45) = 880 \geq 10$.

Now the EV for $\hat{p}$ is p, which is 0.45. As for the SD of $\hat{p}$, it is:

```
sqrt(0.45*(1-0.45)/1600)
```

```
## [1] 0.01243734
```

By the normal approximation, the answer we seek is:

```
pnormGC(c(0.43,0.47),region="between",
        mean=0.45,sd=0.01243)
```

```
## [1] 0.8923859
```

> **Example (5):** 50% of all OSU students favor stricter gun laws, and 40% of all UK students favor stricter gun laws. A statistician plans to take a SRS of 400 OSU students and an independent SRS of 900 UK students, and to ask the sample students if they favor stricter gun laws. Fill in the blanks, and explain: there is about a 95% chance that the difference in sample proportions (OSU minus UK) will be between ________ and ________.

Solution: First we find the EV and the SD of $\hat{p}_1 - \hat{p}_2$. The EV is just

$$p_1 - p_2 = 0.50 - 0.40 = 0.10,$$

and the SD is:

```
sqrt(0.50*(1-0.50)/400+0.40*(1-0.40)/900)
```

```
## [1] 0.02986079
```

Next, we note that the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal. This is because:

- $n_1 p_1 = (400)(0.50) = 200 \geq 10$;
- $n_1(1-p_1) = (400)(0.50) = 200 \geq 10$;
- $n_2 p_2 = (900)(0.40) = 360 \geq 10$;
- $n_2(1-p_2) = (900)(0.60) = 540 \geq 10$.

By the 68-95 Rule, there is about a 95% chance that the approximately-normal $\hat{p}_1 - \hat{p}_2$ will land within 2 SDs of its EV. So we fill in the blanks as follows:

- The blank for the lower bound is:

```
0.10-2*0.02986
```

```
## [1] 0.04028
```

or about 4.03%.

- The blank for the upper bound is:

```
0.10+2*0.02986
```

```
## [1] 0.15972
```

or about 15.97%.

8.7 Inference With Estimators

8.7.1 We Are Not Deities

So far in this chapter, we have been imagining that we are deities who somehow can know everything about a population. This knowledge, together with a little statistical theory, allowed us to answer questions concerning the chances for a sample from the population to take on certain given values.

But now let's turn it around, and take the point of view of a statistician, a mere mortal. Such a person does NOT know everything about the population. However, she has taken a sample from the population, and she can compute any number that is based on her sample. She would like to use statistics from her sample to estimate parameters in the population.

This statisticians has two primary questions:

1. Based on my sample, what's the best single guess at the population parameter?
2. By how much is my best guess liable to differ from the actual value of the population parameter?

In the case of the Basic Five, the answer to Question 1 should be fairly clear: your best guess at a Basic Five parameter is the *estimator* for that parameter: e.g., your best guess at a population mean μ is the sample mean $\bar{x}$, and so on for the other four members of the Basic Five.

How about Question #2? This is a bit tougher. The statistician has the same access to statistical theory that a deity has: she knows that the estimator is liable to differ from the parameter by an SD or so. However, the formulas for the SDs all involve population parameters, so she cannot compute the numerical value of the SD.

For example, the SD of $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$. The statistician knows n, the sample size, but she does NOT know σ, the SD of the population. Hence it would seem that she cannot say anything about *how much* her estimate $\bar{x}$ is liable to differ from the the target parameter μ.

There is a way around this, however. Since the statistician has the sample she can compute the *sample standard deviation* s. Since s is an estimate of σ, it stands to reason that the quantity

$$\frac{s}{\sqrt{n}}$$

could serve as an estimate of

$$\frac{\sigma}{\sqrt{n}}.$$

The quantity $\frac{s}{\sqrt{n}}$ is known as the *standard error* of $\bar{x}$, and it is often written $SE(\bar{x})$ for short.

8.7.2 Practice With Estimator and SE

The SD for every Basic Five estimator has a corresponding SE. A mere mortal who knows only the sample can use the SE to estimate the SD of the estimator of the parameter of interest:

- One Mean: $SE(\bar{x}) = \frac{s}{\sqrt{n}}$, where s is the SD of the sample;
- Difference of Two Means: $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where s_1 and s_2 are the SDs of the first and second sample, respectively;
- One Proportion: $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $\hat{p}$ is the sample proportion.
- Difference of Two Proportions: $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, where $\hat{p}_1$ and $\hat{p}_2$ are the two sample proportions.
- Mean of Differences: $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$, where s_d is the standard deviation of the differences in the sample.

We will use the estimator and its SE together to say what we think the parameter is, and by how much we think our estimate might be off.

> **Example (1):** Using the `mat111survey` data, estimate the mean height of the population of all GC students. Also, give a figure that indicates the amount by which your estimate might differ from the population mean.

Solution: We are interested in one population mean:

> $\mu =$ mean height of all GC students.

We therefore want $\bar{x}$ and $SE(\bar{x})$. For $SE(\bar{x})$, we will need the SD of the sample, and we will need n, the sample size. So we really need three things, and we can get them quickly with **favstats**:

```
favstats(~height,data=m111survey)[6:8]
```

```
##      mean       sd  n
##  67.98662 5.296414 71
```

So:

- $\bar{x} = 67.987$,
- $s = 5.296$,
- $n = 71$.

Now we can get $SE(\bar{x})$ by applying the formula:

```
5.296/sqrt(71)
```

```
## [1] 0.6285196
```

At last, we can answer the question that was originally asked: *The mean height of all GC students is about 67.99 inches, give or take 0.63 inches or so.*

> **Example (2)**: In a random sample of 2500 adults in the year 1995, 70% said they favored the death penalty. An independent random sample of 1600 adults in the year 2013 had 55% percent of them favoring the death penalty. Fill in the blanks, and explain: The difference in the proportions of adults who favor the death penalty (Year 1995 minus Year 2013) is about __________, give or take __________ or so.

Solution: It's easier to answer this sort of question using summary data than with raw data (like we had in the previous example). We just need to plug into the formulas for the estimator and the SE of the estimator. The parameter of interest is a difference of two proportions, so we just plug into the formulas.

The estimator $\hat{p}_1$ and $\hat{p}_2$ is:

```r
0.70-0.55
```

```
## [1] 0.15
```

That's 15%. Now for the SE of $\hat{p}_1 - \hat{p}_2$:

```r
sqrt(0.70*(1-0.70)/2500)+0.55*(1-0.55)/1600
```

```
## [1] 0.009319839
```

This is about 0.93%.

Now we can fill in the blanks: *The difference in the percentage of adults who favor the death penalty (Year 1995 minus Year 2013) is about 15%, give or 0.93% or so.*

8.7.3 The 68-95 Rule for Estimation

Say that you are about to flip a fair coin. You don't know how it will land, but you do know that there is a 50% chance that it will land Heads. Next, imagine that a friend has flipped a coin, and he is hiding it in his fist. You don't know whether it's Heads or Tails, but you feel 50% *confident* that it is Heads. Why? Because *before* it was flipped there was a 50% chance that it would land heads.

This idea about confidence carries over to more complicated random processes, including processes involving the random collection of data.

For example, say that you are about to take a large simple random sample from a population. Statistical theory, together with the 68-95 Rule, says that there is about a 68% chance that the sample mean $\bar{x}$ will land within one SE of the mean μ of the population. Therefore, if you have already taken a sample, you are justified in feeling 68% *confident* that $\bar{x}$ *actually did* land within one SE of μ. Another way of saying the same thing is that you feel 68% confident that μ lies within one SE of the $\bar{x}$ that you got.

We encapsulate this idea in the

> **68-95 Rule for Estimation**: If an estimator for a population parameter has a roughly bell-shaped probability distribution, then:
>
> - we can be about 68%-confident that the parameter is within one standard error of the estimator;
>
> - we can be about 95%-confident that the parameter is within two standard errors of the estimator;

- we can be about 99.7%-confident that the parameter is within three standard errors of the estimator;

Here is an example:

> **Example:** You sample 36 people at random from a population, and find that their mean height is 68 inches, and that the SD of their heights is 3 inches. Fill in the blanks: you can be about 95%-confident that the population mean μ is between ________ and ________.

For you, the value of $\bar{x}$ is 68, and the SE of $\bar{x}$ is:

$$3/\sqrt{36} = 3/6 = 0.5$$

inches. So, two SEs is 1 inch.

Hence you can feel 95% confident that μ, the mean height of all people in the population, is within 1 inch of 68, that is, between 67 and 69 inches.

People say that the interval of real numbers from 67 to 69 inches is a *95%-confidence interval for μ.* Formulas for the confidence intervals we have produced using the 68-95 Rule for Estimation are often written out as follows:

- The "68" part of the Rule gives us the following approximately 68%-confidence interval for μ:

$$\bar{x} \pm 1 \times SE(\bar{x}).$$

- The "95" part of the Rule gives us the following approximately 95%-confidence interval for μ:

$$\bar{x} \pm 2 \times SE(\bar{x}).$$

- The "99.7" part of the Rule gives us the following approximately 99.7%-confidence interval for μ:

$$\bar{x} \pm 3 \times SE(\bar{x}).$$

In each of the above formulas, the sample mean $\bar{x}$ — the estimator for μ — is smack in the middle of the confidence interval. The quantity after the ±-sign — the number you subtract to make the lower bound of the interval and add to make the upper bound — is called the *margin of error.* As you can see the margin of error is the product of two further numbers:

- the standard error $SE(\bar{x})$, and
- a number — 1,2 or 3 — that determines your level of confidence that μ actually lies inside the interval that you are forming. This number is called a *multiplier.*

The sample estimate determines the center and the margin of error determines the width of the confidence interval. Since the margin of error is determined by the multiplier and the standard error, we can see the role that sample size and confidence level play in the width of the confidence interval:

- The larger the sample size, n, the smaller the SE. So, *larger sample sizes produce more narrower confidence intervals.*

- The higher level of confidence we have that an interval contains the true parameter value, the bigger the multiplier will have to be. So, *higher confidence levels produce wider confidence intervals.*

The confidence intervals formed by the 68-95 Rule for Estimation are somewhat "rough", in the sense that they have only approximately the level of confidence that they advertise. This is due to two approximations:

- the 1,2 and 3 numbers give only roughly the areas under the normal curve that they claim. (For example, the area under the standard normal curve between -2 and 2 is a bit more than 0.95. If you want to capture 95% of the area, you should really look between -1.96 and 1.96.)
- we replaced $SD(\bar{x})$ — which a working statistician would likely not know – with the approximation $SE(\bar{x})$ that she could actually compute from her sample.

We will explore the idea of confidence intervals in more depth in the next Chapter: in particular, we will look for ways to make them less "rough."

8.8 Summary of Formulas

- For one mean μ:
  - Estimator is $\bar{x}$
  - EV is μ
  - SD is $\frac{\sigma}{\sqrt{n}}$
  - SE is $\frac{s}{\sqrt{n}}$
- For the difference of two means $\mu_1 - \mu_2$:
  - Estimator is $\bar{x}_1 - \bar{x}_2$
  - EV is $\mu_1 - \mu_2$
  - SD is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
  - SE is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- For one proportion p:
  - Estimator is $\hat{p}$
  - EV is p
  - SD is $\sqrt{\frac{p(1-p)}{n}}$
  - SE is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- For the difference of two proportions $p_1 - p_2$:
  - Estimator is $\hat{p}_1 - \hat{p}_2$
  - EV is $p_1 - p_2$
  - SD is $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - SE is $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- For the mean of differences μ_d:
  - Estimator is $\bar{d}$
  - EV is μ_d
  - SD is $\frac{\sigma_d}{\sqrt{n}}$
  - SE is $\frac{s_d}{\sqrt{n}}$

8.9 Thoughts on R

In this chapter:

- You have really seen how useful it can be to use R as a calculator.
- You have also got a lot more practice with `pnormGC()` and a bit more practice with `pbinomGC()`. It's important to become very familiar with these functions.

You have also seen some examples in which we use `favstats()`, but only need to see a few columns of the output. For example, to see only columns 6 and 7 (the columns that give the mean and the SD), you simply type:

```
favstats(Formula,data=MyData)[6:7]
```

Chapter 9

Confidence Intervals

9.1 Introduction

Let's begin by going over what we have learned recently.

One of the primary goals in statistics is *inference* — the art of using the knowledge we obtain from a sample to infer something about the population from which the sample was collected. Numbers that describe a particular aspect of the sample are called *statistics* and numbers that describe an aspect of the population are called *parameters*. Typically we do not know the value of a parameter, but a well-chosen statistic may serve as our single best guess at it.

In Chapter Eight you learned about the Basic Five Parameters and the statistics — the Basic Five Estimators — that are commonly used to estimate them. Due to the randomness involved in selecting a sample, a statistic has variation associated with it. This means that if we take a multiple samples from the same population, the value of the statistic computed from each sample won't always be the same value.

The variation in the value of a statistic is what gives us a *distribution* for the statistic. Each of the Big Five Estimators has an associated distribution with a center, spread, and shape:

- **Center.** The center of the distribution is what we call the expected value, EV — the average value of the statistic that we would get if we could somehow repeat the sampling procedure many, many times. This figure also represents the number we think the statistic would turn out to be *around*, for any single sample.
- **Spread**. The spread of the distribution is what we call the standard deviation of the estimator, its SD. Recall that the SD of an estimator represents how far off the actual value of the statistic might be from the value of the parameter, for a single sample. You can think of the SD as a "give or take" figure. In pratical applications we usually cannot compute the value of the SD of the estiamator since it often depends on features of the populations that we don't know, so we estiamate it with a quantity that we called the *standard error*, or SE for short.
- **Shape.** We saw that at large sample sizes, the shape of each of the distributions for the Basic Five Estimators was approximately normal (bell-shaped).

For easy reference, here at the summary of Chapter Eight formulas regarding the Basic Five Parameters:

- For one mean μ:
  - Estimator is $\bar{x}$
  - EV is μ
  - SD is $\frac{\sigma}{\sqrt{n}}$

  - SE is $\frac{s}{\sqrt{n}}$

- For the difference of two means $\mu_1 - \mu_2$:
  - Estimator is $\bar{x}_1 - \bar{x}_2$
  - EV is $\mu_1 - \mu_2$
  - SD is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
  - SE is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- For one proportion p:
  - Estimator is $\hat{p}$
  - EV is p
  - SD is $\sqrt{\frac{p(1-p)}{n}}$
  - SE is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- For the difference of two proportions $p_1 - p_2$:
  - Estimator is $\hat{p}_1 - \hat{p}_2$
  - EV is $p_1 - p_2$
  - SD is $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - SE is $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- For the mean of differences μ_d:
  - Estimator is $\bar{d}$
  - EV is μ_d
  - SD is $\frac{\sigma_d}{\sqrt{n}}$
  - SE is $\frac{s_d}{\sqrt{n}}$

A statistic is our single best guess at the value of a parameter. But the fact of chance variation in sampling that we considered in Chapter Eight leads naturally to a new question:

> *How SURE are we that the value of the statistic that we've calculated is actually close to the value of the parameter?.

Instead of reporting a single number for our estimate — even if it is our best guess — perhaps we would be better off giving a *range* of values such that we have some degree of confidence that this range contains the true value of the parameter. Reporting a range of reasonable values for the parameter is a better approach to giving an estimate of a population parameter than giving one single best guess. Reporting a single value as our estimate is like throwing all of our eggs into one basket — "This is our best guess and that's it!". We, as researchers, would be failing to recognize the chance variation that is an essential part of our process of estimation. People reading the results of such a study also need to be made aware of the "give or take" that is associated with any estimate.

The range of values we seek is called by statisticians a *confidence interval.*

Confidence Interval A *confidence interval* is an interval of values for the population parameter that could be considered reasonable, based on the data at hand.

Confidence intervals in this course will be calculated using the following general equation:

$$\text{Sample Estimate } \pm \text{ Margin of Error}$$

where

$$\text{Margin of Error} \;=\; \text{Multiplier} \;\times\; \text{Standard Error.}$$

The sample estimate, multiplier, and standard error depend on the parameter being estimated, but the general form of the confidence interval for any parameter will be as above. Let's take a look at the parts of the gneral formula:

- The Sample Estimate is our single best guess at the value of the population parameter. It is the center of the confidence interval.
- The Standard Error is the estimate of the standard deviation of the population parameter. It is affected by the sample size, n. Recall from Chapter Eight that larger sample sizes yield smaller standard deviations. The same holds true for standard error, since the formulas for SE have the same structure as the formulas for SD. (Look at the formulas in the summary shown above.) **Larger sample sizes produce smaller standard errors**.
- The Multiplier is determined by the desired *confidence level.* The confidence level gives the probability that the our method of constructing confidence intervals will produce an interval that contains the parameter value. Commonly used confidence levels include:
  - 90%
  - 95% (by far the most common)
  - 99%.

The confidence level aids in the interpretation of a confidence interval. For example, suppose you want to construct a 95% confidence interval for the estimation of a population parameter. A correct interpretation of the confidence interval would be: *If the study were repeated many times, and on each occasion a 95% confidence interval were calculated, then about 95% of these intervals would contain the true parameter value and about 5% would not.*

(As you can see, the general formula for a confidence interval is just an extension of the "rough" confidence intervals produced at the end of Chapter Eight by the 68-95 Rule for Estimation.)

9.2 Chapter Outline

The next two chapters will follow the same format. We will go through each of the Basic Five parameters, one by one, doing the following:

- State a Research Question that involves the parameter of interest.
- Write out an analysis of the Research Question, using an R-function to do most of the computations that we will need.
- Look "under the hood" of the procedure to see how the R-function is computing certain values.
- Work a couple of examples involving the same parameter.

As you work through the confidence interval construction for each of the Basic Five parameters, put most of your focus on the *conceptual understanding*, i.e, seeing how what the confidence interval can be used to shed light on the Research Question that you originally had in mind.

9.3 One Population Mean μ

Let's start looking at the confidence interval for the example from `imagpop` we looked at above. We will go through the Four Steps for this first example, using the an R-function to compute the interval for us. Then, we will go through the construction of the confidence interval step by step in detail.

Let's continue to consider the research question posed at the beginning of this chapter. The `imagpop` dataset is a nice one to use for purposes of demonstration because it represents an entire population, so we can have values for parameters and statistics.

> **Research Question**: Is the average annual income for the folks from the `imagpop` population more than 50,000 dollars?

9.3.1 The Four Steps

Whne you construct confidence intervals and use them to answer a given Research Question, it will be helpful to include four "steps" in your analysis. We'll number the steps now, since we are just now getting familiar with the topic, but later on in practical writing about statistics, you will won't number them. You'll simply make sure that your analysis includes each one of them — you will weave them together to form a clear and coherent argument for your reader.

Step One: Definition of Parameter(s)

> Let μ = mean annual income for the `imagpop` population.

For this example, we will go back to pretending that we are a very powerful being and we know the true value of μ. We are going to watch as a statistician, who does not know μ, constructs a 95% confidence interval to estimate μ. Our goal with this example is to understand how a confidence interval is constructed and how it produces a range of believable values for the population parameter.

Let's start by having our statistician draw a simple random sample of size 50 from this population and calculate the sample mean from it, as we did above:

```r
set.seed(138)
mysample<-popsamp(50,imagpop)
xbar <- mean(~income,data=mysample)
xbar
```

```
## [1] 40610
```

The sample data is contained in the variable `mysample`. The sample mean is 40,610.

Step Two: Safety Check and Calculation of the Confidence Interval

In Chapter Eight, we learned that the sampling distribution of $\bar{x}$ looks normal if two conditions hold. Interestingly, these two conditions are also important — for reasons that we will learn soon — in determining whether confidence intervals have approximately the level of confidence that they advertise.

- Condition 1: The population is roughly normal in shape or the sample size, n, is large enough. Since our sample is size 50, it is probably big enough, but it is always a good idea to take a look anyway at a graph of your sample. Let's make a quick density plot of it (see Figure [Imagpop Sample]:

```
densityplot(~income,data=mysample,
        xlab="Sample Incomes (dollars/year)")
```

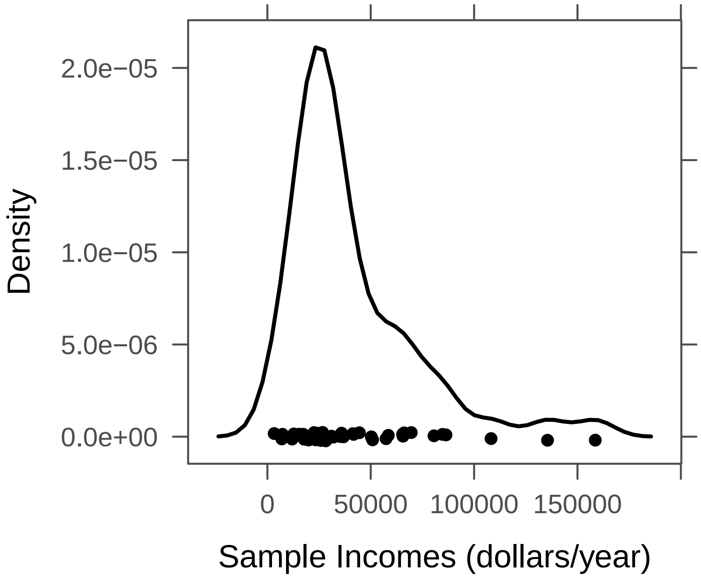


Figure 9.1: Density plot of the sample incomes. Since the random sample is probably a good cross-sectional representation of the population, we can use the plot to get some idea of the distribution of the population itself.

The sample is rather right-skewed, and so the population is probably right-skewed as well, but we saw from Chapter Eight that at such a large sample size the distribution of $\bar{x}$ as a random variable would be approximately normal. (Interestingly, this is connected with the ability to produce reliable confidence intervals by the methods of this chapter.) So we have passed the first part of the Safety Check.

The Safety Check has a second part:

- Condition 2: The sample is like a SRS from the population, at least in regard to the variables of interest. *This is the really important condition!* Since our sample was drawn using the `popsamp` function, we know that it was a simple random sample. This assures us that the sample is probably representative of the population at large.

Now that the Safety Check is passed, we can go ahead and compute a 95% confidence interval using the R-function `ttestGC()`. The rationale behind this function will be explained when we "look under the hood."

```
ttestGC(~income,data=mysample)
```

```
##
##
## Inferential Procedures for One Mean mu:
##
##
## Descriptive Results:
##
##  variable  mean    sd  n
##    income 40610 31691 50
```

```
##
##
## Inferential Results:
##
## Estimate of mu:   40610
## SE(x.bar):    4482
##
## 95% Confidence Interval for mu:
##
##           lower.bound         upper.bound
##           31603.509177        49616.490823
```

From the output above we see that the 95% confidence interval for the population mean μ is:

$$(\$31603.51, \$49616.49).$$

Step Three: Interpretation of the Confidence Interval

We are 95% confident that if the true population mean were known, the interval (\$31603.51,\$49616.49) would contain it. In other words, this interval gives the most believable values for μ.

(Recall that in this case because we actually do know the value of μ, so we can check to see if μ is, in fact, contained in the interval we computed. The population mean is:

```
mean(~income,data=imagpop)
```

```
## [1] 40316.72
```

The mean is indeed contained in the 95% confidence interval that was computed in Step Two.)

It is important to keep in mind that once the confidence interval is computed, it either contains μ or it does not. There is no "probability" associated with any specific confidence interval. The probability is associated instead with the *method* that R used used to create such the interval. What this means is that if we re-computed this 95% confidence interval for many different samples, we could expect that about 95% of those intervals would contain μ and about 5% would not. Let's do this now, for twenty samples:

```
##       lower    upper included
## 1  32559.54 46916.46      Yes
## 2  27952.18 42191.82      Yes
## 3  35363.80 46812.20      Yes
## 4  29621.25 44938.75      Yes
## 5  31280.17 48579.83      Yes
## 6  38426.30 55777.70      Yes
## 7  27975.14 39472.86       No
## 8  31776.47 48699.53      Yes
## 9  35712.63 51959.37      Yes
## 10 34833.14 51314.86      Yes
## 11 29482.53 47961.47      Yes
## 12 31691.08 45988.92      Yes
## 13 35735.36 53580.64      Yes
## 14 27302.87 44509.13      Yes
## 15 30764.31 44579.69      Yes
## 16 36232.91 51815.09      Yes
## 17 29142.10 46101.90      Yes
## 18 39001.92 54974.08      Yes
## 19 36136.50 50455.50      Yes
## 20 27582.89 45265.11      Yes
```

Observe that about 95% of the 100 intervals contain $\mu = \$40316.72$. In other words, about $0.95 \cdot 20 = 19$ intervals should contain μ and about $0.05 \cdot 20 = 1$ do not.

The following app will help you to explore this idea visually:

```
require(manipulate)
CIMean(~income,data=imagpop)
```

What you see is that in repeated sampling the 95%-confidence intervals contain the mean of the population about 95% of the time.

As a side-light, you should experiment with changing the sample size and confidence level in the app. Take note of what happens to the yellow confidence interval.

- In general, what happens to the width of the yellow confidence interval as the sample size gets bigger?
- In general, what happens to the width of the yellow confidence interval as the confidence level increases?

Warning: A common misinterpretation of the 95% confidence interval computed above would be to say that about 95% of the people in the `imagpop` population make an annual salary between \$31603.51 and \$49616.49. Don't fall into this trap! The confidence interval only gives us an interval of believable values for the population mean. It does not give us any information about the range of individual's incomes.

Step Four: Write a Conclusion.

The conclusion should be a non-technical statement about what the confidence interval tells us about the original Research Question.

We can be reasonably sure that the average annual income of the folks from the `imagpop` population is between \$31603.51 and \$49616.49. Note, however, that the numbers in this interval are all less that 50,000. Hence it would not be reasonable, in the face of this data, to believe that the mean income of the `imagpop` population is more 50,000 dollars or more.

Of course, we already know for sure that the mean income in the population is less that 50,000. What is interesting here is that a statistician could become quite sure of this, simply on the basis of a random sample of 50 individuals. That's the beauty of confidence intervals, and of other techniques in inferential statistics.

9.3.2 Under the Hood

Understanding the construction of the confidence interval explains a couple of things.

- It explains the 95% confidence level.
- It explains why we use a *t*-test (that's the extra "t" in `ttestGC()`).

In order to understand the confidence intervals produced by `ttestGC()`, it's best to start with the "rough" 95%-confidence intervals produced by the 68-95 Rule for Estimation, back in Chapter Eight. In fact, let's compute this interval for our example. We will find the needed values by looking at `favstats()`:

```
favstats(~income,data=mysample)
```

```
##   min    Q1 median    Q3    max  mean       sd  n missing
##  3300 21475  29050 50750 158600 40610 31691.03 50       0
```

We are now equipped with all the information we need to compute the approximate 95% confidence interval for the population mean.

- $\bar{x} = 40610$
- $s = 31691.03$ *Note* that s is the standard deviation of the sample.
- $n = 50$
- $SE = \dfrac{s}{\sqrt{n}} = \dfrac{31691.03}{\sqrt{50}} = 4481.7883184$
- CI $= (\bar{x} - 2 \cdot SE,\ \bar{x} + 2 \cdot SE)$ =(40610 −2· 31691.03, 40610 +2· 31691.03) = (31646.42, 49573.58)
- *Note:* The $2 \cdot SE$ is the *margin of error.*

Based on the 68-95 Rule, we can say that we are *about* 95% confident that the population mean income for the folks in `imagpop` falls somewhere in the interval ($31646.42, $49573.58).

This seems to make sense, but this is not the confidence interval that the `ttestGC` function gave us originally. The interval from `ttestGC` was ($31603.51,$49616.49). The two intervals differ by a little bit.

So how does R compute the confidence interval in `ttestGC()`? R's approach is essentially the same as the 68-95 Estimation approach. so we should we should first recall how that Rule came about: then we'll look at how R tweaks it a bit.

When we employ the 68-95 Rule for Estimation, we say that we are about 95% confident that

$$\bar{x} - 2SE(\bar{x}) < \mu < \bar{x} + 2SE(\bar{x}),$$

We feel justified in saying this because before the sample was taken,

$$P(\bar{x} - 2SE(\bar{x}) < \mu < \bar{x} + 2SE(\bar{x})) \approx 0.95.$$

(This came from the 68-95 Rule for Probability.)

Let's reason a bit with the above probability statement. To say that:

$$\bar{x} - 2SE(\bar{x}) < \mu < \bar{x} + 2SE(\bar{x})$$

is the same as saying that

> μ is within 2 SE's of $\bar{x}$.

But this means the same thing as:

> $\bar{x}$ is within two SE's of μ.

Now the way we measure how many SE's $\bar{x}$ is from μ is to take the difference $\bar{x} - \mu$ and divide it by $SE(\bar{x})$, thus:

$$\frac{\bar{x} - \mu}{SE(\bar{x})}.$$

This measure — the number of standard errors by which the sample mean differs from the mean of the population – is so important that we give it a special notation, the symbol t:

$$t = \frac{\bar{x} - \mu}{SE(\bar{x})}.$$

Using this notation, we can see that saying that $\bar{x}$ is within two SE's of μ is the same thing as saying that

$$-2 < t < 2.$$

Now let's apply our logic to the probability assertion. Saying that

$$P(\bar{x} - 2SE(\bar{x}) < \mu < \bar{x} + 2SE(\bar{x})) \approx 0.95$$

amounts to the same thing as saying that

$$P(-2 < t < 2) \approx 0.95.$$

But how good is this approximation? And can we do any better? The answer to this question depends upon realizing that t depends on:

- $\bar{x}$ and s, which depend on ...
- the sample, which depends on ...
- ... **chance!**

Hence t is a random variable: since it depends upon a sample, we call it the *t-statistic*. As a random variable, t has some sort of probability distribution, which we might hope to learn about and to approximate.

In fact some things are known about the distribution of the t-statistic. Early in the twentieth century, the statistician William Sealy Gossett discovered the following:

- if you take a random sample from a population, and
- if the population in perfectly normal, and
- your sample is of size n

then the probability distribution of t is given by a *t-density curve* with $n - 1$ *degrees of freedom.*

But what are t-curves? well, the fact is that:

- There is a t-curve for each degree of freedom $df = 1, 2, 3, \ldots$.
- They are symmetric and centered around 0.
- They have fatter tails than the standard normal curve does.
- But the bigger the degree of freedom is, the more the t-curve resembles the standard normal curve.

You can see this in the following app:

```
require(manipulate)
tExplore()
```

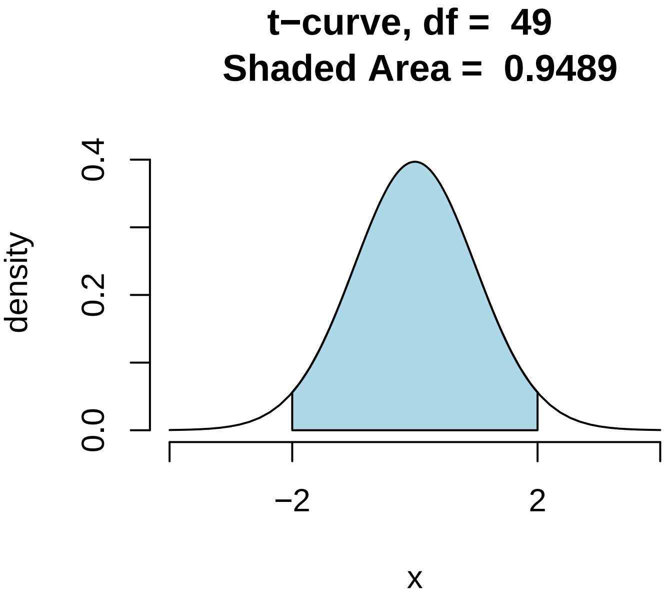
Now t-curves are determined by known mathematical formulas. Hence we can get a machine to compute areas underneath them, thus obtaining the probability for the t-statistic to fall within various ranges. To find probabilities for t-random variables, we will use `ptGC()`, a probability calculator that works very much like `pnormGC()`.

Say, for example, that you are going to take a SRS of size $n = 50$ from a population. What is

$$P(-2 < t < 2)?$$

To find out, we note that if the sample size is 50 then the degrees of freedom is 49, so we run the following code:

```
ptGC(c(-2,2),region="between",
     df=49,graph=TRUE)
```



```
## [1] 0.9489409
```

The probability is quite close to 95%. This means that at sample size size $n = 50$ (when the sample is a random sample drawn from a perfectly normal population) rough 95% intervals deliver almost exactly the level of confidence that they advertise.

For smaller, samples, though, the situation is different. For example, say that you are going to take a simple random of size $n = 4$ from a normal population. What is

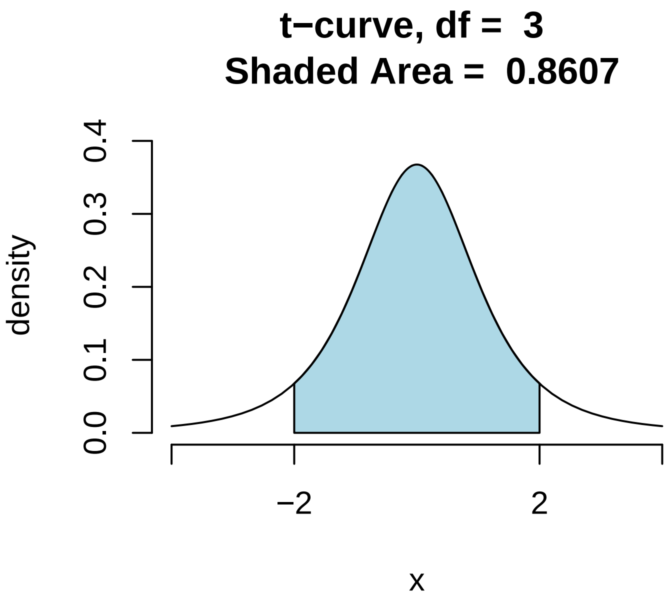
$$P(-2 < t < 2)?$$

When the sample size is 4, the degrees of freedom is 3, so the code to run is:

```
ptGC(c(-2,2),region="between",
     df=3,graph=TRUE)
```

```
## [1] 0.860674
```

This time the actual probability is only about 86%, pretty far from 95%. Hence at size $n = 4$, the 95%-confidence intervals delivered by the 68-95 Rule for Estimation are not very reliable: they advertise a



confidence level of 95%, but in reality they will cover the population mean only about 86% of the time in repeated sampling.

The problem with 68-95 Estimation rule confidence intervals comes down to their rigid use of multipliers: 95%-confidence intervals, for example use the multiplier 2, no matter what. What we really ought to do is to adjust this "rough" multiplier so that the interval will have the level of confidence that it advertises.

Suppose that:

- you have taken a random sample of size $n = 4$ from a population;
- you know the population is normally distributed,
- you don't know μ or σ;
- you want to make an *exact* 95%-confidence interval for μ.

You know now that using 2 as a multiplier just won't do. You need a different multiplier, one that we will denote t^*. That is, you want your interval to look like:

$$\bar{x} \pm t^* SE(\bar{x}),$$

where t^* is exactly the right the multiplier for a 95%-confidence interval for μ, at sample size $n = 4$.

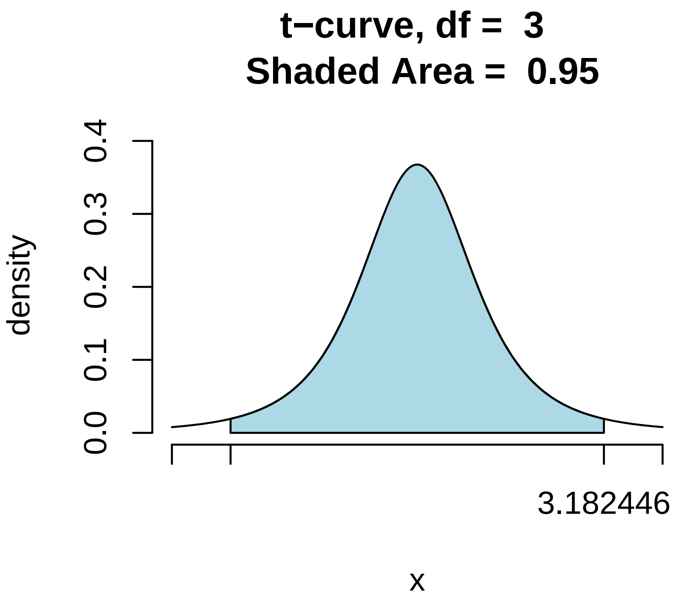
According to the statistical theory we have built up, what we need is that

$$P(-t^* < t < t^*) = 0.95.$$

R has very quick ways to find the right t^*. For our purposes, it will suffice simply to note that it can be found quickly, and that it is about 3.182446. We can see this using the following code:

```
ptGC(c(-3.182446,3.182446),region="between",
     df=3,graph=TRUE)
```

```
## [1] 0.95
```

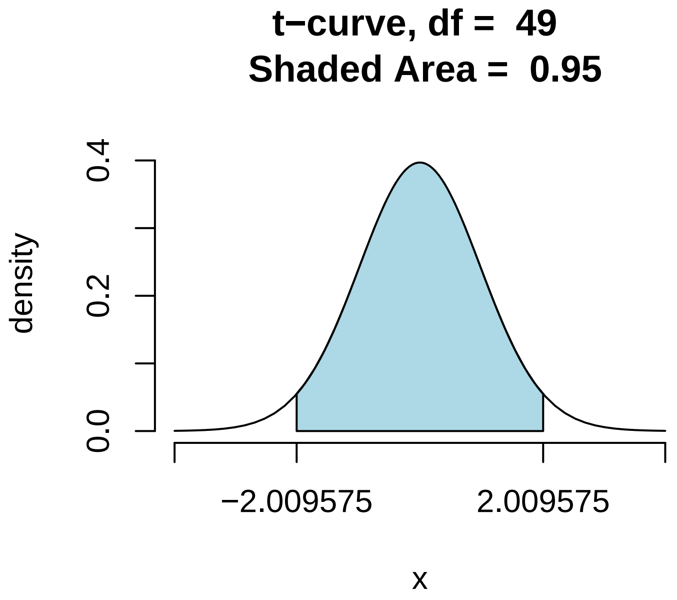


So at sample size $n = 4$, R computes a 95%-confidence interval using the formula:

$$\bar{x} \pm 3.182446 \times SE(\bar{x}).$$

At sample size 50 (the size of our original example), R would compute a different t^* multiplier, about 2.009575. We can see verify this in the following code:

```
ptGC(c(-2.009575,2.009575),region="between",
     df=49,graph=TRUE)
```



```
## [1] 0.95
```

Since $t = 2.009575$ is the correct multiplier for sample size 50, let's recompute the confidence interval with it and compare to the interval given by R:

$$(\bar{x} - 2.009575 \cdot SE,\ \bar{x} + 2.009575 \cdot SE),$$

$$(40610 - 2.009575 \cdot 4481.788,\ 40610 + 2.009575 \cdot 4481.788),$$

$$(31603.5, 49616.5).$$

This agrees with the interval computed previously by `ttestGC()`.

Summary: In general, the formula to compute a 95% confidence interval to estimate a population mean is

$$\text{CI} = (\bar{x} - t^* \cdot SE,\ \bar{x} + t^* \cdot SE),$$

where t is calculated from the t-distribution based on the appropriate degrees of freedom. Keep in mind that the packaged function, `ttestGC()`, in R does all of this work for you.

Since the construction of confidence intervals for means depends on the t-distribution, Condition 1 of the Safety Check is actually be a condition that verifies that the sampling distribution of the t-statistic approximately follows a t-distribution. Above (and in Chapter Eight) we said that $n \geq 30$ provides a basis for hope that $\bar{x}$ is approximately normal, as a random variable. It turns out that $n \geq 30$ also provides some basis for hope that that the t-statistic approximately follows a t-curve. You can investigate this idea with the following app.

```
require(manipulate)
tSampler(~income,data=imagpop)
```

For small sample sizes, you can see how the distribution of the t-statistic differs substantially from the t-distribution. However, for sample sizes around 30 you can't really tell the difference. This is why 30 is so often used as a cut-off figure in Condition 1 of the Safety Check.

9.3.3 Additional Example, and Further Ideas

> **Research Question**: What is the average height of *all* GC students?

Step One: *Definition of parameter.*

Let μ = mean height of GC student population.

Step Two: *Safety Check and Calculation of the Confidence Interval*

For this problem, we are using the observations in our `m111survey` data set as our sample. The population is **all** GC students.

Let's check the two conditions of the safety check.

- Condition 1: **The population is roughly normal in shape or the sample size, n, is at least 30.** To check this, take a look at `favstats()`.

```
##  min Q1 median    Q3 max     mean       sd  n missing
##   51 65     68 71.75  79 67.98662 5.296414 71       0
```

There are 71 people in the survey data, so our sample size $n = 71$ is so large that the population would have to be VERY far from normal in ourder for our confidence interval not to be reliable. (In fact, a graph of the sample indicates very little in the way of departure from normaility, so the population was probably pretty close to normal after all.)

- Condition 2: **The sample is like a SRS from the population, at least in regard to the variables of interest.**

Our sample of students in this survey consists of all students enrolled in MAT 111 in a particular semester. You might very well question if this constitutes a simple random sample of all GC students: after all, students certainly decide to enroll in MAT 111 for specific reasons such as it being a requirement for their major. We might hope, though that their decision to enroll in MAT 111 has little or nothing to do with their height. (If we knew of a some strong association between height and choice of a major for which MAT 111 is required, we would abandon this hope!) If there is no such relationship, then as far as the variable **height** is concerned, the students in the survey are may be thought of as *like* a group that would typically be produced by a random sampling procedure.

Since the safety check is passed (admnittedly with some queasy feelings about randomness in Condition 2), we'll go ahead and compute a 95% confidence interval.

```
##
##
## Inferential Procedures for One Mean mu:
##
##
## Descriptive Results:
##
##  variable  mean    sd  n
##    height 67.99 5.296 71
##
##
## Inferential Results:
##
## Estimate of mu:   67.99
## SE(x.bar):    0.6286
##
## 95% Confidence Interval for mu:
##
##           lower.bound         upper.bound
##           66.732979           69.240260
```

Our 95% confidence interval for the estimation of the population mean, μ, is (66.73298,69.24026) inches.

Step Three: *Interpretation of the Confidence Interval*

We are 95% confident that if the average height of the GC student population were known, the interval (66.73298,69.24026) would contain it. Put another way, (66.73298,69.24026) gives the most believable values for the average height of the GC student population.

Step Four: *Write a Conclusion.*

We can be reasonably sure that the average height of the GC student population is between 66.73298 inches and 69.24026 inches.

Additional Note: A confidence interval with any other confidence level can be calculated easily by just adding an extra argument, `conf.level`, into the `ttestGC()` function. We could calculate a 99% confidence interval as follows.

```
ttestGC(~height,data=m111survey,conf.level=0.99)
```

```
##
##
## Inferential Procedures for One Mean mu:
##
##
## Descriptive Results:
##
##  variable  mean    sd  n
##    height 67.99 5.296 71
##
##
## Inferential Results:
##
## Estimate of mu:   67.99
## SE(x.bar):    0.6286
##
## 99% Confidence Interval for mu:
##
##           lower.bound         upper.bound
##           66.322230           69.651010
```

We are 99% confident that if the average height of the GC student population were known, the interval (66.32223,69.65101) would contain it.

Check out a 68% confidence interval:

```
ttestGC(~height,data=m111survey,conf.level=0.68)
```

```
##
##
## Inferential Procedures for One Mean mu:
##
##
## Descriptive Results:
##
##  variable  mean    sd  n
##    height 67.99 5.296 71
##
##
## Inferential Results:
##
## Estimate of mu:   67.99
## SE(x.bar):    0.6286
##
## 68% Confidence Interval for mu:
##
##           lower.bound         upper.bound
##           67.357063           68.616177
```

Let's compare the widths of these three confidence intervals - 68%, 95%, and 99%. For the mean height of the GC student population,

- (67.35706,68.61618) is a 68% CI and has width 1.2591138.
- (66.73298,69.24026) is a 95% CI and has width 2.5072813.
- (66.32223,69.65101) is a 99% CI and has width 3.3287797.

You can see that higher confidence levels produce wider intervals. Wider intervals cover a broader range of numbers so you're more confident that the population parameter is in that interval.

Let's make sure we understand why this is so. As we have said before, there are two quantities that affect the width of a confidence interval:

- multiplier
- sample size

In this case, the sample size is the same, so the multiplier is entirely responsible for the differing widths of the intervals. Recall that the multiplier is the number from the t distribution that captures a specified percentage of the area between the multiplier and the multiplier's negative. As the multiplier increases it captures more of the area under the curve. This is illustrated in the three t-distributions with $df = 70$ shown below. See Figure[68% t-Distribution], Figure[95% t-Distribution], and Figure[99% t-Distribution].

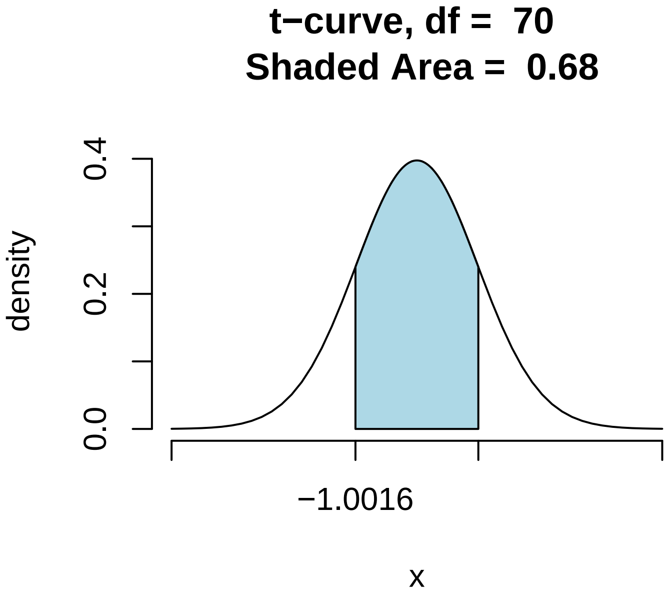


Figure 9.2: 68% t-Distribution: Visualization of the t-multiplier used in the construction of a 68% confidence interval for one population mean.

A larger area underneath the t-curve correspond to a higher probability that the confidence interval formula will produce intervals that contain the parameter, which in turn leads to a higher level of confidence that any one interval does, in fact, contain the population parameter. At the same time, though, the larger multiplier makes for a larger margin of error. Thus when the sample size is held constant, higher confidence levels are associated with wider confidence intervals.

9.4 Difference of Two Population Means, $\mu_1 - \mu_2$

Research Question: Do GC males sleep more at night, on average, than GC females?

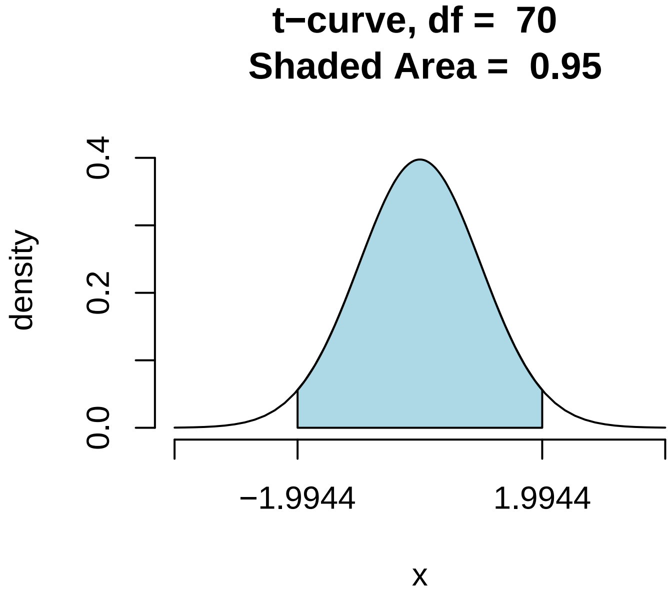


Figure 9.3: 95% t-Distribution: Visualization of the t-multiplier used in the construction of a 95% confidence interval for one population mean.

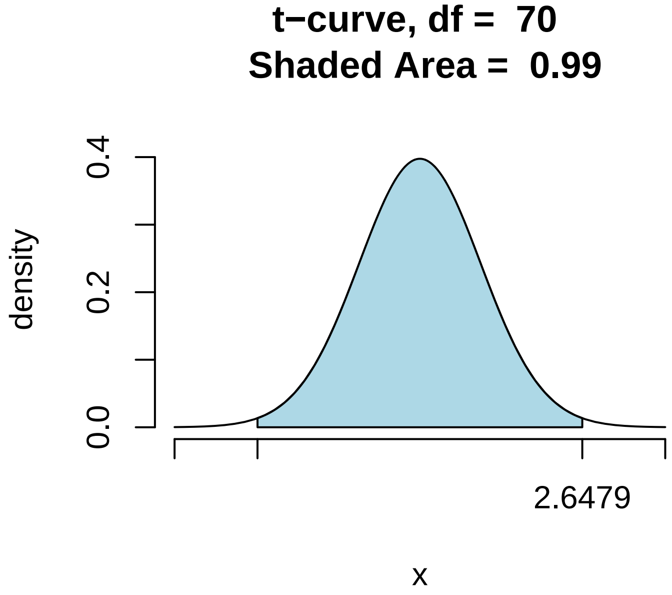


Figure 9.4: 99% t-Distribution: Visualization of the t-multiplier used in the construction of a 99% confidence interval for one population mean.

9.4.1 The Four Steps

Step One: *Definition of Parameter(s)*

For this problem, we are dealing with two populations - all GC males and all GC females - for which we don't know the mean hours of sleep per night of either one.

Let μ_1 = the mean hours of sleep per night for all GC females.

Let μ_2 = the mean hours of sleep per night for all GC males.

We are interested in the difference, $\mu_1 - \mu_2$.

Step Two: *Safety Check and Calculation of the Confidence Interval*

For this parameter, we also have two conditions for our safety check. However, the conditions are slightly different.

- Condition 1: **Both populations are roughly normal in shape or the sample sizes, n_1 and n_2, are both at least 30.** To check this condition, let's take a look at `favstats`.

```
##   .group min Q1 median    Q3 max     mean       sd  n missing
## 1 female   2  5   6.75 7.125   9 6.325000 1.619394 40       0
## 2   male   4  5   7.00 7.000  10 6.483871 1.557155 31       0
```

The sample size for females is $n_1 = 1.6193937$ and the sample size for males is $n_2 = 1.5571548$, so both samples are large enough.

Since our sample of males was barely 'large enough', it's rather important to take a look at the distribution of the samples just to make sure we aren't dealing with outliers or extreme skewness. See Figure[Sleep Histograms].

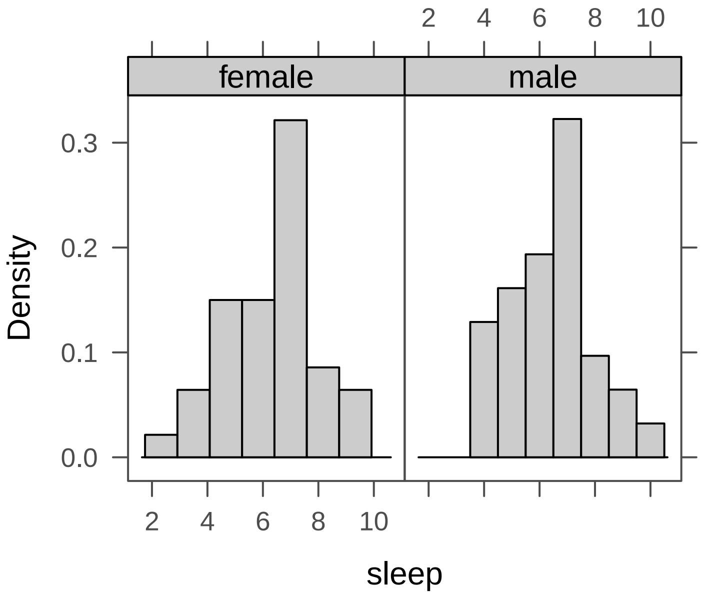


Figure 9.5: Sleep Histograms: The histogram on the left shows the hours of sleep that females in the sample get. The histogram on the right shows the hours of sleep that males in the sample get.

There do not seem to be any outliers and neither histogram looks extremely skewed. Couple that with the fact that both sample sizes were, in fact, large enough, this part of the safety check is okay.

- Condition 2: One of two things is true:
  - **We did a completely randomized experiment and the two samples are the two treatment groups in the experiment.**
  - **We took two independent random samples from two populations. This means that the samples have nothing to do with one another.**

Since this research question is a result of an observational study (the explanatory variable was simply observed, not assigned), then we need to verify the second part of this condition. Since the hours of sleep per night that an individual gets is not related to their being enrolled in MAT 111, we can say that with regard to the variable `sleep`, the sample is representative of the population of GC students. We could extend this statement to say that the sample of males and the sample of females are like simple random samples from their respective populations, at least in terms of the variable `sleep`.

We will now compute the confidence interval. We could choose to compute a confidence interval for any desired confidence level, but let's just stick to the typical 95%.

We will use the same `ttestGC` as we did for one population mean. The argument `first` should indicate which group is being considered for the first mean. In our assignment of the parameter, we let μ_1 represent the means of the female, so we need to set `first="female"` in the function.

```r
ttestGC(sleep~sex,data=m111survey,first="female")
```

```
##
##
## Inferential Procedures for the Difference of Two Means mu1-mu2:
##  (Welch's Approximation Used for Degrees of Freedom)
##   sleep grouped by sex
##
##
## Descriptive Results:
##
##   group  mean    sd  n
##  female 6.325 1.619 40
##    male 6.484 1.557 31
##
##
## Inferential Results:
##
## Estimate of mu1-mu2:  -0.1589
## SE(x1.bar - x2.bar):  0.3792
##
## 95% Confidence Interval for mu1-mu2:
##
##           lower.bound         upper.bound
##           -0.915971           0.598229
```

Our 95% confidence interval for the estimation of the population difference in the average number of hours a GC female sleeps and the average number of hours a GC male sleeps at night, $\mu_1 - \mu_2$, is (-0.9159705,0.5982286) hours.

Step Three: *Interpretation of the Confidence Interval*

We are 95% confident that if the average difference in hours of sleep per night between GC females and GC males were known, the interval (-0.9159705,0.5982286) would contain it. Put another way, (-0.9159705,0.5982286) gives the most believable values for this difference.

Step Four: *Write a Conclusion.*

We are reasonably sure that the difference in the average hours of sleep per night of GC females and GC males is between -0.9159705 and 0.5982286.

The original Research Question asked whether GC males sleep more, on average, than GC females do. If this were so, then $\mu_1 - \mu_2$ would be negative. Observe, however, that the confidence interval for $\mu_1 - \mu_2$ contains zero and some positive values, as well as negative ones: hence we cannot rule out the view that females sleep as much as males do ($\mu_1 - \mu_2 = 0$) or even more than males do ($\mu_1 - \mu_2 > 0$).

9.4.2 Under the Hood

9.4.2.1 Calculation of CI

Let's check out the actual calculation of this interval. The formula follows the same ideas as what we discussed for one population mean.

$$\Big((\bar{x_1} - \bar{x_2}) \ \pm\ t \cdot SE(\bar{x_1} - \bar{x_2})\Big)$$

where $SE(\bar{x_1} - \bar{x_2}) = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

All of these values, except for t can be found from `favstats`:

```
##   .group min Q1 median    Q3 max     mean       sd  n missing
## 1 female   2  5   6.75 7.125   9 6.325000 1.619394 40       0
## 2   male   4  5   7.00 7.000  10 6.483871 1.557155 31       0
```

- $\bar{x_1}$ = sample mean hours of sleep per night for females = 9
- $\bar{x_2}$ = sample mean hours of sleep per night for males = 10
- s_1 = sample standard deviation for females = 6.325
- s_2 = sample standard deviation for males = 6.483871
- n_1 = sample size of females = 1.6193937
- n_2 = sample size for males = 1.5571548

To find the multiplier t, recall that the t-distribution relies on $df = n - 1$. In this case, we have two sample sizes, n_1 and n_2. There are various methods for finding df for this case, but we won't go into detail here. (If you want to learn more, consult GeekNotes.) To construct this interval by hand, we can look to see what `t.test` used for df and perform the calculation from there. From the test results above, $df = 65.81$. So, t can be found as we did before.

```
qt(0.975,df=65.81)
```

```
## [1] 1.996672
```

Our 95% confidence interval for the difference of means can be computed as follows:

$$\Big((\bar{x_1} - \bar{x_2}) \ \pm\ t \cdot SE(\bar{x_1} - \bar{x_2})\Big),$$

which becomes

$$\left((9-10) \pm 1.997 \cdot \sqrt{\frac{6.325^2}{1.6193937} + \frac{6.483871}{1.5571548}}\right),$$

which in turn reduces to

$$\left(-15.359306, 13.359306\right).$$

9.4.2.2 Order of Groups

Our original interpretation stated that (-0.9159705,0.5982286) gives the most believable values for the difference between the average amount of sleep a GC female and a GC male get per night. We can actually say more than that.

Notice how the interval covers 0. In other words, zero is a plausible value for the population parameter. If the true mean difference in hours of sleep between females and males were 0, this would tell us that GC females and males get the same amount of sleep per night, on average.

Negative numbers are also included in the interval. If $\mu_1 - \mu_2$ (females - males) were a negative value, this would mean that males get more sleep on average than females.

Positive numbers are included in our confidence interval, as well. If $\mu_1 - \mu_2$ were a positive number, this would mean that females get more sleep on average than males.

Since our confidence interval contains the most believable values for $\mu_1 - \mu_2$, all of these values are reasonable. In other words, because our confidence interval spans 0, it does not give us a good idea which sex actually gets more sleep on average. When you are examining the difference of means, it is important to know which mean you are treating as μ_1 and which you are treating as μ_2, so that you can correctly interpret positive and negative numbers.

9.4.3 Additional Example

Let's consider a research question that depends on summary data rather than a built-in dataset in R.

> **Research Question**: For a randomly selected sample of 36 men from a certain population, the mean head circumference is 57.5 cm with a standard deviation equal to 2.4 cm. For a randomly selected sample of 36 women from the population, the mean head circumference is 55.3 cm with a standard deviation equal to 1.8 cm. What is the average difference in head circumference between men and women for this population?

Step One: *Definition of Parameter(s)*

Since we are dealing with two populations - men and women - the parameter of interest is the difference of means.

> Let μ_1 = the mean head circumference of the females in the population.
>
> Let μ_2 = the mean head circumference of the males in the population.

Step Two: *Safety Check and Calculation of the Confidence Interval*

- Condition 1: **Both populations are roughly normal in shape or the sample sizes, n_1 and n_2, are both at least 30.**

Since we are dealing with summary data for this example, we are not able to plot histograms or look at `favstats`. We will have to rely only on the information given in the problem. Since we are told that both samples are at least of size 30, we have some hope that this Condition 1 is statisfied (but really we would like to be able to look at graphs of the data as well).

- Condition 2: **One of two things is true**:
  - We did a completely randomized experiment and the two samples are the two treatment groups in the experiment.
  - We took two independent random samples from two populations. This means that the samples have nothing to do with one another.

Again, in this example we are dealing with an observational study, not an experiment. We can also assume that since the samples were drawn randomly from a population of men and a population of women, they have nothing to do with one another. This condition is satisfied as well.

Let's compute a 95% confidence interval. We will have to input our summary data. Be sure to input the summary consistent with the order that the parameters were assigned in Step 1.

```
ttestGC(mean=c(55.3,57.5),sd=c(1.8,2.4),n=c(36,36))
```

```
##
##
## Inferential Procedures for the Difference of Two Means mu1-mu2:
##  (Welch's Approximation Used for Degrees of Freedom)
##  Results from summary data.
##
##
## Descriptive Results:
##
##    group mean  sd  n
##  Group 1 55.3 1.8 36
##  Group 2 57.5 2.4 36
##
##
## Inferential Results:
##
## Estimate of mu1-mu2:  -2.2
## SE(x1.bar - x2.bar):  0.5
##
## 95% Confidence Interval for mu1-mu2:
##
##           lower.bound         upper.bound
##           -3.198595           -1.201405
```

The 95% confidence interval for the population mean difference in head circumference (females - males) is (-3.1985952, -1.2014048) centimeters.

Step Three: *Interpretation of the Confidence Interval*

We are 95% confident that if the population mean difference in head circumference (females - males) were known, it would lie in the interval (-3.1985952, -1.2014048).

Since all of the numbers in this interval are negative, we can say with 95% confidence that the average head circumference for males in this population is larger than the average head circumference for females. We could go even further to say that males are likely to have at least a 1.2 cm larger head circumference, on average, than females.

Step Four: *Write a Conclusion*

We are reasonably sure that the average head circumference for males in this population is larger than the average head circumference for females. The difference between the average head circumferences for females and males is likely to be a number in the range (-3.1985952, -1.2014048) centimeters.

9.5 Mean of Differences, μ_d

> **Research Question**: Does the ideal height of GC students differ, on average, than their actual height?

We will use the `m111survey` dataset to answer this research question.

9.5.1 The Four Steps

Step One: *Definition of Parameter(s)*

This was a *repeated-design* study - we are comparing the answers that each individual gave to two questions. You cannot think of this as two populations—actual height and ideal height—because these two populations are not independent. The actual height and ideal height are coming from the same person! Thus, this is not a difference of means, but it is a **mean of differences**.

Let μ_d = the mean of the difference between the ideal height and actual height of all students in the GC population.

Step Two: *Safety Check and Calculation of the Confidence Interval*

Since we are again dealing with one mean from one population, the safety check will be the same as it was for the *one population mean* case that we studied above.

- Condition 1: **The population is roughly normal in shape or the sample size, n, is at least 30.**

Let's look again at `favstats` for this sample.

```
##  min Q1 median Q3 max     mean       sd  n missing
##   -4  0      2  3  18 1.945652 3.205828 69       2
```

The sample size for this variable is 69 which is larger than 30, so it's big enough.

- Condition 2: **The sample is like a simple random sample from the population, at least in regard to the variables of interest.**

As we've said before, the variable `height` is unrelated to an individual's decision to enroll in MAT 111. For this reason, our sample from MAT 111 students can be regarded as a simple random sample of all GC students.

Since the safety check is passed, let's go ahead and compute a 95% confidence interval.

```
ttestGC(~ideal_ht-height,data=m111survey)
```

```
##
##
## Inferential Procedures for the Difference of Means mu-d:
##   ideal_ht minus height
##
##
## Descriptive Results:
##
##         Difference mean.difference sd.difference  n
##  ideal_ht - height           1.946         3.206 69
##
##
## Inferential Results:
##
## Estimate of mu-d:     1.946
## SE(d.bar):    0.3859
##
## 95% Confidence Interval for mu-d:
##
##           lower.bound         upper.bound
##           1.175528            2.715776
```

Our 95% confidence interval for the estimation of the GC mean difference between ideal and actual heights, μ_d, is (1.175528,2.715776) inches.

Step Three: *Interpretation of the Confidence Interval*

We are 95% confident that if the true population mean difference in ideal and actual heights were known, the interval (1.175528,2.715776) would contain it. In other words, this interval gives the most believable values for μ_d.

Since this entire interval lies above 0, the reasonable values for μ_d are all positive. Since we took `ideal height - actual height`, this indicates that an individual's ideal height is likely to be greater than their actual height.

Step Four: *Write a Conclusion.*

In the GC population, it is likely that the mean difference between a student's ideal height and their actual height is between 1.175528 inches and 2.715776 inches. Since the interval lies entirely above the number 0, we are pretty sure that GC students wish that they were taller than they are, on average.

9.5.2 Under the Hood

The calculation of a confidence interval for the **mean of differences** works exactly like it did for **one population mean**.

9.5.3 Additional Example

A simple random sample of 30 college women was taken. Each woman was asked to provide her own height, in inches, and her mother's height, in inches. The difference in heights was computed for each woman. Here are the summary numericla statistics:

```
##    min     Q1 median    Q3  max     mean       sd  n missing
##  -1.99 -0.105   1.61 3.885 7.56 1.852667 2.507909 30       0
```

Research Question: Are college women taller, on average, than their mothers?

Step One *Define the Parameter(s)*

It might seem that you are dealing with two populations in this case - mothers' heights and daughters' heights. However, these two groups are not independent. An individual's height is very much dependent on the heights of their parents.

For this situation, we are dealing with *matched pairs*, so we are interested in a **mean of differences**. Each mother and daughter pair constitute one matched pair. T

Let μ_d = the mean difference in heights of college women and their mothers.

Step Two: *Safety Check and Construction of the Confidence Interval*

- Condition 1: **The population is roughly normal in shape or the sample size, n, is at least 30.**

Since we have 30 mother/daughter pairs, our sample size is $n = 30$. Whoever provided us with the summary data really out to give us a graph of the sample differences so we can check it for outliers and skewness. Let's say that we are able to get a boxplot of the data (see Figure[Boxplot of Differences]).

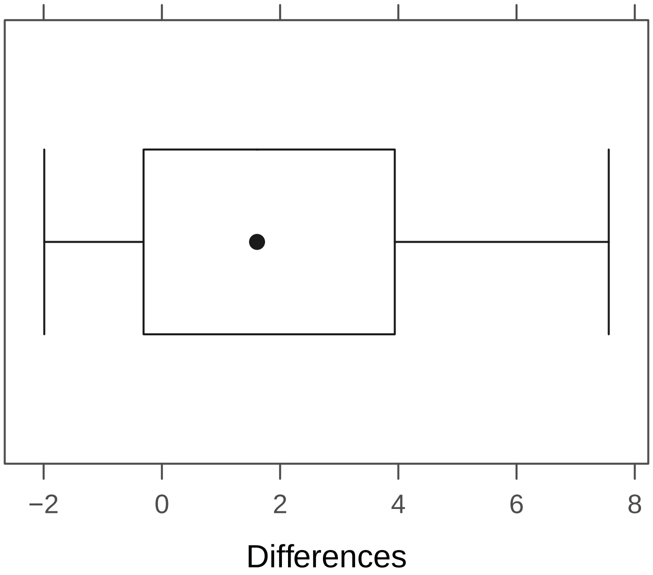


Figure 9.6: Boxplot of Differences: The distribution of the differences of college women's heights and their mother's heights.

The distribution looks fairly symmetric without any outliers.

- Condition 2: **The sample is like a simple random sample from the population, at least in regard to the variables of interest.**

We are told that a simple random sample was taken, so we will accept that and continue on to the construction of the confidence interval.

```
ttestGC(mean=1.852667,sd=2.507909,n=30)
```

```
##
##
## Inferential Procedures for One Mean mu:
##
##
## Descriptive Results:
##
##   mean    sd  n
##  1.853 2.508 30
##
##
## Inferential Results:
##
## Estimate of mu:   1.853
## SE(x.bar):    0.4579
##
## 95% Confidence Interval for mu:
##
##           lower.bound         upper.bound
##           0.916198            2.789136
```

The 95% confidence interval for the mean difference in heights of college women and their mothers is (0.9161984,2.789136) inches.

Step Three *Interpretation of the Confidence Interval*

We are 95% confident that if the true population mean of differences of heights between college women and their mothers were known, it would fall in the interval (0.9161984,2.789136). Since this is an interval of all positive numbers, we can say with 95% confidence that college women are taller than their mothers, on average.

Step Four *Write a Conclusion*

The 95%-confidence interval for μ_d consists of all the numbers between 0.9161984 inches and 2.789136 inches. If college women and their mothers were, on average, equally tall, then μ_d would be zero. However, all of the numbers in the confidence interval for μ_d are positive! Hence it seems that on average college women are taller than their mothers.

9.6 One Population Proportion, p

> **Research Question**: What is the proportion of GC students that believe in love at first sight?

Let's use the `m111survey` dataset to answer this research question.

9.6.1 The Four Steps

Here, the parameter of interest is a single population proportion.

Step One: *Definition of Parameter(s)*

> Let $p =$ the proportion of all GC students that believe in love at first sight.

Step Two: *Safety Check and Calculation of the Confidence Interval*

The statistic we will be using to estimate the population proportion is:

$$\hat{p} = \frac{X}{n},$$

where:

- X = number of students in the sample that believe in love at first sight, and
- n = the total number of students in the sample.

Notice that X is a binomial random variable. A "success" is defined as believing in love at first sight.

We know from Chapter 8 that when the number of trials, n, is large enough, then the distribution of a binomial random variable looks very much like a normal curve. We saw that our sample was "big enough" if there are least 10 successes and at least 10 failures.

Therefore, the safety check for proportions is only slightly different than the safety checks that we have been doing for means. Again, we want to check that the distribution for the statistic, $\hat{p}$, is approximately normal. This is important because, as you may recall, our beginning discussions of the construction of a confidence interval depended on the normal curve and the 68-95 Rule.

Condition 1: **The sample size is large enough if there are at least 10 successes and at least 10 failures in the sample.**

Let's take a look at the table of counts and table of percents:

```
## love_first
##  no yes
##  45  26
```

So, there are 26 successes (students in the sample that believe in love at first sight) and there are 45 failures (students in the sample that do not believe in love at first sight).

Let's check the second condition.

Condition 2: **The sample is like a simple random sample from the population, at least in regard to the variables of interest.**

Again, this is debatable since we are dealing with a sample of only those students enrolled in MAT 111 in a given semester. However, believing in love at first sight seems to be totally unrelated to an individual enrolling in MAT 111. For this reason, our sample can be regarded as a SRS from the population of GC students, in regards to the variable `love_first`.

Let's construct the confidence interval. The function that we will use for this parameter is `binomtestGC`. It works similarly to `ttestGC` that we used for means. The main difference is that we need to input what a "success" is defined to be.

```
binomtestGC(~love_first,data=m111survey,success="yes")
```

```
## Exact Binomial Procedures for a Single Proportion p:
##  Variable under study is love_first
##
## Descriptive Results:  26 successes in 71 trials
##
## Inferential Results:
```

```
##
## Estimate of p:    0.3662
## SE(p.hat):    0.0572
##
## 95% Confidence Interval for p:
##
##           lower.bound         upper.bound
##           0.254958            0.488976
```

The 95% confidence interval for the population proportion is (0.254958 , 0.488976).

Step Three: *Interpretation of the Confidence Interval*

If we knew the true proportion of GC students who believe in love at first sight, it would likely fall between 0.254958 and 0.488976.

Let's put this another way. If we computed a large number of these 95% confidence intervals to estimate the population proportion, we would expect that about 95% of the intervals would contain the true population proportion and about 5% would not. You can investigate this idea further with the following app.

```
require(manipulate)
CIProp()
```

As you use this app, consider the following question:

- Leaving the value of p and confidence level the same, what happens when you increase the number of trials?
- Leaving the value of p and the number of trials the same, what happens when you increase the confidence level?

Step Four: *Write a Conclusion.*

It is likely that the true proportion of GC students that believe in love at first sight is between 0.254958 and 0.488976. We are reasonably sure that the percentage of GC students that believe in love at first sight is somewhere in the range (25.4958%, 48.8976%).

9.6.2 Under the Hood

9.6.2.1 How the Test Works

The concept of a confidence interval is the same with proportions as it was with means. The construction is similar, as well. We won't bore you with the details again, but we should point out a couple of things about our test.

The `binomtestGC` finds the multiplier for construction of the confidence interval from the **binomial distribution**. When the sample size, n, is large enough, we know that the binomial distribution can be well approximated using the normal curve. If we did not have access to R and our sample size was large enough, we could compute an approximate 95% confidence interval using a multiplier from the normal curve. It would look like

$$\left(\hat{p} \pm z \cdot \sqrt{n \cdot \hat{p} \cdot (1-\hat{p})}\right),$$

where z would be found using the normal curve.

However, this is not the calculation that R's `binomtestGC` does! The multiplier for the confidence interval output from `binomtestGC` is found using the binomial distribution. This is nice because it does not rely on having large sample sizes so that we have a good approximation to the normal curve. In other words, we don't have the sample size restriction when we use `binomtestGC`. When you use `binomtestGC`, you do not need to worry about Condition 1 of the safety check.

There is a function, `proptestGC`, that computes this confidence interval based on the normal approximation to the binomial distribution. You will see this function again in Chapters 10, so we will mention it here so you are not alarmed when the confidence interval from `proptestGC` is different than the confidence interval in `binomtestGC`.

```r
proptestGC(~love_first,data=m111survey,success="yes")
```

```
##
##
## Inferential Procedures for a Single Proportion p:
##  Variable under study is love_first
##  Continuity Correction Applied to Test Statistic
##
##
## Descriptive Results:
##
##  yes  n estimated.prop
##   26 71         0.3662
##
##
## Inferential Results:
##
## Estimate of p:    0.3662
## SE(p.hat):    0.05717
##
## 95% Confidence Interval for p:
##
##           lower.bound         upper.bound
##           0.254136            0.478258
```

Notice that this interval is different than the one computed by `binomtestGC`. The `proptestGC` function constructs the confidence interval using the normal approximation which is only good if the sample size is "big enough". For this reason, we suggest that you stick to `binomtestGC` when you are looking for a confidence interval for one proportion.

We will not go into any more detail about this now. The important thing is knowing how to interpret the confidence interval.

9.6.2.2 Values in the Confidence Interval

As you compute more confidence intervals for one population proportion, you will notice that you will never see values less than 0 or greater than 1 included in the interval. This is because a proportion is always a number between 0 and 1. Always keep in mind that a confidence interval reports the most believable values for a parameter, so the interval *should* only include legitimate values for that parameter.

9.6.3 Additional Example

> **Research Question**: In a simple random sample of 1000 American adults, a researcher reported that 59% of the people surveyed said that they believe the world will come to an end. What is the proportion of all American adults that believe the world will come to an end?

Step One: *Definition of Parameter(s)*

> Let $p =$ the proportion of all American adults that believe the world will come to an end.

Step Two: *Safety Check and Calculation of the Confidence Interval*

Since we plan to use `binomtestGC`, we do not need to worry about the sample size since this test calculates the confidence interval from the binomial distribution. Thus, the only condition we need to verify for the safety check is Condition 2.

Condition 2: The sample is like a simple random sample from the population, at least in regard to the variables of interest. For this problem, we are told that the sample was a SRS, so let's proceed with the calculation of the confidence interval.

We first need to find the number of successes. Let's define "success" as believing that the world will come to an end. Out of the 1000 people surveyed, 59% said they believe that the world will come to an end. This means that the number of successes is 590 because $\frac{590}{1000} = 0.59$.

```
binomtestGC(x=590, n=1000)
```

```
## Exact Binomial Procedures for a Single Proportion p:
##  Results based on Summary Data
##
## Descriptive Results:  590 successes in 1000 trials
##
## Inferential Results:
##
## Estimate of p:    0.59
## SE(p.hat):    0.0156
##
## 95% Confidence Interval for p:
##
##           lower.bound         upper.bound
##           0.558788            0.620680
```

Our 95% confidence interval for the population proportion is (0.558788, 0.62068).

Step Three: *Interpretation of the Confidence Interval*

We are 95% confident that if the true proportion of American adults who believe the world is going to end was known, it would fall between 0.5588 and 0.6207.

Step Four: *Write a Conclusion*

We are reasonably sure that the percentage of American adults who believe the world is going to end is in the range (55.8788%, 62.068%).

9.7 Difference of Two Population Proportions, $p_1 - p_2$

> **Research Question**: In the GC population, are females more likely than males to believe in extra-terrestrial life?

We will use `m111survey` to answer this question.

9.7.1 The Four Steps

Step One: *Definition of Parameter(s)*

We are dealing with two independent populations - the population of all GC females and the population of all GC males. We are interested in computing the proportion that believe in extra-terrestrial life for each of these populations and then finding the difference. Thus, the parameter is the difference of proportions, $p_1 - p_2$.

Let p_1 = proportion of all GC females that believe in extra-terrestrial life.

Let p_2 = proportion of all GC males that believe in extra-terrestrial life.

For both populations, a "success" is believing in extra-terrestrial life.

Step Two: *Safety Check and Calculation of the Confidence Interval*

We will use the function `proptestGC` to compute the confidence interval for the difference between two population proportions. Since this function relies on the normal approximation to the binomial distribution, we will need to verify both conditions of the safety check.

Condition 1: The sample size is large enough if $n \cdot \hat{p_1} \geq 10$, $n \cdot (1 - \hat{p_1}) \geq 10$, $n \cdot \hat{p_2} \geq 10$ and $n \cdot (1 - \hat{p_2}) \geq 10$.

Since a success is believing in extra-terrestrial life, we can verify that we have at least 10 successes and at least 10 failures for each of the samples.

```
##         extra_life
## sex       no yes
##   female 30  10
##   male   11  20
```

For the females, ther are 10 successes and 30 failures. For the males, there are 20 successes and 11 failures. Our sample size is big enough that we can use the normal approximation to the binomial, i.e. we can use the `proptestGC` function to compute the confidence interval.

Condition 2: The sample is like a simple random sample from the population, at least in regard to the variables of interest.

We've talked about using the `m111survey` data as a sample of the population of all GC students before. We need to be sure that one's belief about extra-terrestrial life is independent of their being enrolled in MAT 111. If these two things were related, this would introduce bias into our study. For example, if there was some connection between believing in extra-terrestrial life and needing to take MAT 111, then our sample would not be representative of the entire population of GC students.

Since it seems reasonable to believe that these two things are not related, we can think of our MAT 111 survey data to be like a simple random sample.

```
proptestGC(~sex+extra_life,data=m111survey,success="yes")
```

```
##
##
## Inferential Procedures for the Difference of Two Proportions p1-p2:
##   extra_life grouped by sex
##
##
## Descriptive Results:
##
##        yes  n estimated.prop
## female  10 40         0.2500
## male    20 31         0.6452
##
##
## Inferential Results:
##
## Estimate of p1-p2:    -0.3952
## SE(p1.hat - p2.hat):  0.1099
##
## 95% Confidence Interval for p1-p2:
##
##           lower.bound         upper.bound
##           -0.610510           -0.179812
```

The 95% confidence interval for the difference in two population proportions is (-0.6105, -0.1798).

Step Three: *Interpretation of the Confidence Interval*

We are 95% confident that if the true difference in proportions of GC females and GC males that believe in extra-terrestrial life were known, it would be included in the interval (-0.6105,-0.1798).

Notice that all of the values in this interval are negative. This means that we are 95% confident that the difference in population proportions, $p_1 - p_2$, is negative. This difference will be negative if $p_2 > p_1$. This means that we are pretty confident, based on the data we have collected, that the population proportion of GC males that believe in extra-terrestrial life is GREATER than the proportion of GC females that believe in extra-terrestrial life.

Step Four: *Write a Conclusion.*

We are reasonably sure that the proportion of GC males that believe in extra-terrestrial life is greater than the proportion of GC females that believe in extra -terrestrial life. Moreover, it is likely that about 0% to 17.9812% more males believe in extra-terrestrial life than females.

9.7.2 Under the Hood

In a confidence interval for one population proportion, we said that you should only get values between 0 and 1. Do not be alarmed that you may get negative values in a confidence interval for the **difference** in two population proportions. Always keep in mind what parameter you are estimating and what values are legitimate for that parameter!

9.7.3 Additional Example

> **Research Question**: A study was done to determine whether there is a relationship between snoring and the risk of heart disease. A simple random sample selected a total of 2484 American adults. Among 1105 snorers in the study, 85 had heart disease, while only 24 of the non-snorers

had heart disease. In this population, are snorers more likely than non-snorers to have heart disease?

Step One: *Definition of Parameter(s)*

Let p_1 = population proportion of snorers that have heart disease.

Let p_2 = population proportion of non-snorers that have heart disease.

We are interested in estimating $p_1 - p_2$.

Step Two: *Safety Check and Calculation of Confidence Interval*

Since we are dealing with the difference of proportions, we will have to use the `proptestGC`, so both conditions of the safety check need to hold.

Condition 1: The sample size is large enough if we have at least 10 successes from the sample of snorers, at least 10 failures from the sample of snorers, at least 10 successes from the sample of non-snorers, and at least 10 failures from the sample of non-snorers.

For the sample of 1105 snorers, there are 85 successes (snorers with heart disease) and $1105 - 85 = 1020$ failures (snoreres without heart disease).

For the sample of 1379 non-snorers, there are 24 successes (non-snorers with heart disease) and $1379 - 24 = 1355$ failures (non-snorers without heart disease).

Condition 2: The sample is like a simple random sample from the population, at least in regard to the variables of interest. We can assume that since the study participants were selected via a simple random sample, they constitute a representative sample of the population.

Let's compute a 99% confidence interval.

```
proptestGC(x=c(85,24), n=c(1105,1379), conf.level=0.99)
```

```
##
##
## Inferential Procedures for the Difference of Two Proportions p1-p2:
##  Results taken from summary data.
##
##
## Descriptive Results:
##
##         successes    n estimated.prop
## Group 1        85 1105        0.07692
## Group 2        24 1379        0.01740
##
##
## Inferential Results:
##
## Estimate of p1-p2:    0.05952
## SE(p1.hat - p2.hat):  0.008756
##
## 99% Confidence Interval for p1-p2:
##
##           lower.bound         upper.bound
##           0.036966            0.082072
```

The 99% confidence interval for the difference in the two population proportions is (0.0369663, 0.082072).

Step Three: *Interpretation of the Confidence Interval*

We are 95% confident that if the true difference in proportions of American adult snorers and non-snorers that have heart disease were known, it would be included in the interval (0.0369663, 0.082072).

Notice that all of the values in this interval are positive. This means that we are 99% confident that the difference in population proportions, $p_1 - p_2$, is positive. This difference will be positive if $p_1 > p_2$, which means that the population proportion of snorers with heart disease is greater than the population proportion of non-snorers with heart disease.

Step Four: *Write a Conclusion.*

We are reasonably sure that the proportion of American adult snorers with heart disease is greater than the proportion of non-snorers with heart disease. Moreover, it is likely that about 3.6966341% to 8.2071981% more snorers have heart disease than non-snorers.

9.8 Thoughts on R

9.8.1 New R Functions

Know how to use this functions:

- `ttestGC`
- `binomtestGC`
- `proptestGC`

Chapter 10

Tests of Significance

10.1 Introduction

Confidence intervals aim to answer the question:

> *Given the data at hand, within what range of values does the parameter of interest probably lie?*

Now we will turn to another important type of question in statistical inference:

> *Based on the data at hand, is it reasonable to believe that the parameter of interest is a particular given value?*

When we address such a question, we perform a *test of hypothesis* (also called a *test of significance*).

In this Chapter we will look at significance tests for each of the Basic Five parameters, one by one. Although the parameters will vary from one research question to another, a test of hypothesis always follows the standard five-step format we studied back in Chapter 4. In what follows, we will go through each of the Basic Five parameters one by one. Our procedure will be as follows:

1. State a Research Question that turns out to involve the parameter of interest.
2. Write out all five steps of the test, calling a "packaged" R-function, the output of which we will use to fill in some of the steps. These functions produce confidence intervals for the parameter in question, so you will have met them already in the previous chapter on confidence intervals.
3. Look "under the hood" of the test just a bit, to see how the packaged function is getting the test statistic and the P-value. We may also introduce some concepts or terminology that apply to all tests of significance, or to all tests involving one of the Basic Five.
4. We will try one or two other examples involving the same parameter.

After each parameter has been covered, we will discuss some general issues in hypothesis testing. These considerations will apply to all of the tests introduced in this chapter, as well as to the chi-square test from Chapter 4.

Finally, we will undertake a "Grand Review", in which we practice moving from a Research Question to the proper inferential procedures to address that question.

10.2 One Population Mean μ

10.2.1 Introductory Research Question

Consider the following Research Question concerning the `mat111survey` data:

> *Does the data provide strong evidence that the mean fastest speed for all Georgetown College exceeds 100 mph?*

10.2.2 The Five Steps

Clearly, the parameter of interest is a single population mean. Unlike the situation with the chi-square test, the debate is about the value of a parameter. Hence Step One should include a clear definition of the parameter of interest, followed by a statement of Null and Alternative Hypotheses that lay out views concerning the value of this parameter.

Step One: *Definition of parameter and statement of hypotheses.*

We define the parameter first. Let

> $\mu =$ the mean fastest driving speed of ALL Georgetown College students.

Now that we have defined the parameter of interest, we can state our hypotheses in terms of that parameter, using the symbol we have defined:

> H_0: $\mu = 100$.
>
> H_a: $\mu > 100$.

Step Two: *Safety Check and Reporting the Value of the Test Statistic*

Just as `chisqtestGC()` was the work-horse for the χ^2 test for relationship between two factor variables, so in this Chapter there is an R-function that does much of the computational work we need for tests involving population means. In this case the function is the very same one that produced confidence intervals for us, namely: `ttestGC()`. We simply need to provide a couple of extra arguments in order to alert it to our need for a test of significance:

```r
ttestGC(~fastest,data=m111survey,mu=100,
          alternative="greater")
```

The `mu` argument specifies the so-called “Null value”—the value that the Null hypothesis asserts for the parameter μ. The `alternative` argument indicates the “direction” of the Alternative Hypothesis—in this case, the Alternative asserts that μ is *greater* than 100.

When we run the code above, we get results that will help us through Steps Two through Four in a test of significance.

The first thing we will do is the “safety check”. For each of the Basic Five parameters, the test of significance is built upon the same mathematical ideas that are used in the construction of a confidence interval. Hence the conditions for reliability of the tests are the same as for those of confidence intervals. In this case, we ask that the sample be roughly normal in shape, or that the sample size be at least 30. Hence we pay attention to the “Descriptive Results” first:

```
### Descriptive Results:
###
###  variable  mean    sd  n
###   fastest 105.9 20.88 71
```

We see that the sample size was $n = 71$, quite large enough to trust the approximation-routines that R uses for inferential procedures involving a mean.

The other element of the safety check is to ask whether our sample is like a simple random sample from the population. As we have seen with the `mat111survey` data, this is a debatable point: the sample consists of all students in MAT 111 from a particular semester, so it's not really a random sample. On the other hand, for a question like "fastest speed ever driven", it might be equivalent to a random sample. At any rate, we don't have much reason to believe that there is a relationship between one's fastest driving speed and whether or not one enrolls in MAT 111.

So much for the safety check. Now let's go for the test statistic.

The test statistic in this test is called t. From the `ttestGC()` output, we see:

```
###   Test Statistic:       t = 2.382
```

Apparently the test statistic is $t = 2.382$.

Step Three: *Statement and Interpretation of the P-value*

Again we read, from the function output, that:

```
###   P-value:      P = 0.009974
```

The p-value is about 0.01, or 1%.

In this test, the way to interpret the P-value is as follows:

> **P-value Interpretation**: *If the mean fastest speed of all GC students is 100 mph, then there is about a 1% chance that the test statistic will be at least as big as the test statistic that we actually got.*

Step Four: *Decision*

As we see from the interpretation of the P-value, a low P-value makes the Null appear to be implausible. Since $\mathrm{P} < 0.05$ for us, we *reject the Null Hypothesis.*

Step Five: *Write a Conclusion*

As in Chapter Three, this should be a complete sentence in nontechnical language that says how much evidence the data provided for the Alternative Hypothesis. Our conclusion this time is:

> *This data provides strong evidence that the mean fastest speed for all GC students is greater than 100 mph.*

10.2.3 Under the Hood, and Further Ideas

10.2.3.1 t-Statistic Formula

The formula for the t-statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where:

- $\bar{x}$ is the sample mean,
- μ_0 is what the Null hypothesis thinks that μ is, and
- $s/\sqrt{n}$ is the SE for $\bar{x}$, the same standard error that we have met in previous chapters.

Remember that the EV of $\bar{x}$ is just μ, the mean of the population from which the sample was drawn. The Null believes that μ is μ_0, so the Null expects $\bar{x}$ to be about μ_0, give or take some for chance error. The numerator in the formula, $\bar{x} - \mu_0$, is often called the *observed difference*, because it gives the difference between the sample mean that was actually observed and what a believer in the Null would expect that sample mean to be.

The denominator of the t-statistic gives the SE for $\bar{x}$, which means that is gives the amount by which $\bar{x}$ is liable to differ from the mean of the population. So if the Null is right, then $\bar{x}$ says about how big the observed difference is liable to be.

The t-statistic divided the observed difference by the standard error. Hence the t-statistic is *comparing* the observed difference with how big the difference should be, if the Null is right. Another way of putting it is:

> *The t-statistic tells us how many stnadard errors the sample mean is above or below the value that the Null expects.*

Consider the test statistic in our first example:

$$t \sim 2.38.$$

If we had wanted to, we could have computed the t-statistic from other elements of the test output. Part of the output gives:

```
### Estimate of mu:   105.9
### SE(x.bar):   2.478
```

So the observed difference is

$$\bar{x} - \mu_0 = 105.9 - 100 - 5.9$$

miles per hour. Dividing this by the standard error 2.478, we get:

$$t = \frac{5.9}{2.478} \approx 2.38.$$

The results tell us that our sample mean was 2.38 SEs above the value (100 mph) that the Null was expecting it to be. A little bit of give-and-take—due to chance error in the sampling process—is understandable, but these results are more than two SEs above what the Null would expect. Even before we look at the P-value, a large t-statistic tells us that the Null Hypothesis is implausible!

A t-statistic functions analogously to the z-score that we met back in Chapter 2. Just as a z-score tells us how many SDs an individual is above or below the mean for its groups, so the t-statistic tells us how many SEs the estimator is above or below what the Null expects it to be.

10.2.3.2 How the P-value is Computed

The P-value in this test is supposed to be the probability – given that the Null Hypothesis is true—of getting a t-statistic at least as big as 2.38, the test statistic that we actually got. Therefore, in order to compute this probability we have to know something about the distribution of the test statistic when the Null is true (sometimes this is called the *Null distribution* of t).

The key to the null distribution of t lies in a new family of distributions called the *t-family*. There is one member of the t-family for each positive integer, and this positive integer is called the *degrees of freedom* of the distribution. The distributions are all continuous, so they all have density curves. The following app lets you explore them:

```
require(manipulate)
tExplore()
```

Over a hundred years ago, William Gossett, the person who discovered the t-statistic, was able to relate the t-statistic to the t-curves.

He found that if:

- you take a random sample of size n, at least 2, and
- the population you sample from is normally distributed, and
- the Null Hypothesis is true

then

- the t-statistic has the same distribution as the t-curve whose degrees of freedom is one less than the sample size n.

So to find the P-value for your test, we only have to find the area under a t-curve with 70 degrees of freedom (one less than our sample size of 71), past our t-value 2.3818. We can do this with the r-function `ptGC()`:

```
ptGC(2.3818,region="above",df=70)
```

```
## [1] 0.009975159
```

The shaded area is the P-value!

If you would like to see a graph of the P-value along with the results of the t-test, then in the `ttestGC()` function, simply set the argument `graph` to `TRUE`:

```
ttestGC(~fastest,data=m111survey,mu=100,
         alternative="greater",graph=TRUE)
```

10.2.3.3 Importance of Safety Checks

According to the mathematics that Gosset worked out, the P-value given by R's routine is exactly correct only when we have a taken a random sample from a population that is EXACTLY normal in shape. Of course this never occurs in practice.

However, the P-value given by the `ttestGC()` is approximately correct if the population from which the sample is drawn is approximately bell-shaped. We can't examine the entire population, but we could look at a histogram or a boxplot of the sample to see if *the sample* is approximately bell-shaped. If a histogram or boxplot of the sample reveals not too much skewness and no terribly large outliers, we take it as evidence that the population is roughly bell-shaped, in which case the Null distribution of the test statistic has approximately a t-distribution and so R's P-value is approximately correct.

In general, the larger the sample size, the better the t-curve approximation is, even if the underlying population is far from normal. As a rule of thumb: when the sample size is more than 30, we accept the approximation given by the `ttestGC()` function, even if a histogram of the sample shows skewness and/or outliers.

In order to judge the effect of using the t-curve to approximate P-values, try the app `tSampler()` on a skewed population:

```r
require(manipulate)
tSampler(~income,data=imagpop)
```

Try first at a very small sample size, such as $n = 2$, building the density curve for the t-statistics. After you have had enough, compare the density curve for your t-statistics with the t-curve that would give the correct distribution of the t-statistic if the underlying population were normal. Try again for a larger sample size, such as $n = 30$.

10.2.3.4 Further Examples

10.2.3.4.1 A One-Sided "Less Than" Test Consider the following:

> **Research Question**: *The mean GPA of all UK students is known to be 3.3. Does the `mat111survey` data provide strong evidence that the mean GPA of all GC students is lower than this?*

Step One: Definition of parameter and statement of hypotheses.

Let

$\mu =$ the mean GPA of all GC students.

Then our hypotheses are:

H_0: $\mu = 3.3$

H_a: $\mu < 3.3$

Note: This time, the alternative involves a "less than" statement. Both the original example and this example are "one-sided" tests, but the "side" differs.

Step Two: Safety Check and Test Statistic.

Again we run the test to extract information needed for steps Two through Four:

```r
ttestGC(~GPA,data=m111survey,
        mu=3.3,alternative="less")
```

For the safety check, let's first see if we have more than 30 students in the sample (if we do, we don't have to make a histogram or boxplot of the sample).

```
###  variable  mean     sd  n
###       GPA 3.195 0.5067 70
```

Only one student did not report a GPA. The sample size is 70, which is bigger than 30, so we are safe to proceed.

Next we get the test statistic:

```
###   Test Statistic:       t = -1.733
```

The test statistic is -1.73, so our sample mean of 3.19 was 1.73 SEs below what the Null was expecting.

Step Three: Statement of the P-value

```
###   P-value:      P = 0.04382
```

The P-value was 0.044. It's important to interpret it:

> **Interpretation of the P-value**: *If the mean GPA of all GC students is the same as it is at UK, then there is only about a 4.4% chance of getting a test statistic less than or equal to the one we actually got.*

Note: When the alternative hypothesis involves "less than" then the P-value is the chance of getting results *less* than what we actually got, rather than greater than what we got. The P-value appears as a shaded area in Figure [One-Sided Less-Than P-Value].

```
ptGC(-1.73,region="below",df=70,graph=T)
```

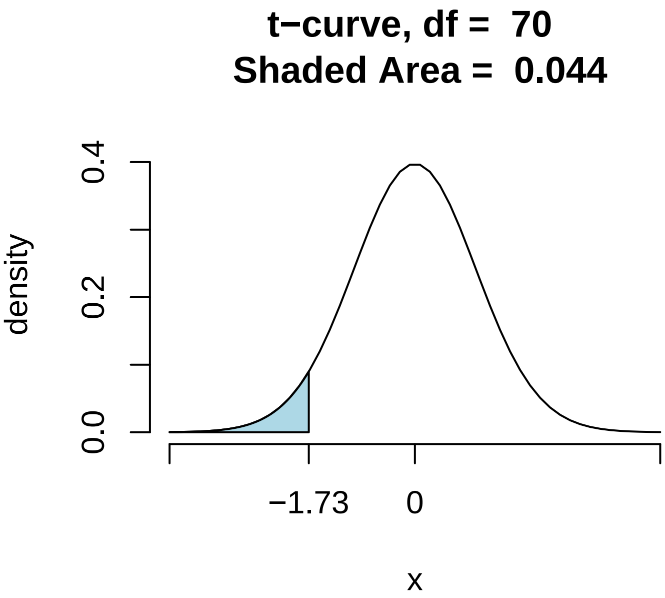


Figure 10.1: One-Sided Less-Than P-Value.

```
## [1] 0.04401866
```

Step Four: Decision

Since the P-value was < 0.05, we reject the Null Hypothesis.

Step Five: Conclusion

This data provided strong evidence that the mean GPA at GC is less than it is at UK.

10.2.3.4.2 A Test From Summary Data Sometimes you do not have access to the raw data in a data frame, and only summary statistics are available. You can still perform tests, if you are given sufficient summary information.

Suppose, for example, that the mean length of 16 randomly-caught Great White sharks is 14 feet, and that the standard deviation of the lengths is 3 feet. You are told that a histogram of the 16 weights looked fairly bell-shaped. You are asked to say whether the data provide strong evidence that the mean length of all Great Whites is less than 15 feet.

Step One: Definition of parameter and statement of hypotheses.

Let

$\mu =$ the mean length of ALL Great White sharks.

Then our hypotheses are:

H_0: $\mu = 15$

H_a: $\mu < 15$

Step Two: Safety Check and Test Statistic

The sample size is 16 which is less than 30, but you are told that the sample is roughly normal. On that basis you hope that the population is roughly normal, so you decide that the P-value provided by `ttestGC()` is approximately correct.

The protocol for entering summary data is the same as it was for confidence intervals, except that you add the `mu` and `alternative` arguments:

```
ttestGC(mean=14,sd=3,n=16,,mu=15,alternative="less")
```

```
##
##
## Inferential Procedures for One Mean mu:
##
##
## Descriptive Results:
##
##  mean sd  n
##    14  3 16
##
##
## Inferential Results:
##
## Estimate of mu:   14
## SE(x.bar):    0.75
##
```

```
## 95% Confidence Interval for mu:
##
##          lower.bound         upper.bound
##          -Inf                15.314788
##
## Test of Significance:
##
##  H_0:  mu = 15
##  H_a:  mu < 15
##
##  Test Statistic:     t = -1.333
##  Degrees of Freedom:   15
##  P-value:        P = 0.1012
```

We see that the test statistic is t = -1.33

Step Three: P-value.

The P-value is 0.10.

> **Interpretation of P-value**: *If the mean length of all Great White sharks is 15 feet, then there is about a 10% chance of getting a t-statistic of -1.33 or less, as we got in our study.*

Step Four: Decision

Since P > 0.05, we do not reject the Null Hypothesis.

Step Five: Conclusion

The data did not provide strong evidence that the mean length of all Great Whites is less than 15 feet.

10.3 Difference of Two Population Means, $\mu_1 - \mu_2$

10.3.1 The Difference of Two Means

We will consider the following Research Question concerning the `mat111survey` data:

> **Research Question**: *Does the data provide strong evidence that GC males drive faster, on average, than GC females do?*

10.3.2 The Five Steps

Step One: Define parameters and state hypotheses.

Although we are interested in only one number (the difference of two means) we have to define both population means in order to talk about that difference. Here we go:

Let

> μ_1 = the mean fastest speed of all GC females.
>
> μ_2 = the mean fastest speed of all GC males.

Now we can state the Hypotheses:

H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 < 0$

Step Two: Safety Check and Test Statistic.

As usual, we run the test in order to gather information needed for steps Two through Four:

```
ttestGC(fastest~sex,data=m111survey,mu=0,alternative="less")
```

This time, the `mu` argument specifies the Null value of $\mu_1 - \mu_2$.

Just as with confidence intervals, the results of the packaged test that R performs are trustworthy only if the following are true:

1. Either we took two independent random samples from two populations, or we did a completely randomized experiment and the two samples are the two treatment groups in that experiment
2. Both populations are roughly normal, or else the sample sizes are both "large enough" (30 or more will do).

We'll assume that the sample of males and the sample of females are like SRSs from their respective populations (at least as far as fastest speed is concerned),so Part 1 is OK. As for Part 2, it is doubtful that the two populations of fastest speeds are normal, but from the Descriptive Results (below) we see that our samples are indeed "big enough":

```
###   group  mean    sd  n
###  female 100.7 17.82 38
###    male 113.9 22.81 30
```

Both sample sizes exceed 30. We squeaked by! If one of them had been less than 30 then we would have made a histogram of it and checked for strong skewness/outliers.

As for the test statistic, it is:

```
###   Test Statistic:       t = -2.725
```

We see that the value of the t-statistic is $t = -2.735$.

Step Three: P-Value

From the output, we see that:

```
###   P-value:      P = 0.004289
```

The P-value is about 0.0043, or 0.43%.

> **Interpretation of P-value**: *If males and females at GC drive equally fast on average, then there is only about a 0.43% chance of getting a t-statistic less than or equal to -2.735, as we did in our study.*

Step Four: Decision about Null Hypothesis

Since P $<$ 0.05, we reject H_0.

Step Five: Conclusion

This data provided strong evidence that GC females drive slower than GC males do, on average.

10.3.3 Under the Hood

10.3.3.1 The Formula for the t-statistic

When we are testing the difference between two means, the formula for the t-statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where:

- $\bar{x}_1 - \bar{x}_2$ is the difference between the sample means
- $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ is the SE for the difference

Remember that if the Null is right, then $\mu_1 - \mu_2$ is 0, so the EV of $\bar{x}_1 - \bar{x}_2$ is 0. Hence the Null is expecting that the numerator of the t-statistic will work out to around 0, give or take a SE or two.

Therefore, just as in the case of one mean, the t-statistic compares the observed difference with the likely size of the observed difference, by dividing the former quantity by the latter.

This is a common pattern for the test statistics for Basic Five parameters:

$$\text{test statistic} = \frac{\text{observed difference}}{\text{SE for the difference}}.$$

Therefore when the test statistic is big, we know that the observed difference is many standard errors away from what the Null expects, and that looks bad for the Null Hypothesis.

10.3.3.2 Order of Groups

In the foregoing example, note that we have a one-sided "less than" alternative hypothesis, even though the research question asked if there was strong evidence that the mean speed of all GC males is GREATER than the mean speed of all GC females. This is because we have defined our first mean μ_1 as the mean speed of all females, so males being faster would imply that $\mu_1 - \mu_2$ is negative. R has certain defaults, usually involving alphabetical order, for which population to consider as the "first" one. If we wish to override these defaults we can do so, using the `first` argument:

```
ttestGC(fastest~sex,data=m111survey,
         mu=0,alternative="less",
         first="male")
```

10.3.3.3 The Degrees of Freedom in Two Samples

Look at the stated Degrees of Freedom in the output for our test:

```
###   Degrees of Freedom:     55.49
```

It is not a whole number, nor it does not have any clear relationship to the sizes of the two samples (40 females, 31 males). What is going on, here?

The explanation is that when we are dealing with the two-sample t-test, then even if both populations are both perfectly normal the t-statistic does not necessarily have a distribution that is exactly the same as one of the t-curves. Instead, R searches for a t-curve that statistical theory suggests will have a distribution fairly similar to the Null distribution of the t-statistic, and it computes the P-value on the basis of that particular t-curve. The "closest" t-curve in this case turns out to have 55.49 degrees of freedom. (Yes, t-curves can have fractional degrees of freedom!)

Fortunately this is not an issue with which we will have to concern ourselves very much. We only need to pay attention to it if we plan to make pictures of the P-values, like the one in Figure [P-Value in Two-Sample Test]:

```
ptGC(-2.735,region="below",df=55.49,graph=T)
```

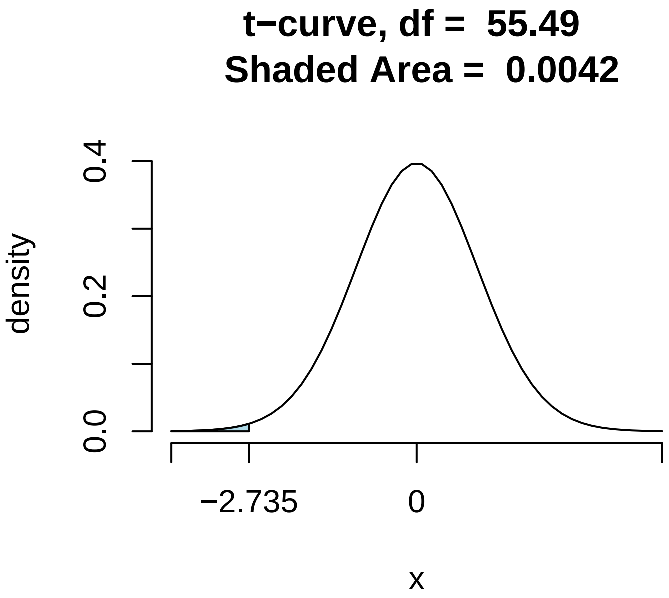


Figure 10.2: P-Value in Two-Sample Test

```
## [1] 0.004179397
```

10.3.4 Additional Examples

10.3.4.1 A Randomized Experiment

Recall the `attitudes` experiment. One of the Research Questions we considered back in Chapter Six was the question

> **Research Question**: *Does the data provide stronge evidence that the suggested race of the defendant was a factor in the sentence recommended by survey participants?*

First, let's take a look at the data again:

```
data(attitudes)
View(attitudes)
help(attitudes)
```

Let's also remind ourselves of the relationship we found in the data:

```
favstats(sentence~def.race,data=attitudes)
```

```
##   .group min Q1 median Q3 max     mean       sd   n missing
## 1  black   3 15     25 40  50 27.77500 14.94891 120       1
## 2  white   4 10     25 40  50 25.84694 15.75602 147       0
```

The defendant whose name suggested he was Black got sentenced to about two years more, on average, than the defendant with the White-sounding name. This goes along with some suspicion we might have that there was some lingering racial prejudice among at least some of the (mostly White) survey participants.

Let us now address our Research Question with a test of significance.

Step One: Define parameters and state hypotheses

As for the parameters, let:

> $\mu_1 =$ the mean sentence recommended by all 267 survey participants, if all of the them were to look at a form in which the suggested race of the defendant was Black

> $\mu_2 =$ the mean sentence recommended by all 267 survey participants, if all of the them were to look at a form in which the suggested race of the defendant was White

(You may wish to Review Chapter 8 on how to define parameters of interest when an experiment has been conducted).

Now for the hypotheses:

> H_0: $\mu_1 - \mu_2 = 0$

> H_a: $\mu_1 - \mu_2 > 0$

Step Two: Safety Check and Test Statistic

Once again, we begin by running the test code:

```
ttestGC(sentence~def.race,data=attitudes,
         mu=0,alternative="greater")
```

Safety Check: It was a matter of chance which participant was assigned which type of survey form, so we conducted a randomized experiment. As for the underlying population being roughly normal, we don't have to worry about that because the descriptive results in the output shows us that both sample sizes were well above 30:

```
###  group  mean    sd   n
###  black 27.77 14.95 120
###  white 25.85 15.76 147
```

As for the test statistic, we see:

```
###   Test Statistic:       t = 1.023
```

The observed difference between the treatment group means is about 1 SD above the value of 0 that the Null expects.

Step Three: P-value

From the output we find that the P-value is 0.1536.

> **Interpretation of P-Value**: *If the sentences recommended by the students in the study were, on average, unaffected by the suggested race of the defendant, then there is about a 15.36% chance of getting a test statistic at least a big as the one we got.*

Step Four: Decision

Since P $>$ 0.05, we do not reject the Null Hypothesis.

Step Five: Conclusion

This study did not provide strong evidence that the students in the survey were affected by race when they recommended a sentence.

10.3.4.2 A Two-Sided Test

So far all of the tests we have conducted have been *one-sided*: either the Alternative Hypothesis states that the value of the parameter is GREATER than the Null's value. or else it says that the parameter is LESS than the Null's value.

A one sided-test comes along with a one-sided approach to computing a P-value.

For example, when the alternative says GREATER, then the P-value is the chance of the test statistic being equal to or GREATER THAN the value we actually observed for it. On the other hand, when the alternative says LESS, then the P-value is the chance of the test statistic being LESS THAN or equal to the value we actually observed for it.

As a rule, people work with a one-sided Alternative Hypothesis if they have some good reason to suspect—prior to examining the data in the study at hand—that the value of the parameter lies on the side of the Null's value that is specified by the Alternative. For example, in the "race and recommended sentence" study in the first example of this section, we might have had some prior data or other knowledge that suggests that GC students might be inclined to mete out heavier sentences to a black defendant, and we are using the study to check that view. In that case the one-sided hypothesis would be an appropriate choice.

But suppose that we really don't have any prior reason to believe that a black defendant would be treated worse, and that we are conducting the study just to see if race makes a difference in sentencing, *one way or the other*. In that case, a *two-sided* test of significance might be more appropriate.

Let's re-do the race and sentencing test, but this time in a two-sided way:

Step One: Define parameters and state hypotheses

This is the same as before. Let

> μ_1 = the mean sentence recommended by all 267 survey participants, if all of the them were to look at a form in which the suggested race of the defendant was Black

> μ_2 = the mean sentence recommended by all 267 survey participants, if all of the them were to look at a form in which the suggested race of the defendant was White

Now for the hypotheses:

H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 \neq 0$

Step Two: Safety Check and Test Statistic

Safety Check: Same as before. Randomized experiment, large group sizes: we are safe.

For the test statistic, we run the test with the `alternative` argument set to "two.sided":

```
ttestGC(sentence~def.race,data=attitudes,
       mu=0,alternative="two.sided")
```

```
##
##
## Inferential Procedures for the Difference of Two Means mu1-mu2:
##  (Welch's Approximation Used for Degrees of Freedom)
##   sentence grouped by def.race
##
##
## Descriptive Results:
##
##  group  mean    sd   n
##  black 27.77 14.95 120
##  white 25.85 15.76 147
##
##
## Inferential Results:
##
## Estimate of mu1-mu2:  1.928
## SE(x1.bar - x2.bar):  1.884
##
## 95% Confidence Interval for mu1-mu2:
##
##           lower.bound         upper.bound
##           -1.782671           5.638793
##
## Test of Significance:
##
##  H_0:  mu1-mu2 = 0
##  H_a:  mu1-mu2 != 0
##
##  Test Statistic:     t = 1.023
##  Degrees of Freedom:   259.1
##  P-value:        P = 0.3072
```

Again see that t is about 1.02. The observed difference between the treatment group means is about 1 SD above the value of 0 that the Null expects.

Step Three: P-value

This time P-value is 0.3072, as compared to 0.1536 in the one-sided test. What has happened?

In a two-sided test, the P-value is the probability (assuming that the Null is true) of getting a test statistic *at least as far from zero* (either on the positive or the negative side) as the test statistic that we actually got. Graphically, it looks like Figure [Two-Sided P-Value]:

```
ptGC(c(-1.02,1.02),region="outside",
     df=259.052,graph=TRUE)
```

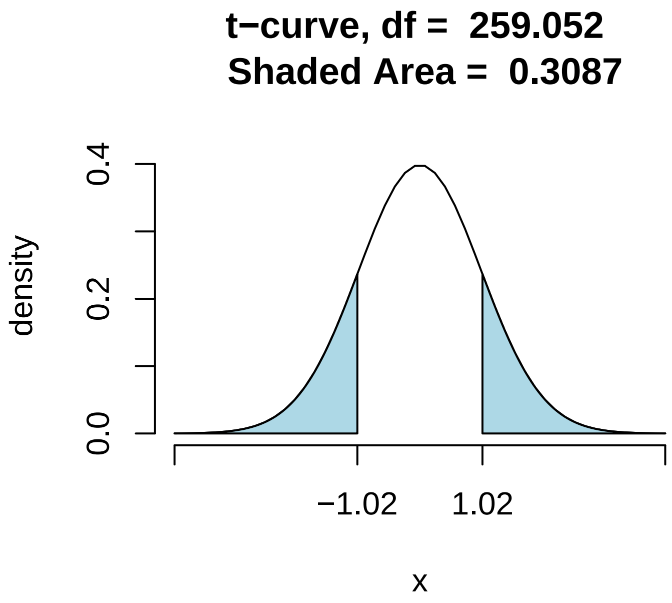


Figure 10.3: Two-Sided P-Value

```
## [1] 0.3086801
```

Due to the symmetry of the t-curve, a two-sided P-value will generally be twice the corresponding one-sided P-value.

The reason we incorporate both "sides" of the shaded area into the P-value is that at the outset we were indifferent as to whether $\mu_1 = \mu_2$ was positive or negative. The one-sided hypothesis is offered presumably by someone who already has some prior evidence suggesting that the test statistic should turn out a particular way—either positive or negative. Since the P-value measures strength of evidence against the Null, with smaller P-values providing more evidence against the Null, it makes sense that for the same test statistic a one-sided test should return a smaller P-value than a two-sided test. The one sided-test has the evidence from the current data *together with some prior evidence*, whereas the two-sided test has only the evidence from the current data.

How do we interpret the P-value in a two-sided test? We simply talk about "distance from zero." So for this study we say:

> *If the suggested race of defendant has no effect on recommended sentence, then there is about a 30.72% chance of getting a t-statistic at least as far from zero as the one we got in our study.*

Step Four: Decision

Since P $>$ 0.05, we do not reject the Null Hypothesis.

Step Five: Conclusion

This study did not provide strong evidence that the students in the survey were affected by race when they recommended a sentence.

10.4 Mean of Differences μ_d

In this case we usually have paired data.

10.4.1 Introductory Research Question

Recall the `labels` data, where participants were asked to rate peanut butter from a jar labeled Jiff, and to rate peanut butter from a jar labeled Great Value, when—unknown to the subjects—both jars contained the same peanut butter.

```
data(labels)
View(labels)
help(labels)
```

Recall that subjects tended to rate the Jiff-labeled jar higher:

```
diff <- labels$jiffrating-labels$greatvaluerating
favstats(diff)
```

```
##  min Q1 median Q3 max     mean       sd  n missing
##   -5  1    2.5  4   8 2.366667 2.809876 30       0
```

But does this data provide strong evidence that the GC population as a whole (not just this sample of 30 students) would rate Jiff higher? In order to address this question, we perform a test of significance:

10.4.2 The Five Steps

Step One: Define parameters and state hypotheses.

The study had a repeated measures design (every participant was measured twice). The parameter of interest is

$$\mu_d = \text{mean difference in ratings (Jiff minus Great Value) for ALL Georgetown College students}$$

As for hypotheses:

H_0: $\mu_d = 0$ (jar makes no difference)

H_a: $\mu_d > 0$ (Jiff jar rated higher, on average)

We chose the one-sided alternative because Jiff is the more expensive brand, and we have some reasons—based on prior studies about price and perception of quality—to believe that pricier brands are perceived to be of higher quality.

Step Two: Safety Check and Test Statistic.

Hopefully the 30 students in the study were like a random sample from the GC population. The sample size is right at the boundary-value of 30, so let's go ahead and look at a graph of the sample of differences, checking for strong skewness or big outliers.

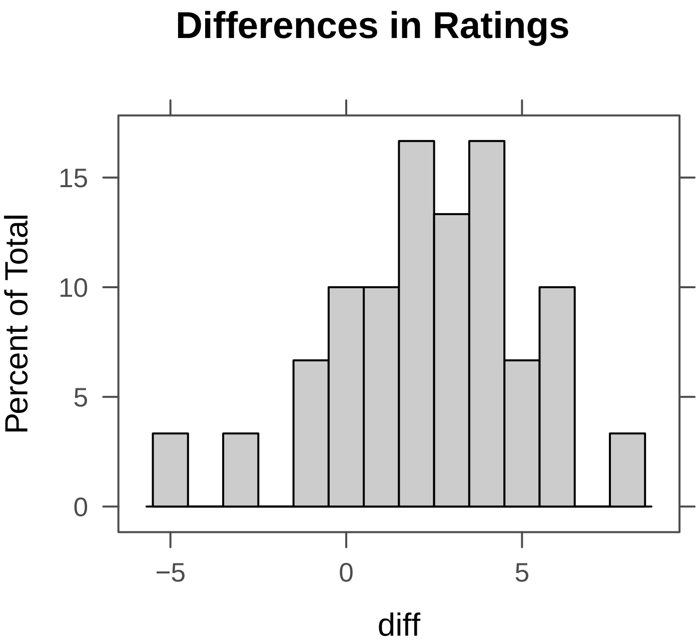


Figure 10.4: Rating Differences

A histogram of the differences is shown in Figure [Rating Differences]. There is a little skewness to the left, perhaps, but not nearly enough to worry about, especially considering our reasonably large sample size. We are safe.

For the test statistic and other information in steps Two through Four, we run `ttestGC()`, using a special formula-style for matched pairs or repeated-measures data:

```
ttestGC(~jiffrating-greatvaluerating,
        data=labels,mu=0,alternative="greater")
```

```
##
##
## Inferential Procedures for the Difference of Means mu-d:
##   jiffrating minus greatvaluerating
##
##
## Descriptive Results:
##
##                     Difference mean.difference sd.difference  n
##  jiffrating - greatvaluerating           2.367          2.81 30
##
##
## Inferential Results:
##
## Estimate of mu-d:     2.367
## SE(d.bar):    0.513
##
## 95% Confidence Interval for mu-d:
##
##           lower.bound         upper.bound
##           1.494996            Inf
```

```
##
## Test of Significance:
##
##  H_0:  mu-d = 0
##  H_a:  mu-d > 0
##
##  Test Statistic:     t = 4.613
##  Degrees of Freedom:   29
##  P-value:        P = 3.711e-05
```

The t-statistic is about 4.6133. The mean of the sample differences in ratings is 4.6 SDs above what the Null expects it to be!

Step Three: P-value

Once again, R finds an approximate P-value by using a t-curve (with degrees of freedom one less than the number of pairs in the sample). As we can read from the test output, the P-value is 3.7×10^{-5}, or 0.000037 or about 0.00004, which is very small indeed. The interpretation is:

> **Interpretation of P-Value**: *If ratings are unaffected by labels, then there is only about 4 in 100,000 chance of getting a t-statistic at least as big as the one we got in our study.*

Step Four: Decision

Since P $<$ 0.05, we reject the Null.

Step Five: Conclusion

This data provided very strong evidence that on average people will rate the peanut butter with a more expensive brand-label more highly than the same peanut butter labeled with a less-expensive brand name.

10.4.3 Under the Hood

The formula for the t-statistic is:

$$t = \frac{\hat{d} - \mu_{d,0}}{s_d/\sqrt{n}},$$

where

- $\hat{d}$ is the mean of the sample differences
- $\mu_{d,0}$ is what the Null Hypothesis believes μ_d is (usually this is 0)
- $s_d/\sqrt{n}$ is the SE for $\hat{d}$

So once again the t-statistic follows the familiar pattern:

$$\text{test statistic} = \frac{\text{observed difference}}{\text{SE for the difference}}.$$

It tells us how many SEs the estimator $\hat{d}$ is above or below what the Null expects it to be.

This familiar pattern—which applies to each of the three Basic Five parameters involving means—-deserves a name. Let's call it a *z-score-style* statistic, since it works like a z-score, measuring the number of SEs the estimator is above or below what the Null expects the parameter to be.

10.4.4 Additional Examples

10.4.4.1 Height and Ideal Height

Recall that in the `m111survey` participants were asked their height, and the ideal height that they wanted to be. Let's consider the following

> **Research Question**: *Does the study provide strong evidence that, on average, the ideal height of GC students differs from their actual height?*

Step One Define Parameter and State Hypotheses

Let

> μ_d = the mean difference (ideal height minus actual height) for all GC students

Then our hypotheses are:

> H_0: $\mu_d = 0$
>
> H_a: $\mu_d \neq 0$

Note that we have here a two-sided test. Apparently we did not have any prior idea as to whether students want to be taller than they are, on average, or shorter than they are on average.

Step Two Safety Check and Test Statistic

We get the information we need:

```
ttestGC(~ideal_ht-height,data=m111survey,
         mu=0,alternative="two.sided")
```

```
##
##
## Inferential Procedures for the Difference of Means mu-d:
##   ideal_ht minus height
##
##
## Descriptive Results:
##
##         Difference mean.difference sd.difference  n
##  ideal_ht - height           1.946         3.206 69
##
##
## Inferential Results:
##
## Estimate of mu-d:     1.946
## SE(d.bar):    0.3859
##
## 95% Confidence Interval for mu-d:
##
##           lower.bound         upper.bound
##           1.175528            2.715776
```

```
##
## Test of Significance:
##
##  H_0:  mu-d = 0
##  H_a:  mu-d != 0
##
##  Test Statistic:     t = 5.041
##  Degrees of Freedom:   68
##  P-value:        P = 3.652e-06
```

There are 69 people who gave usable answers, so the sample is large enough that we don't have to verify that it looks roughly normal. Also, we are assuming that the sample is like a simple random sample, as far as variables like height are concerned, so we are safe to proceed.

The t-statistic is about 5.04. The sample mean of differences is more than 5 SEs above what the Null expected it to be!

Step Three: P-value

The P-value is 3.65×10^{-6}, about 3 in a million.

> **Interpretation of P-Value**: *If on average GC students desire to be no taller and no shorter than they actually are, then there is only about a 3 in one million chance of getting a test-statistic at least as far from 0 as the one we got in this study.*

Step Four: Decision

P $<$ 0.05, so we reject the Null.

Step Five: Conclusion

This study provided very strong evidence that on average GC students want to be taller than they actually are.

10.4.4.1.1 Tests and Confidence Intervals In the output from the previous example, look at the confidence interval for μ_d that is provided. (Recall that by default it is a 95%-confidence interval.) Notice that it does not contain 0, so according to the way we interpret confidence intervals we are confident that μ_d does not equal 0. Of course 0 is what the Null Hypothesis believes μ_d is, so we could just as well say that we are confident that the Null is false.

This is an example of a relationship that holds quite generally between two-sided tests and two-sided confidence intervals:

> **Test-Interval Relationship**: When a 95% confidence interval for the population parameter does not contain the Null value μ_0 for that parameter, then the P-value for a two-sided test
>
> $H_0\ \mu = \mu_0$
>
> $H_a\ \mu \neq \mu_0$
>
> will give a P-value less than 0.05, and so the Null will be rejected when the cut-off value α is 0.05.

Also:

> When a 95% confidence interval for the population parameter DOES contain the Null value μ_0 for that parameter, then the P-value for a two-sided test will give a P-value greater than 0.05, and so the Null will not be rejected when the cut-off value α is 0.05.

This all makes, sense, because a confidence interval gives the set of values for the population parameter that could be considered reasonable, based on the data at hand. When a particular value is inside the interval, it is reasonable. When outside, it is not reasonable.

The relationship also holds for other confidence levels, provided that you adjust the cut-off value α. In general:

> When a $100 \times (1 - \alpha)\%$ confidence interval for the population parameter does not contain the Null value μ_0 for that parameter, then the P-value for a two-sided test will give a P-value less than α.

> When a $100 \times (1 - \alpha)\%$ confidence interval for the population parameter DOES contain the Null value μ_0 for that parameter, then the P-value for a two-sided test will give a P-value greater than α.

You may also have noticed that in one sided tests, the `ttestGC()` function produces one-sided confidence intervals. Recall for example:

```
t.test(fastest~sex,data=m111survey,mu=0,alternative="less")
```

The one-sided 95% confidence interval extended from negative infinity to -5.17. This interval does not contain zero, which is the Null's value for $\mu_1 - \mu_2$, and so the Null is rejected with a P-value of less than 0.05.

The relationship between tests and confidence intervals holds not only for two-sided tests and two-sided intervals, but also for one-sided tests and the corresponding one-sided confidence intervals.

10.4.4.2 Repeated Measures, or Two Independent Samples?

Let's look again at the `labels` data. By now we are convinced that the data provide strong evidence that GC students (the population from which the data was drawn) are inclined to rate the same product more highly when it is packaged as an "expensive" product that when it as packaged as "cheap". But notice that in addition to recording the ratings, researchers also recorded the sex of each participant in the study. This raises another interesting research question:

> **Research Question**: *Who will be more affected, on average, by the label on a product: a GC female or a GC male?*

The original study had a repeated measures design, which resulted in the parameter of interest being a difference of two means. However, when we take sex into consideration things change substantially. Although we still think about the difference between the ratings, we are primarily interested in whether the mean difference for all GC females and the mean differences for all GC guys differ. Thus, the parameter of interest will be a differences of means, where each mean is itself a mean of differences!

Here are the five steps:

Step One: Definition of Parameters and Statement of Hypotheses

Let

> μ_d^f = the mean difference in rating (Jiff minus Great Value) for all GC females, if all of them could have participated in this study.

μ_d^m = the mean difference in rating (Jiff minus Great Value) for all GC males, if all of them could have participated in this study.

The hypotheses are:

H_0: $\mu_d^f - \mu_d^m = 0$

H_0: $\mu_d^f - \mu_d^m \neq 0$

We chose a two-sided test because we do not have any prior evidence indicating that the difference is positive, or that it is negative.

Step Two Safety Check and Test Statistic

This time we are dealing with two independent (and hopefully random) samples from our two populations. The samples sizes, though, are not so large:

```
diff <- labels$jiffrating - labels$greatvaluerating
```

[Note that the sample mean difference for the females was a good bit larger than the sample mean difference for the males.]

We have only 15 students in each sample. Since these sample sizes are less than our "safe level" of 30, we really ought to examine them to see if they show signs of strong skewness or outliers.

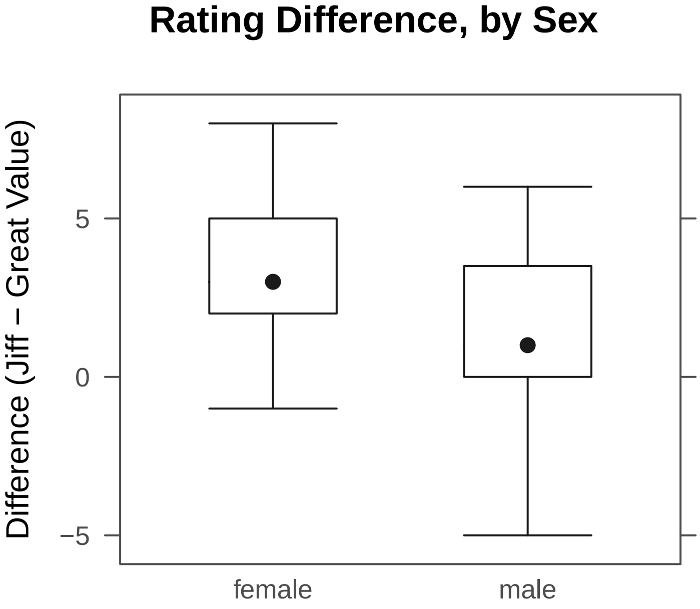


Figure 10.5: Rating Differences for Males and for Females

The results appear in Figure [Rating Differences for Males and Females]. There is some left skewness in the sample of males. We will go ahead and perform the test, but the P-value should be considered a bit dodgy.

For the test, we need to add the differences to our data frame. We accomplish this step as follows:

```
labels$diff <- diff
```

Now R will recognize **diff** and process it appropriately, with the usual formula-data input:

```
ttestGC(diff~sex,data=labels,
         mu=0,alternative="two.sided")
```

```
##
##
## Inferential Procedures for the Difference of Two Means mu1-mu2:
##  (Welch's Approximation Used for Degrees of Freedom)
##   diff grouped by sex
##
##
## Descriptive Results:
##
##   group  mean    sd  n
##  female 3.533 2.264 15
##    male 1.200 2.883 15
##
##
## Inferential Results:
##
## Estimate of mu1-mu2:  2.333
## SE(x1.bar - x2.bar):  0.9465
##
## 95% Confidence Interval for mu1-mu2:
##
##           lower.bound         upper.bound
##           0.389583            4.277084
##
## Test of Significance:
##
##  H_0:  mu1-mu2 = 0
##  H_a:  mu1-mu2 != 0
##
##  Test Statistic:     t = 2.465
##  Degrees of Freedom:   26.51
##  P-value:        P = 0.02047
```

The value of the t-statistic is 2.47. The observed difference between the sample mean differences for the females and the males is 2.47 SEs above 0 (the value the Null expected it to be).

Step Three: P-value.

The two-sided P-value is 0.02047.

Step Four: Decision

Since $P < 0.05$, we reject the Null.

Step Five: Conclusion

This data provided strong evidence that GC females are more affected by labels than GC males are.

We might wonder what this all means. Do GC females pay more attention to their surroundings—in particular, to labels on products?? Who knows!

10.5 One Population Proportion p

Recall the distribution of the variable **sex** in the `m111survey` data:

```
rowPerc(xtabs(~sex,data=m111survey))
```

```
##
## sex female  male Total
##      56.34 43.66   100
```

56% of the sample—-more than half—were females. This suggests the following:

> **Research Question**: *Does this data constitute strong evidence that a majority of the GC student body are female?*

10.5.1 The Five Steps

Step One: Definition of Parameter and Statement of Hypotheses

Let

> p = the proportion of females in the GC student population

Then our hypotheses are:

> H_0: $p = 0.50$

> H_a: $p > 0.50$

We chose 0.50 as the Null value for p, because we wonder if females are in the majority (proportion bigger than 0.5) at GC. Choosing the Null in this way allows the Alternative to express the idea that females are in the majority. If we had let the Null say something else ($p = 0.48$ for example) then the Alternative ($p > 0.48$) would leave a bit of room for females NOT be in the majority.

Step Two: Safety Check and Test Statistic

As with confidence intervals for one population proportion, the safety check for significance tests includes thinking about whether our sample is like a simple random sample—at least with respect to the variables we are measuring. In this study, our sample consists of all the students taking MAT 111 in a given semester. Now people often take MAT 111 for reasons connected to the requirements of their prospective major. If the two sexes differ with regard to what majors they prefer, then this sample could be a biased one. On these grounds, our test is going to be a bit untrustworthy.

Let's go ahead and run the test. As with confidence intervals, when we are interested in a single proportion we can use `binomtestGC()`.

```
binomtestGC(~sex,data=m111survey,p=0.5,
            alternative="greater",success="female")
```

```
## Exact Binomial Procedures for a Single Proportion p:
##  Variable under study is sex
##
## Descriptive Results:  40 successes in 71 trials
##
## Inferential Results:
##
## Estimate of p:    0.5634
## SE(p.hat):    0.0589
##
## 95% Confidence Interval for p:
##
##           lower.bound         upper.bound
##           0.458929            1.000000
##
## Test of Significance:
##
##  H_0:  p = 0.5
##  H_a:  p > 0.5
##
##  P-value:        P = 0.1712
```

Notice what we had to do in order to get all that we needed for the test:

- we had to specify the Null value for p by using the `p` argument;
- we had to set `alternative` to "greater" in order to accommodate our one-sided Alternative Hypothesis;
- we had to say what the proportion is a proportion of: is it a proportion of females or a proportion of males? We accomplished this by specifying what we considered to be a success when the sample was tallied. Since we are interested in the proportion of females, we counted the females as a success, so we set the `success` argument to "female".

Looking at the output, you might wonder what the test statistic is. In `binomtestGC()`, it's pretty simple: it's just the number of females: 40 out of the 71 people sampled.

Step Three The P-value

This is 0.1712. It is the probability of getting at least 40 females in a sample of size 71, if the population consists of 50% females. So we might interpret the P-value as follows:

> **Interpretation of P-Value**: *If only half of the GC population is female, then there is about a 17% chance of getting at least as many females (40) as we actually got in our sample.*

Step Four Decision

We do not reject the Null, since the P-value is above our standard cut-off of 5%.

Step Five: Conclusion

This survey data did not provide strong evidence that females are in the majority at Georgetown College.

10.5.2 Under the Hood

Note that the test statistic did not follow the "z-score" pattern of the three previous tests involving means. The test statistic is simply the observed number of successes.

binomtestGC() gets its P-value straight from the binomial distribution. If we wanted, we could get the same P-value using the familiar **pbinomGC** function:

```
pbinomGC(39,region="above",size=71,prob=0.5)
```

It may be worth recalling that, as with any test function in the tigerstats package, you can get a graph of the P-value simply by setting the argument graph to TRUE:

```
binomtestGC(~sex,data=m111survey,p=0.5,
            alternative="greater",success="female",
            graph=TRUE)
```

Now we know from Chapter 7 that when the number of trials n is large enough, then the distribution of a binomial random variable looks very much like a normal curve. In fact, when this is so, binomtestGC() will actually use a normal curve to approximate the P-value!

There is another test that makes use of the normal approximation in order to get the P-value, and it is encapsulated in the proptestGC() function:

```
proptestGC(~sex,data=m111survey,p=0.50,
        alternative="greater",success="female")
```

```
##
##
## Inferential Procedures for a Single Proportion p:
##  Variable under study is sex
##  Continuity Correction Applied to Test Statistic
##
##
## Descriptive Results:
##
##  female  n estimated.prop
##      40 71         0.5634
##
##
## Inferential Results:
##
## Estimate of p:    0.5634
## SE(p.hat):    0.05886
##
## 95% Confidence Interval for p:
##
##          lower.bound         upper.bound
##          0.466564            1.000000
##
## Test of Significance:
##
##  H_0:  p = 0.5
##  H_a:  p > 0.5
##
##  Test Statistic:     z = 0.9571
##  P-value:        P = 0.1692
```

Here the test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}},$$

where

- $\hat{p}$ is the sample proportion,
- p_0 is the Null value of p,
- the denominator is the standard error of $\hat{p}$.

Like the test statistics for means, it is "z-score style": it measures how many standard errors the sample proportion is above or below what the Null Hypothesis expects it to be.

When the sample size is large enough and the Null is true, the distribution of this test statistic is approximately normal with mean 0 and standard deviation 1. in other words, it has the standard normal distribution. Therefore the P-value comes from looking at the area under the standard normal curve past the value of the test statistic. (We also apply a small *continuity correction* in order to improve the approximation; consult GeekNotes for details.)

Many people will say that the sample should contain at least 10 successes and 10 failures in order to qualify as "big enough." (The `binomtestGC()` gives an "exact" P-value and is not subject to this restriction.)

For tests involving a single proportion, you may use either `proptestGC()` or `bimom.testGC()`. The choice is up to you, but we have a slight preference for `binom.testgC()`, since it delivers "exact"" P-values regardless of the samples size.

10.5.3 Additional Examples

10.5.3.1 An ESP Experiment: Summary Data, and Care in Drawing Conclusions

The following example deals with a study that was actually conducted at the University of California at Davis around the year 1970.

The researcher, a para-psychologist named Charles Tart, was looking for evidence of extrasensory perception (ESP) in human beings. Tart designed a machine that he called the "Aquarius", which could generate numbers. He set the machine to generate either 1, 2, 3, or 4. For each subject in the study, the Aquarius machine generated 500 numbers, and after each number was generated the subject was asked to guess the number. After each guess was submitted, the machine would show the subject the "correct" number by flashing one of four numbered lights, and then would generate the next number.

One study involved fifteen subjects, who made a total of

$$15 \times 500 = 7500$$

guesses. Out of this total, 2006 of the guesses were correct.

Let's think about these results. If the subjects have no ESP at all, then it would seem that they would be reduced to guessing randomly, in which case the chance of a correct guess should be 1 in 4, or 0.25. Then the number X of correct guesses in these 7500 tries would be a binomial random variable, with $n = 7500$ trials and chance of success $p = 0.25$. The expected number of correct guesses would then be:

$$np = 0.25 \times 7500 = 1875.$$

The observed number of correct guesses was 2006, which is 131 more than expected. We wonder if this observation constitutes strong evidence that at least some of the subjects had at least some ESP powers.

Let's try a test of significance to find out:

Step One: Define Parameter and State Hypotheses.

Let

$p =$ probability that a subject will guess correctly.

Then our hypotheses are:

H_0: $p = 0.25$

The Null expresses the view that the subjects have no ESP; they are just randomly guessing.

H_a: $p > 0.25$

The Alternative expresses the view that the subjects can do something better than guess randomly.

Step Two: Safety Check and Test Statistic.

The test statistic is easy: it's the 2006 correct guesses. As for the safety check: we are not picking from a population so we don't have to check whether we have taken a simple random sample from a population. If the Null is right, the subjects are just guessing randomly, so the number of correct guesses would be `binom(7500,0.25)`, as explained above.

Step Three: P-value

Let's run the test. we have summary data, so we don't have to specify what counts as a success; we only need to enter the number of successes and the number of trials:

```
binomtestGC(2006,n=7500,p=0.25,alternative="greater")
```

```
## Exact Binomial Procedures for a Single Proportion p:
##  Results based on Summary Data
##
## Descriptive Results:  2006 successes in 7500 trials
##
## Inferential Results:
##
## Estimate of p:    0.2675
## SE(p.hat):    0.0051
##
## 95% Confidence Interval for p:
##
##           lower.bound         upper.bound
##           0.259061            1.000000
##
## Test of Significance:
##
##  H_0:  p = 0.25
##  H_a:  p > 0.25
##
##  P-value:        P = 3e-04
```

Hmm, the P-value is very small: about 0.0003.

> **Interpretation of P-Value**: *If all of the subjects had only a 25% chance of guessing correctly, then there is only about a 3 in 10,000 chance that they would get a total of at least 2006 correct guesses, as they did in this study.*

Step Four: Decision

We reject the Null, since the P-value was so very small.

Step Five: This study provided very strong evidence that at least some of the subjects in the study had more than a 1-in-4 chance of guessing correctly.

Notice how carefully the conclusion was framed, in terms of the probability of a correct guess rather than whether or not anyone has ESP. The reason for this caution is that the test of significance only addresses the mathematical end of things: in particular, it does not address whether the study was designed well enough so that a p being larger than 0.25 would be *the very same thing as* some subjects possessing ESP powers.

It turned out, in fact, that there was a problem with the design of the study: the machine's random-number generation program was defective, in such a way that it rarely generated the same number twice in a row. Some of the subjects probably recognized this pattern and adopted the strategy of guessing randomly one of the three numbers that had not been generated in the previous round. These subjects would then improve their chance to 1-in-3 for the rest of their trials.

There is an important moral to this story:

> *A significance test only tells you whether or not it is reasonable to believe that the results could be obtained by chance if the Null is true. It does NOT tell you how the Alternative Hypothesis should be interpreted. When the study has a flawed design, the results of the test will not mean what you expect them to mean!*

10.6 Difference of Two Proportions $p_1 - p_2$

Recall that in the `mat111survey`, participants were asked their sex, and were also asked whether they believed in love at first sight. The results were as follows:

```
SexLove <- xtabs(~sex+love_first,data=m111survey)
rowPerc(SexLove)
```

```
##         love_first
## sex          no    yes  Total
##   female  55.00  45.00 100.00
##   male    74.19  25.81 100.00
```

In the sample, the females were much more likely than the males to believe in love at first sight (45% vs. 25.81%), but these figures are based on fairly small samples. We wonder whether they provide strong evidence that in the GC population at large females are more likely than males to believe in love at first sight.

Now this is a question about the relationship between two categorical variables, so we could address it with the chi-square test from Chapter 3:

```
chisqtestGC(SexLove)
```

However, *when the explanatory and reponse variables both have only two values*, then it is possible to construe the question as a question about the difference between two proportions. A two-proportions test confers some advantages in terms of how much information we extract from the data.

10.6.1 The Five Steps

Step One: Definition of Parameters and Statement of Hypotheses

Let

p_1 = the proportion of all GC females who believe in love at first sight

p_2 = the proportion of all GC males who believe in love at first sight

The our hypotheses are:

H_0: $p_1 - p_2 = 0$

H_a: $p_1 - p_2 > 0$

Step Two: Safety Check and Test Statistic

The Safety Check is the same as for confidence intervals:

- We should have taken two independent simple random samples from two populations, OR we should have done a completely randomized experiment with two treatment groups;
- the sample size should be "large enough" (see below).

In this current case, we have two samples from two populations (the GC gals and the GC guys). There is some concern as to whether the samples are really "like" simple random samples, but since one's decision to take MAT111 (and hence get into the survey) is probably unrelated to whether one believes in love at first sight, maybe we can count ourselves as "safe" on this point.

As for a "large enough" sample size, we plan to use `proptestGC()`, just as we used it to make confidence intervals for $p_1 - p_2$. Hence the safety criteria are the same: if the number of yesses and nos in both samples exceeds 10, then we can trust the approximation to the P-value that the test provides. If in one of the samples there are not at least ten yesses, and not at lest ten nos, then the computer will issue a warning.

```
proptestGC(~sex+love_first,data=m111survey,
        success="yes",p=0,alternative="greater")
```

The descriptive results are given first:

```
###        yes  n estimated.prop
### female  18 40         0.4500
### male     8 31         0.2581
```

From these results we can see that there are fewer than ten males who answered "yes", and sure enough the routine delivers its warning:

```
### WARNING:  In at least one of the two groups,
### number of successes or number of failures is below 10.
### The normal approximation for confidence intervals
### and P-value may be unreliable.
```

For the sake of seeing the entire example, we will proceed anyway.

When we come to the inferential results, we see the estimator $\hat{p}_1 - \hat{p}_2$, and the standard error of this estimate:

```
### Estimate of p1-p2:   0.1919
### SE(p1.hat - p2.hat):     0.1112
```

Note that the estimate is not even two standard errors above the value of 0 that the Null expects it to be. The results of this study are not very surprising, if the Null is in fact correct.

The formula for the test statistic is:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

so once again it has "z-score style", telling us how many SEs the estimator is above or below the 0 that the Null expects. In this case its numerical value, according to the output, is:

```
###   Test Statistic:       z = 1.726
```

Step Three: P-Value

The output says that the P-value for the one-sided test is:

```
###   P-value:      P = 0.04216
```

According to this test, if in the GC population females are equally likely to believe in love at first sight, then there is about a 4.2% chance for the differences between the sample proportions (45%-25.8%) to be at least a big as it was observed to be.

Step Four: Decision

Since $P = 0.042 < 0.05$, we reject the Null.

Step Five: Conclusion

This data provided strong evidence that GC females are more likely than GC males to believe in love at first sight.

We stress, though, that we completed the test in the presence of a warning from R. There do exist other routines that attempt to provide better approximations to the P-value when samples are small, and they result in somewhat different conclusions. (One such test is `prop.test()`; see the GeekNotes.)

10.6.2 Working With Summary Data

Imagine an experiment in which there are 2000 subjects who are randomly divided into two groups of size 1000 each. All subjects get a shot in mid-October. Member of Group A (control) gets a shot that feels just like getting a flu vaccine but which actually delivers an inert substance, whereas every member of group B gets a shot containing the current flu vaccine. (Note the blinding.) Doctors monitor the subjects until Spring, and record whether or not each subject caught the flu. Of the 1000 members of Group A, 80 got the flu, whereas of the 1000 members of Group B, 40 got the flu. We are interested in the following:

> **Research Question**: *Do the results provide strong evidence that the flu vaccine was effective in reducing the risk of flu that year?*

We have summary data, so we use `proptestGC()` for summary data just as we did for confidence intervals.

Step One: Define parameters and state hypotheses.

Let:

> p_1 = the proportion of all 2000 subjects who would have caught the flu, if all of them had been given the inert substance

> p_2 = the proportion of all 2000 subjects who would have caught the flu, if all of them had been given the real flu vaccine

Then the hypotheses are:

> H_0: $p_1 - p_2 = 0$

> H_a: $p_1 - p_2 > 0$

Step Two: Safety Check and Test Statistic

Let's run the test:

- We need to tell R the number of successes in the two samples, and we do this by putting `c(80,40)` into an argument called `x`.
- We also need to tell R the sample sizes, so we put `c(1000,1000)` into an argument called `n`.
- We tell R that the Null thinks $p_1 - p_2$ is 0, by setting an argument `p` to 0.
- Finally, we tell R what the alternative hypothesis looks like, by setting `alternative` to `two.sided`:

```
proptestGC(x=c(80,40),n=c(1000,1000),
          p=0,alternative="two.sided")
```

```
##
##
## Inferential Procedures for the Difference of Two Proportions p1-p2:
##  Results taken from summary data.
##
##
## Descriptive Results:
##
##         successes    n estimated.prop
## Group 1        80 1000           0.08
## Group 2        40 1000           0.04
##
##
## Inferential Results:
##
## Estimate of p1-p2:    0.04
## SE(p1.hat - p2.hat):  0.01058
##
## 95% Confidence Interval for p1-p2:
##
##           lower.bound         upper.bound
##           0.019258            0.060742
```

```
##
## Test of Significance:
##
##  H_0:  p1-p2 = 0
##  H_a:  p1-p2 != 0
##
##  Test Statistic:     z = 3.78
##  P-value:        P = 0.0001571
```

There was no warning, so the sample sizes were big enough. Also, we did a randomized experiment, so we are safe to proceed.

The test statistic is $z = 3.78$, indicating that the difference in sample proportions (8% in the control group vs. 4% in the vaccine group) was about 3.78 standard errors bigger than the Null expected it to be.

Step Three: P-value

The P-value was quite small, about 0.00016.

> **Interpretation of P-Value**: *If the vaccine is doing no good, then there is only about a 16 in 100,000 chance of getting a test statistic at least as far from 0 as the one we got in our study.*

Step Four: Decision

Since $P = 0.00016 < 0.05$, we reject the Null.

Step Five: Conclusion

This experiment provided strong evidence that the vaccine helps to prevent flu.

By the way: if the vaccine was effective, why was it not perfect? 40 out of 1000 people using the vaccine got the flu anyway! There are two reasons why flu vaccines are not completely effective:

1. The flu vaccine is based on a "dead virus", so it takes a while—usually about two weeks—for the body to recognize its existence and to develop the appropriate antibodies. During that period a person who is exposed to the flu could easily "catch" it.
2. The flu vaccine is designed each year by scientists at the Center for Disease Control, who make the best prediction they can as to which strains of flu are most likely to be prevalent in the coming winter. The vaccine is based on these strains, and is not guaranteed to work against all possible strains to which a person might be exposed.

10.7 Further Considerations About Significance Tests

As we have covered each of the Basic Five in this Chapter and in Chapter Three, we have run across ideas that apply to all significance tests. Here is a quick rundown of what we have learned so far:

- The Five-Step logic applies to all significance tests.
- If your test concerns one or more parameters, then you need to define your parameters and then use the symbols for those parameters to state your hypotheses.
- The hypotheses always talk about what's going on in the population(s), not the samples. (If we did an experiment, then the hypotheses talk about what would happen if all the subjects were given each of the treatments.)
- The Null is the hypothesis that says that there is no pattern in the population.
- The Alternative always involves an inequality.

- Tests involving one of the Basic Five can be either one-side or two-sided. This affects the computation of P-values.
- It's always useful to stop and think about the test statistic. With the exception of `binomtestGC()` the test statistic is "z-score style", and it tells you how many SEs the estimator is above or below what the Null expects. When it is far from 0, that's bad for the Null!
- P-values can always be interpreted according to following format: *If [the Null is True], then there is about a [P-value] chance of getting a test statistic at least as extreme as the one we actually got in this study.*
- We can run tests straight from a data frame using formula-data input, or we can enter summary data.
- Tests only check the Null Hypothesis in mathematical form. They don't check whether the study was well-designed (see the ESP example).
- There is a correspondence between tests and confidence intervals that basically says: when the interval does not contain the Null's belief about the parameter, then the test will reject the Null, and when the interval contains the Null's belief the test will not reject the Null.

There are some additional concepts that apply to all significance tests, and in this section we will consider some of them.

10.7.1 Types of Error

There is no guarantee that Step Four (the Decision) in a test of significance will be correct. That's because the sample is based on chance. No matter how much the sample estimator for the population parameters differs from what the Null expects it to be, the difference COULD POSSIBLY be due to our having drawn a really rare, unlikely sort of sample. So even when we reject the Null based on a low P-value, the Null could be correct anyway. The error that we make in such a situation is called a *Type-I Error.*

It goes the other way around, too: we might fail to reject the Null when it is in fact false. Such an error is called a *Type-II Error.*

The fact that a test can deliver an error is not a sign that the testing procedure is bad. On the contrary, just as a well-designed formula for a 95%-confidence interval SHOULD fail to contain the population parameter 5% of the time in repeated sampling, so a well-designed test of significance with a 5% cut-off standard SHOULD commit a Type-I error 5% of the time in repeated sampling, if the Null is in fact correct.

The only way never to commit a Type-I error is to NEVER reject the Null. But if that's your plan, then you might as well not collect any data at all: it won't affect your decision about the Null Hypothesis!

When the Null is false, you are correct when you reject the Null. The probability of rejecting the Null when it is, in fact, false is called the *power* of the test. The power depends on "how false"" the Null actually is. The more the Null's belief about the parameter differs from what the parameter actually is, the higher the power should be.

Sample size makes a difference, too. If the Null is, in fact, false then you are more likely to be able to detect that fact using a large sample than you are using a small one. Tests that are based on large samples are therefore more powerful than tests based on small samples.

The app below illustrates the idea we have broached concerning Type-I and Type-II errors. The app draws random samples from a population that is normally distributed, computes a confidence interval based on the sample, and performs a two-sided test for the mean of the population, too.

In all scenarios, the true mean of the population is $\mu = 170$.

When you choose to draw one sample per mouse click, you will see a histogram of of the sample in blue. You will also see a confidence interval in yellow, with the sample mean as a blue dot in the middle of that interval. In the console, you will get the results of a two-sided test for where the null says that μ is 170 (that is, $\mu_0 = 170$). You can vary the true population mean μ with a slider.

```
require(manipulate)
Type12Errors()
```

To start out, keep the true population mean μ set at 170, so that the null is actually true. This means that whenever the tests rejects the Null, a Type-I error has been made. Click through a number of samples one at a time. Then try many samples (a hundred, or a thousand). What proportion of the time do you get a Type-I error?

Next, move the μ slider a bit away from 170. Now the Null is false, so whenever you reject it you have not made an error. Try lots of samples: what happens to the proportion of rejections, in the long run, after many samples?

Next. keeping μ where it is, move up the sample size n, and run one sample at a time. What happens to the confidence intervals: are they wider than before, or narrower? What proportion of the time do you correctly reject the Null? Is it higher or lower than it was when the sample size was smaller?

Some Basic Ideas Learned From the the App:

- The test is designed so that if the Null is true and your cut-off value (aka *level of significance*) is α, then the test will incorrectly reject the Null (i.e. commit a Type-I Error) 100α percent of the time.
- The further the Null's belief is from the actual value of the population parameter, the more likely it is that the test will reject the Null (i.e., the more powerful the test is).
- The bigger the sample size, the more powerful the test will be when the Null is false. When the sample size is very large, even a Null that is only a "little bit" false is quite likely to be rejected.

The last point is especially important. When the sample size is very large, the test is very likely to detect the difference between a false Null belief and the true parameter value, even when the two are very close together. A large-sample test can thus provide strong evidence for what amounts to a very weak pattern in the population. Hence you should always keep in mind:

> *Strong evidence for a relationship is not the same thing as evidence for a strong relationship.*

10.7.2 The Dangers of Limited Reporting

Suppose that you have twenty friends, and you are interested in knowing whether any of them possess powers of telekinesis (TK for short). Telekinesis (if such a thing exists at all) is the ability to move objects by mental intention alone, without touching them.

You perform the following test with each of your friends: you flip a fair coin 100 times, after instructing your friend to attempt to make the coin land Heads simply by concentrating on it.

You get the following results:

```
CoinData
```

```
##    heads
## 1     57
## 2     54
## 3     57
## 4     61
## 5     45
## 6     52
## 7     60
## 8     52
## 9     48
## 10    57
## 11    58
## 12    58
## 13    51
## 14    46
## 15    49
## 16    53
## 17    51
## 18    51
## 19    48
## 20    46
```

Hmm, one of your friends (Friend Number Four) got 61 Heads in 100 flips. That seems like a lot. If your friend has no TK—so that the coin has the usual 50% chance of landings heads—then the chance of 61 or more heads is pretty small:

```
pbinomGC(60,region="above",size=100,prob=0.5)
```

```
## [1] 0.0176001
```

If you run a `binomtestGC()` on your friend's results, you get the same information:

Step One: Define Parameter and State the Hypotheses

Let

$p =$ chance that coin lands Heads when Friend Number Four concentrates on it.

The hypotheses are:

H_0: $p = 0.50$ (Friend #4 has no TK powers)

H_a: $p > 0.50$ (Friend #4 has some TK powers)

Step Two: Safety Check and Test Statistic

We flipped the coin randomly so we are safe. The test statistic is 61, the number of Heads our friend got.

Step Three: P-value

```
binomtestGC(61,n=100,p=0.5,alternative="greater")
```

```
## Exact Binomial Procedures for a Single Proportion p:
##  Results based on Summary Data
##
## Descriptive Results:  61 successes in 100 trials
##
## Inferential Results:
##
## Estimate of p:    0.61
## SE(p.hat):    0.0488
```

```
##
## 95% Confidence Interval for p:
##
##           lower.bound         upper.bound
##           0.523094            1.000000
##
## Test of Significance:
##
##  H_0:  p = 0.5
##  H_a:  p > 0.5
##
##  P-value:        P = 0.0176
```

Sure enough, the P-value is 0.0176.

Step Four: Decision

> Since P < 0.05, we reject the Null.

Step Five: Conclusion

> Our data (the 100 coin flips in the presence of Friend Number Four) provide strong evidence that when Friend Four concentrates on this coin he/she has more than a 50% chance of making it land Heads.

There is nothing wrong with any step in this test, and yet it seems quite wrong to conclude that Friend Number Four actually has some TK powers.

Why? Because your friend was only one of twenty people tested. When you test a lot of people, then even if none of them have TK you would expect a few of them to get an unusually large number of heads, just by chance, right?

Just as with any test of significance, `binomtestGC()` considers ONLY the data that you present to it. When it is allowed to consider ONLY the results from Friend Number Four, R leads you to a perfectly reasonable conclusion: those 100 flips, *when considered in isolation from other data*, provide strong evidence for some TK powers in Friend Four.

But it's not right to consider the 100 flips in isolation. Instead, should we not at least take into account that our friend's results were the maximum out of a study involving twenty friends?

Perhaps a more relevant P-value for this situation—one that takes more of the data into account—is the probability of at least one person getting 61 or more heads when 20 people, none of whom have TK, participate in a study like this.

If you run the following R code, you will teach R a function that conducts the entire study and returns the maximum number of heads achieved by any friend

```r
HeadMax <- function(friends=20,flips=100,coin=0.5) {
  return(max(rbinom(friends,size=flips,prob=coin)))}
```

Now run the function:

```r
HeadMax(friends=20,flips=100)
```

Re-run the function a number of times. Do you notice that it is not such a rare thing at all for at least one person to get at least 61 heads?

In fact, if you were to run the function many times and keep track of the results, you would find that—even if there is no such thing as TK—there is about a 30% chance that at least one friend will get 61 or more heads. When you think of it this way, you see that 61 heads is not strong evidence for TK!

The moral of the story is this:

> *Tests of significance consider only the data presented to them.*

If you use a test that is based on only a limited portion of the data, the test does not deliver meaningful results. Instead:

- Report all relevant data you have collected.
- Look for ways to incorporate all of the relevant data when you are testing for statistical significance.

10.7.3 The Dangers of Data-Snooping

The example in the previous section may seem somewhat artificial. After all, you are an honest-to-good scientist—or at least you are planning to take some honest-to-good science classes—and you know that real, honest scientists do not knowingly limit their reporting to just those portions of the data that point in an interesting direction.

But it's amazing how much we can limit our reporting without even knowing it.

For example, consider the `attitudes` data. If you need a reminder, run:

```
data(attitudes)
View(attitudes)
help(attitudes)
```

As you will recall, there was a limited number of primary research questions that guided the design of the `attitudes` study:

- Does suggested race of defendant affect the sentence recommended?
- Does suggested race of victim affect the sentence recommended?
- Does the way one lost one's money affect the decision about whether to attend the concert anyway?

But the researchers could not resist the urge to collect some other information beyond what was needed to narrowly address the primary research questions. For example, the researcher got information on the sex, class rank and intended major of each participant.

There is nothing wrong at all with the urge to collect extra data. In fact it's a wonderful human trait, is it not?

And of course there is nothing at all wrong with the urge to explore the data thus collected. It's a rich data set (there are a lot of variables), so there are many research questions one could raise.

For example, we might explore the

> **Research Question**: *Who is harder on crime: a GC female or a GC male?*

When you think about the kinds of variables involved in this question (explanatory = **sex** is a factor, response = **sentence** is numerical) you see that `favstats()` will provide useful descriptive statistics:

```
favstats(sentence~sex,data=attitudes)
```

```
##   .group min   Q1 median Q3 max     mean       sd   n missing
## 1 female   3 15.0     25 40  50 27.30793 15.46790 164       1
## 2   male   3 12.5     25 40  50 25.76699 15.31832 103       0
```

Hmm, nothing much seems to be going on here.

Let's try another

> **Research Question**: *Who is more likely to decide to attend the rock concert anyway: a GC female or a GC male*

Again, we work our way from research question to variable to type of each variable to exploration technique and we hit upon the following R code:

```
SexRock <- xtabs(~sex+conc.decision,data=attitudes)
SexRock
```

```
##         conc.decision
## sex      buy not.buy
##   female 112      53
##   male    64      39
```

```
rowPerc(SexRock)
```

```
##         conc.decision
## sex          buy not.buy  Total
##   female  67.88   32.12 100.00
##   male    62.14   37.86 100.00
```

Hmm, apparently there not much going on in the way of a relationship between sex and decision to attend a rock concert.

Still another

> **Research Question**: *Who is harder on crime: A GC student who would decide to atttend the rock concert or a GC student who decide not to attend?*

Again we check:

```
favstats(sentence~conc.decision,data=attitudes)
```

```
##    .group min Q1 median Q3 max     mean       sd   n missing
## 1     buy   3 15     25 45  50 28.15143 15.70951 175       1
## 2 not.buy   3 10     20 30  50 23.97826 14.48453  92       0
```

Interesting: in the survey those who chose to attend (let's call them "Big Spenders") recommended sentences for the defendant that were, on average, five years longer than the sentences recommended by the Thrifty types—the ones who chose not to attend.

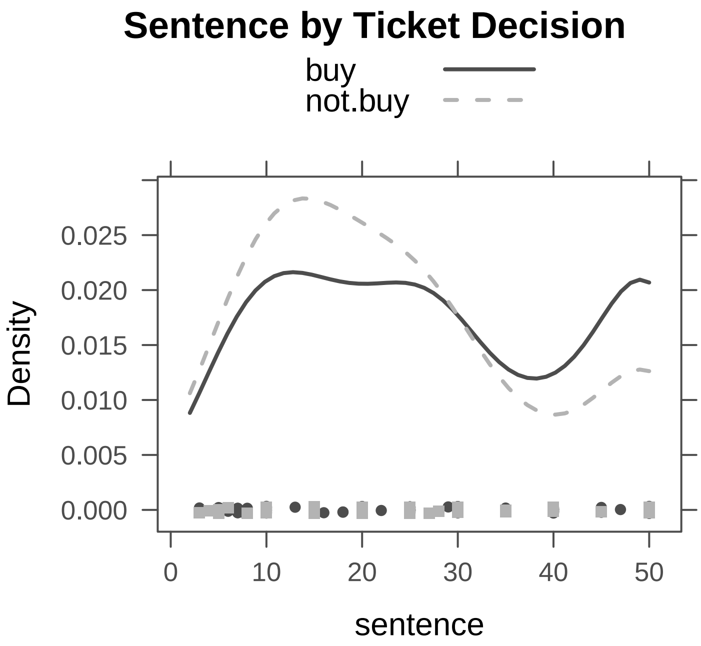


Figure 10.6: Sentence and Ticket Decision

Let's check it out graphically: Figure [Sentences and Ticket Decision] shows the results. The plot thickens! The distributions are fairly similar in shape, but it appears that the Big Spender group had a substantially higher proportion of people who recommended the maximum sentence of 50 years: that pulled up their mean sentence.

Let's test to see if these results are statistically significant. We won't write out all five steps, but instead let's just peek at the test output. (We'll choose a two-sided test, because we had no advance evidence that Big Spenders should be harder on crime that Thrifty people.)

```r
ttestGC(sentence~conc.decision,data=attitudes,
        mu=0,alternative="two.sided")
```

```
##
##
## Inferential Procedures for the Difference of Two Means mu1-mu2:
##  (Welch's Approximation Used for Degrees of Freedom)
##   sentence grouped by conc.decision
##
##
## Descriptive Results:
##
##    group  mean    sd   n
##      buy 28.15 15.71 175
##  not.buy 23.98 14.48  92
##
##
## Inferential Results:
##
## Estimate of mu1-mu2:  4.173
## SE(x1.bar - x2.bar):  1.921
##
```

```
## 95% Confidence Interval for mu1-mu2:
##
##           lower.bound         upper.bound
##           0.384772            7.961563
##
## Test of Significance:
##
##  H_0:  mu1-mu2 = 0
##  H_a:  mu1-mu2 != 0
##
##  Test Statistic:     t = 2.172
##  Degrees of Freedom:   198.6
##  P-value:        P = 0.03102
```

Hmm, the P-value is about 0.03. It appears that we have some strong evidence here that GC Big Spenders are harder on crime than GC Thrifties are.

Do you see what's going on here? We have followed a natural and perfectly commendable human urge to paw through the available data, formulating research questions as we go and investigating them with appropriate descriptive techniques, deciding to run an inferential test only when descriptive techniques turned up an interesting pattern in the data. Although our intentions were quite honest, we are acting very much like the person in the previous section who decided to run `binomtestGC()` on the one friend who "produced" an intriguingly-large number of heads! This time, though, instead of testing on a limited amount of relevant data, we are testing on a *limited number of aspects* of the data. We are looking at the data from many different angles, as it were, and choosing to test on the angles that look "pretty".

Even a totally ordinary piece of furniture (for example) can look "pretty" from a particular point of view, in a particular light, etc.

This practice—pawing through a rich data set, examining it from different points of view and performing tests on the basis of interesting patterns that we happen to notice when looking at the data from some of those vantage points—is called *data-snooping.*

Data Snooping *Data Snooping* is the practice of looking through a data set for patterns that interest us and performing inferential procedures on patterns that we notice, rather than limiting our tests to specific Research Questions that we had in mind prior to collecting the data.

Data snooping is not an inherently evil practice: on the contrary, it is an expression of two excellent human traits: curiosity, and the ability to detect many types of patterns. The problem is that the more curious we are, and the better we are at noticing patterns, the more likely it is that our data-snooping will lead us to report results as statistically significant when in fact they are only due to random quirks in the data.

The problem is especially acute because—unlike the "limited reporting" example of the previous section, where we can correct ourselves by taking all of the data into account—we don't really know what *kinds* of patterns we as humans are disposed to find intriguing: we don't know what constitutes the "set of all aspects" that have to be taken into account when one attempts to adjust the P-values for the tests on those particular aspects that, for one reason or another, got us excited.

This is really interesting: we have identified a problem in how to assess rationally what our observations entail, but this problem does not arise from our not knowing something about how the data was collected, nor does it arise from our not knowing something about the population. It arises from our not knowing something about *ourselves.*

There is no fool-proof solution to the problem of data-snooping. As a general rule of thumb, though, we should keep clearly in mind the distinction between:

- those research questions that we had in mind prior to collecting the data (*Primary Questions*) versus

- those research questions that occur to us as we examine the collected data in all of its richness (*Secondary Questions*).

Results of tests on Primary Questions should be taken more seriously than results of tests on Secondary Questions. The Results for Secondary Questions may be worth reporting to others (especially if they provide exceedingly strong evidence for a pattern), but we should be clear when we report them that the results were based on patterns that we noticed AFTER having examined the data. Perhaps in the future other researchers could treat them as Primary Questions, should they conduct a study similar to ours. If the same pattern persists in study after study, then we can be more and more certain that it is not due to chance variation in the data collection process, but reflects instead a pattern in the population itself.

10.8 Thoughts About R

10.8.1 New Code to Learn?

Functions for tests are the same as for confidence intervals! However, you have to consider some additional arguments. You need to pay attention to:

- the argument that sets the Null value for the parameter:
    - `mu` in tests involving means
    - `p` in tests involving proportions
    - `prob` in `binomtestGC()`
- the `alternative` argument that specifies the "side" of the test
    - "greater"
    - "less"
    - "two.sided"

The default value of `alternative` is "two.sided", so if you want a two-sided test you need not mention `alternative` in your call to the function. For example, the following two functions calls will deliver exactly the same results:

```
ttestGC(~fastest,data=m111survey,
       mu=100)
ttestGC(~fastest,data=m111survey,
       mu=100,alternative="two.sided")
```

10.8.2 Old Descriptive Friends

When you perform safety checks in the tests involving means and you have access to the original data rather than just a summary of it, then you may have to make graphs of your samples in order to check for skewness and outliers. For this you will revisit some old friends from Chapter Two. You should be fine with:

- `histogram()`
- `bwplot()`

10.8.3 Adding a Variable to a Data Frame

On rare occasions you may need to add a variable to a data frame prior to running a test involving that variable. This occurred to us once, in the Research Question about whether women or men are more affected by the label on a jar of peanut butter. In order to add a variable **Var** to a data frame `MyData`, simply write:

```
MyData$Var <- Var
```

The length of **Var** must be the same as the number of rows in the data frame.

Chapter 11

Goodness of Fit

At the beginning of Chapter Two our very first foray into Descriptive Statistics concerned exploring the distribution of a single factor variable. In this Chapter we will introduce a way to address inferential aspects of Research Questions about one factor variable.

11.1 The Gambler's Die

Imagine that you are the Sheriff of a small town in the Wild, Wild West. A professional gambler comes to town, and sets up shop in the local saloon. He claims to play with a fair die, but he wins so much that the locals have begun to mutter that his die is loaded. Since loaded dice are illegal in your town, it is up to you to investigate. You impound the die.

Ideally, you would take the die apart to see if it is weighted in one direction or another, but the gambler claims that it is his "lucky" die, and he begs you to do it no harm. You decide to roll it a few times, instead.

There are other pressing law enforcement matters in the town—gunfights, and the like—so you have time only for sixty rolls. The results of the rolls are in the Table [60 Rolls of a Die].

Spots	One	Two	Three	Four	Five	Six
Freq	8	18	11	7	9	7

Table 11.1: 60 Rolls of a Die

At once your eye is drawn to the rather large number of two-spots. After all, if the six-sided die is really fair then the chance of a two-spot on any given roll is only 1 in 6, so you would expect only about ten twos in 60 rolls—give or take a few, of course, for chance variation.

The locals notice the large number of twos as well. They pull out their computers, fire up R, and perform a quick One-Proportion test with `binomtestGC()`:

```
binomtestGC(x=18,n=60,p=1/6)
```

```
## Exact Binomial Procedures for a Single Proportion p:
##  Results based on Summary Data
##
## Descriptive Results:  18 successes in 60 trials
```

```
##
## Inferential Results:
##
## Estimate of p:    0.3
## SE(p.hat):    0.0592
##
## 95% Confidence Interval for p:
##
##           lower.bound         upper.bound
##           0.188451            0.432083
##
## Test of Significance:
##
##  H_0:  p = 0.1666667
##  H_a:  p != 0.1666667
##
##  P-value:        P = 0.0089
```

The results show that if the die is really fair, then the chance of 18 or more twos in sixty rolls is only about 0.009, which is less than 1%. The locals prepare to ride the gambler out of town on a rail.

You, on the other hand, know about Data Snooping, the dangerous practice of performing a test on the basis of a pattern you notice in your data. You know that it's not really fair to concoct a test on the proportion of twos, simply because the twos grab your attention: after all, the tally of the 60 rolls has six counts—one for each side of the die—and it stands to reason that one or more of the counts might depart quite a bit from its expected value of 10, just by chance variation.

What you need is a test that takes all of the data, not just the count of twos, into account in an even-handed way. This need puts you in mind of the χ^2-statistic from Chapter Three:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

You table of counts has six cells, and if the die is fair then the expected count for each cell is 10—one-sixth of the 60 rolls. You compute the χ^2-statistic for the tally of the sixty rolls:

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(18-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(7-10)^2}{10} = 8.8.$$

The result is 8.8. But you wonder: is this value too big to be explained as just due to chance variation in the sixty rolls of a fair die?

Fortunately you have access to a simulator that will perform the following task:

- it will roll a fair die 60 times;
- it will tally the results, getting a table of observed counts;
- it will compute the χ^2-statistic for the table thus obtained;
- it will repeat the preceding three steps as many times as you like, keeping track of the results and checking to see how many of them exceed the 8.8 value that you got in your actual study of the gambler's die.

In order to use the simulator, you need to put together the table of observed counts:

```
throws <- c(one=8,two=18,three=11,
            four=7,five=9,six=7)
```

You also need to make a list of the probabilities for getting each of the six possible sides. These probabilities are based on the temporary assumption that the die really is fair, so they are all equal to 1/6:

```
NullProbs <- rep(1/6,6)
```

Now you are ready to put this information into the simulator, the function `SlowGoodness()`:

```
SlowGoodness(throws,NullProbs)
```

Try the simulator for a good number of re-samples. It does not seem terribly unlikely that a fair die would produce a 60-roll tally with a χ^2-statistic of 8.8 or more. Certainly the chance appears to be much larger than the 1% figure you were getting from `binomtestGC()`. Maybe the gambler is honest, after all, and the large number of twos was just a fluke.

11.2 chisqtestGC() for Goodness of Fit

The above procedure can be formulated in terms of a test of significance, which is called a *goodness-of-fit* test because it tests whether the observed values of some factor variable can reasonably be said to "fit" a given distribution.

11.2.1 The Gambler's Die as a Test of Significance

In the foregoing study the factor variable at issue was **spots**, the number of spots on the upward-facing die, after it is thrown. The given distribution is the set of probabilities for each side that amounts to the gambler's claim that the die is fair. These "Null" probabilities appear in Table [Fair Die Probabilities].

Spots	One	Two	Three	Four	Five	Six
Prob	1/6	1/6	1/6	1/6	1/6	1/6

Table 11.2: Fair Die Probabilities

Now we are ready for the test.

Step One: Statement of Hypotheses

H_0: The die is fair (all sides have probability 1/6).

H_a: The die is weighted (at least one side has probability $\neq 1/6$)

Step Two: Safety Check and Test Statistic

For safety, all that is needed is to know that the 60 tosses of the die were random.

As for the test statistic and other information needed in future steps, we run `chisqtestGC()`. We must provide

- the table of observed counts, and
- an argument p that gives the probabilities that go with the Null Hypothesis.

We also plan to perform simulations in order to compute the P-value, so:

- We must set the argument simulate.p.value to TRUE;
- We should choose the number of re-samples, indicated by the argument B. We'll set B to 2500.
- For reproducibility of results, we should set a seed in advance with the set.seed() function. Quite arbitrarily, we'll set the seed to 12345.

The function call is:

```
set.seed(12345)
throws <- c(one=8,two=18,three=11,
            four=7,five=9,six=7)
NullProbs <- rep(1/6,6)
chisqtestGC(throws,p=NullProbs,
             simulate.p.value=T,B=2500)
```

```
## Pearson's chi-squared test with simulated p-value
##   (based on 2500 resamples)
##
##       Observed counts Expected by Null Contr to chisq stat
## one                 8               10                0.4
## two                18               10                6.4
## three              11               10                0.1
## four                7               10                0.9
## five                9               10                0.1
## six                 7               10                0.9
##
##
## Chi-Square Statistic = 8.8
## Degrees of Freedom of the table = 5
## P-Value = 0.1152
```

Note that the output gives the χ^2-statistic as 8.8.

Step Three: P-value

The simulated approximation to the P-value is 0.1148.

> **Interpretation of P-Value**: *If the die is fair, then there is about an 11.48% of getting a χ^2-statistic at least as large as the one we got in our study of the gambler's die.*

Step Four: Decision

> Since $P = 0.1148 > 0.05$, we do not reject the Null.

Step Five Conclusion

> This study did not provide strong evidence that the gambler's die was loaded.

11.2.2 Facts About χ^2 Goodness of Fit

Mathematicians have studied the χ^2-statistic under conditions where sample size is “big enough”, and they have discovered the following:

- When the Null Hypothesis is true, its EV is df, the “degrees of freedom.”
- In the goodness-of-fit situation—i.e., when one factor variable is under investigation—the df figure is:

$$df = \text{number of cells} - 1.$$

- The standard deviation of the χ^2-statistic is:

$$SD(\chi^2) = \sqrt{2 \times df}.$$

Thus, in the Gambler’s Die study,

$$df = 6 - 1 = 5,$$

so if the die is fair then we would expect the χ^2-statistic to be around 5, give or take

$$\sqrt{2 \times 5} = \sqrt{10} \approx 3.16$$

or so. From this point of view, the 8.8 value that we got does not seem very unlikely, if the Null is right: it is not much more than one SD above the EV.

As for what sample size counts as “big enough”" the results we stated above are quite accurate if the Null’s expected cell counts are all at least five. Mathematicians have even discovered a family of random variables, known as the χ^2 family, such that at large sample sizes the χ^2-statistic behaves like one of the members of this family, the member with the degrees of freedom given by the “cells minus one” formula. Figure [Chi-Square df=5] shows a density curve for the χ^2 curve with 5 degrees of freedom, with the area under the curve after 8.8 shaded in. Note that the area is quite close to the P-value that we obtained through simulation.

Since at large sample sizes the χ^2 density curve will deliver approximately the right P-value, we usually don’t ask for simulation. For the Gambler’s Die study, the function call would be:

```
chisqtestGC(throws,p=NullProbs)
```

If you would like to produce a graph of the P-value, you can do so by setting the `graph` argument to `TRUE`.

```
chisqtestGC(throws,p=NullProbs,graph=T)
```

If you have chose to simulate, you can get a graph of the re-sampled χ^2-statistics:

```
chisqtestGC(throws,p=NullProbs,
            simulate.p.value=T,B=2500,
            graph=TRUE)
```

If any of the expected cell counts are below five, `chisqtestGC()` will issue a warning, in which case you should re-run the test using simulation.

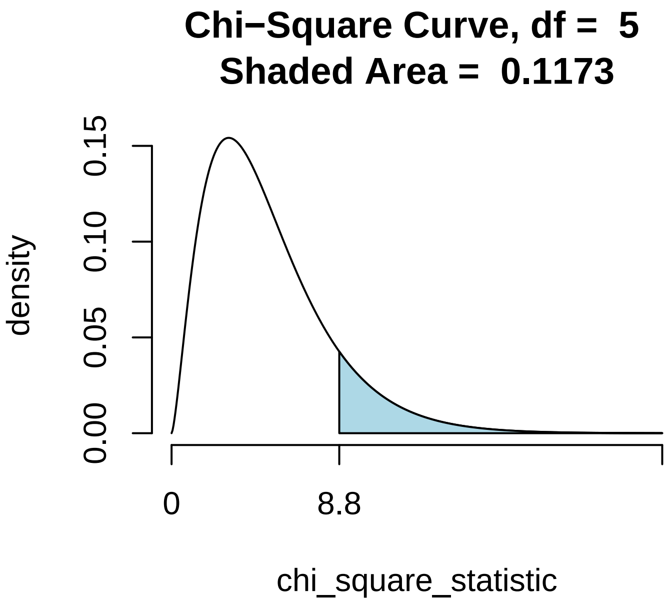


Figure 11.1: Chi-Square df=5

11.3 Further Example: Seating Preference

Recalling the seating-preference variable **seat** in the `m111survey` data frame, one might ask the following

> **Research Qeestion**: *In the Georgetown College population, is there equal perference for each type of seat: front, middle and back?*

In this Research Question, we are interested in the distribution of a single factor variable: **seat**. We can use the χ^2-goodness of fit test to see whether the data in the `m111survey` provide strong evidence that the GC population is not indifferent to seat-location.

Step One: Statement of Hypotheses

If the GC population is indifferent, then the proportion of the populations that prefers each type of seat will be 1/3, for each of the three possible values of **seat**.

The Null Hypotheses may then be stated as:

> H_0: Each of the three types of seat is preferred by one-third of the GC population.

The Alternative disagrees:

> H_a: At least one of the three proportions claimed by the Null is wrong.

Step Two: Safety Check and Test Statistic

For safety, we must assume that the students in `m111survey` are like a simple random sample from the GC population, at least as far as seating preference is concerned.

For the test statistic and other information needed in subsequent steps, we can use formula-data input to call `chisqtestGC()`:

```
chisqtestGC(~seat,data=m111survey,
            p=c(1/3,1/3,1/3))
```

```
## Chi-squared test for given probabilities
##
##          Observed counts Expected by Null Contr to chisq stat
## 1_front               27            23.67                0.47
## 2_middle              32            23.67                2.93
## 3_back                12            23.67                5.75
##
##
## Chi-Square Statistic = 9.1549
## Degrees of Freedom of the table = 2
## P-Value = 0.0103
```

We see that the P-value is about 0.01, which is less than our "cut-off" of 0.05, so we reject the Null and conclude that this data provided strong evidence that GC students are not equally likely to prefer any of the three seating locations. (The middle appears to be most desired.)

11.4 Further Example: Nexus Attendance

A Nexus event held at Georgetown College in April of the Spring semester was attended by 200 students. The class breakdown of the students is given in Table [Nexus Attendance].

Class	Fresh	Soph	Junior	Senior
Freq	62	27	33	80

Table 11.3: Nexus Attendance

Suppose that is is known that the distribution of class rank at Georgetown College is as follows:

- 30% freshmen,
- 25% sophomores,
- 25% juniors,
- 20% seniors.

We are interested in the following

> **Research Question**: *Is the attendance at April Nexus events like a random sample of GC students?*

Note that in this Research Question the distribution of the variable **Class Rank** in the GC population in NOT at issue: everyone knows that it is given by the percentages above. What is at issue is whether the data on this Nexus event show that students are not attending randomly, in the sense that one's likelihood of attending a Nexus event in April might be associated with one's class rank.

In this study the hypotheses should be stated as:

H_0: Attendance at the event was random.

H_a: Likelihood of attendance depended on class rank.

Information needed for the remaining four steps comes from the following R-code:

```
Nexus <- c(fresh=62,soph=27,junior=33,senior=80)
NullClass <- c(0.30,0.25,0.25,0.20)
chisqtestGC(Nexus,p=NullClass)
```

```
## Chi-squared test for given probabilities
##
##        Observed counts Expected by Null Contr to chisq stat
## fresh               62             60.6                0.03
## soph                27             50.5               10.94
## junior              33             50.5                6.06
## senior              80             40.4               38.82
##
##
## Chi-Square Statistic = 55.8482
## Degrees of Freedom of the table = 3
## P-Value = 0
```

The χ^2 statistic was about 55.8. If the Null is true, then it should have been about 3 (the degrees of freedom), give or take

$$\sqrt{2 \times 3} = \sqrt{6} \approx 2.45$$

or so. A value of 55.8 looks very bad for the Null!

In fact, if the Null is right then the chance of getting a χ^2-statistic of 55.8 or more is about 4.5×10^{-12}, which is a very tiny number indeed. We reject the Null. This event provided strong evidence that, in April, Nexus attendance depends on class rank.

In fact, from the test output we see that the Null expected about 40.4 seniors to attend, whereas 80 in fact did. Apparently GC seniors are scrambling for last-minute Nexus credits!

11.5 Further Example: Too Good to be True?

Imagine that a statistics professor gives a student the following Homework assignment:

> **Assignment**: *Familiarize yourself with chance variation by rolling a fair die 6000 times. Turn in a tally of your results.*

Most students would find the assignment rather onerous!

Now suppose that the next day the student hands in the results shown in Table [6000 Rolls of a Fair Die?].

Spots	One	Two	Three	Four	Five	Six
Freq	1003	998	999	1002	1001	997

Table 11.4: 6000 Rolls of a Fair Die?

The professor observes that each cell count is quite close to what would be expected in 6000 rolls of a fair die. But are the counts perhaps TOO close to what would be expected? In other words:

> **Research Question**: *Is the observed fit "too good to be true"?*

Consider running a χ^2-goodness of fit test on the data:

```
Alleged Rolls <- c(one=1003,two=998,three=999,
            four=1002,five=1001,six=997)
FairProbs <- rep(1/6,6)
chisqtestGC(AllegedRolls,p=FairProbs)
```

```
## Chi-squared test for given probabilities
##
##       Observed counts Expected by Null Contr to chisq stat
## one              1003             1000                0.01
## two               998             1000                0.00
## three             999             1000                0.00
## four             1002             1000                0.00
## five             1001             1000                0.00
## six               997             1000                0.01
##
##
## Chi-Square Statistic = 0.028
## Degrees of Freedom of the table = 5
## P-Value = 1
```

As one might have guessed from the close agreement between observed and expected counts, the χ^2 statistic is quite small, so the P-value—the chance of getting at least this value in 6000 rolls of a fair die—is quite large: it is nearly 100%, in fact.

But think of it the other way around: if the table turned in by the student really was the result of tossing a fair die 6000 times, then what is the chance of getting such a small χ^2-statistic? It would have to be 1 minus the P-value given in the test output:

```
1-0.9999931
```

```
## [1] 6.9e-06
```

This is a very small chance indeed! It would be reasonable to infer that the student made up the tally table, but did not allow for a realistic amount of chance variation away from the expected cell counts!

11.6 Thoughts About R

11.6.1 Names for Elements in a List

When you are working with summary data, you have to provide `chisqtestGC()` with a tally of the observed counts. The output for the test is most readable if you also provide names for each of the possible values of the factor variable under study. You can do so as you create the lists of observed counts. For example, if your variable has three possible values called "a", "b" and "c", with counts 13, 12 and 27 respectively, then you can create a named tally as follows:

```
ObsCounts <- c(a=13,b=12,c=27)
```

11.6.2 Quick Null Probabilities

For goodness of fit tests, `chisqtestGC()` requires that the `p` argument be set to the list of proportions claimed by the Null for the factor variable under study. If these values are all the same, then you can write them quickly using the `rep()` function.

`rep()` repeats a given value a given number of times. For example to get 7 threes, just type:

```
rep(3,7)
```

```
## [1] 3 3 3 3 3 3 3
```

If there are eight null probabilities and they are all the same, then each would be 1/8, so you could set them as follows:

```
MyNulls <- rep(1/8,8)
MyNulls
```

```
## [1] 0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125
```

Chapter 12

Geek Notes

12.1 Chapter 2

12.1.1 More on Structure

Everything in R is an object. Every object has a structure. In FDN 111 , we learn that the structure of an object consists of its parts and the way that the parts relate together. R can show us the structure of an object using the **str** function. We have already seen this for data frames:

```
str(m111survey)
```

```
## 'data.frame':    71 obs. of  12 variables:
##  $ height         : num  76 74 64 62 72 70.8 70 79 59 67 ...
##  $ ideal_ht       : num  78 76 NA 65 72 NA 72 76 61 67 ...
##  $ sleep          : num  9.5 7 9 7 8 10 4 6 7 7 ...
##  $ fastest        : int  119 110 85 100 95 100 85 160 90 90 ...
##  $ weight_feel    : Factor w/ 3 levels "1_underweight",..: 1 2 2 1 1 3 2 2 2 3 ...
##  $ love_first     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ extra_life     : Factor w/ 2 levels "no","yes": 2 2 1 1 2 1 2 2 2 1 ...
##  $ seat           : Factor w/ 3 levels "1_front","2_middle",..: 1 2 2 1 3 1 1 3 3 2 ...
##  $ GPA            : num  3.56 2.5 3.8 3.5 3.2 3.1 3.68 2.7 2.8 NA ...
##  $ enough_Sleep   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 2 1 2 ...
##  $ sex            : Factor w/ 2 levels "female","male": 2 2 1 1 2 2 2 2 1 1 ...
##  $ diff.ideal.act.: num  2 2 NA 3 0 NA 2 -3 2 0 ...
```

The parts of a data frame are the variables. The way they relate together to make an actual data frame is that all have the same length (71 in this case). This allows R to combine the variable in columns, and to interpret the rows as individuals.

You can think of a data frame as being like a book. The “chapters” of the book are the variables.

If a data frame is like a book, then a package, such as tigerstats, is a like collection of books. The authors of R must take this analogy pretty seriously, because one way to load is package is as follows:

```
library(tigerstats)
```

The `library()` function takes all of the books in `tigerstats` out of storage and puts them on the shelves of R’s library.

Just like you, R is a reader, so R reads for structure, too. Look at the following code (and see Figure [A Simple Histogram] for the results):

```
histogram(~fastest, data=m111survey)
```

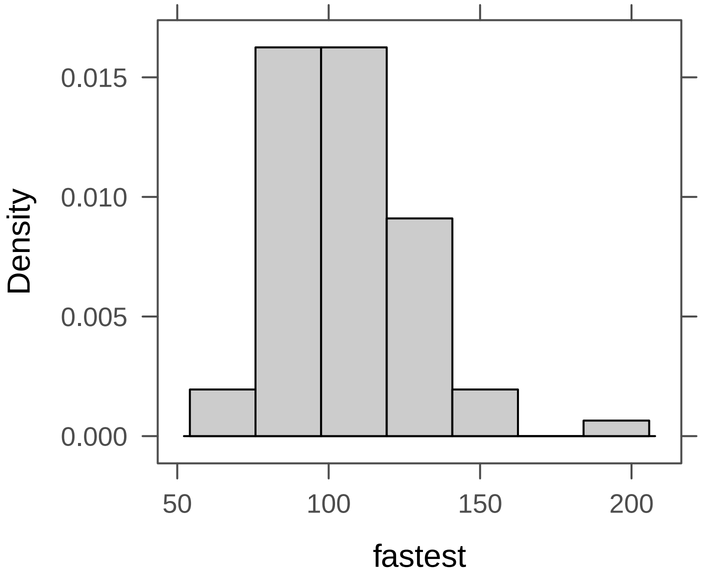


Figure 12.1: A Simple Histogram.

You can think of the code as saying to R: “Put on your `histogram()` glasses. Then take up the book named `m111survey`, turn to chapter **fastest**, and read that chapter with your `histogram()` glasses.”

When R gets interprets that code, it “reads” **fastest** with histogram glasses. It can do so because of the structure of fastest:

```
str(m111survey$fastest)
```

```
##  int [1:71] 119 110 85 100 95 100 85 160 90 90 ...
```

R sees that is **fastest** is a numerical vector. It can use histogram glasses to read that vector and produce the histogram you see on the screen.

Suppose you were to ask R to make a histogram of **sex**. The result appears in Figure [Bad Histogram].

```
histogram(~sex,data=m111survey)
```

```
## Warning in mean.default(evalF$right[, 1], ...): argument is not numeric or
## logical: returning NA
```

You don’t get a histogram; you get something that looks like a cross between a density histrogram and a barchart. R was programmed to look at the structure of the input variable. If it’s a factor rather than a numerical vector and a histogram was requested, then R looks turns the factor into a numerical variable, as best it can. In this case, “female” was turned into a 1 and “male” was turned into a 2. The rectangle over female extends from 0.5 to 1.5, and the recangle over “male” extends from 1.5 to 2.5 Very kindly, R prints the values “female” and “male”, rather than 1 and 2, so it’s doing the best it can to give you a pair of “glasses” through which you can read the data.

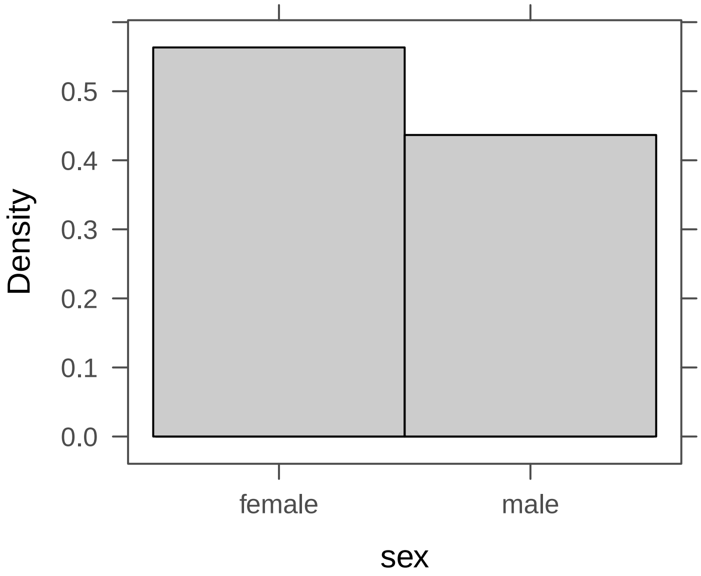


Figure 12.2: Bad Histogram. You should not try to make a histogram from a factor variable.

We said that everything in R is an object, and every object has a structure. Therefore, even graphs have a structure. Try this:

```
FastHist <- histogram(~fastest,data=m111survey,
                     main="Fastest Speed Ever Driven",
                     xlab="speed in mph",
                     type="density")
```

Where's the graph? Well, we didn't ask for it to go the screen; instead we asked for it to be stored as an object named `FastHist`. Let's look at the structure of the object:

```
str(FastHist)
```

Run the chunk above. It's an enormous list (of 45 items). When you look through it, you see that it appears to contains the information need to build a histogram.

The "building" occurs when we '`print()` the object:

```
print(FastHist)
```

The `print()` function uses the information in `FastHist` to produce the histogram you see on in Figure [Now we get the histogram]. (When you ask for a histogram directly, you are actually asking R to print the histogram object created by the `histogram()` function.)

Of course when we read a histogram, we usually read the one we see on the screen, so we think of its structure differently than R does. In general, we think of the structure of a graph as:

- the axes
- the panel (the part that is enclosed by the axes)
- the annotations (title, axis labels, legend, etc.)

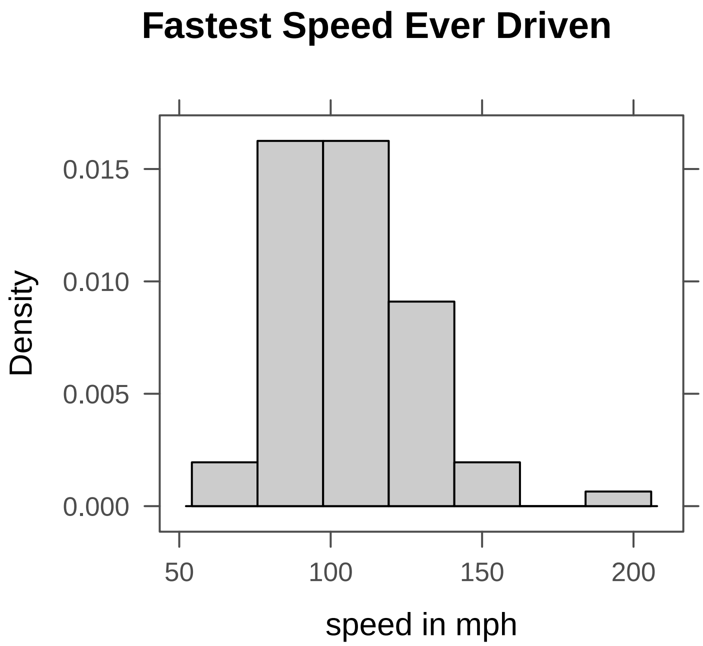


Figure 12.3: Now we get the histogram.

12.1.2 Fancier Histograms

In a density histogram, it can make a lot of sense to let the rectangles have different widths. For example, look at the tornado damage amounts in `tornado`:

```
histogram(~damage,data=tornado,
           main="Average Annual Tornado\nDamage, by State",
           xlab="Damage in Millions of Dollars",
            type="density")
```

The distribution (see Figure [Tornado damge, with default breaks]) is very right-skewed, but most of the states suffered very little damage. Let's get a finer-grained picture of these states by picking our own breaks:

```
data(tornado)
histogram(~damage,data=tornado,
           main="Average Annual Tornado\nDamage, by State",
           xlab="Damage in Millions of Dollars",
            type="density",
            breaks=c(0,2,4,6,10,15,20,25,30,40,50,60,70,80,90,100))
```

Figure [Tornado damage, with customized rectangles] shows the result. You should play around with the sequence of breaks, to find one that "tells the story" of the data well.

12.1.3 Combined Plots

If you would like to make violin plot combined with a box-and-whisker plot, here is how to do it:

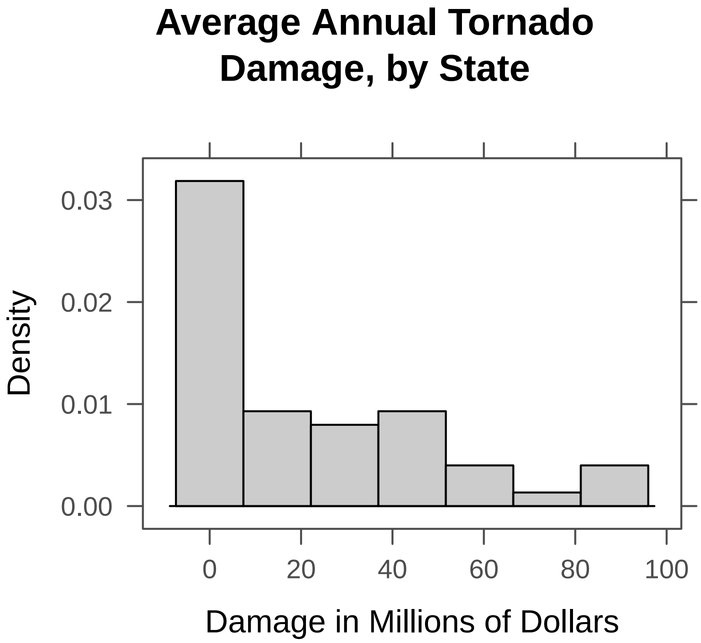


Figure 12.4: Tornado damge, with default breaks. All rectangles have the same width.

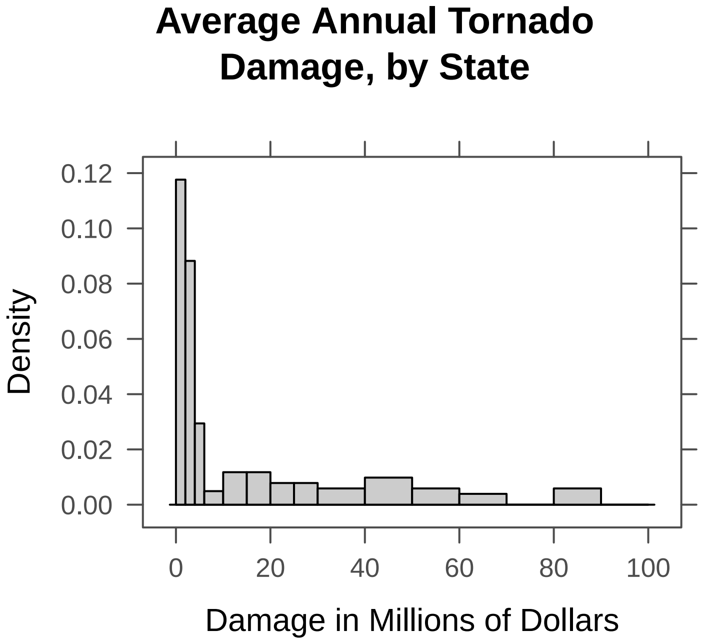


Figure 12.5: Tornado damage, with customized rectangles.

```
bwplot(GPA~seat,data=m111survey,
       main="Grade Point Average,\nby Seating Preference",
       xlab="Seating Preference",
       ylab="GPA",
       panel = function(box.ratio,...) {
                  panel.violin(..., col = "bisque",
                               from=0,to=4)
                  panel.bwplot(..., box.ratio = 0.1)
                })
```

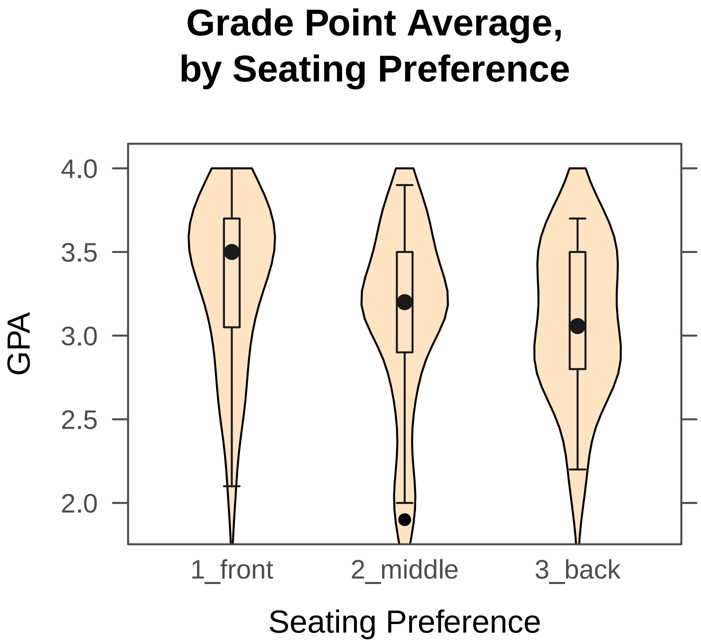


Figure 12.6: Combined Plot. Box-and-Whisker plots combined with violin plots are very cool.

The result is shown in Figure [Combined Plot].

In order to get more than one graph into the "panel" area of a plot, you modify something called the "panel" function. In advanced courses (or own your own) you canlearn more about how R's graphics systems work, but for now just try copying and modifying the code you see in the Course Notes.

12.1.4 Adding Rugs

Adding the argument `panel.rug` to the panel function gives a "rug" of individual data values along the x-axis.

```
bwplot(~damage,data=tornado,
           main="Average Annual Tornado\nDamage, by State",
           xlab="Damage in Millions of Dollars",
           panel=function(x,...) {
             panel.violin(x,col="bisque",...)
             panel.bwplot(x,...)
             panel.rug(x,col="red",...)
            }
          )
```

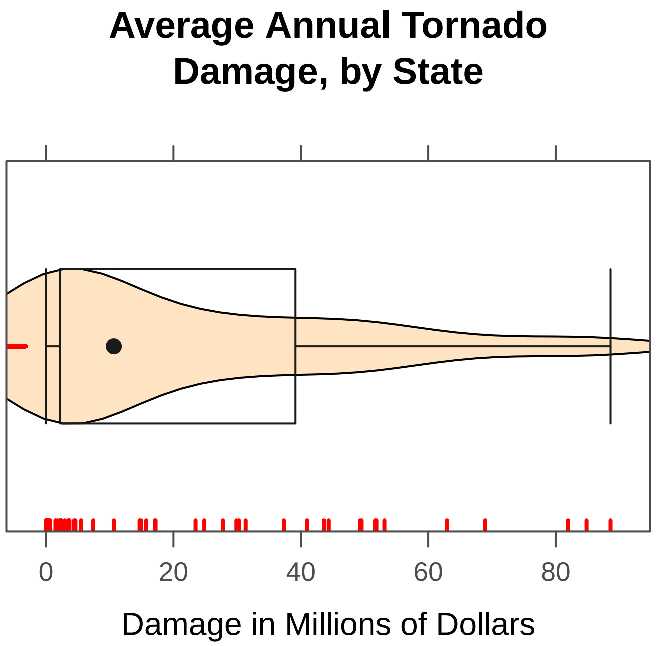


Figure 12.7: Damage with Rug. We added a 'rug' to this plot.

The result appears in Figure [Damage with Rug].

12.1.5 Tuning Density Plots

Adding a list of *density arguments* fine tunes features of the density plot. For example, `bw` specifies how "wiggly" the plot will be; `from` and `to` tell R where to begin and end estimation of the density curve.

Here is an example of what can be done (see Figure [Setting Bandwidth] for the results):

```
histogram(~damage,data=tornado,
           main="Average Annual Tornado\nDamage, by State",
           xlab="Damage in Millions of Dollars",
           type="density",
           breaks=c(0,2,4,6,10,15,20,25,30,40,50,60,70,80,90,100),
           panel=function(x,...) {
             panel.histogram(x,...)
             panel.rug(x,col="red",...)
             panel.densityplot(x,col="blue",
                          darg=list(bw=3,from=0,to=100),...)
            }
          )
```

R constructs a density plot by combining lots of little bell-shaped curves (called *kernals*), one centered at each point in the data. The bandwidth `bw` tells R how spread out these kernals should be: the bigger the bandwidth, the shorter and wider the kernal, and the stiffer the density curve will be. With a small bandwidth, the kernals are skinny and tall, giving the density plot a wiggly appearance, especially near isolated data points.

How do you know what the bandwidth should be? For now, you just have to try various values. The following `manipulate()` app helps you experiment with different values of the bandwidth.

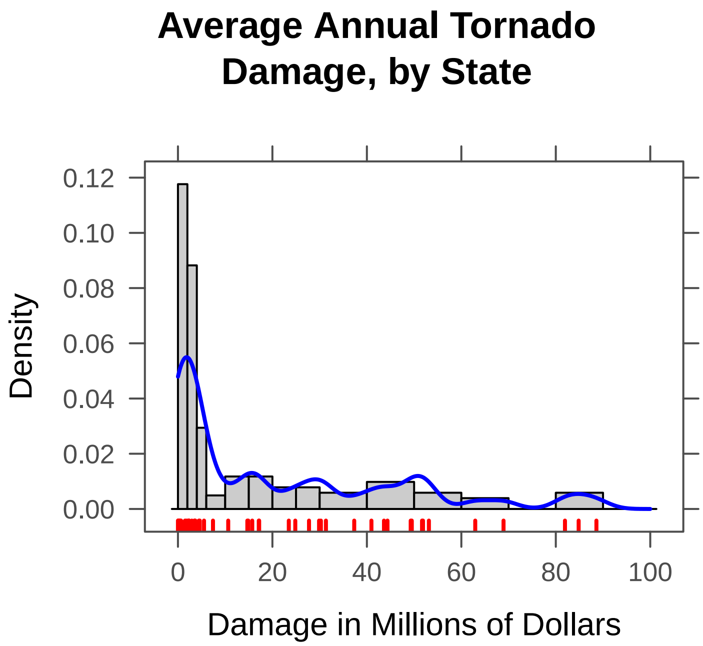


Figure 12.8: Setting Bandwidth. The bandwidth of the density plot was set to 3.

```
require(manipulate)
manipulate(
  bandwidth=slider(0.5,20,init=5,label="Bandwidth (1 = wiggly, 20 = stiff)"),
  histogram(~damage,data=tornado,
           main="Average Annual Tornado\nDamage, by State",
           xlab="Damage in Millions of Dollars",
           type="density",
           breaks=c(0,2,4,6,10,15,20,25,30,40,50,60,70,80,90,100),
           panel=function(x,...) {
             panel.histogram(x,...)
             panel.rug(x,col="red",...)
             panel.densityplot(x,col="blue",
                        darg=list(bw=bandwidth,from=0,to=100),...)
            }
          )
)
```

When the bandwidth is set too low, the wiggles in the density plot are too sensitive to chance clusters of data points – clusters that probably would not appear in the same place in a repeated study. When the bandwidth is set too high, the density plot is not able to capture the overall shape of the distribution.

12.1.6 More on Standard Deviation

Recall that when we compute the sample standard deviation, we don't quite average the squared deviations. Instead, we divide by one less than the number of data values:

$$s = \sqrt{(\sum (x_i - \bar{x})^2)/(n-1)}.$$

What if we have the entire population? Then the SD is called σ, and it is computed like this:

$$\sigma = \sqrt{(\sum (x_i - \mu)^2)/N},$$

where μ is the mean of the population and N is the number of individuals in the population. So you might well ask: "Why do we divide by one less than the number of items when we have a sample, but not when we have the entire population?"

To answer this, we first have to back up to the idea of variance. The sample variance is:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1},$$

and the population variance is

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}.$$

The formula for the population variance makes perfect sense. Although the $n-1$ in the formula for sample variance does not appear to make good sense, it has been cleverly designed so that the sample variance will be a good estimate of the population variance.

What do we mean by "good estimate"? Let's suppose that a statistician wants to estimate the variance of the heights of `imagpop`, but she only has time to take a sample of size, say, $n = 4$. Unknown to her the population variance is:

```
sigmasq <- var(imagpop$height)*(9999/10000)
sigmasq
```

```
## [1] 15.26323
```

Her sample might, on the other hand, might look like this:

```
HerSamp <- popsamp(n=4,pop=imagpop)
HerSamp
```

```
##         sex math income cappun height idealheight diff kkardashtemp
## 8248 female   no  71800 oppose   69.9          71  1.1           11
## 3223   male   no  18700 oppose   73.4          76  2.6           99
## 8173 female   no   2500 oppose   66.2          67  0.8           18
## 9448   male   no  76700 oppose   73.4          76  2.6           94
```

Then her estimate of the variance would be

```
var(HerSamp$height)
```

```
## [1] 11.8225
```

Her estimate might be high or low: it depends on the luck of the draw. But suppose that many, many statisticians (10,000 of them, let's say) were to each take a random sample of size 4 and compute the sample variance. Then the results would be like this:

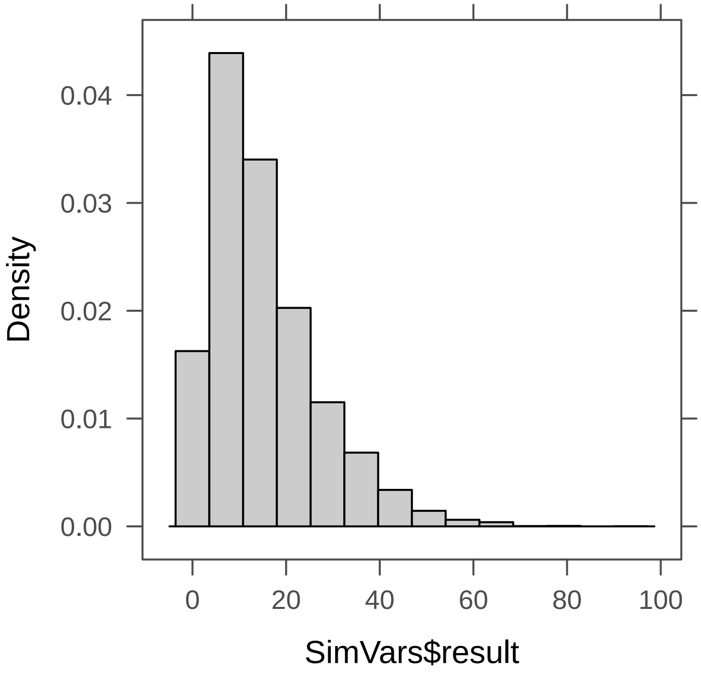
```
SimVars <- do(10000)*var(popsamp(n=4,pop=imagpop)$height)
```

Individually, their estimates would be all over the place (see Figure[Variance Estimates] for the plot:

```
head(SimVars,n=10)
```

```
##       result
## 1   1.396667
## 2  24.102500
## 3  21.510000
## 4   5.886667
## 5   4.033333
## 6  24.549167
## 7   8.609167
## 8  19.776667
## 9  16.070000
## 10  3.150000
```

```
histogram(~SimVars$result)
```



But on average, they would get:

```
mean(SimVars$result)
```

```
## [1] 15.26376
```

Notice that this about the same as the population variance $\sigma^2 = 15.2632308$.

On average, over many, many samples, the sample variance equals the population variance. We say that the sample variance is an *unbiased* estimator of the population variance. On the other hand, if the statisticians were to compute the sample variance by dividing by $n = 4$ instead of dividing by $n - 1 = 3$, then they would get results that are, on average, *too small*:

```
BadVars <- SimVars$result*3/4  #so that there is now a 4 on the bottom
mean(BadVars)
```

```
## [1] 11.44782
```

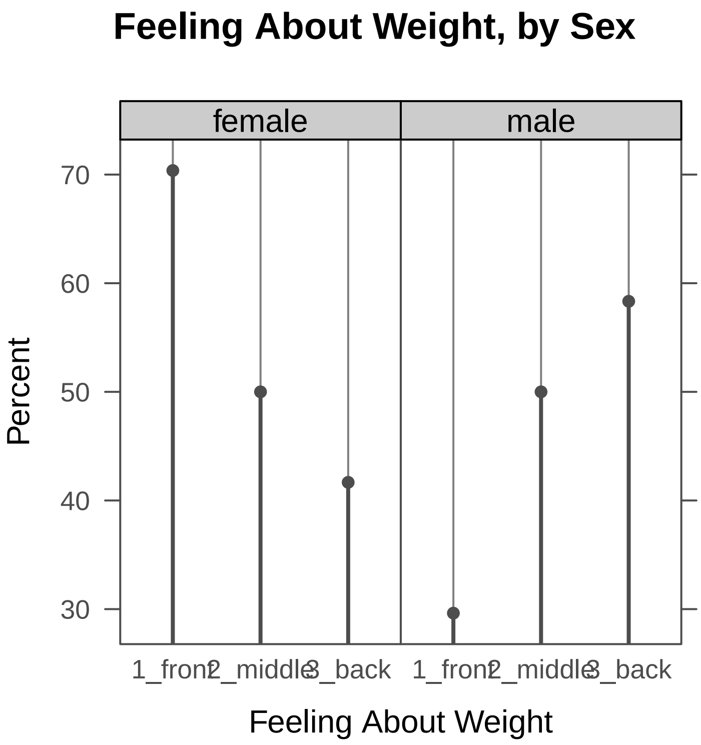
Sure enough, the results, on average are only about 3/4th the size of the true σ^2. Dividing by n in the sample variance would give you a biased estimator of population variance!

12.2 Chapter 2

12.2.1 Cleveland Dotplots

Barcharts are very popular for investigating categorical variables, but modern statisticians believe that the *Cleveland dot plot* is more useful in most situations.

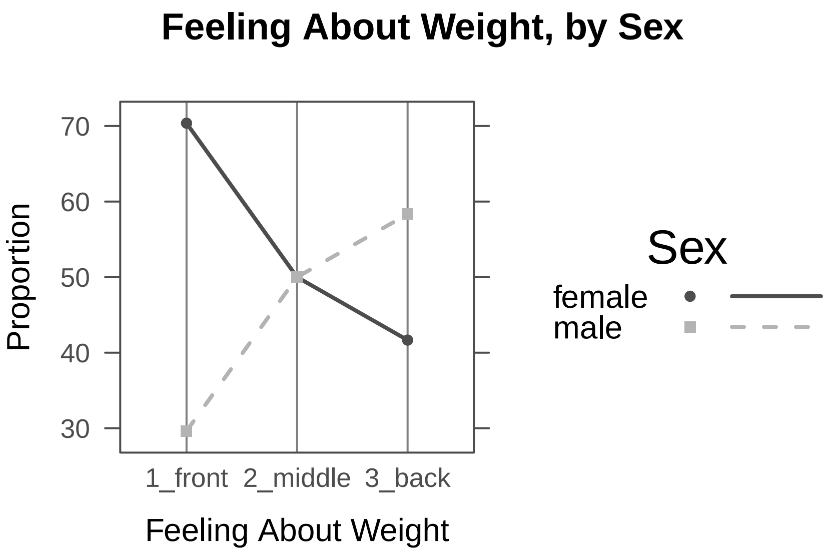
```
SexSeatrp <-100*prop.table(xtabs(~seat+sex,data=m111survey),margin=1)
dotplot(SexSeatrp,groups=FALSE,horizontal=FALSE,type=c("p","h"),
        ylab="Percent",xlab="Feeling About Weight",
        main="Feeling About Weight, by Sex")
```



The resulting plot appears as Figure [Cleveland Plot]. The first line of code above constructs a twoway table and computes row percentages for it, using the `prop.table()` function to prevent having to deal with the extraneous column of total percentages. Note that in the twoway table the explanatory variable comes second. Reverse the order to see the effect on the layout of the plot.

The second line constructs the dot plot itself. Whereas barcharts indicate percentages by the tops of rectangles, the Cleveland dot plot uses points. Setting the `type` argument to `c("p","h")` indicates that we want points, but also lines extending to the points. The lines are helpful, as the human eye is good at comparing relative lengths of side-by-side segments. The `groups` argument is FALSE by default; we include it here to emphasize how the plot will change when it is set to TRUE, as in the next example. The results appears in Figure [Cleveland Plot 2].

```
dotplot(SexSeatrp,groups=TRUE,horizontal=FALSE,type=c("p","o"),
        ylab="Proportion",xlab="Feeling About Weight",
        auto.key=list(space="right",title="Sex",lines=TRUE),
        main="Feeling About Weight, by Sex")
```



Setting `groups` to TRUE puts both sexes in the same panel. Setting `type=c("p","o")` produces the points, with points in the same group connected by line segments. The `lines` argument in `auto.key` calls for lines as well as points to appear in the legend.

12.3 Chapter 3

12.3.1 Fixed and Random effects in Simulation

When we used the ChisqSimSlow apps during the ledgejump study, we set the `effects` argument to “fixed.” Later on, in the **sex** and **seat** study, we set `effects` to “random”. What was all that about?

Try the ChisqSimSlow app in the ledgejump study again, and this time pay careful attention to each twoway table as it appears.

```
require(manipulate)
ChisqSimSlow(~weather+crowd.behavior,data=ledgejump, effects="fixed")
```

Now try it again, but this time with `effects` set to “random”:

```
require(manipulate)
ChisqSimSlow(~weather+crowd.behavior,data=ledgejump, effects="random")
```

You might notice that when effects are fixed, the number of cool-weather days is always 9, and the number of warm-weather days is always 12, just as in the original data. On the other hand, when effects are random, although the total number of incidents stays constant at 21, the division of them into cool and warm days varies from one resample to another.

In the ledgejump study, the 21 incidents could not reasonably be regarded as a random sample from some larger "population" of incidents. Most likely, the researcher included in his study all of the incidents for which he could determine the relevant information about weather and crowd behavior. This isn't a *random* sample from among all incidents. Therefore, there is no randomness involved in how many warm-weather and how many cool-weather incidents were involved: if we could go back in time and watch these incidents play out again, 9 of them would still have been in warm weather, and 12 would have been in cool weather.

But chance *is* still involved: in the determination of the value of the response variable. In each incident, factors not associated with the random variable are at play. Such factors – the personalities of the people in the crowd, the length of time the would-be jumper stood on the ledge, etc. – are modeled as "chance" and these chance factors help determine whether the crowd is baiting or polite. Recall that if weather and crowd behavior were unrelated, then our best guess was that for each incident there was a 52.4% chance that the crowd would be polite and a 47.6% chance that it would be baiting. In the resampling with fixed effects, there are 9 cool-weather incidents and 12 warm-weather ones, and each incident is given a 52.4% chance to have a polite crowd.

On the other hand, if our twoway table is based on a random sample from a larger population, as the **sex** and **seat** study was, then we say that the effects are *random*. In the original sex-seat sample, there were 71 individuals: 40 females and 31 males. If we were to repeat the sample again, we would not be guaranteed to have 40 females and 31 males in it. Our best guess, though, based on our sample, is that $\frac{40}{71} \times 100 = 56.3\%$ of the population is female, so in the resampling with random effects, we give each individual a 56.3% chance to be female. Since the resampling is done under the hypothesis that sex and seat are related, the chances for each resample-individual to prefer front, back and middle are the same, regardless of whether the individual is female or male.

Just as the two methods of resampling differ mathematically, so they also differ in the nature and scope of our conclusion in Step Five. In the ledgejump study, fixed effects resampling models the assumption that the 21 incidents themselves would have been the same from sample to sample: the only thing that varies with chance is how the crowd behaves in each incident. Hence your conclusion in Step Five – that the sample data don't quite provide strong evidence for a relationship between weather and crowd behavior – applies only to those 21 incidents. In the **sex** and **seat** study, on the other hand, the random-effects resampling method models the assumption that the the 71 GC students were a random sample from the larger population of all GC students. The conclusions we draw from this data apply to this larger population.

When we set *simulate.p.value* to TRUE in `chisq.testGC`, R does resampling. However, it takes a third approach: the the row sums (tallies of the various values of the X variable) are fixed, as in our fixed effects, but the column sums are also fixed to be the same as the column sums of the original data. In our terminology, you could say the resampling is "double-fixed." R has its own reasons for the double-fixed approach that we will not cover here.

If you want to fixed effects simulations, just use set the `simulate.p.value` argument in `chisq.testGC()` to "fixed". For random effects, set the argument to "random".

Be assured that, as sample size increases, all three methods — fixed, random and double-fixed — yield approximations that agree more and more nearly with each other. At small sample sizes, though, they can differ by a few percentage points.

Note to Instructors: Our use of the terms "fixed effects" and "random effects" is not quite standard, but is analogous to the use of these terms in mixed-effects linear modeling.

12.4 Chapter 4

12.4.1 Point Shapes in Scatterplots using `pch`

The plot character, `pch`, is an integer between 1 and 25 that controls how the points on your scatterplot appear. A summary can be seen in Figure[Plot Characters].

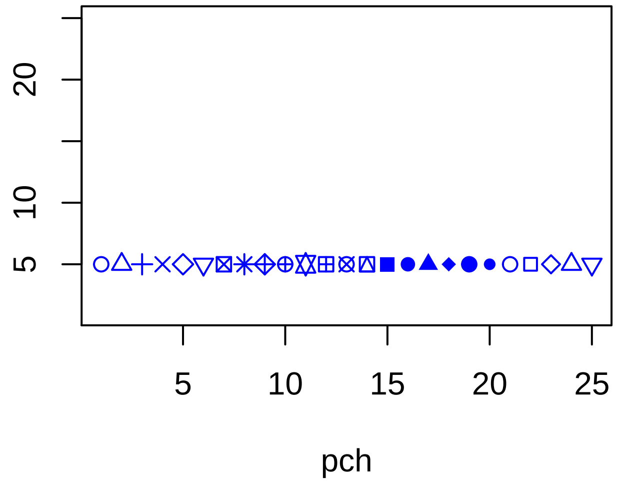


Figure 12.9: Plot Characters.

12.4.2 Scatterplot Matrix

Given several numerical variables, R can produce a group of scatterplots, one for each pair of variables – all in one graph. Such a graph is called a *scatterplot matrix.* We can create a matrix of scatterplots using the following `pairs` function in R. You only need to enter in the variables that you want plotted and the dataset that contains them. R will create a square matrix of plots for every combination of variables. See FigureScatterplot Matrix.

```
pairs(~height+sleep+GPA+fastest,data=m111survey)
```

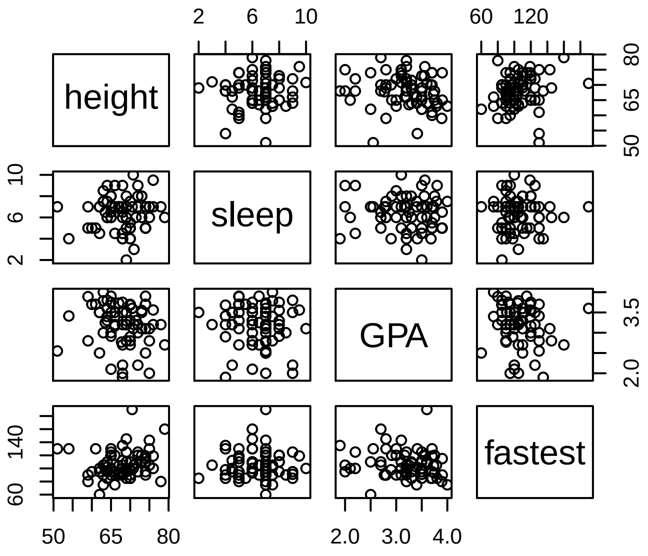


Figure 12.10: Scatterplot Matrix.

Of course, you can always make this look nicer by changing the colors and plot characters. Notice that the scatterplots are arranged in a somewhat symmetric way across the main diagonal (the boxes with the variable names). A scatterplots mirror image uses the same variables, but the explanatory and response variables are reversed.

You can also plot only the upper (or lower) panels. See Figure[Upper Panel] and Figure[LowerPanel].

```
pairs(~height+sleep+GPA+fastest,data=m111survey,pch=19,col="red",lower.panel=NULL)
```

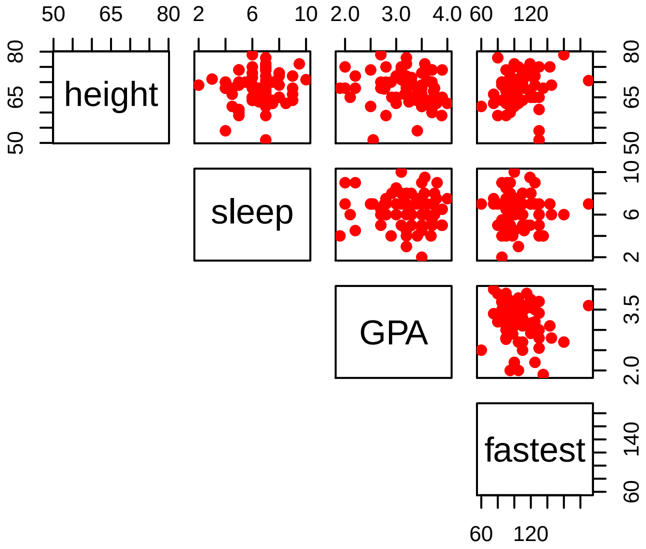


Figure 12.11: Upper Panel. Scatterplot matrix showing only the upper panel of scatterplots.

```
pairs(~height+sleep+GPA+fastest,data=m111survey,pch=19,col="red",upper.panel=NULL)
```

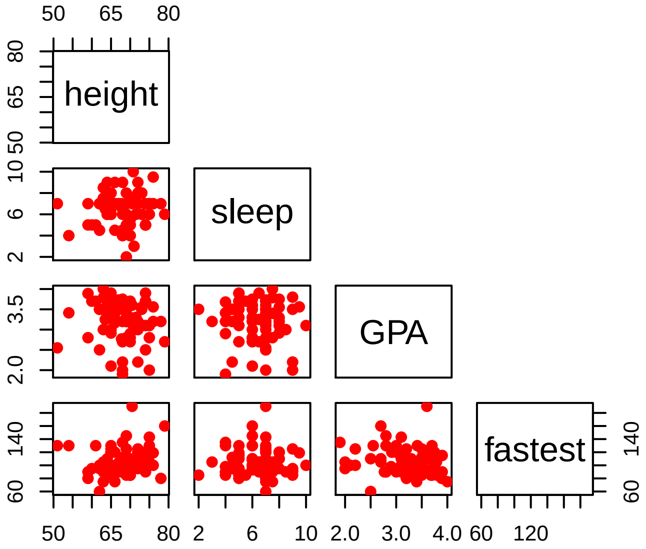


Figure 12.12: Lower Panel. Scatterplot matrix showing only the lower panel of scatterplots.

12.4.3 The Rationale for Values of the Correlation Coefficient, r

Let's consider why variables with a positive linear association also have a positive correlation coefficient, r. Consider what value of r you might *expect* for **positively correlated** variables. Let's recall how we plotted the two "mean" lines to break a scatterplot into four "boxes". See Figure[Four Boxes].

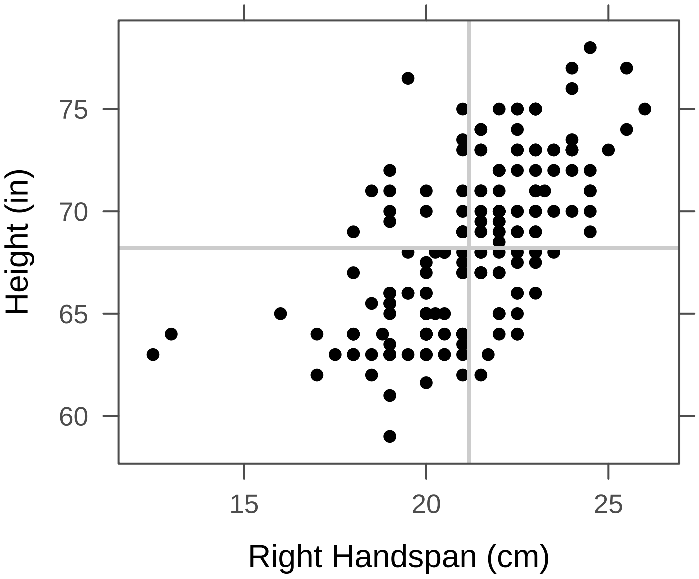


Figure 12.13: Four Boxes. Scatterplot of Right Handspan (cm) versus Height (in). The lines marking the mean of the handspans and the mean of the heights have been plotted to break the scatterplot into four boxes.

We've looked at this scatterplot before, and determined that it indicates a positive association between **RtSpan** and **Height**. Now, let's think carefully about how the points in the scatterplot contribute to the value of r. Check out the formula again:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- When an x-value lies *above* the mean of the x's, it's z-score is **positive**. Likewise, a y-value that lies *above* the mean of the y's has a **positive** z-score. Every ordered pair in the upper right box has an x and y-coordinate with **positive** z-scores. Multiplying 2 positive z-scores together gives us a **positive** number. So, every point in the upper right box contributes a positive number to the sum in the formula for r.

- When an x-value lies *below* the mean of the x's, it's z-score is **negative**. Likewise for y. Every ordered pair in the lower right box has an x and y-coordinate with **negative** z-scores. Multiplying 2 negative z-scores together gives us a **positive** number. So, every point in the lower left box has a positive contribution to the value of r.

Following the same rationale, the points in the upper left box and lower right box will contribute negative numbers to the sum of r.

- When an x-value lies *above* the mean of the x's, it's z-score is **positive**. A y-value that lies *below* the mean of the y's has a **negative** z-score. Every ordered pair in the lower right box has an x-coordinate with a **positive** z-score and a y-coordinate with a **negative** z-score. Multiplying a positive and a negative z-score together gives us a **negative** number. So, every point in the lower right box contributes a negative number to the sum in the formula for r.

- When an x-value lies *below* the mean of the x's, it's z-score is **negative**. A y-value that lies *above* the mean of the y's has a **positive** z-score. Every ordered pair in the upper left box has an x-coordinate with a **negative** z-score and a y-coordinate with a **positive** z-score. Multiplying a positive and a negative z-score together gives us a **negative** number. So, every point in the upper left box contributes a negative number to the sum in the formula for r.

Since **positively associated** variables have *most* of their points in the upper right and lower left boxes, *most* of the numbers being contributed to the summation are **positive**. There are some negative numbers contributed from the points in the other boxes, but not nearly as many. When these values are summed, we end up with a **positive** number for r. So we say that these variables are **positively correlated**!

In a similar manner, we can argue that since *most* of the points in a scatterplot of **negatively associated** variables are located in the upper left and lower right boxes, most of the products being contributed to the sum of r are negative (with a few positive ones sprinkled in). This gives us a **negative** number for r. So we say that these variables are **negatively correlated**!

12.4.4 Computation of the Coefficients in the Regression Equation

The regression equation is $\hat{y} = a + bx$. You might be wondering... how are a and b calculated? The formula for the slope b is:

$$\text{slope } = b = r \cdot \frac{s_y}{s_x},$$

where

- r is the correlation coefficient,
- s_y is the SD of the y's in the scatterplot, and
- s_x is the SD of the x's in the scatterplot.

The formula for the intercept a is:

$$\text{intercept } = a = \bar{y} - b \cdot \bar{x},$$

where

- b is the slope calculated above,
- $\bar{y}$ is the mean of the y's in the scatterplot, and
- $\bar{x}$ is the mean of the x's in the scatterplot.

Before interpreting these formulas, let's look at a little late 19th century history. Sir Francis Galton, a half-cousin of Charles Darwin, made important contributions to many scientific fields, including biology and statistics. He had a special interest in heredity and how traits are passed from parents to their offspring. He noticed that extreme characteristics in parents are not completely passed on to their children.

Consider how fathers' heights is related to sons' heights. See Figure[Galton].

It seems reasonable to think that an average height father would probably have an average height son. So surely our "best fit" line should pass through the *point of averages*, $(\bar{x}, \bar{y})$. See Figure [Point of Averages]

Intuitively, it might also seem that a reasonably tall father, say, 1 standard deviation taller than average would produce a reasonably tall son, also about 1 standard deviation taller than average. The line that would "best fit" this assumption would have slope equal to $\frac{s_y}{s_x}$.

However, this *not* the "best fit" line. It does not minimize the Sum of Squares! Check out how the *regression* line looks in comparison to this *standard deviation* line.

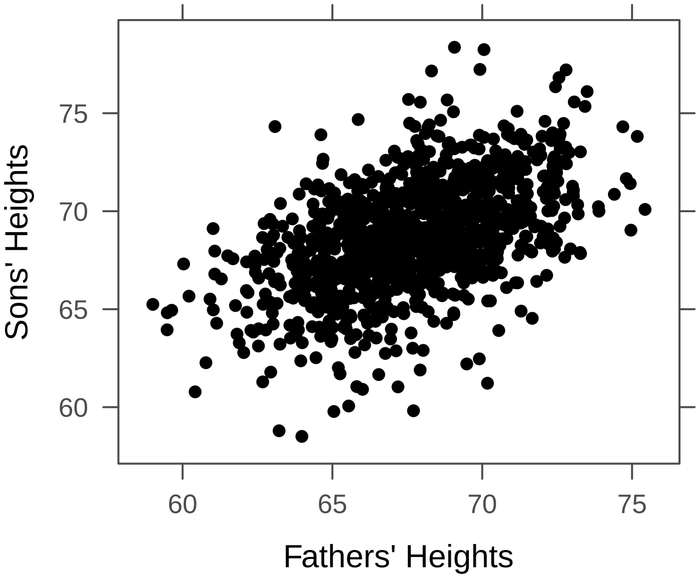


Figure 12.14: Galton. Relationship Between Father and Sons' Heights

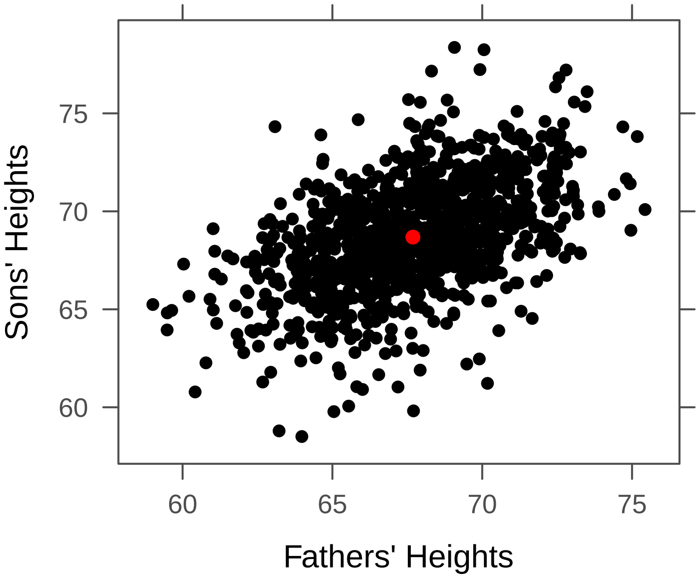


Figure 12.15: Point of Averages. Galton data with the point of averages plotted.

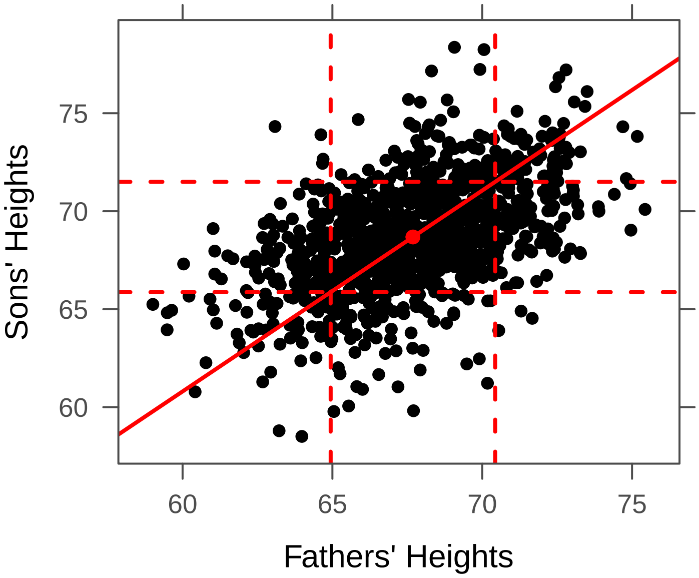


Figure 12.16: SD Line. Galton Data with SD line

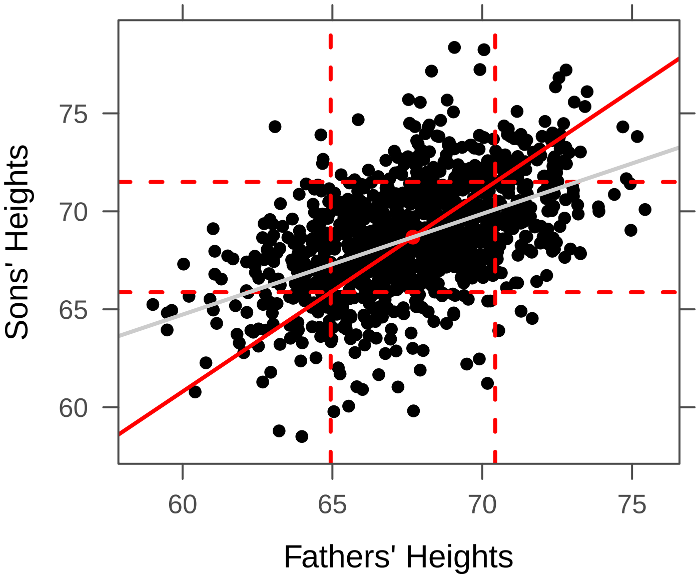


Figure 12.17: Regression. Galton data with SD line and regression line.

The slope of the SD line is $b = \frac{s_y}{s_x}$. The slope of the regression line is $b = r \cdot \frac{s_y}{s_x}$. Since r is a value between -1 and 1, you can see why this causes the regression line to be more *shallow*.

This is what is known as the **regression effect** or **regression to the mean**. Extremely tall fathers do tend to have taller than average sons, but the sons don't tend to be as extreme in height as their fathers. Likewise for short fathers.

Check out the following app to explore this idea further!

```
require(manipulate)
ShallowReg()
```

12.5 Chapter 5

12.5.1 The rbind Function

The rbind function combines objects in R by rows. (It is called rbind to stand for "rowbind".) If you have several lists stored and you want to combine them into one object, you can use rbind.

```
list1=c(1,2,3)
list2=c(5, 6, 7)
list3=c(100, 200, 300)
rows=rbind(list1,list2,list3)
rows
```

```
##       [,1] [,2] [,3]
## list1    1    2    3
## list2    5    6    7
## list3  100  200  300
```

Essentially, you have created a matrix. You can access objects out of rows similar to how you would access a value out of a list.

```
rows[1,2] #gives the number in the 1st row and 2nd column
```

```
## list1
##     2
```

```
rows[2,1] #gives the number in the 2nd row and 1st column
```

```
## list2
##     5
```

12.5.2 The cbind Function

The cbind function is very similar to rbind. It combines objects in R by columns. (It is called cbind to stand for "columnbind".)

```
columns=cbind(list1,list2,list3)
columns
```

```
##      list1 list2 list3
## [1,]     1     5   100
## [2,]     2     6   200
## [3,]     3     7   300
```

You can use `cbind` and `rbind` to combine objects other than numbers, such as characters.

```
list4=c("A", "B", "C", "D")
list5=c("E", "F", "G", "H")
rows=rbind(list4,list5)
rows
```

```
##       [,1] [,2] [,3] [,4]
## list4 "A"  "B"  "C"  "D"
## list5 "E"  "F"  "G"  "H"
```

```
columns=cbind(list4,list5)
columns
```

```
##      list4 list5
## [1,] "A"   "E"
## [2,] "B"   "F"
## [3,] "C"   "G"
## [4,] "D"   "H"
```

12.6 Chapter 6

12.6.1 The Role of Limits in Density Plots

Recall the grouped density plots, for example:

```
densityplot(~sentence,data=attitudes,
            groups=def.race, plot.points=FALSE,
            main="Race and Recommended Sentence",
            xlab="Recommended Sentence",
            auto.key=list(space="right",title="Suggested\nRace"),
            from=2,to=50)
```

The result is shown in Figure [Race and Sentence]. Density plots for different are especially effective when overlaid, because differences in the modes (the “humps”) of the distribution are readily apparent.

In the case of this data, we know that the minimum possible sentence is 2 and the maximum possible is 50. (These limits were specified on the survey forms.) Hence we should communicate these limits to R by means of the `from` and `to` arguments. R then constructs the kernel density estimators with these limits in mind.

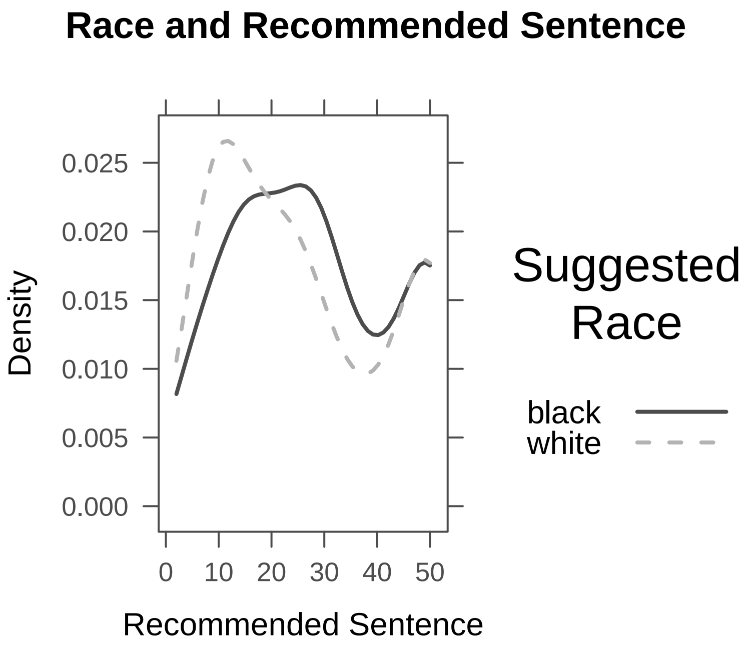


Figure 12.18: Race and Sentence. We set limits for the density curves.

12.6.2 More about Legends

There are many ways to modify the legend provided by the `auto.key` argument. These modifications are communicated by setting the values of certain other arguments and combining them in a list. The `space` argument is set by default to "top", in which case the legend appears above the graph. It may also be set, to "left", "right", or "bottom". A legend title may also be supplied through the argument title. `Finally, settings` acolumns‘ argument controls the layout of the elements in the legend (see Figure [Sentence by Defendant's Race]:

```
densityplot(~sentence,data=attitudes,
            groups=def.race, plot.points=FALSE,
            main="Race and Recommended Sentence",
            xlab="Recommended Sentence",
            auto.key=list(space="top",title="Suggested Race",columns=2),
            from=2,to=50)
```

12.6.3 More on Strip-plots

Strip-plots are most effective when the groups sizes are small: when groups are large, many data values may equal one another, and *overplotting* will result. There are some techniques available to alleviate the effects, of over-plotting, though, provided the dataset is not too large. The two primary techniques are *jittering* and *translucence.*

See Figure [Sentence by Major] for the result of the following code:

```
stripplot(sentence~major,data=attitudes,
         main="Sentence By Major",xlab="Major",col="red",
         jitter.data=TRUE,alpha=0.5,
         panel= function(...){
```

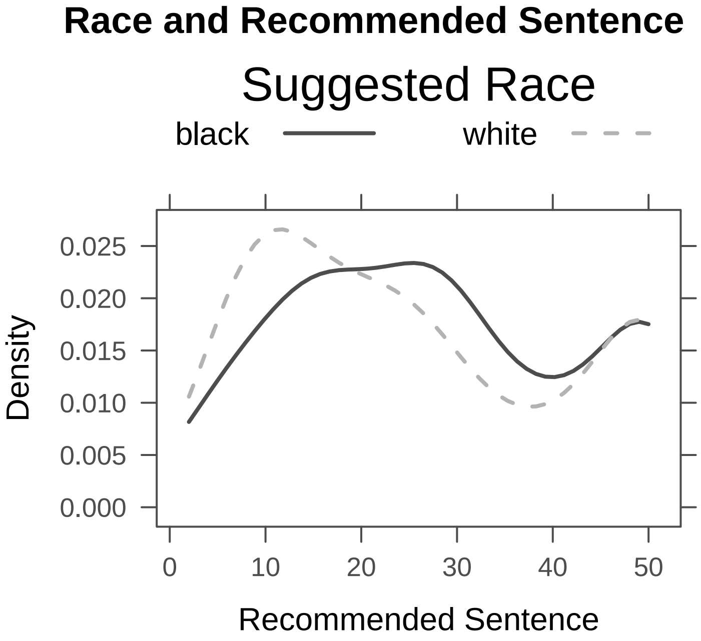


fig-1.pdf

```
            panel.violin(...)
            panel.stripplot(...)
          })
```

In the code above, setting the argument `jitter.data` to `TRUE` has the effect, in a strip-plot, of moving each point randomly a bit in the direction perpendicular to the axis along which the groups are ordered, thus separating the points from one another. The `alpha` argument has a default value of 1. When set to a value between 0 and 1, it introduces a degree of translucence to the points. At value 0.5, for example, two over plotted points would appear as dark as a single point would when alpha is set at 1.

12.6.4 Assessing Statistical Significance

Recall that in a randomized experiment, chance is always involved in the collection of the data, simply because it is involved in the assignment of subjects to treatment groups. Thus we can always ask the question of statistical significance. Let's investigate that question for the Knife-or-Gun study.

When the consent problem restricts us from applying the results of an experiment to a larger population, we think about the problem in terms of the set of subjects themselves. We adopt what is known as the *ticket model*.

In the ticket model, we imagine that every subject has a magical ticket. Values of the response variable for that subject under the various possible treatments are written on fixed areas of the ticket. In the Knife or Gun study, we imagine that on the left-hand side of the ticket is written the volume of the dying screams he or she would emit, should he or she be killed with a knife. In the right-hand side of the ticket is written the volume of screams he/she would emit if being killed by a gun.

In the ticket model, the question of whether or not the explanatory variable makes a difference in the response variable boils down to what is on these tickets. For a subject with a ticket like

(Knife 65, Gun 65),

the means of slaying makes no difference: she would yell at volume 65 regardless of whether she was killed by knife or by gun. For a subject with a ticket reading

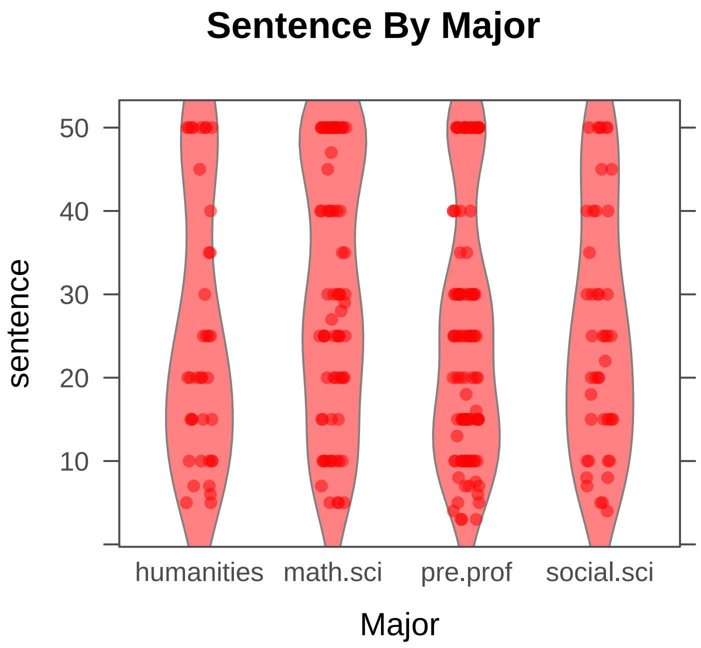


Figure 12.19: Sentence by Major. Strip-plot comined with violin plot.

> (Knife 70, Gun 67),

the means of slaying makes a difference: being killed by a knife would make her yell louder.

The tickets are truly magical, because researchers are allowed to read at most one part of any person's ticket. That's because each subject is assigned to just one treatment group. Competing hypotheses about the effect of the means of slaying on the volume of yells can be stated in terms of the ticket model as follows:

> H_0 [Means of slaying makes no difference, on average, for the subjects]: The mean of the Knife-side of the tickets of all subjects equals the mean of the Gun-side of the tickets.

> H_a [On average, dying by gun makes the subjects yell louder]: The mean of the Knife-side of the tickets of all subjects is greater than the mean of the Gun-side of the tickets.

We have stated our hypotheses. That was Step One of a test of significance.

Now for Step Two: computing a test statistic. A reasonable test statistic would be the difference of sample means:

```
compareMean(volume~means,
            data=knifeorgunblock)
```

```
## [1] 20.13
```

The difference of means is 20.13, indicating that on average the Knife subjects yelled 20.13 decibels louder than the Gun subjects did.

Next, Step Three: computing the P-value. We would like to know the probability of getting a difference in sample means at least as big as the one we actually got, if H_0 is actually true.

To find this probability we imagine—temporarily and for the sake of argument only—that the NUll is really true. In fact, we'll make the extra-strong assumption that the Null is super-duper true: that means of slaying makes no difference for ANY subject. In that case, for every the number on the Knife-side equals the number on the Gun-side. If that's true, then we actually know all of the numbers on all of the tickets. (Reading one side—that, is, killing the subject—tells us what the other side says.)

This neat fact puts us in the happy position of being able to simulate what would happen if we were to repeat the experiment many, many times. We would have the same 20 subjects each time: only the group assignments would differ. But no matter the group assignment, we can tell what each person's dying screams will be.

For convenience, we'll write a function that pretends to run the who experiment all over again, with blocking, computing the difference in the mean volumes of yells for each group, each time, and recording the difference:

```
set.seed(12345)
KnifeGunSim <- do(500)*compareMean(volume~treat.grp,
                    data=RandomExp(knifeorgunblock,
                      sizes=c(10,10),groups=c("k","g"),
                      block="hoghollerer"))
```

Let's look at the first few simulations::

```
head(KnifeGunSim,n=5)
```

```
##   result
## 1  -0.99
## 2   5.59
## 3  -1.13
## 4  -9.55
## 5  -3.31
```

Remember: these differences are all based on the assumption that means of slaying has no effect at all on the volume of dying screams. So, about how big are the differences, when the Null is right? Let's see:

```
favstats(~result,data=KnifeGunSim)
```

```
##     min     Q1 median    Q3   max     mean       sd   n missing
##  -16.73 -4.025   -0.7 3.285 15.81 -0.48924 5.228157 500       0
```

As you might expect, the typical difference is quite small: about 0, give or take 5.5 or so. The difference we saw in the study (20.13) was about four SDs above what the Null would expect.

In fact, the maximum of the simulated differences was only 12.73: not once in our 500 simulations did the test statistic exceed the value of the test statistic that we got in the actual study.

This gives us Step Four in a test of significance: the P-value is very small, probably less than one in 500, so we reject H_0.

This study provided very strong evidence that, *for these 20 subjects*, slaying with a knife evokes louder yells than slaying with a gun does.

12.6.5 Interaction

There is one other important concept that often applies in experiments, that wee think bears a leisurely discussion: it is the concept of *interaction*.

```
data(ToothGrowth)
View(ToothGrowth)
help(ToothGrowth)
```

```
bwplot(len~as.factor(dose)|supp,data=ToothGrowth)
```

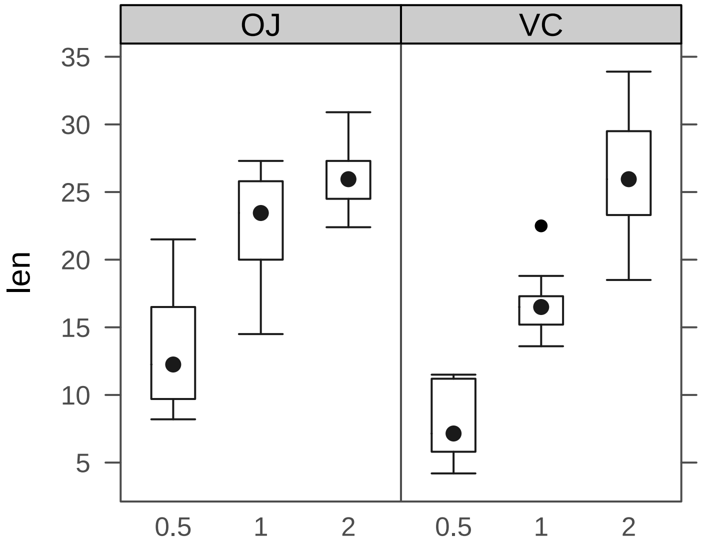


Figure 12.20: Tooth growth.

Figure [Tooth growth] shows boxplots of the data. In both panels, the boxes rise as you read to the right. Hence, for both values of the explanatory variable **supp**, the length of tooth increases as dosage (also an explanatory variable) increases. However, the increase in length as dosage of Vitamin c increases from 1 to 2 is greater when the dosage method is by ascorbic acid (VC) than when the Vitamin C is administered in the form of orange juice (OJ). Hence, the effect of *dose **on** len **differs with differing values of the other explanatory variable** supp**. Because of this difference, the variables** dose **and** supp** are said to be *interact*. The formal definition follows:

Interaction Two explanatory variables X_1 and X_2 are said to *interact* when the relationship between X_1 and the response variable Y differs as the values of the other variable X_2 differ.

Practice: In each of the situations below, say whether there is a confounding variable present, or whether there is interaction. In the confounding case, identify the confounding variable and explain why it is a confounder. In the interaction case, identify the two explanatory variables that interact.

(1). In a study of the effect of sports participation and sex on academic performance, it is found that the mean GPA of male athletes is 0.7 points less than the mean GPA of female athletes, but the mean GPA of male non-athletes is only 0.2 points lower than the mean GPA of female non=athletes.

> (2). In a study of the effect of alcohol on the risk of cancer, it is found that heavy drinkers get cancer at a higher rate than moderate drinkers do. However, it is known that smokers also tend to drink more than non-smokers, and that smoking causes various forms of cancer.

As another example, consider the `pushups` data frame:

```
data(pushups)
View(pushups)
help(pushups)
```

Play with the data using the a Dynamic Trellis app:

```
require(manipulate)
DtrellScat(pushups~weight|position,data=pushups)
```

The relationship between weight and push-ups varies depending on position: for Skill players the relationship is somewhat positive (the scatterplot rises as you read to the right), but for Skill players the relationship is somewhat negative (scatterplot falls as you move to the right). Thus, variables **weight** and **position** appear to interact. One might wonder, though, whether the observed interaction is statistically significant: ater all, there weren't many Line players in the study to begin with.

12.7 Chapter 8

12.7.1 We Lied About the SD Formulas!

Recall the SimpleRandom app: let's play with it one more time:

```
require(manipulate)
SimpleRandom()
```

This time, pick one of the variables and move the slider up to the sample size 10,000. Click on the slider several times, keeping it set at 10,000. Watch the output to the console.

You probably noticed that the sample statistics did not change from sample to sample, and that they were equal to the population parameters every time. This makes sense, because when the sample size is the same as the size of the population, then simple random sampling produces a sample that HAS to be the population, each and every time!

But wait a minute: if the sample statistic is ALWAYS equal to the population parameter, then the likely amount by which the statistic differs from the parameter is ZERO. Hence the SD of the estimator should be zero. Fro example, if we are estimating the mean height of `imagpop`, then the SD of $\bar{x}$ should be zero. But the formula we gave for the SD is:

$$\frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{10000}} = \frac{\sigma}{100},$$

which has to be BIGGER than zero. Therefore the formula is wrong.

Well, it is wrong for simple random sampling. It is correct for random sampling *with replacement* form the population. The correct formula for the SD of $\bar{x}$, when we are taking a simple random sample – sampling without replacement – is:

$$Sd(\bar{x}) = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}},$$

where n is the sample size and N is the size of the population. The quantity

$$\sqrt{\frac{N-n}{N-1}}$$

is called the *correction factor*.

As you can see, at sample size $n = 10000$ and population size $N = 10000$ the quantity $N - n$ will be zero, forcing the correction factor to be zero, and thus forcing the SD of $\bar{x}$ to be zero as well.

Usually we don't bother with the correction factor in practice, because usually n is small in comparison to N. For example, when we take a SRS of size $n = 2500$ from the population of the United States ($N \approx 312000000$), then the correction factor is equal to:

```
N <- 312000000
n <- 2500
CorrFac <- sqrt((N-n)/(N-1))
CorrFac
```

```
## [1] 0.999996
```

The correction factor is approximately 0.999996,which so close to 1 that it is rounded to one in the Knitted version of this document. We know that multiplying a number by 1, won't change the original number, so multiplying the "Wrong" SD formula by the correction factor barely changes the number at all.

If you happen to know the population size, however, there is no harm in using the correct SD formula, with the correction factor.

The same correction factor shows up in the correct SD formulas for $\hat{p}$ and for $\bar{d}$, and there are correction factors for the SDs of the other two Basic Five parameters, too.

12.7.2 Are We Only Ever Interested in Population Parameters?

We have spent the whole chapter on population parameters and the statistics that we use to estimate them. But are we only ever interested in population parameters? Are statistics never used to estimate anything else?

The quick answer is No, there are times when the number we want to estimate is not a parameter for a population. For example, sometimes we want to estimate a probability:

- If we would like to know the probability p for a coin to land Heads, then we might toss the coin many times, compute the proportion $\hat{p}$ of times that the coin landed Heads, and use this to estimate p. We weren't actually taking a sample, because there isn't really a "population" of coin tosses to sample from.
- Another example: we often estimate a P-value by simulation. Again the P-value is a number – the probability of getting data as extreme as the data we actually got, if the Null is true – and we estimate it by simulating the study on the computer many times with a set-up in which the Null is true. Here again, we are simulation many times, but not sampling from some "population" of all possible simulations.

When we estimate a probability by simulation, we still call the probability a "parameter". It's just not a *population* parameter.

On the other hand, some problems that do not appear to be about population parameters really are problems about populations parameters, in disguise. A good example would be a question about the relationship between two categorical variables, for example:

Research Question: At Georgetown College, is sex related to seating preference?

We could frame the question about relationship as a question about some populations proportions. Let:

- $p_{male,front}$ = the proportion of all GC males who prefer to sit in front;
- $p_{female,front}$ = the proportion of all GC females who prefer to sit in front;
- $p_{male,middle}$ = the proportion of all GC males who prefer to sit in the middle;
- $p_{female,middle}$ = the proportion of all GC females who prefer to sit in the middle;
- $p_{male,back}$ = the proportion of all GC males who prefer to sit in back;
- $p_{female,back}$ = the proportion of all GC females who prefer to sit in back;

Then someone who believes that sex and seat are unrelated at GC believes three things

- $p_{male,front} = p_{female,front}$
- $p_{male,middle} = p_{female,middle}$
- $p_{male,back} = p_{female,back}$

Someone who thinks that the two variables are related believes that at least one of the three equalities above is incorrect.

12.8 Chapter 9

12.8.1 Distinction Between t and z in Confidence Intervals for Means

In order to more fully understand the distinction between using the t-multiplier (from the t-distribution) and using the z-multiplier (from the normal distribution) in the construction of confidence intervals for **means**, let's first remind ourselves of the statement of the Central Limit Theorem.

Central Limit Theorem: For any population with a finite mean μ and finite standard deviation σ, the distribution of the sample mean $\bar{x}$ gets closer and closer to

$$norm(\mu, \frac{\sigma}{\sqrt{n}})$$

as the sample size n gets larger and larger.

Now, let's consider four cases:

Case 1: The population standard deviation, σ, is **known** and the population is normally distributed.

When we know σ, we do not have to approximate the SD with the SE in the formula for the confidence interval. In addition, we can find our z-score $= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ exactly.

Furthermore, if the population from which we are drawing our sample is normal, then our sample estimate, $\bar{x}$, is also going to exactly follow a normal distribution, *regardless of the sample size.* This means that the z-score comes from the normal curve.

So, $\dfrac{\bar{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ exactly follows a standard normal distribution, regardless of the sample size.

In this situation, the z-multiplier is actually the correct multiplier to use. However, this situation rarely crops up. It is very unlikely that we would know the distribution of our population and know the population standard deviation.

Case 2: The population standard deviation, σ, is **known** and the population is not normally distributed.

The difference here is that we either don't know if the population is normally distributed or we know that it is not. For this reason, we would be unable to say that the estimator, $\bar{x}$, follows a normal distribution. However, the Central Limit Theorem ensures that for large enough sample sizes, $\bar{x}$ is *approximately* normally distributed. So, in this case, even though we are not making an approximation to the $\text{SD}(\bar{x})$ (since we know σ), we are making an approximation when we use a z-multplier (since we don't know that $\bar{x}$ follows a normal distribution exactly).

So, $\dfrac{\bar{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ approximately follows a standard normal distribution for large sample sizes. The Central Limit Theorem guarantees nothing about small sample sizes.

For this situation, it is still acceptable to use the z-multiplier as long as your sample is not too small.

Case 3: The population standard deviation, σ, is **unknown** and the population is normally distributed.

For this case, $\bar{x}$ is normally distributed regardless of the sample size. However, since we do not know σ, we must use the t-multiplier, $\dfrac{\bar{x}-\mu}{\dfrac{s}{\sqrt{n}}}$, with $n-1$ degrees of freedom. The numerator of this ratio is normally distributed, but the denominator is not. Since s is a random variable, the denominator is a random variable. Thus, we have a ratio of two random variables and this has a t distribution.

So, $\dfrac{\bar{x}-\mu}{\dfrac{s}{\sqrt{n}}}$ exactly follows a t- distribution with $n-1$ degrees of freedom, regardless of the sample size.

For this situation, you should always use the t-distribution with $n-1$ degrees of freedom.

Case 4: The population standard deviation, σ, is **unknown** and population is not normally distributed.

Here again, we must rely on the Central Limit Theorem. For large sample sizes, $\dfrac{\bar{x}-\mu}{\dfrac{s}{\sqrt{n}}}$ will approach a standard normal distribution.

The decision on which multiplier to use for this situation is somewhat ambiguous. For small sample sizes, you can't do anything without *assuming* that the population is normally distributed. Even if this assumption is not really correct, the t-distribution is likely to be approximately right. For large sample sizes, you have the Central Limit Theorem to ensure your assuption of normality. Regardless of whether you decide to use the t or z-multiplier, you are still using an approximation. If you use the z-mutliplier, you are assuming that the sample size is big enough that $\dfrac{\bar{x}-\mu}{\dfrac{s}{\sqrt{n}}}$ is well approximated by the standard normal distribution, i.e., that the Central Limit Theorem holds. If you use the t-mutliplier, you are assuming that the population can be well approximated by the normal distribution.

The likelihood of us knowing the real value of σ are slim to none, being that σ is a population parameter. Chances are, we will be dealing with Case 4. We are relying on an assumption for this case, regardless of the multiplier we choose to use. So which one should we choose?

Since the t-distribution carries more weight in it's tails than the normal distritbution, the t-multipliers are always a little bigger than the z-mutlipliers (for the same confidence level). For this reason, the confidence interval that is calculated using a t-multiplier will be slightly wider than the confidence interval calculated using the z-mutliplier. Using the t-mutliplier makes our confidence interval estimate more conservative. This is one reason why we choose to always stick to using the t-distribution in the calculation of confidence intervals for means.

12.8.2 How Does R Find df?

When you are dealing with a situation where you have sampled from two independent populations, degrees of freedom is more difficult to calculate. We can't just take $n-1$ because we have two sample sizes, n_1 and n_2. There are different methods for calculating df. One method is to use one less than the smaller of the two sample sizes for the degrees of freedom. In other words,

$$df = min(n_1 - 1, n_2 - 1).$$

If the standard deviations of the two samples are equal, another method is to use two less than the sum of the two sample sizes for the degrees of freedom. In other words,

$$df = n_1 + n_2 - 2.$$

In fact, by setting `var.equal=TRUE` in the `ttestGC` function, R will use this formula for df.

By default, the function `ttestGC` in R uses the Welch-Satterthwaite equation to calculate degrees of freedom for the 2 sample test of means.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^2}{n_2^2(n_2-1)}}$$

Research Question: Do GC males sleep more at night, on average, than GC females?

The `ttestGC` function gives us the following:

```
ttestGC(sleep~sex,data=m111survey)
```

```
##
##
## Inferential Procedures for the Difference of Two Means mu1-mu2:
##  (Welch's Approximation Used for Degrees of Freedom)
##   sleep grouped by sex
##
##
## Descriptive Results:
##
##   group  mean    sd  n
##  female 6.325 1.619 40
##    male 6.484 1.557 31
##
##
## Inferential Results:
##
## Estimate of mu1-mu2:  -0.1589
## SE(x1.bar - x2.bar):  0.3792
```

```
##
## 95% Confidence Interval for mu1-mu2:
##
##           lower.bound         upper.bound
##           -0.915971           0.598229
```

The $df = 65.81$. Let's use the Welch-Satterthwaite equation to verify this.

```
##    .group min Q1 median    Q3 max     mean       sd  n missing
## 1 female   2  5   6.75 7.125   9 6.325000 1.619394 40       0
## 2   male   4  5   7.00 7.000  10 6.483871 1.557155 31       0
```

The following statistics will be used in our calculation of df:

- s_1 = standard deviation of the females amount of sleep = 6.325.
- s_2 = standard deviation of the males amount of sleep = 6.483871.
- n_1 = sample size of females = 1.6193937.
- n_1 = sample size of females = 1.5571548.

$$\text{So, } df = \frac{\left(\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right)^2}{\dfrac{s_1^4}{n_1^2(n_1-1)}+\dfrac{s_2^2}{n_2^2(n_2-1)}} = \frac{\left(\dfrac{6.325^2}{1.6193937}+\dfrac{6.483871^2}{1.5571548}\right)^2}{\dfrac{6.325^4}{1.6193937^2(1.6193937-1)}+\dfrac{6.483871^4}{1.5571548^2(1.5571548-1)}} = 1.1654895$$

.