

---

# **APPLIED STATISTICS**

## **An Introduction to Statistical Analysis**

---

**Christopher Wetherill**  
Virginia Polytechnic Institute and State University

### Statement of Purpose

The laws of most jurisdictions throughout the world automatically confer exclusive Copyright and Related Rights (defined below) upon the creator and subsequent owner(s) (each and all, an “owner”) of an original work of authorship and/or a database (each, a “Work”).

Certain owners wish to permanently relinquish those rights to a Work for the purpose of contributing to a commons of creative, cultural and scientific works (“Commons”) that the public can reliably and without fear of later claims of infringement build upon, modify, incorporate in other works, reuse and redistribute as freely as possible in any form whatsoever and for any purposes, including without limitation commercial purposes. These owners may contribute to the Commons to promote the ideal of a free culture and the further production of creative, cultural and scientific works, or to gain reputation or greater distribution for their Work in part through the use and efforts of others.

For these and/or other purposes and motivations, and without any expectation of additional consideration or compensation, the person associating CC0 with a Work (the “Affirmer”), to the extent that he or she is an owner of Copyright and Related Rights in the Work, voluntarily elects to apply CC0 to the Work and publicly distribute the Work under its terms, with knowledge of his or her Copyright and Related Rights in the Work and the meaning and intended legal effect of CC0 on those rights.

1. Copyright and Related Rights. A Work made available under CC0 may be protected by copyright and related or neighboring rights (“Copyright and Related Rights”). Copyright and Related Rights include, but are not limited to, the following:
  - i. the right to reproduce, adapt, distribute, perform, display, communicate, and translate a Work;
  - ii. moral rights retained by the original author(s) and/or performer(s);
  - iii. publicity and privacy rights pertaining to a person’s image or likeness depicted in a Work;
  - iv. rights protecting against unfair competition in regards to a Work, subject to the limitations in paragraph 4(a), below;
  - v. rights protecting the extraction, dissemination, use and reuse of data in a Work;
  - vi. database rights (such as those arising under Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, and under any national implementation thereof, including any amended or successor version of such directive); and
  - vii. other similar, equivalent or corresponding rights throughout the world based on applicable law or treaty, and any national implementations thereof.
2. Waiver. To the greatest extent permitted by, but not in contravention of, applicable law, Affirmer hereby overtly, fully, permanently, irrevocably and unconditionally waives, abandons, and surrenders all of Affirmer’s Copyright and Related Rights and associated claims and causes of action, whether now known or unknown (including existing as well as future claims and causes of action), in the Work (i) in all territories worldwide, (ii) for the maximum duration provided by applicable law or treaty (including future time extensions), (iii) in any current or future medium and for any number of copies, and (iv) for any purpose whatsoever, including without limitation commercial, advertising or promotional purposes (the “Waiver”). Affirmer makes the Waiver for the benefit of each member of the public at large and to the detriment of Affirmer’s heirs and successors, fully intending that such Waiver shall not be subject to revocation, rescission, cancellation, termination, or any other legal or equitable action to disrupt the quiet enjoyment of the Work by the public as contemplated by Affirmer’s express Statement of Purpose.
3. Public License Fallback. Should any part of the Waiver for any reason be judged legally invalid or ineffective under applicable law, then the Waiver shall be preserved to the maximum extent permitted taking into account Affirmer’s express Statement of Purpose. In addition, to the extent the Waiver is so judged Affirmer hereby grants to each affected person a royalty-free, non transferable, non sublicensable, non exclusive, irrevocable and unconditional license to exercise Affirmer’s Copyright and Related Rights in the Work (i) in all territories worldwide, (ii) for the maximum duration provided by applicable law or treaty (including future time extensions), (iii) in any current or future medium and for any number of copies, and (iv) for any purpose whatsoever, including without limitation commercial, advertising or promotional purposes (the “License”). The License shall be deemed effective as of the date CC0 was applied by Affirmer to the Work. Should any part of the License for any reason be judged legally invalid or ineffective under applicable law, such partial invalidity or ineffectiveness shall not invalidate the remainder of the License, and in such case Affirmer hereby affirms that he or she will not (i) exercise any of his or her remaining Copyright and Related Rights in the Work or (ii) assert any associated claims and causes of action with respect to the Work, in either case contrary to Affirmer’s express Statement of Purpose.
4. Limitations and Disclaimers.
  - a. No trademark or patent rights held by Affirmer are waived, abandoned, surrendered, licensed or otherwise affected by this document.
  - b. Affirmer offers the Work as-is and makes no representations or warranties of any kind concerning the Work, express, implied, statutory or otherwise, including without limitation warranties of title, merchantability, fitness for a particular purpose, non infringement, or the absence of latent or other defects, accuracy, or the present or absence of errors, whether or not discoverable, all to the greatest extent permissible under applicable law.
  - c. Affirmer disclaims responsibility for clearing rights of other persons that may apply to the Work or any use thereof, including without limitation any person’s Copyright and Related Rights in the Work. Further, Affirmer disclaims responsibility for obtaining any necessary consents, permissions or other rights required for any use of the Work.
  - d. Affirmer understands and acknowledges that Creative Commons is not a party to this document and has no duty or obligation with respect to this CC0 or use of the Work.

For more information, please see [Creative Commons](#).

# CONTENTS

---

0.1	Introduction	v
0.2	To the Student	vi

## PART I INTRODUCTORY MATERIALS TO THE BOOK

<b>1</b>	<b>Introduction to R</b>	<b>3</b>
1.1	Getting Started	3
1.2	Basic R Syntax	5
1.3	Exercises	8
1.4	Additional Resources	9
<b>2</b>	<b>Measuring Uncertainty</b>	<b>11</b>
2.1	Measures of Central Tendency and Spread	11
2.2	Standard Deviation	14
2.3	Sampling Distributions	17
2.4	Visualizing Data	19
2.5	Exercises	22
2.6	Additional Resources	22
<b>3</b>	<b>Research Design</b>	<b>23</b>

3.1	Relating Design and Analysis	23
3.2	Designing Sound Experiments	25
<b>4</b>	<b>Introduction to Statistical Inference</b>	<b>29</b>
4.1	What Is Statistical Inference	29
4.2	Some Considerations about Statistical Inference	31
4.3	Recap: The Components of a Statistical Test	32

## PART II DESCRIBING RELATIONSHIPS AND PREDICTING VALUES

<b>5</b>	<b>Correlation</b>	<b>35</b>
5.1	Scatterplots	35
5.2	Correlation: Measuring Linear Association	35
5.3	Correlation Matrices (Multiple Correlation)	38
5.4	Partial Correlations	39
5.5	Some Considerations: Causality and Linearity	40
<b>6</b>	<b>Simple Linear Regression</b>	<b>47</b>
6.1	Overview of Simple Linear Regression	47
6.2	Cautions and Considerations	51
6.3	Case Study: [STUDY]	52
6.4	Exercises	52
6.5	Additional resources	52
<b>7</b>	<b>Multiple Regression</b>	<b>53</b>
<b>8</b>	<b>Logistic Regression</b>	<b>55</b>
<b>9</b>	<b>Model Refinement and Variable Selection</b>	<b>57</b>

## PART III COMPARISONS AMONG MULTIPLE SAMPLES

<b>10</b>	<b><math>t</math>-Tools</b>	<b>61</b>
<b>11</b>	<b>Alternatives to <math>t</math>-Tools</b>	<b>63</b>
<b>12</b>	<b>Multifactor Studies</b>	<b>65</b>
<b>13</b>	<b>Multifactor Studies</b>	<b>67</b>

<b>14</b>	<b>Multifactor Studies with Replication</b>	<b>69</b>
-----------	---	-----------

**PART IV INFERENCES FROM  
UNUSUAL DATA STRUCTURES**

<b>15</b>	<b>Time-Series Data and Serial Correlation</b>	<b>73</b>
-----------	--	-----------

<b>16</b>	<b>Rank-Ordered Data</b>	<b>75</b>
-----------	--------------------------	-----------

<b>17</b>	<b>Tables of Counts, Proportions, and Odds</b>	<b>77</b>
-----------	--	-----------

<b>18</b>	<b>Factor Analysis and Principle Component Analysis</b>	<b>79</b>
-----------	---	-----------

**PART V SUPPLEMENTAL MATERIALS**

<b>19</b>	<b>Choosing the Right Analysis: A Cheat Sheet</b>	<b>83</b>
-----------	---	-----------

<b>20</b>	<b>Getting Outside Help: How to Share Data</b>	<b>85</b>
-----------	--	-----------



# PREFACE

---

## 0.1 Introduction

### About the Author

Christopher Wetherill is currently a PhD student in Virginia Tech's [Translational Biology, Medicine, and Health](#) program. He received his undergraduate degree from John Carroll University. He has consulted with individuals and nonprofits on experimental design and statistical analyses since 2012. He recently edited [Data. Design. A simple introduction to preparing and visualizing information](#), a Creative Commons-licensed ebook detailing best practices for survey design and data collection and presentation.

### Structure of the Book

This book is split into four main sections: an introduction to statistical inference; regression tools; comparisons of means among several samples; and working with unusual data structures. By and large, no part or chapter requires any existing knowledge of any previous part of chapter. The one exception to this is the introductory material: if you are not comfortable with the basic ways in which we quantify uncertainty, then you will struggle with the rest of the material. If you do not have a firm grip of the basics of statistical inference, or if you haven't thought about this in a while, you should probably take an afternoon to refresh yourself.

Each chapter in this book follows the same general structure: we begin with an introduction to the topic and considerations of when and how a given test should be applied, what are its limitations, and what conditions it assumes are met. We then provide a sample case study, including all relevant background information and a complete analysis and write-up of the data. Finally, we include a summary of how to

implement the test using the statistical software R and conclude with a number of comprehension exercises and data sets to be analyzed.

## 0.2 To the Student

### What You Will Learn

You shouldn't expect this book to make you an experienced statistician. That isn't what it's meant to do. Rather, this book aims to explain statistical analyses and their output in a way that is easily-digested by individuals without a strong mathematical or statistical background or who have previously had bad experiences with statistics and have shied away from the subject. The hope is that, after going through this book, you will be able to (1) understand when to apply a number of the most commonly-used statistical tests; (2) accurately construct and run the tests; and (3) easily interpret and report the output of these analyses to both technical and non-technical audiences.

There may still be a point at which you do need to consult a professional statistician: in fact, this book tries to emphasize recognizing your limits and when it may be appropriate to seek outside help. When this is the case, you should still be better-equipped to present your data and whatever analyses have already been run, and to explain the central problems that you are trying to solve easily and effectively.

### Case Studies and Exercises

Each statistical analysis presented is accompanied by one or more case studies in which we perform a complete analysis of a given data set, highlight the relevant output and how to interpret it, and give a full APA-style write-up of the analysis and results. All data used are real (unless otherwise specifically noted), and where possible we link to the original study or other supporting materials. The full data used in the analyses will always be provided so that the reader is able to perform the analysis alongside us and can confirm his or her results.

Likewise, all exercises are taken from real-world problems and all original supporting materials that can be provided are. For data analysis problems, we do not provide a full stepwise write-up of the analysis: only a summary of the results that should be obtained. If your results do not agree with our answers, you are always welcome to open an [issue on GitHub](#). If you do, please properly format any R code that you choose to include: [GitHub-flavored Markdown](#) makes this easy to do.

### Mathematical Background

This book doesn't assume that the reader has any strong mathematical background: the emphasis of this book is on identifying, running, and interpreting appropriate statistical analyses for the data that you are working with, and not on the statistical theory that informs these tests. However, to have a good understanding of these analyses, the reader will occasionally need to learn and understand a few mathematical formulas. When these are presented, they are always accompanied with a non-technical explanation.

### Recommended Software

All analyses presented in this book are implemented by the author using [The R Project for Statistical Computing](#) and [R Studio](#), a graphical user interface for R. We chose R because, unlike SAS, SPSS, STATA, MiniTab, etc., it is entirely open-source, meaning that it is free for anyone to use and even contribute back to. Further, analyses run in R are easily reproducible: that is, it is very easy to take the



code that we use to run an analysis in a case study, paste it into R, and get the exact same results that we do.

We understand that different statistical packages have different interfaces and very different ways of presenting analysis results. And although we only document R in this book, we do our best to give you links to additional resources that explain how to run and interpret the analysis using other popular commercial stats packages. If you're ever desperate to run an analysis in another statistical package and we don't provide links to help you, you're welcome to open an [issue on GitHub](#), although we can't promise that we will have a good answer for you.

## Common Statistical Terms

<b>Statistic</b>	Any quantity that can be calculated from a set of data.
<b>Population</b>	The collection of all possible measurements taken. For example, “the heights of every 12-year-old boy in America” or “all automobiles with three wheels.”
<b>Sample</b>	A subset of a population. For example, “the heights of 100 12-year-old boys from Ohio” or “20 tricycles bought from Toys-R-Us.”
<b>Parameter</b>	A statistic referring to a population. This is usually denoted by a Greek letter.
<b>Estimate</b>	A statistic referring to a sample. This is a “best guess” at the parameter for the population. Estimates are denoted by either a hat or a Roman letter.
<b>Mean</b>	The arithmetic mean of a set of data. When referring to the parameter, we denote this $\mu$ ; when referring to the estimate, we denote this $\bar{x}$ (pronounced “x-bar”) or $\hat{\mu}$ (pronounced “mu-hat”).
<b>Standard Deviation</b>	A measure of how far dispersed observations are from their mean. Denoted $\sigma$ when referring to the parameter or $s$ for the estimate.
<b>Standard Error</b>	The standard deviation of the sampling distribution of a statistic. Denoted s.e. or SE.

## Additional Resources

Additional resources are provided at the end of every chapter. These will usually include:

1. A link to the answer key for the chapter's exercises;
2. Links to resources for performing the chapter's analyses in other statistical packages;
3. Our references for writing the chapter and further readings if you would like a more advanced knowledge of the analyses covered;
4. A link to all data sets and other supporting materials used and referenced in the chapter and end-of-chapter exercises;
5. And any other miscellaneous material that we think might be helpful to you.



## PART I

---

# INTRODUCTORY MATERIALS TO THE BOOK

---



# CHAPTER 1

---

## INTRODUCTION TO R

---

### 1.1 Getting Started

#### Installing R and R Studio

Although this software isn't required, we do recommend that you use R. It's what we will be using to carry out analyses in all of our examples and what we will use to analyze data in the end-of-chapter exercises. Although there won't be any major differences using other statistical software, occasionally you may obtain different results than we do.

To install R, select any of the [available download mirrors](#) from its website. Select the version that is appropriate for your operating system—Windows, Mac OS, or a Linux distribution. You may also wish [to install RStudio](#). You will be fine if you don't install it; however, it pretties up R's graphical interface and makes working with data a bit easier for newcomers to the software. To use RStudio, you must have R installed on your computer already: installing RStudio does not also install R.

#### Reading in Data

There are two strategies used for reading in data: the command line or RStudio's graphical interface. To load data via the command line, we will type a command such as:

```
# Here, myData is what we will name our data once we load it.
# We use the function read.csv to tell R to load our data
# from the path specified. If you aren't sure where your
# working directory is, type the function getwd()
# and hit Enter.
```

```
myData <- read.csv("/Path/to/File.csv", headers=TRUE)
```

Alternately, in RStudio's Environment tab, we can click the `Import Dataset` button and choose a local file or web URL to read our data from.

Whichever option you choose, be sure to use a .txt, .dat, .csv, or similar file: R can't read many proprietary file formats (E.g., Excel's .xlsx; SPSS .sav; etc.). Usually, whatever program you use to manage your data will have an option to allow you to save to CSV that you can access through the `File -> Save As` menu.

## Creating a Script

Working in R, there are two ways that you can execute commands: through the console and through a script. The console is what first opens when you start R. It will usually contain some text like:

```
R version 3.1.0 (2014-04-10) -- "Spring Dance"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.1.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
```

When you're working in the console, all you have to do is type in your command and hit Enter; the code will be automatically evaluated. Alternately, however, you can use a script by clicking `File -> New File -> R Script`. This allows you to write multiple lines of R code at once and then selectively run them by holding `Ctrl/Cmd + Enter` (depending on whether you are using a Mac or PC keyboard). By default, this will execute whatever line is currently selected by your pointer; however, you can also highlight multiple lines, hit `Ctrl/Cmd + Enter`, and evaluate everything highlighted at once.

We will generally recommend using an R script as this can be saved and referenced later. It also makes it easier to go back and rerun old code or to reference old work. (Opposed to the console, where you have a much more limited history.)

## 1.2 Basic R Syntax

### Constants

Constants are just what they sound like: things. They don't do anything; they can't take in one number and spit out another; they're just static objects. This can mean a number or a string of text, or something more complex like a vector, matrix, or data frame. So let's first define the different data types that R supports. (NOTE: throughout the rest of the book, we will use the term data type rather than constant.)

**Vectors** A vector, in this case, you can think of as being a column of data in Excel. It's just a series of items that are all stored in one object. For instance, let's say that we want a column of numbers from 1 through 10:

```
# NOTE: We use the notation OBJECT <- DATA
# to tell R to assign data to a variable.
# This way, we can type in, for example,
# myNumbers and R will know what data
# we're talking about.

myNumbers <- c(seq(from=1, to=10, by=1))
moreNumbers <- c(1,2,3,4,5,6,7,8,9,10)
```

In both of the above examples, we create a column of data with the numbers 1 through 10. But these don't have to only be numeric. For instance:

```
# a character vector:
characters <- c("one", "two", "three", "four", "five")

# a boolean (logical) vector:
boolean <- c(TRUE, FALSE, FALSE, FALSE, TRUE)
```

We can see that there are three basic types of vectors you can construct: numeric; character; and logical. Numeric vectors are composed of arabic numerals; character vectors of strings of text; and logical vectors of TRUE/FALSE booleans.

**Matrices** You can think of a matrix in R as either (1) a single vector split up into multiple rows/columns or (2) multiple vectors of the same length piled up next to one another. It's roughly analogous to a spreadsheet. However, the one caveat is that, just like with vectors, all elements in a matrix must have the same type (i.e., all numeric, all character, or all logical). There's no mixing and matching here. The basic syntax for creating a matrix is:

```
myMatrix <- matrix(vector, nrow=r, ncol=c, byrow=FALSE,
  dimnames=list(c("rownames"), c("colnames")))
```

Here, `vector` is a vector of elements that you would like to divide into rows and columns; `nrow` and `ncol` respectively refer to the number of rows and columns in your matrix; `byrow` takes a TRUE/FALSE boolean and indicates whether the matrix should be filled in by rows (TRUE) or by columns (FALSE). Finally, `dimnames` is an optional argument that specifies the column and row names.

As an example, let's create a 4x3 matrix:

```
matrixVector <- c(21, 5.9, 147,
                  28, 5.7, 168,
                  18, 5.5, 126,
                  24, 6.1, 195)
myMatrix <- matrix(matrixVector, nrow=4, ncol=3, byrow=TRUE,
                   dimnames=list(c("Chris", "John", "Amy", "Max"),
                                c("Age", "Height", "Weight")))
```

This will result in a matrix that looks like this:

	Age	Height	Weight
Chris	21	5.9	147
John	28	5.7	168
Amy	18	5.5	126
Max	24	6.1	195

**Data Frames** A data frame in R is a generalized instance of a matrix: this you can truly think of as a page of an Excel spreadsheet. Each column represents a vector of a single type; however, each column can be of a different data type. So, for instance, if we wanted to turn the matrix above into a data frame but also add in a column for hair color, we would do something like:

```
name <- c("Chris", "John", "Amy", "Max")
age <- c(21, 28, 18, 24)
height <- c(5.9, 5.7, 5.5, 6.1)
weight <- c(147, 168, 126, 195)
hair <- c("brown", "blonde", "red", "brown")

myData <- data.frame(name, age, height, weight, hair)
colnames(myData) <- c("name", "age", "height", "weight", "hair")
```

Just like the matrix, this will give us the following:



Name	Age	Height	Weight	Hair
Chris	21	5.9	147	brown
John	28	5.7	168	blonde
Amy	18	5.5	126	red
Max	24	6.1	195	brown

## Factors

Oftentimes in experiments we will have factors: nominal variables that indicate levels. For example, we might have a placebo group and an experimental group. In this case, these are important variables for R to properly interpret; however, they are non-numeric and don't make sense to interpret as a string. In these cases, we can indicate to R that they should be interpreted as factors. For example:

```
# The variable "drugCondition" contains 20 experimental trials
# and 20 control trials. Currently, these values are stored
# as characters. You can check this by running str(drugCondition)
# You should see "chr" indicating the elements are characters

drugCondition <- c(rep("experimental",20), rep("control",20))

# We will now convert the elements to factors.
# If you run str(drugCondition) again, you should see:
# Factor w/ 2 levels "control","experimental"

drugCondition <- factor(drugCondition)
```

## Functions

In addition to static data types, detailed above, R also uses functions. These take the form `functionName(argument1, argument2)` (although there can be any number of arguments passed to the function—not necessarily only 2!). Functions are chunks of code that *do* something: they may generate a plot or run a statistical analysis or find the mean of a set of data.

Let's say, for example, that we have a numeric vector and we want to find its mean value. How do we know what function does that? Well, we can start by calling the `apropos()` function: this lets us search R's entire library of functions for those that contain some text that we're searching for. So in this case:

```
apropos("mean")
[1] ".colMeans"      ".rowMeans"      "colMeans"      "kmeans"
[5] "mean"           "mean.Date"      "mean.default"  "mean.difftime"
[9] "mean.POSIXct"   "mean.POSIXlt"   "rowMeans"      "weighted.mean"
```

We can see that there are 12 functions R knows of that contain the word “mean.” Specifically, there does seem to be a `mean()` function! That sounds promising. Next, let’s figure out how we can use it properly. To do this, all we ever have to do is type a question mark (?) followed by the function name into the R console and R will return the documentation for that function. So here:

```
?mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

It looks like this function only has one required argument: `x`. So if we want to find the mean of a vector of data called `participantAges`, we would run:

```
# We will also specify na.rm=TRUE
# This way, if there is anyone who didn't
# indicate their age, R won't count them
# when it calculates the mean age

mean(participantAges, na.rm=TRUE)
[1] 21.7      # The mean age is 21.7 years old
```

## 1.3 Exercises

**1.1** Download a data file from the web. To do this, you’ll want to use the code:

```
download.file(
  "http://book.chriswetherill.me/part1/data/introToREx01.csv",
  "sampleData.csv", "wget", extra="--no-check-certificate")
```

Paste that into the R console and hit enter. If it’s successful, you should see something similar to 2014-05-30 20:12:15 (172 KB/s) - ‘sampleData.csv’ saved [8528/8528].

**1.2** Now we will assign the data to an object. This makes it easier to work with once we start plotting it and running analyses on the data. To do this, run the command:

`myData <- read.csv("sampleData.csv", header=TRUE, sep=", ")`. To see if you did this correctly, run the command `head(myData)`. If you don't get an error, then you're good to go.

**1.3** Finally, we will load a file from the desktop. [Click here to download the data](#). Save it to your desktop. Now, in RStudio, click Import Dataset, select the file, make sure that you choose Yes for the Heading option, and import it.

**1.4** Next we will practice working with different data structures in R. Run `head(myData)`; describe what the function does. What about `colnames(myData)`?

**1.5** Make a vector called `myVector` containing the numbers 20 through 30.

**1.6** Now make a matrix with two rows and five columns called `myMatrix` that uses the vector you made in the previous problem.

**1.7** Run the following code:

```
sex <- c(rep("male", 5), rep("female", 5))
age <- c(sample(1:100, 10, replace=TRUE))
weight <- c(sample(95:200, 10, replace=TRUE))
```

And now combine the three vectors above with `myVector` to create a data frame called `myDataFrame`.

**1.8** Almost there! Finally, convert the column `sex` into a factor. (HINT: to reference a single column of a data frame, try `myDataFrame$sex`.)

## 1.4 Additional Resources

For a more complete tutorial on using R, we would recommend:

1. [Intro to R](#) by Google Developers (YouTube)
2. [Data Science](#) by Johns Hopkins University (Coursera)
3. [An Introduction to R](#) by The R Project (CRAN)
4. [R Resources](#) by IDRE (UCLA)



## CHAPTER 2

---

# MEASURING UNCERTAINTY

---

## 2.1 Measures of Central Tendency and Spread

### Mean, Median, and Mode

Mean, median, and mode represent the three common ways of describing important features of a data set at a glance. The **arithmetic mean** of a sample,  $x_1, x_2, \dots, x_n$ , is denoted by  $\bar{x}$  (pronounced “x-bar”). For a sample of  $n$  observations, it is calculated:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Alternately, it may be represented as:

$$\bar{x} = \frac{\sum_i^n x_i}{n} \quad (2.1)$$

This means we add up every item in a set of  $n$  items and then divide by however many items we added together. Here, the capital sigma ( $\Sigma$ ) indicates that we take a sum (that is, we add up all the values). This is the most commonly-used measure of central tendency, although it should never be used when dealing with ordinal or categorical data. (That is, ranked data—such as “rate these from best to worst”—or categories—such as male/female or religion. Although technically possible, the result will be difficult to interpret, if not impossible.)

The **median** of a set of data is, quite literally, the middle value. If we take a series of numbers, order them from smallest to largest, and pick the number that’s right in the middle (think of a see-saw balancing

at its fulcrum), then we have the median. This can be calculated in many of the same cases as the mean and is more resistant to outliers (values that are unusually far away from most other data points) than is the mean.

Finally, the **mode** is the most frequently-occurring number in a series of data. Usually, we will compute the mode when working with categorical data: for instance, it could tell us that the majority of survey respondents were female or were Jewish or lived in the Midwest.

If we want to calculate these in R, we would do something like this:

```
# First we will generate 200 random values
# with mean 0 and standard deviation 1

> set.seed(0)
> data <- rnorm(200, mean=0, sd=1)

# Now we will find the mean:

> mean(data)
[1] -0.01144155

# We can see that the mean of -0.011 is
# close to the value (0) we specified above.
# Next let's calculate the median

> median(data)
[1] -0.1061751

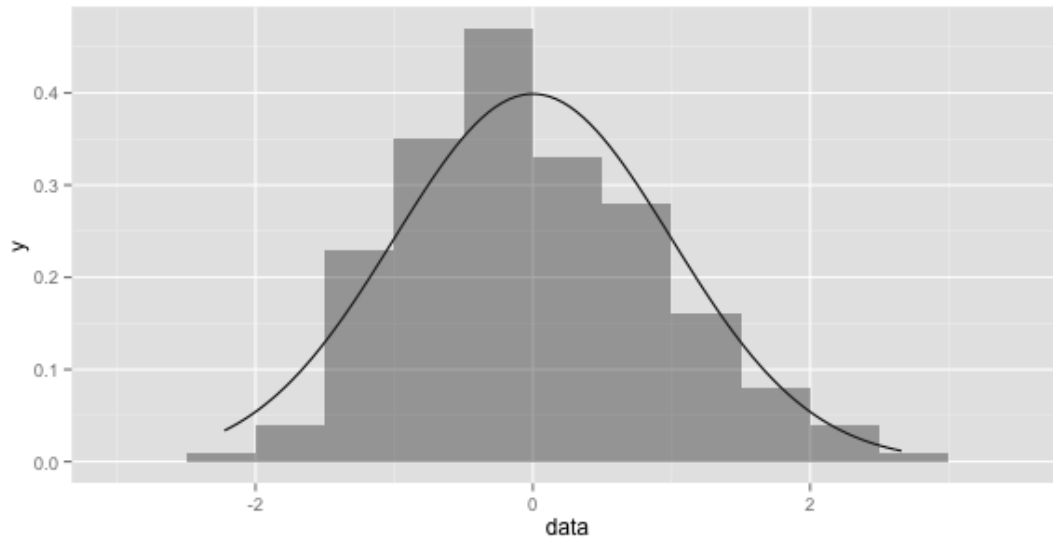
# We see that the median differs from the mean!
# That means that our data are a little skewed
# We'll talk about skewness in the next section.
```

## Skewness and Mean vs. Median

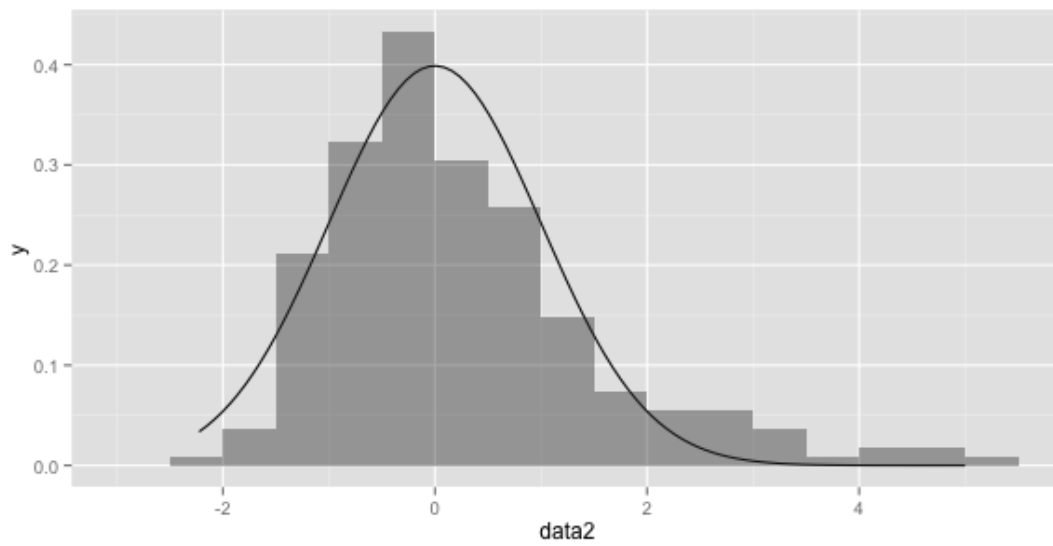
To get an idea of skewness, let's make a histogram of those data that we just generated. If you aren't familiar with a histogram, it takes your data and counts how many data points fall within a certain range, giving you something like Figure 2.1.

The smooth line that you see there is the **normal distribution**; the columns represent our actual observations. As you can see, the two don't line up perfectly. Rather, our graph is stretched out to the right a bit. We call this stretching **skewness**: the idea that, rather than being perfectly symmetrical about the mean, our data are stretched out to the right or to the left. In cases where our data are stretched out to the right, we say that they are skewed right; when they are stretched to the left, we say they are skewed left.

Many times, as in this example, there isn't very much skewness and our data, even though they aren't perfect, are still fairly symmetrical. However, sometimes we will have strong outliers that really mess up that symmetry. For instance, take Figure 2.2.



**Figure 2.1** Histogram of approximately normal data with a slight rightward skew



**Figure 2.2** Histogram of data heavily skewed right

There, we have a strong and obvious rightward skew. Now, let's go ahead and compare the means and medians of our two datasets:

```
# First we'll compare the means
> mean(data)
[1] -0.01144155

> mean(data2)
[1] 0.2489018

# And now the medians
> median(data)
[1] -0.1061751

> median(data2)
[1] -0.005767173
```

As you can see, both the mean and the median change. However, the median changes much less than the mean does. (There's a change of 0.26 in the means versus a change of 0.10 in the medians.) In cases where you have strongly skewed data, it will often be better to describe them using the median rather than the mean: specifically, the median is what we call **resistant to outliers**. In other words, when you have a few outliers (numbers that are far away from every other data point), the median will be changed much less than the mean will.

## 2.2 Standard Deviation

**Standard deviation** (represented as  $s$  or  $\sigma$ ) is a measure of dispersion: that is, it tells us how far from or close to the mean our data tend to be. A data set with a small standard deviation, for instance, tells us that most of our data points are fairly close together and are all near the mean. Alternately, a large standard deviation means that our data points are much more spread out. We can visualize this using three data sets, all with mean  $\mu = 0$  but with different standard deviations, as in Figure 2.3.

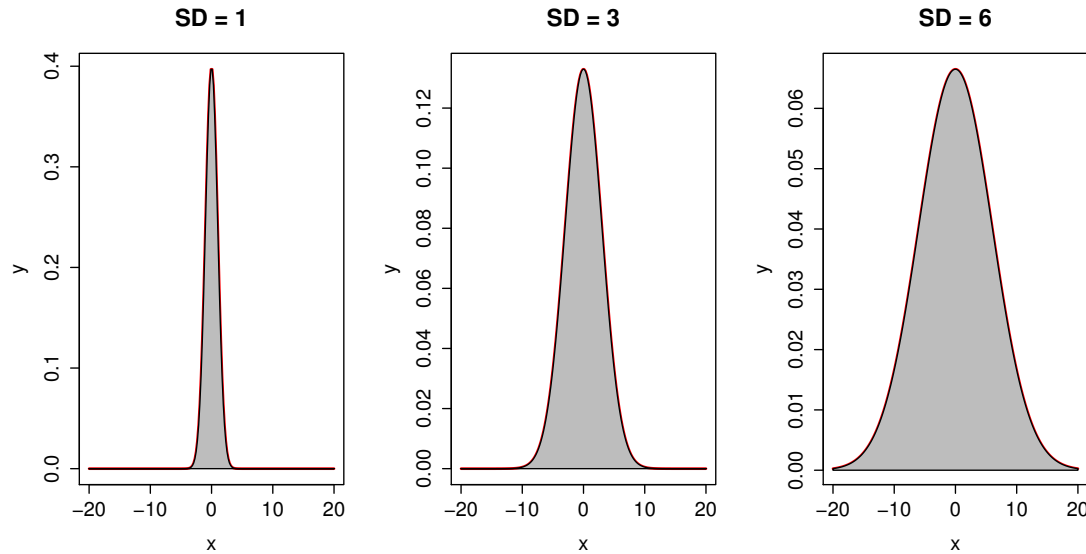
Even though each of these three sets of data has the same mean (0), it's pretty clear that there are some big differences between each of them. This is why standard deviation is important to know and report: without it, you aren't going to be getting a complete picture of what your data look like. They could all be close to the mean, as in the graph on the far left, or they could be much more spread out, as in the graph on the right.

Now, before we can use standard deviations, we have to calculate them. The formula for standard deviation looks like:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}, \text{ where } \mu = \frac{1}{n} \sum_{i=1}^n (x_i) \quad (2.2)$$

Let's unpack that: it's saying that we start by finding the mean ( $\mu$ ) of the data. Next, we'll go ahead and take every data point and subtract the mean from it (the  $x_i - \mu$  part of the equation). So what we're doing is basically finding out how many units away from the mean each data point is. But there's a problem: some data points might give us a negative number, others a positive number. So we'll take that difference and raise it to the second power (remember, a negative number times a negative number is always equal to a positive number). Now we just repeat that for every other data point in our sample and add them all together.





**Figure 2.3** Three samples of data with mean  $\mu = 0$  and standard deviations of 1, 3, and 6, respectively.

So now we have summed up all of those squared differences, right? Next, we will divide by the number ( $n$ ) of observations (just like when we calculated the mean) to get the **average distance from the mean**. But there's one last step before we're done: since we raised everything to the second power a couple steps ago, we have to undo that operation. To do this, we'll take the square root of everything, leaving us at last with the standard deviation.

## Variance

In many ways, variance and standard deviation are two sides of the same coin. If you remember, standard deviation is calculated:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}, \text{ where } \mu = \frac{1}{n} \sum_{i=1}^n (x_i)$$

Variance, then, is:

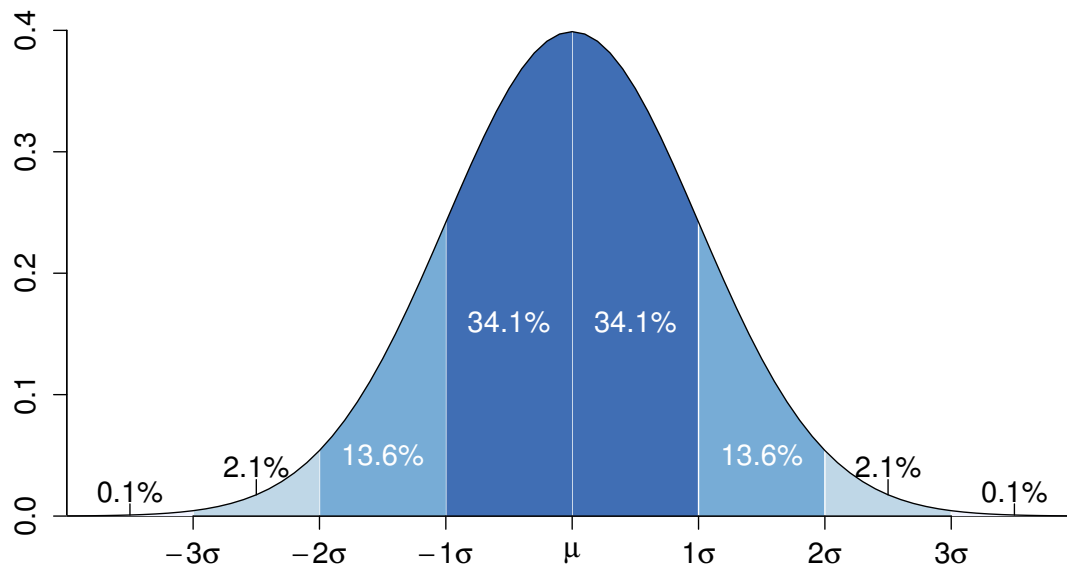
$$\begin{aligned} \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \\ &= \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \mu^2, \text{ where } \mu = \frac{1}{n} \sum_{i=1}^n (x_i) \end{aligned}$$

So, we can see that population variance is calculated exactly the same as is standard deviation, except we never take a square root at the end of it all. But if they're calculated so similarly and they both measure essentially the same thing, why would we use one over the other and why do we have both?

One big reason for using standard deviation instead of variance, at least in applied statistics, is that the standard deviation is in the same units as your data. (E.g., if you're looking at weights, the standard

deviation is in pounds.) Variance, on the other hand, is in squared units. (So instead of pounds, you'd have a variance in pounds squared. That's less easy to interpret, isn't it?) Beyond this, standard deviation also gives us the 68-97-99.5% rule.

This rule tells us that 68% of all observations in a normal distribution will fall within 1 standard deviation of the mean; 97% will fall within 2 standard deviations of the mean; and 99.5% will fall within 3 standard deviations of the mean. We can illustrate this in Figure 2.4:



**Figure 2.4** The 68-97-99.5% rule states that nearly all observed data points will be within 3 standard deviations of the mean. This is often useful for flagging outliers or inappropriate distributions to fit your data to. Source: [Wikipedia](#)

So, for instance, if we wanted to find what proportion of our data would be more than 1 standard deviation above the mean, we would add 13.6%, 2.1%, and 0.1% to end up with 15.8% of our data being more than 1 standard deviation above the mean.

### Remember:

**Standard deviation** is the average distance of a data point from the mean of your data. **Variance** is the squared average distance of a data point from the mean. Standard deviation is always in the same units as your data; variance is in units<sup>2</sup>.

### Interquartile Range

The last measure of spread that we will discuss is the interquartile range. This is most commonly used when the median is used rather than the mean. This number is based off of what is known as the “5-number summary,” defined as:

<b>Minimum</b>	The smallest value in your data set
<b>Q1</b>	The value in your data set that is larger than 25% of all other values
<b>Median</b>	The median value in your data set

<b>Q3</b>	The value in your data set that is larger than 75% of all other values
<b>Maximum</b>	The largest value in your data set

You'll quickly recognize this information if you're familiar with boxplots (if not, we cover them below). To derive the interquartile range (IQR) from these, we simply take  $Q3 - Q1$ . From the definitions above, we know that exactly 50% of our data points lie between these two values. Given this, the IQR can be said to be another measure of the dispersion of our data: instead of a 65-97-99.5% rule, here we have a 50-25-25% rule: 50% of our observations lie within the range given by the IQR; 25% lie above it; and 25% lie below it. As with standard deviation, a smaller IQR indicates a tighter spread of data (i.e., more of the data points are squished together) and a larger IQR indicates more variance of the data.

Finally, the IQR is often used to identify outliers in a data set. To do this, we will usually find our upper and lower cutoffs by:

$$\begin{aligned}\text{Upper} &= Q3 + 1.5 \cdot IQR \\ \text{Lower} &= Q1 - 1.5 \cdot IQR\end{aligned}$$

That is, any data point more than 1.5 times the IQR above  $Q3$  or more than 1.5 times the IQR below  $Q1$  can be considered suspect and possibly an outlier. Important to note, however, is that this isn't a hard-and-fast rule: there is no real reason that we use 1.5 times the IQR as a cutoff. (Actually, the statistician who came up with this measure chose 1.5 times because 1 seemed too small and 2 seemed too large.)

## 2.3 Sampling Distributions

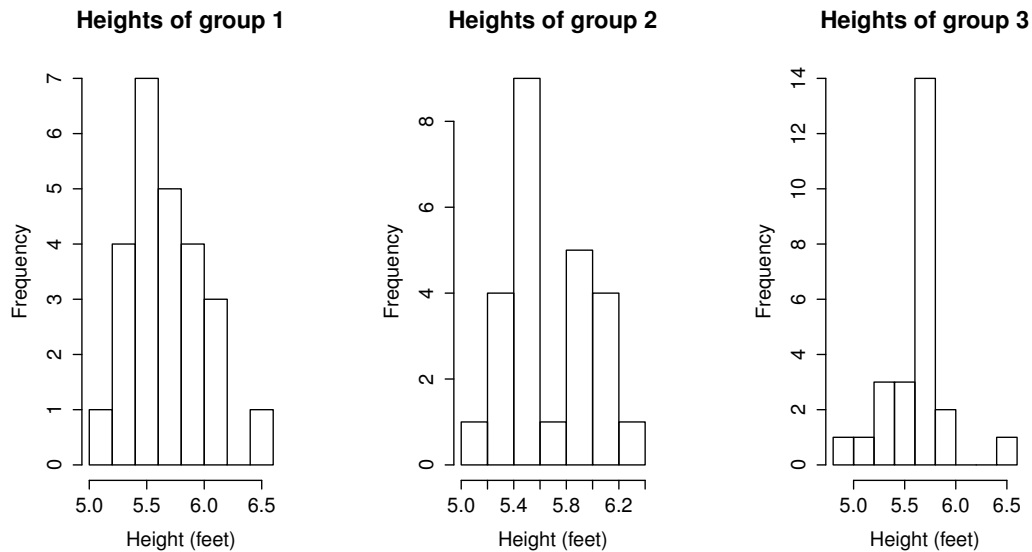
### Central Limit Theorem and Sampling Distributions

Does the central limit theorem ring any bells? Probably not. But it's still probably something that you've heard of before. Specifically, you may have heard of it in the context of **regression to the mean**. This is the idea that (let's use height as an example) if you randomly choose one adult male in the U.S., his height could very easily be 7'3". That's obviously a lot taller than most people. But the next person you choose will probably have a height less than the first guy's. And if you keep choosing more and more people and measuring their heights, pretty soon, the average height of everyone you chose will be pretty darned close to the actual national average height for adult males. More generally, the idea is that every now and then there will always be extreme scores. But every time you get one of those strong outliers, the next observation is almost certainly going to be less extreme and it is going to pull things back down to the population average.

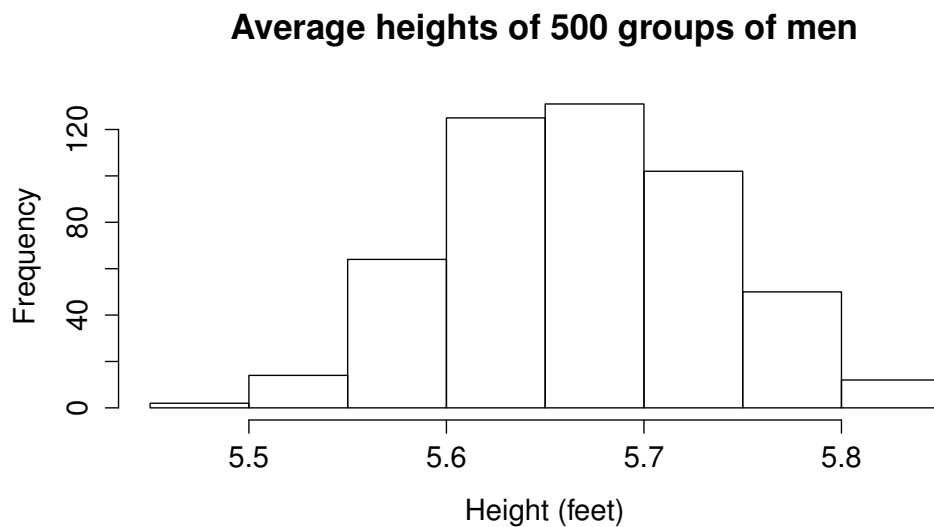
The **central limit theorem** is sort of similar: it states that the distribution of an average tends to be normal, even if the population is not itself normal.

Make sense? Maybe not (that's okay!), so let's break it down a bit. Whenever we're talking about a distribution of averages, we call that a **sampling distribution**. To continue with our height example from above, let's say that we randomly choose 25 guys and measure their heights. Then we choose another 25 guys and measure their heights. Then we choose another 25 guys and measure their heights. Assuming that the mean male adult height is 5'8" with a 4" standard deviation (we have to represent that with decimals in R, so it becomes a mean of 5.67 feet and a standard deviation of 0.33 feet), our three samples might look like Figure 2.5.

We can see that there's a lot of variation among our three groups, right? And groups 2 and 3 don't really look all that normal. But, let's say that we take the average male height of each of our three samples. Now, let's repeat that with another 497 batches of 25 men each. If we make a histogram of the average height from each of those 500 groups, it will look like Figure 2.6.



**Figure 2.5** Heights of three randomly-selected groups of men



**Figure 2.6** Average height of 500 groups of randomly selected men

All of a sudden, we have a normal distribution with a mean height of 5.67 feet (or 5'8") and a standard deviation of 0.068 feet (or 0.81 inches). We see that the mean of our sampling distribution is exactly equal to the mean of our population: that's the central limit theorem in action.

## Standard Error

You might have noticed that the standard deviation of the sampling distribution above was much smaller than the standard deviation of the actual population (0.81 inches versus 4 inches). This arises because we're looking at the variation of a set of means (which themselves are the average value of a larger set of data): these will always vary less than the individual data points that we sample from. And actually, when we're talking about the variation of sample means, instead of standard deviation, we will want to use what's called the **standard error**, a specific term referring to the standard deviation of a sampling distribution. This will always be smaller than the standard deviation of the population and can be estimated by:

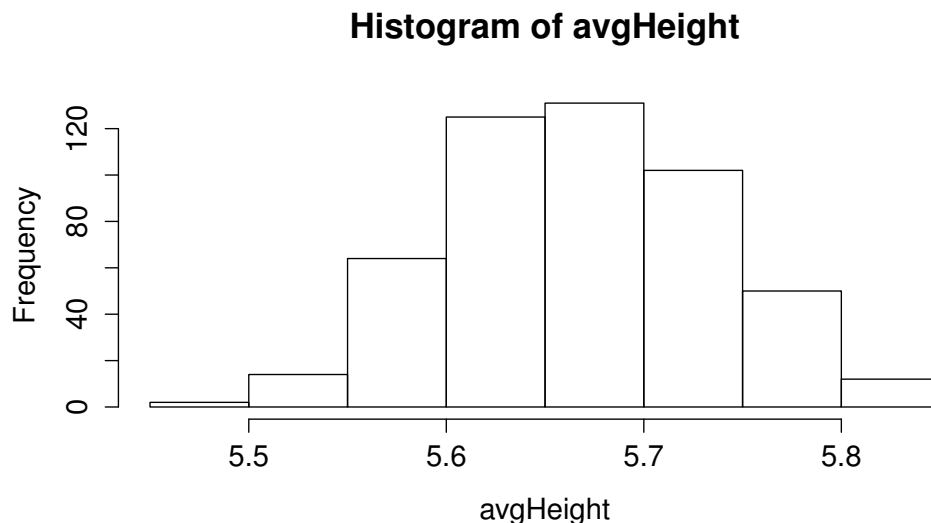
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \quad (2.3)$$

Here,  $s$  is the sample standard deviation (here, 0.81 inches) and  $n$  is the number of observations (that is, the number of samples—here, 500). As you can see, the standard error is always going to be smaller than the population's actual standard deviation. In short, the **standard error** is a measure of how far a sample mean is likely to be from the population mean; the **standard deviation** is a measure of how far an individual observation is likely to be from the sample mean.

## 2.4 Visualizing Data

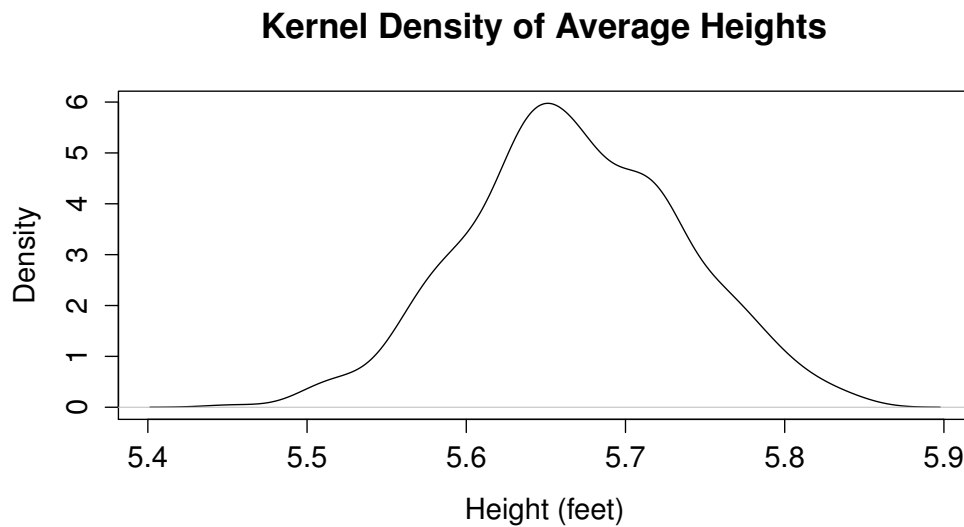
### Histograms

A **histogram** displays the number of observations that fall within a given range. For instance, in the height examples above, we would count the number of men in our sample between 5 feet and 5.2 feet and display them all in one column. Then we'd do the same for 5.2 to 5.4 feet, for 5.4 to 5.6 feet, and so on. This gives us something like:

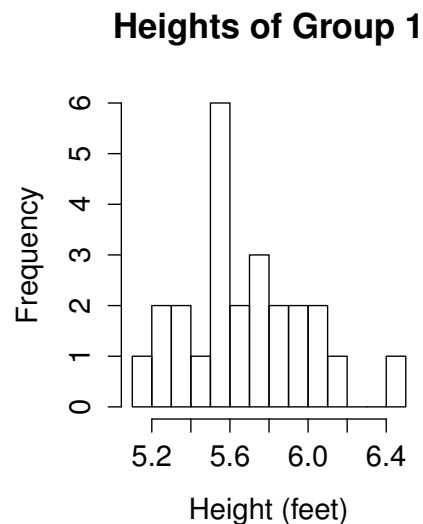
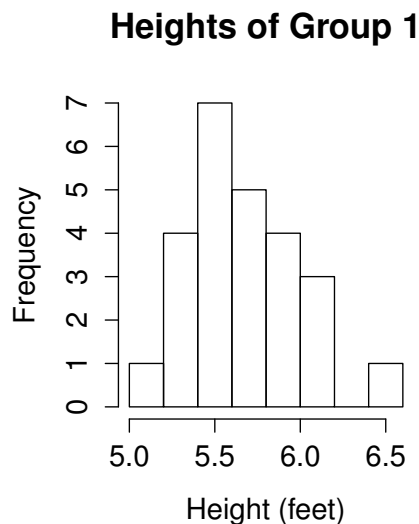


### Kernel Density Plots

A kernel density plot is similar to a histogram in that it shows us about what proportion of our sample is likely to be at any certain value. The big difference between kernel density plots and histograms is that density plots are continuous: whereas histograms take all of your data and lump them together in ranges, a density plot displays these data continuously. For instance, taking the same data that we used in our histogram example:

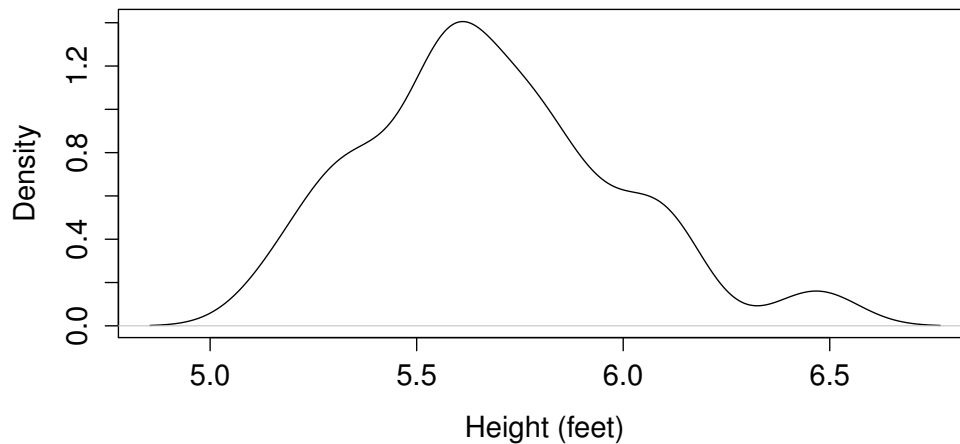


Comparing the two graphs, they tell about the same story. However, the kernel density plot isn't subject to issues of binning. Specifically, when using histograms, the width of the bins that you use (or the range of each column) can drastically impact your interpretation of the data. For instance, if we increase the number of bins, we get something like:



As the range of each bin varies, we can go from a nice-looking normal distribution to a much worse looking non-normal distribution. But with a kernel density plot, you remove those issues in interpretation:

### Kernel Density of Heights of Group 1

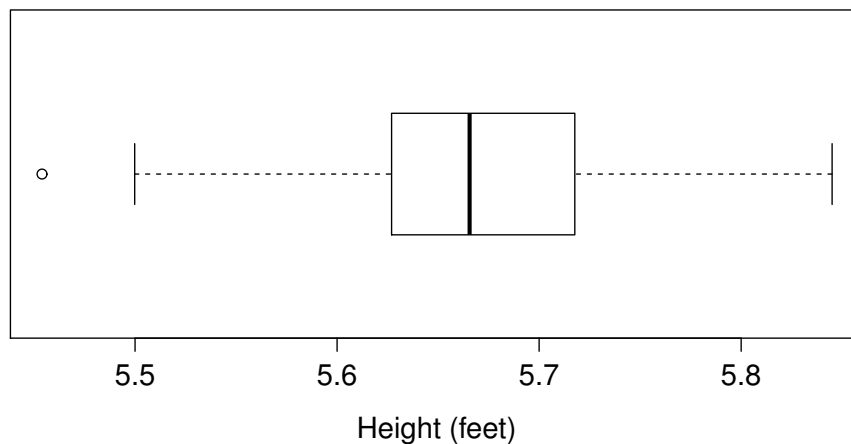


As we can see, in reality, it actually is a roughly normal distribution with some funky skewness over to the right.

### Boxplots

As discussed previously, boxplots are good to use when you're working with the median, interquartile range, and 5-number summary. These take that 5-number summary and present the data visually:

### Boxplot of Average Male Heights



Here, the line on the far left is the minimum value, the left edge of the box is Q1, the line bisecting the box is the median, the right edge of the box is Q3, and the line to the far right is the maximum value. Any circles indicate the presence of outliers based on the IQR.

## 2.5 Exercises

**2.1** In R, execute the following code:

```
set.seed(0)
height1 <- rnorm(25, 5.6667, 1/3)

avgHeight <- NULL

for(i in 1:500) {
  height <- rnorm(25, 5.6667, 1/3)
  avgHeight <- rbind(avgHeight, mean(height))
}
```

**2.2** What are the mean and median values of `height1` and `avgHeight`? What are their standard deviations?

**2.3** What would we expect the standard error of `avgHeight` to be if we had 500 samples and a population standard deviation of 0.33 feet?

**2.4** Prepare a histogram of `height1` and `avgHeight`. Do they differ? Does either look problematic? (E.g., departures from normality)

**2.5** Make a kernel density plot of `height1`. Does it differ from the histogram? Which gives a better picture of the data?

**2.6** Make a boxplot of `avgHeight`. Are there any outliers present?

## 2.6 Additional Resources

1. [R script](#) containing all commands used in examples throughout this chapter
2. [Answer key](#) for the chapter's exercises



## CHAPTER 3

---

# RESEARCH DESIGN

---

### 3.1 Relating Design and Analysis

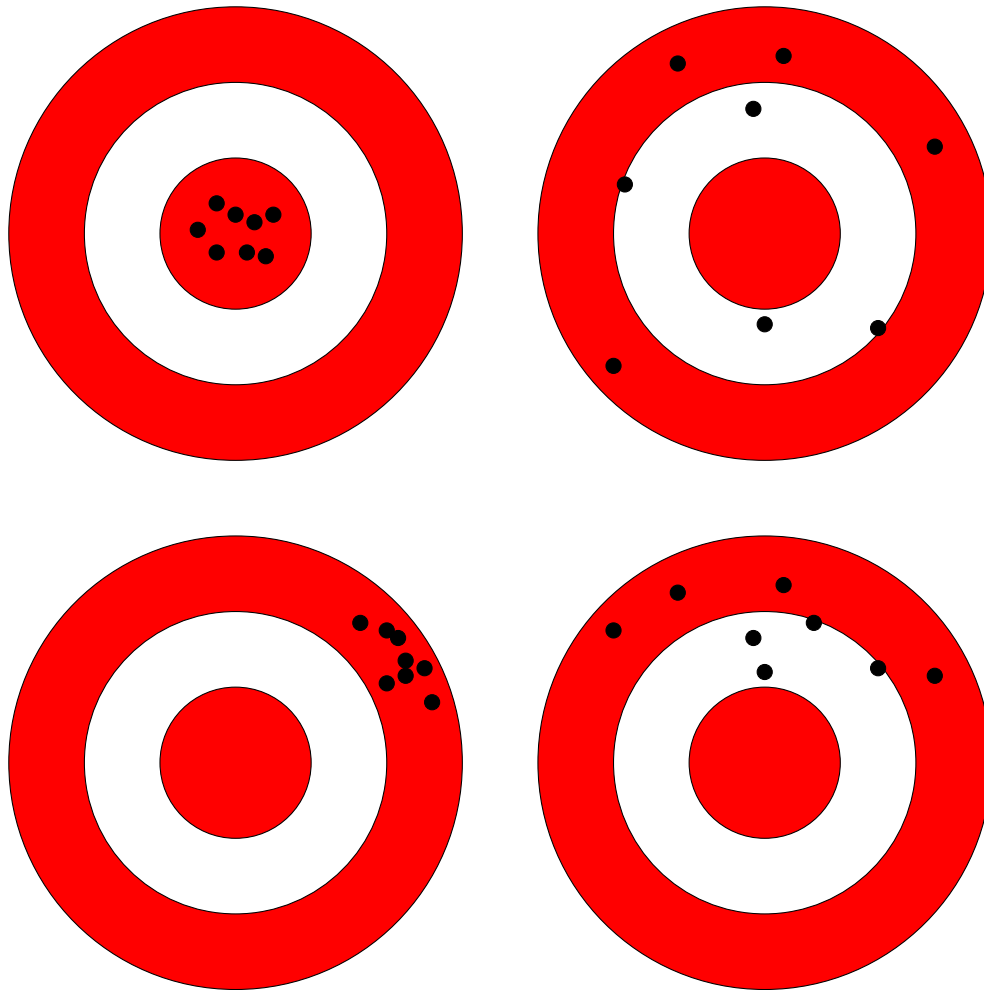
#### Garbage In, Garbage Out

This is a book about applied statistical analysis. So why are we including a whole chapter on considerations for designing your experiments?

It's simple: if your experimental design is bad, or your data collection methods are bad, then your results will be bad. Choosing an appropriate statistical analysis, although that's the crux of this book, does not guarantee that you will walk away with interpretable, generalizable results. For that, you need a well-designed experiment.

One of the biggest reasons that experiments are ever conducted is to evidence (or at least plausibly suggest) a causal link between one or more variables and a particular outcome. To do this, we design experiments with various control and experimental groups, all of which we assume are **probabilistically equivalent**. This means that we know the probability of finding a difference between any of our groups on the basis of pure chance alone. That is, this allows us to believe that just because we find a difference between two groups on one variable, we cannot infer anything about differences among other variables. In turn, this (1) assures us that our findings are due to differences in the treatment and not in the participants and (2) let's us generalize our results to larger populations.

However, to do this, we need to ensure that our studies are both internally and externally valid. **Internal validity** refers to our ability to justify our conclusions: it is what enables a researcher to establish a causal link between a treatment and an effect. **External validity** is an extension of this: it is our ability to take our causal inferences and apply them to populations (rather than only to the sample used in a study).



**Figure 3.1** Accuracy and precision in a game of darts. High accuracy and high precision (upper left); high accuracy and low precision (upper right); low accuracy and high precision (lower left); low accuracy and low precision (lower right). NOTE: The upper right circle is “accurate” because each of the darts **averages out** to near the bullseye.

### Accuracy and Precision

In research, both accuracy and precision are a concern to the researcher: accuracy being how well you are measuring what you actually are trying to measure and precision being how often you measure that trait. Let’s take the example of a game of darts: your accuracy is how close to the bullseye you are on any given throw; your precision is how much your aim varies throw-to-throw (Figure 3.1).

Another way of phrasing this is that your accuracy is how far away you are from something’s true value and your precision is how reproducible your measurement is. When talking about research, accuracy translates to **validity** and precision translates to **reliability**.

Understandably, both are important to good research; however, we will only focus on validity in this chapter. Measures of reliability differ widely discipline-to-discipline and a discussion of inter-rater reliability isn’t going to translate into the context of a physics experiment. If you have questions about reliability, talk to others in your field of study or search around on the internet!

## 3.2 Designing Sound Experiments

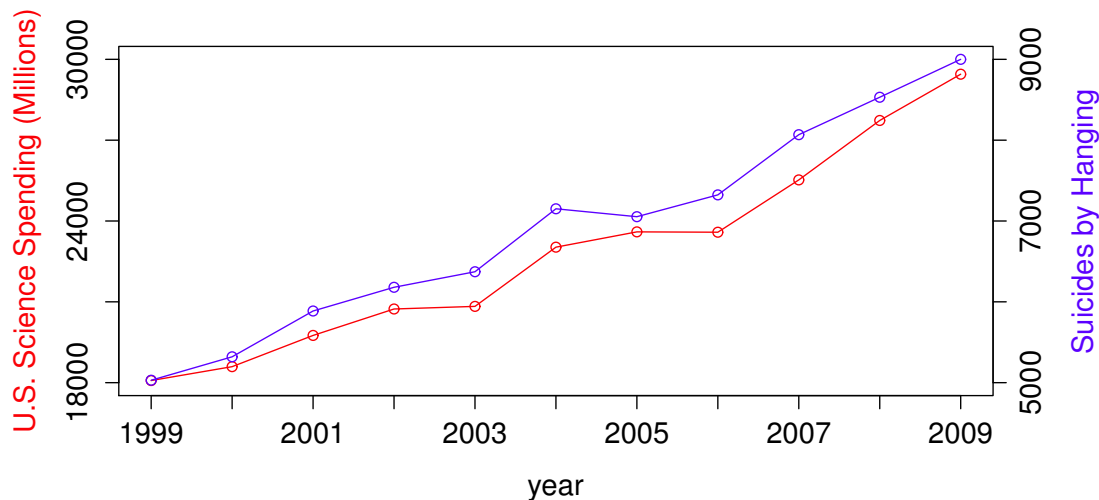
### Internal Validity

Again, **internal validity** reflects the ability of a researcher to draw causal conclusions from a study. Generally, for a study to be internally valid, it must broadly satisfy three conditions:

1. Temporal precedence (the cause precedes the effect);
2. Covariation (the cause and effect are related); and
3. Nonspuriousness (there are no other plausible explanations for the observed effect).

Unfortunately, there are a number of practices that can threaten a study's internal validity. A short list of these is included below.

**Confounding** A confounding variable is one that is not controlled for by the experimenter, yet that correlates with both the dependent and independent variables. This may result in the appearance of a relationship between the two variables when in actuality neither impacts the other, but both are mediated by a third variable. This is termed a **spurious relationship**. For instance, if we look at U.S. spending on science and technology and the number of suicides by hanging, strangulation, and suffocation from 1999 through 2009:



we see that they correlate. And pretty highly. But that's also an idiotic relationship: U.S. spending on science and technology has no effect on the number of suicides by hanging each year. The fact that these two variables correlate is either purely coincidental or mediated by some third variable that we haven't considered. (For instance, over that 10-year span, the U.S. has experienced both inflation—about 28%—and population growth—about 13%—that would each cause the increases that we see in our two variables.)

Yet, if we ignore the absurdity of it and only look at the statistics, we see that there is a highly significant correlation between the two ( $r = 0.99$ ;  $t = 23.90$ ;  $p\text{-value} < 0.001$ ). That's probably the best correlation (or at least the highest Pearson's  $r$ ) you'll see in your natural-born life. Without context, that would be a

highly convincing correlation. Yet, as soon as we stop and think about this, it becomes clear that this is likely a spurious relationship.

Unfortunately, in many cases, spurious relationships will not be so obvious. As such, the researcher must be careful to randomly select his or her study participants and to carefully control for the manipulated factors in the study to prevent the influence of uncontrolled factors.

**Selection Bias** Selection bias is likely one of the more (if not the most) common threats to the internal validity of a study. The idea here is that the sample of participants is not truly a random and representative sample. (I.e., not every possible participant has an equal chance of being selected for the study and certain groups will be overrepresented while other groups may not be represented at all.) Two common practices that create this type of bias are convenience sampling and voluntary response sampling.

**Convenience sampling** refers to the practice of only using as your sample pool participants who are easy to reach; **voluntary response sampling** allows participants to opt in to your study. Unfortunately, these both may skew results, and yet are also common practices in many of the social sciences. Take a psychological experiment, for example: we allow convenience sampling by drawing from the pool of students currently enrolled in an introductory psych course. This pool of participants is neither representative of the university as a whole, nor necessarily generalizable across other universities or to the national population as a whole. Further, these students are usually allowed to choose the studies in which they would like to participate, resulting in voluntary response sampling.

Although neither is necessarily enough to skew the results of a study to the point of it being flatly wrong, considerations such as these do limit the generalizability of one's study. (We'll discuss this more in depth below under the **Sample Limitations** heading.)

When conducting a study, there are various strategies that you can use to help ensure a representative sample is chosen. A brief list of these includes:

1. **Simple Random Sampling** — This implies that everyone in a population has an equal probability of being selected. For instance, say that we want to choose 100 students out of a university with 2000 students. A way to simple randomly sample these students would be to put every student's name into a hat and draw 100 names at random without replacements (i.e., not putting a name back into the hat after it has been drawn). This would give any individual student a 1-in-20 (5%) chance of being selected.
2. **Systematic Sampling** — If you're familiar with Latin, you'll recognize that the English word "decimate" derives from the Latin "decimus" (meaning "tenth"). Originally, this referred to the practice of killing every tenth soldier in a mutinous legion. (Happy, right?) Systematic sampling is similar: given a pool of applicants of size  $N$ , and knowing that you want  $n$  participants in your study, you choose every  $k^{th}$  individual from the pool where  $k = \frac{N}{n}$ . So if we wanted 20 subjects out of a pool of 600, we would arrange them all into a list and pick every 30<sup>th</sup> person ( $\frac{600}{20}$ ) on that list.
3. **Stratified Random Sampling** — Simplifying the actual selection strategies a bit, a stratified random sample is composed of a random sample of participants from different strata of a population (e.g., age group; ethnicity; etc.). Each stratum should be mutually-exclusive (i.e., no individual should be in multiple strata).
4. **Cluster Sampling** — Here, the total population is divided into [clusters](#). A simple random sample of these groups is selected and from these groups a simple random sample of participants is selected.

**History** Unfortunately, history is one of the few threats to internal validity that researchers cannot always control for. Occasionally, events outside the scope of an experiment may impact participants' responses to the experiment. (For example, Americans' attitudes to U.S. clandestine operations would probably have differed dramatically if a researcher asked immediately before 9/11 versus immediately after, regardless of the experimental manipulation.)

**Maturation/Mortality** In longer-term studies, participants may change significantly between measurements (e.g., the concentration of children at 4 years versus 5 years old). Additionally, participants may (and likely will) drop out of longitudinal studies. This is unavoidable; however, should be an important factor for a researcher conducting this type of study to consider and attempt to control for.

**Testing Effects** When participants are given the same (or a similar) test at multiple points in time (e.g., pretesting and post-testing), participant responses on later tests may change as a function of having been already exposed to the measure. For instance, participants may remember questions and their correct answers; they may know what a “good” or “bad” response is; etc. In some cases, steps can be taken to identify whether testing effects have biased participant responses (such as a [Solomon four-group design](#)).

**Experimenter bias** As much as experimenters wish to be objective, we do all have some hypothesis or other that we want to see confirmed or refuted. These personal biases may affect the way in which we interact with participants in experimental and control groups, and by doing so bias their responses or our interpretation of their responses. In many cases this can be eliminated by automating the testing and scoring process and by conducting double-blind experiments in which neither the participant nor the researcher knows which testing group the participant is in.

## External Validity

Given that we can draw causal inferences from our research, we may also want to generalize those results to broader populations. However, researchers often use very specialized populations (college students; psychiatric populations; etc.) that may not be representative of larger groups, or

**Treatment Interactions** Occasionally, an experimental effect may only be observed when multiple treatments interact, or when a treatment interacts with a non-experimental variable such as setting or participant selection. Perhaps the effect only exists in a certain physical location or in an experimental setting (and not in everyday life).

**Order Effects** In addition to randomizing the selection of participants, researchers should be careful, when possible, to randomize the order of treatment. Specifically, if a participant receives Treatment A and then Treatment B, her response may be different than it would have been had she received Treatment B before Treatment A. By randomizing the treatment order, researchers are able to control for this.

**Hawthorne Effect** Also known as the observer effect, this is a phenomenon by which individuals act differently while being observed than they would otherwise.

**Rosenthal Effect** Similar to the Hawthorne Effect, this refers to a tendency for individuals to perform better the more expectation is placed on them to succeed.

**Sample Limitations** This may refer both to the size and makeup of a researcher’s sample. If the number of participants is too small, there is greater risk of sampling error and selecting a non-representative sample from the available pool of participants. Alternately, the participants’ intrinsic characteristics (e.g., age, sex, race, education, etc.) may not be representative of the population that a researcher is attempting to study and to reach general conclusions about.



## CHAPTER 4

---

# INTRODUCTION TO STATISTICAL INFERENCE

---

### 4.1 What Is Statistical Inference

#### Its Purpose

Statistical inference provides formal methods for drawing conclusions about a population from sample data. That is, statistical inference refers to the whole suite of quantitative tools that we have developed to test data-driven hypotheses, accounting for random variation in those data. These are the tools that we will spend the remainder of the book discussing in some degree of depth.

We have developed these statistical models to make predictions about populations and to draw conclusions about differences between populations. These lie in contrast to both descriptive statistics and purely qualitative methods. More correctly, we might say that all three lie on a spectrum with qualitative measures being employed to provide a very general overview of a topic without seeking to address substantive differences among populations or to predict future trends. Oftentimes this type of research is used to inform later quantitative studies by providing new directions for research based on participant opinions, beliefs, and feedback.

Squarely in the middle lie descriptive statistics: these provide a descriptive overview of a set of data; however, they are not (in and of themselves) able to predict trends or quantify difference. It is only when they are applied in the context of statistical models that they can be used to reach decisive conclusions from and about our data.

## Hypotheses for Means

In statistical inference, we often form hypotheses about our data. These will generally take the form of a null and alternative hypothesis. Alternative hypotheses may be one- or two-sided. The null hypothesis (abbreviated  $H_0$ ; pronounced “H-null”) always predicts that there is no difference between our control and treatment groups (i.e.,  $H_0 : \mu = \mu_0$  where  $\mu$  is the mean of our treatment group and  $\mu_0$  is the mean of the control group). Predictably, the alternative hypothesis ( $H_A$ ; pronounced “H-A”) predicts that a difference in our two groups exists. If one-sided, it predicts a direction of the difference (e.g.,  $H_A : \mu > \mu_0$ ; two-sided hypotheses predict only that a difference exists, but are agnostic as to the direction of that difference ( $H_A : \mu \neq \mu_0$ )).

We also operate with the understanding that is easier to prove something as false than as true. Given this, the null hypothesis is always presumed true until sufficient evidence is found supporting the alternative hypothesis. Importantly, though, we do not ever formally accept the null hypothesis: we may only fail to reject it. (Recall again that most inferential statistics seek to disprove a hypothesis rather than prove it.) In this way, we may fail to reject the null hypothesis, but never accept it. (However, we may accept the alternative hypothesis. By disproving the null, we are demonstrating that its converse must, by that fact, be true.)

## Test Statistics and Statistical Significance

Whenever we perform a statistical test, we are usually given two important pieces of data: a test statistic and a p-value. Standardized test statistics can take a number of different forms (z-, t-, and F-statistics are three of the more common) and each has a slightly different interpretation. However, speaking generally, every **test statistic** is a single, quantified measure for assessing patterns in the data that distinguish between the null and alternative hypotheses.

Accompanying this test statistic is the **p-value**. Universally, this measure has a single interpretation: it is the probability (from 0 to 1, inclusive) that the differences observed in the data arise due to chance and natural random variation. As such, smaller p-values indicate that there is a smaller probability of the differences seen being due to chance. Many disciplines impose a “5% rule” on this statistic: that is, they consider any p-value less than 0.05 (or 5%) as being statistically significant.

We use the term **statistically significant** to indicate that we are reasonably confident that the differences observed in the data are in fact due to an experimental manipulation or some actual difference between the groups we’re looking at. There is no specific reason for using a threshold of 5%—and indeed sometimes a 1% or 0.1% cutoff will be used instead—; however, this is the most commonly-agreed-upon threshold for describing a test as significant.

One further consideration when performing statistical tests is that statistical significance is not the same as practical significance. Two sets of data can easily be statistically significant without being practically significant. As an example, consider patient scores on a depression questionnaire: groups from two different clinics may have significantly different scores, but if the mean scores still classify both groups as chronically depressed, a couple-point difference doesn’t make a particularly meaningful impact on either group’s treatment. As such, it may often be important to distinguish between what differences are statistically significant and what differences have a practical, meaningful significance.

Statistical tests will also typically provide a **confidence interval** for the true parameter value. That is, this measure gives a range that, with some percent confidence, should contain the “true” value that the researcher is trying to measure. For example, if a researcher wished to know if there was a difference in average heights of boys and girls at age 8, a 95% CI from 0.3 – 0.9 would indicate that there was a 95% chance that the true mean difference in height (i.e., if we measured the height of every 8-year-old boy and girl in the world) was between 0.3 feet (or 4 inches) and 0.9 feet (or 10.8 inches). Given that this



confidence interval doesn't contain 0 (which would indicate that there is no difference in heights), we can likely assume that there does exist a difference in mean height between our two populations.

## 4.2 Some Considerations about Statistical Inference

### Conditions for Inference

For our conclusions about statistical tests to be valid, there are always assumptions that need to be met about the data. These assumptions differ from test to test; however, they generally require a random sample (or at least a representative sample) of the population to have been selected. Many times the test will also require that the data not be significantly skewed by outliers or that samples have equivalent variance. In any case, with each test presented we will clearly outline the assumptions that must be met as well as those that should be met but that can be violated without necessarily invalidating the test.

### Cautions about Significance Tests

**How Small a  $p$ -value?** As mentioned above, the magnitude of a  $p$ -value needed for a result to be called “significant” is arbitrary: 5% simply happens to be what is widely agreed upon. However, this still means that 1 out of every 20 experiments will tell the researcher that a significant difference exists when there actually isn't one. If we use a 1% threshold, that drops to 1 in every 100 trials. Still more stringent is 1 in 1000 trials, or a 0.1% cutoff. But how small or large a  $p$ -value do we really need to convince us of an effect?

Many times, a much smaller  $p$ -value will be needed to refute a well-established theory. However, if you are exploring a new program of research, larger  $p$ -values may be fine if they are simply exploratory analyses that will be used to indicate potential avenues for further research.

**The Danger of Multiple Tests** When performing statistical analyses, if a researcher analyses the data enough different ways, eventually he or she will obtain statistically significant results. By sheer chance, this is bound to at some point happen. However, one significant result among 20 non-significant results does not constitute strong evidence against a null hypothesis. Rather, this is a process known as  $p$ -hacking: the practice of waiting until a researcher finds an analysis that will produce a favorable result and then reporting that test as definitive evidence in favor of his or her hypothesis.

**Type I and II Errors** Lastly is the issue of errors regarding your conclusion about a statistical test. A **Type I Error** is equivalent to a false positive. This occurs when a researcher achieves a significant  $p$ -value (i.e.,  $p < 0.05$ ) and rejects the null hypothesis when there is actually no evidence for doing so. If you recall from above,  $p = 0.05$  means that every 1 in 20 experiments, the researcher will have a significant  $p$ -value when there actually isn't any difference in the populations. This is a “fluke” that is usually caused by sampling bias or some similar error. Regardless, it ends up that the researcher rejects the null hypothesis when he or she shouldn't have.

Conversely, a **Type II Error** is a false negative, occurring when a researcher fails to reject a null hypothesis despite there actually being a significant difference between the populations being studied. This may be caused by a small sample size, small effect size or power, etc. This distinction can alternately be represented:

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I Error	Correct
Fail to reject $H_0$	Correct	Type II Error

The probability of a test making a Type I error is denoted  $\alpha$ ; the probability of making a Type II error,  $\beta$ .

### 4.3 Recap: The Components of a Statistical Test

To recap, there are 6 main components to any statistical test and write-up: the research question, hypotheses, fundamental and standardized test statistics, the  $p$ -value, and the overall conclusion. As an example, let's say that we are interested in determining whether there is a significant difference in the average heights of 12-year-old boys and girls. Those six steps would look something like:

Item	Example
Research Question	Is there a difference in the average heights of boys and girls at 12 years old?
Hypotheses	$H_0$ : boy height = girl height $H_A$ : boy height $\neq$ girl height
Fundamental Statistics	Average boy height = 4' 10" Average girl height = 4' 11"
Standardized Statistics	This is the t-, z-, F-, etc. statistic that a statistical test will give you.
$p$ -value	The statistical test will also give you a $p$ -value that is between 0 and 1.
Conclusion	If the $p$ -value is less than $p = 0.05$ , most disciplines will consider that significant evidence against the null hypothesis ( $H_0$ ), meaning that we can reject it and accept the alternate hypothesis. In this case, that would mean that there <b>is</b> a significant difference between the average heights of girls and boys at age 12.

Of course, working with real data the second half of the table would look a bit different; however, it would keep that general form. If you're still a little fuzzy on any of the specific elements of a statistical write-up, that's fine: we'll go over it all again each time we present a case study alongside a new statistical test.

## PART II

---

# DESCRIBING RELATIONSHIPS AND PREDICTING VALUES

---



## CHAPTER 5

---

# CORRELATION

---

### 5.1 Scatterplots

Okay, so this is something that we could have presented back a few chapters ago when we talked about data visualization. Unfortunately, that wouldn't have worked with the example we were using, so instead of completely switching gears on you, we decided to wait until now to present it.

Scatterplots are used to display data using cartesian coordinates. For instance, we may wish to visualize the growth of the U.S. population by decade (Figure 5.1).

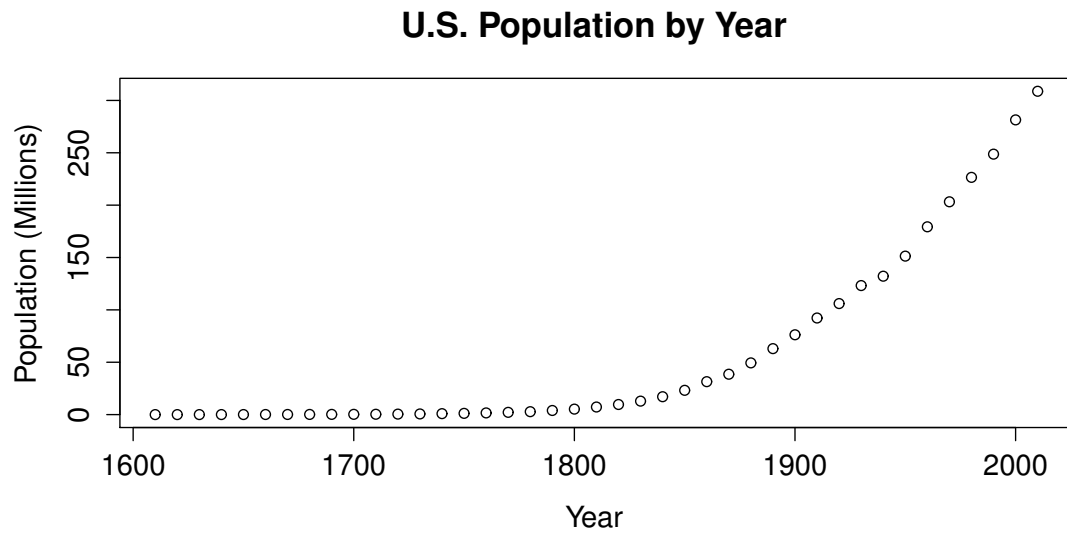
#### Scatterplots with Categorical Variables

Additionally, we can use scatterplots to visualize data with categorical indicators. For example, if we wanted to look at the U.S. population by year, split by males and females (Figure 5.1).

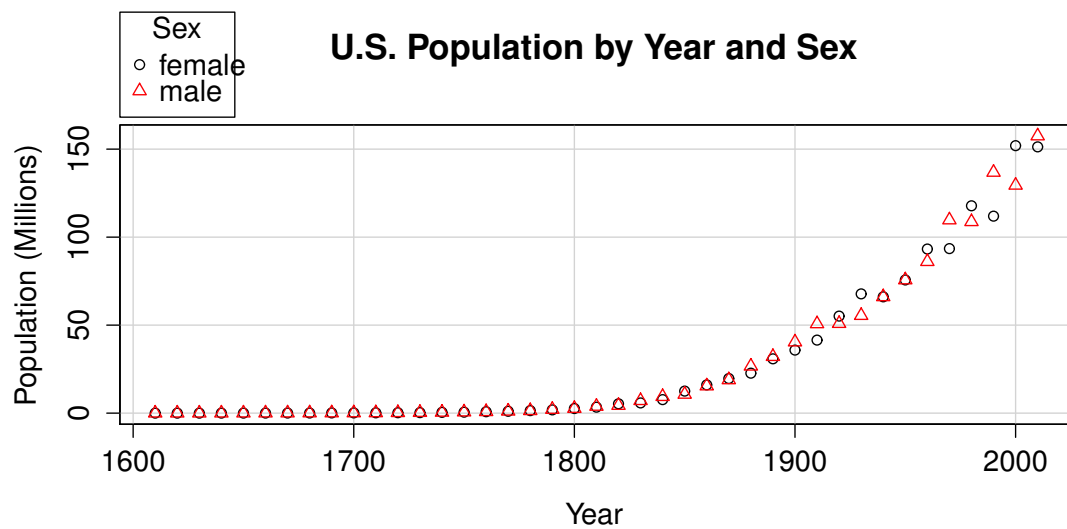
### 5.2 Correlation: Measuring Linear Association

Woah, hold up now: measuring linear association? This sounds like we're heading towards some scary waters. Linear association? I thought this wasn't supposed to be a mathy book!

Don't worry, we didn't lie to you. That's just a bit of a scary term for a not-so-scary concept. **Linear association**, in its most basic form, just means that when you increase one variable by so many units, another variable increases by so many other units. For instance, let's say that for every degree Fahrenheit



**Figure 5.1** United States population by decade: 1600 through 2010



**Figure 5.2** United States population by decade and sex: 1600 through 2010

we increase, we increase the number of daily neighborhood ice cream sales by 10 cones. That's easy, right?

And that's all that correlation is measuring: the degree to which a change in one variable coincides with a change in another variable. Now, importantly, this doesn't mean that one causes the other. As a matter of fact, that's so important that it gets its own section a little ways down the page. But, as long as we keep that in the back of our minds for now, we can go ahead and define correlation as:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{for a population} \quad (5.1)$$

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \quad \text{for a sample} \quad (5.2)$$

So, although this looks like we're going back into the Forbidden Forest, it's a pretty straightforward couple of equations: they're saying that this correlation coefficient  $r$  is just equal to the covariance of two variables,  $x$  and  $y$ , divided by the product of their standard deviations. Or, put another way, it's the amount that two variables change together divided by the amount of that change that we can attribute to chance.

This will then give us a single value,  $r$ , which can be anywhere from -1 to +1. Now, in this case, both negative and positive 1 mean the same thing (or pretty close to the same thing): this coefficient tells us about the strength of the relationship between the two variables. So, if we have  $r = -1$ , that means that for every increase of  $x$  units of one variable, we decrease by  $y$  units of a second variable. Likewise, for  $r = 1$ , this tells us that for every  $x$  units that we increase one variable, we increase the other by exactly  $y$  units. Finally, if we have  $r = 0$ , that tells us that there is absolutely no relationship between our two variables. In other words, you could increase one variable by over 9000 units and absolutely nothing would change in your second variable.

Unfortunately, most of us will never see a correlation that strong in real life: that would mean that there is absolutely no variance of your data. Now, if we're talking about a physical law and measuring it with incredibly high-precision instruments, it's entirely possible that we will have a correlation that strong. However, in most studies, there will be some random variance thrown in that weakens the correlation.

So remember:

$$\begin{aligned} r = \pm 1 &\implies \text{strong correlation} \\ r = 0 &\implies \text{weak correlation} \end{aligned}$$

Using R, if we want to look at the correlation between two variables, we can do so using the `cor()` function. Alternately, if we want to see if there is a **significant** correlation between those two variables, we use the `cor.test()` function. For instance, if we wished to see the correlation between amounts of neck and chest fat, we would run:

```
cor.test(bodyFat$Neck, bodyFat$Chest)

Pearson's product-moment correlation

data: bodyFat$Neck and bodyFat$Chest
t = 18.7782, df = 247, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7102541 0.8136117
sample estimates:
      cor
0.7668596
```

As we can see, there was a correlation between the two variables of  $r = 0.77$  and this reached a level of statistical significance,  $t = 18.78$ ;  $p\text{-value} < 0.001$ . Because our  $p\text{-value}$  was below the 0.05 threshold, we can conclude that the relationship between the two variables actually exists. This means that individuals with more neck fat tend to also have more chest fat, and vice versa.

### 5.3 Correlation Matrices (Multiple Correlation)

We know that correlation is a measure of dependence between two variables. However, there may be times when you want to examine multiple variables at once. In cases like this, it may be useful to create a **correlation matrix**. This is simply a lower-triangular matrix of correlations among multiple variables.

For instance, let's say that we want to see how adiposity in different parts of the body correlate. We may do something like:

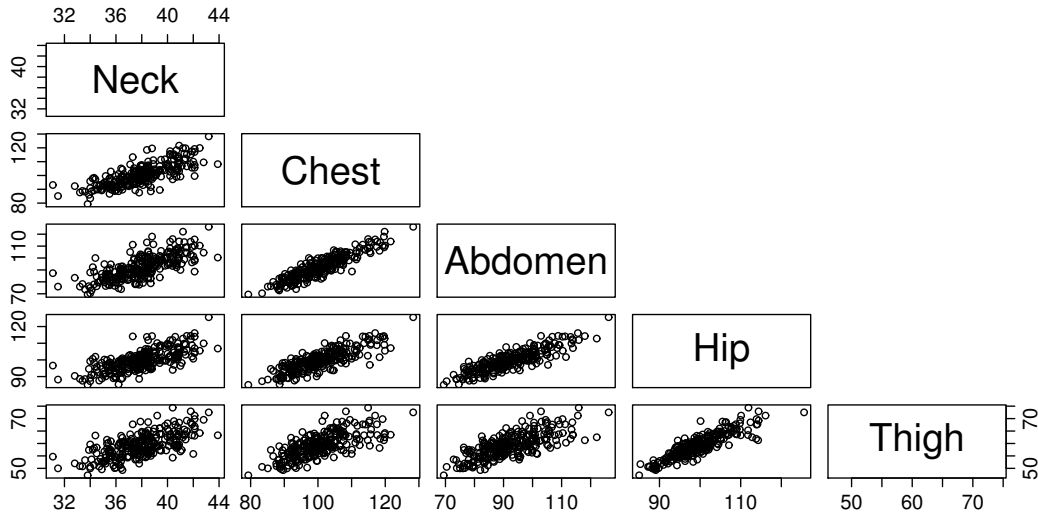
```
as.dist(cor(bodyFat))

      Neck    Chest  Abdomen   Hip
+-----+
Chest | 0.766
Abdomen | 0.728    0.911
Hip | 0.705    0.823    0.860
Thigh | 0.668    0.708    0.736    0.881
```

Here we can see the correlations between Neck, Chest, Abdomen, Hip, and Thigh adiposity—ranging from 0.67 to 0.91. However, what if we wanted to visualize these correlations as scatterplots? We might generate a lower-triangular scatterplot matrix using the `pairs()` function (Figure 5.3):

```
pairs(~Neck+Chest+Abdomen+Hip+Thigh,
      data=bodyFat, upper.panel=NULL)
```





**Figure 5.3** Lower-triangular scatterplot matrix for multiple correlations

Finally, we will likely want to assess the significance of these correlations. To do so, we will have to install the Hmisc package:

```
install.packages("Hmisc")
library(Hmisc)
rcorr(as.matrix(bodyFat))
```

This will return all bivariate correlations and levels of significance for the specified matrix. (NB: The data must be input as a matrix for this function to work. If you are using a data frame, first pass it through the `as.matrix()` function.)

## 5.4 Partial Correlations

In our previous section, we looked at correlations among multiple variables. These were all referred to as **bivariate correlations** because each correlation only looked at exactly two variables. However, we saw that each of those 5 variables correlated significantly with each of the others. So it may be more appropriate to conduct a **partial correlation**: this will take two variables—say, Neck and Chest— and measure their degree of association after controlling for the effect of Abdomen, Hip, and Thigh. Doing so gives the result:

```

source("http://www.yilab.gatech.edu/pcor.R")

neck <- bodyFat$Neck
chest <- bodyFat$Chest
others <- subset(bodyFat, select=c(Abdomen:Thigh))

pcor.test(neck,chest,others)

      estimate p.value statistic
0.344  9.7e-09    5.73    249

```

We can see that the observed correlation drops from  $r = 0.766$  with the bivariate correlation to  $r = 0.344$  with the partial correlation. This means that, when we control for the effects of Abdomen, Hip, and Thigh adiposity, neck and chest still covary significantly with a correlation of about 0.34.

## 5.5 Some Considerations: Causality and Linearity

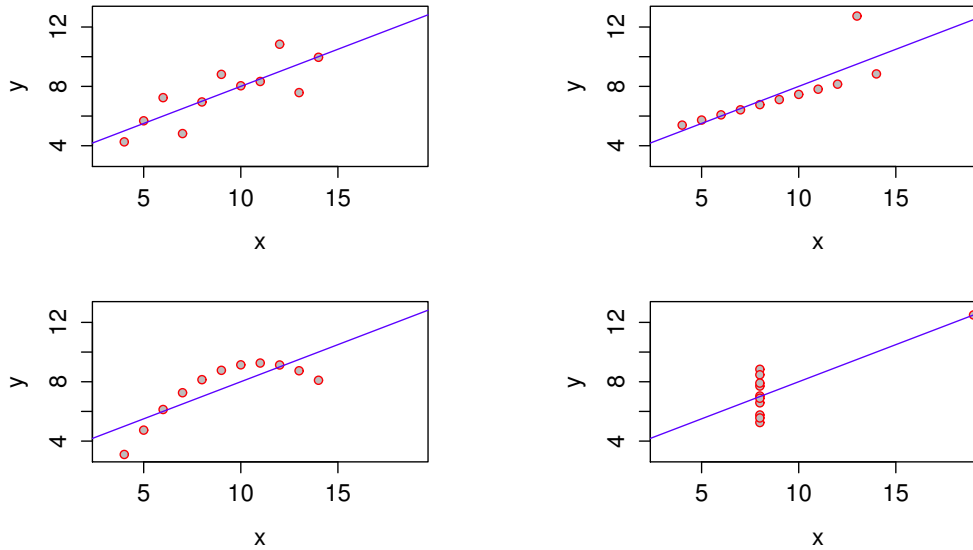
**Causal Inferences** Importantly, one cannot make causal inferences from correlational designs: if you think back to the chapter on research design, for a causal inference to be justified, there must be (1) temporal precedence; (2) covariation; and (3) nonspuriousness. Here, we violate the first and third assumptions. Regarding temporal precedence, neither of the measures that we are looking at clearly precedes the other: the measurements could be made simultaneously; one could always occur before the other; or the order of occurrence could be random at each point of measurement. Additionally, correlational studies don't have any control or experimental conditions in place to ensure that potential confounding variables are controlled for and don't give rise to spurious correlations.

Given these concerns, we are only able to say that Variable A and Variable B tend to covary: that is, when one changes in a certain way, the other is likely to change in a certain way. For instance, consider height and weight. Let's say that there is a positive correlation between the two (i.e., that taller people tend to weigh more and that people who weigh less tend to be shorter). We can say that someone who is 6'1" is likely to weigh more than someone who is 4'10"; however, it becomes silly to say that if someone loses weight, he or she will start shrinking in height. Likewise, by gaining weight, no one will ever grow taller. There is a general trend of the data; however, this does not mean that a change in one ever causes a change in the other.

**Linearity of the Relationship** Another consideration when looking at correlations is the relationship between the two variables of interest. Namely, there must be a linear trend. In this context, a linear trend is going to mean that every time Variable A changes by  $x$  units, Variable B will change by  $y$  units. For instance, take the four plots in Figure 5.5.

Here we have four different relationships between our two variables with the same regression line plotted against all of them. As far as a correlation is concerned, the data follow that solid line.

**Sensitivity to the Distribution** Also important is the distribution of the data. Although the degree of dependence between two variables  $X$  and  $Y$  is unaffected by transformations of the data where both  $X$  and  $Y$  are transformed by constants, the strength of correlation is highly impacted by the range of values sampled. Generally, the wider the range that is sampled, the stronger the correlation will be between the two measures.



**Figure 5.4** Four separate data sets, each with the same mean and correlation. Source: [Wikipedia](#)

Take, for instance, Figure 5.5. In orange are the unrestricted data, giving a correlation of  $r = 0.897$ . Yet, when we restrict our range to only values of  $X$  on the interval  $(0, 1)$ , that correlation drops to  $r = 0.387$  although no transformations or other alterations were made to the data. Given these concerns, there have been attempts to correct for this range restriction; however we will not detail them here. Generally, this is not often a major issue to researchers; however, if you know that your data will be somehow restricted, it is good to keep in mind that this may impact the correlation coefficient.

### Case Study: National Education Trends

First let's go ahead and download our data from GitHub:

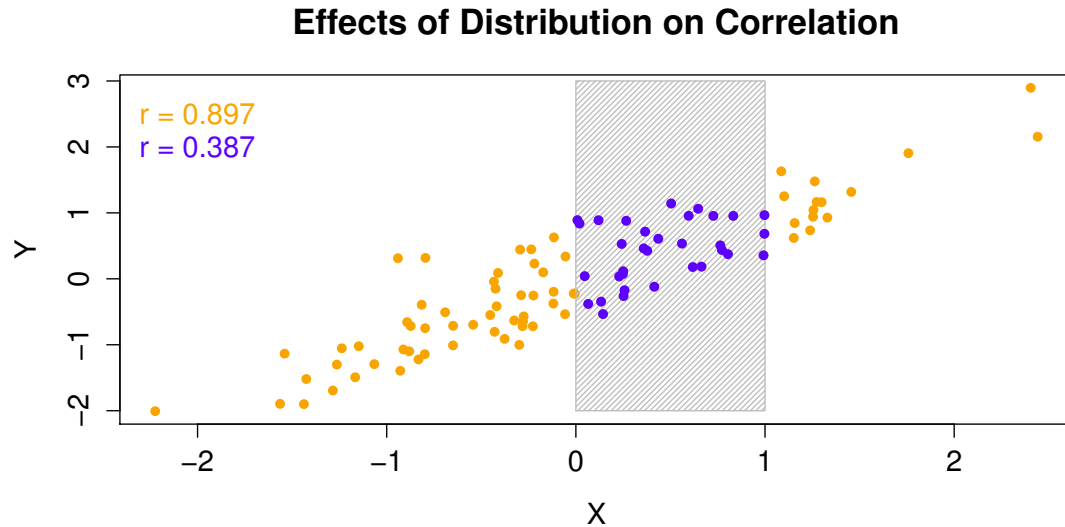
```
# This will let us download the file from a remote URL

download.file(
  "http://book.chriswetherill.me/part2/data/correlationCaseStudy.csv",
  "census.csv", "wget", extra="--no-check-certificate")

# And now we will read the data into R and store it
# in a data frame called "data"

data <- read.csv("census.csv", header=TRUE, sep=",")
```

Now, we have two columns that we're interested in: `highSchoolorHigher` and `perCapitaIncome`. We would like to see if there is a relation between a state's per capita income and the proportion of its residents



**Figure 5.5** Pearson correlation coefficients between X and Y are shown when the two variables' ranges are unrestricted (orange) and when the range of X is restricted to the interval (0,1) (blue)

to have completed high school or higher. So let's start by constructing a scatterplot of the two variables, seen in Figure 5.5:

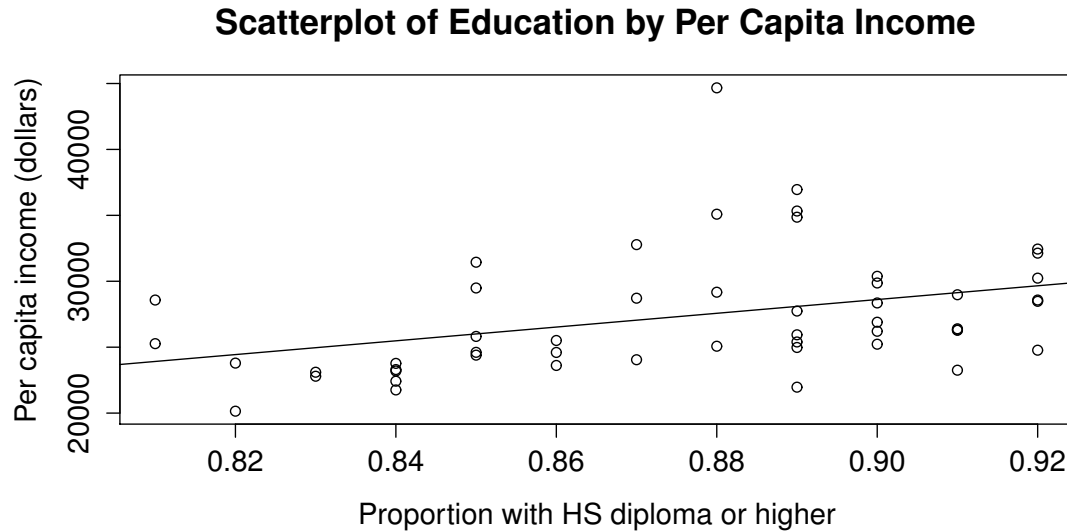
```
# Notice, when we reference a column we use the syntax
# dataset$column. The first half tells R which data set
# we are referencing, the dollar sign indicates that we
# want to reference a specific column, and everything
# after that is the column name itself.

plot(data$HighSchoolOrHigher, data$perCapitaIncome,
     xlab="Proportion with HS diploma or higher",
     ylab="Per capita income (dollars)",
     main="Scatterplot of Education by Per Capita Income")

# We can also plot a line of best fit:

fit <- lm(data$perCapitaIncome~data$HighSchoolOrHigher)
abline(fit)
```

As we can see, there appears to be a positive linear relationship between per capita income and the proportion of a state's residents having a high school diploma or higher. Our next step is then to quantify the strength of this relationship. To do this, we will perform a bivariate correlation.



**Figure 5.6** Scatterplot of perCapitaIncome and HighSchoolOrHigher

```
# The syntax for a correlation test in R is easy:
# it is cor.test(var1, var2) where var1 and var2
# are the two variables that you're interested in.

cor.test(data$perCapitaIncome, data$HighSchoolOrHigher)

Pearson's product-moment correlation

data: data$perCapitaIncome and data$HighSchoolOrHigher
t = 2.7752, df = 49, p-value = 0.007788
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1034750 0.5847427
sample estimates:
cor
0.3685491
```

As we can see, there are several noteworthy items presented in this table. Firstly, we have a  $t$ -statistic of 2.78. This is above the 1.96 threshold that we set, indicating that our correlation is probably going to be significant. From there, we can look at our  $p$ -value (0.008) and see that it is below the 0.05 threshold, indicating that we do indeed have a significant correlation.

If it weren't obvious from the scatterplot above, this is a positive correlation with a Pearson's  $r = 0.37$ , meaning that there exists a fair positive relationship between our two variables. I.e., when one is larger, the other will also tend to be larger.

## Exercises

**5.1** Download the [Patient Satisfaction data set](#) from GitHub. This dataset contains information from patients surveyed at various hospitals following their treatment to assess their satisfaction with the experience. We will be using these data for the following exercises.

**5.2** Do patient ratings for `nursesCommunicateWell` and `doctorsCommunicateWell` correlate with one another? Provide evidence to back up your answer. Include a scatterplot of the data.

**5.3** Now perform a partial correlation between those two same variables, but controlling for `givenInformationAboutRecovery` and `staffExplainedMedications`.

**5.4** Create a lower-diagonal correlation matrix correlating all of the variables included in the dataset (except for the hospital ID). What correlations are significant? Are there any that are non-significant? (NOTE: You will have to remove null values using the `na.omit()` function for this to work properly.

**5.5** Construct a scatterplot matrix of all bivariate correlations using the code:

```
pairs(~nursesCommunicateWell + doctorsCommunicateWell +
      receivedImmediateHelp + painManagedByTreatment +
      staffExplainedMedications + bathroomsAlwaysClean +
      givenInformationAboutRecovery + rateHospitalPositively,
      data=ex01, upper.panel=NULL)
```

Do any of the scatterplots look concerning? Look for outliers, non-linear trends, etc.

**5.6 Test yourself:** Choose two new variables and perform a bivariate correlation test. Do they correlate significantly? Do you think there are any other variables that should be controlled for? If so, perform a partial correlation, controlling for those additional variables. Do the results differ? Explain why they do. Comment on the assumptions made by the correlation tests you have run. Are they met? Are any violated?

**5.7** Download the [Census American Community Survey](#) from GitHub. This dataset, used in the case study above, contains information about employment and other demographic characteristics nationwide.

**5.8** Is there a correlation between `noHighSchoolDiploma` (the proportion of residents without a HS diploma or GED equivalent) and `publicTransit` (the proportion of residents who use public transit to go to and from work)?

**5.9** Is there a correlation between `HighSchoolOrHigher` and `percentOnSNAP`? Justify your findings and include at least one figure.

**5.10** Is there a correlation between `medianRent` and `percentImpoverished`? Are there any variables we might want to control for using a partial correlation?

**5.11 Test yourself:** Choose some (or all) of the variables in this dataset and make a correlation matrix for them. Choose a correlation that looks interesting or surprising and investigate it further. If applicable, perform a partial correlation test rather than a bivariate correlation.

### Additional Resources

1. [All data sets](#) used in the chapter
2. [All R scripts](#) used in the chapter
3. [Answer key](#) to the chapter's exercises





## CHAPTER 6

---

# SIMPLE LINEAR REGRESSION

---

### 6.1 Overview of Simple Linear Regression

#### Uses of Simple Linear Regression

In the last chapter we touched briefly on the idea of using the value of one variable to predict the value of another. Where correlation is used to describe the magnitude of the relationship between two variables, a tool called **simple linear regression** takes that relationship and uses it to predict the values of Variable B, given some arbitrary value of Variable A. In this context, however, we refer to the  $x$  variable as our **explanatory variable** or our **regressor** and to the  $y$  variable as the **response variable**.

So, knowing that we use regression as a predictive tool, what *is* it, exactly? Well, thankfully, it's a pretty straightforward thing that we can define as:

$$\hat{y} = \beta_0 + \beta_1 \cdot x \quad (6.1)$$

So before we explain this equation, we need you to reach way back into those murky memories of your eighth grade math class when you were talking about the equation of a line. Remember that whole “slope-intercept form” lecture? Well, in case you don't, the equation of a line looks something like:

$$y = m \cdot x + b$$

where  $m$  is the slope and  $b$  is the intercept. Sound familiar?

Alright, now let's compare those two equations. Aside from using different variable names, they're identical to one another. All regression boils down to is drawing lines to connect a bunch of points on the page. And here you thought it was going to be scary!

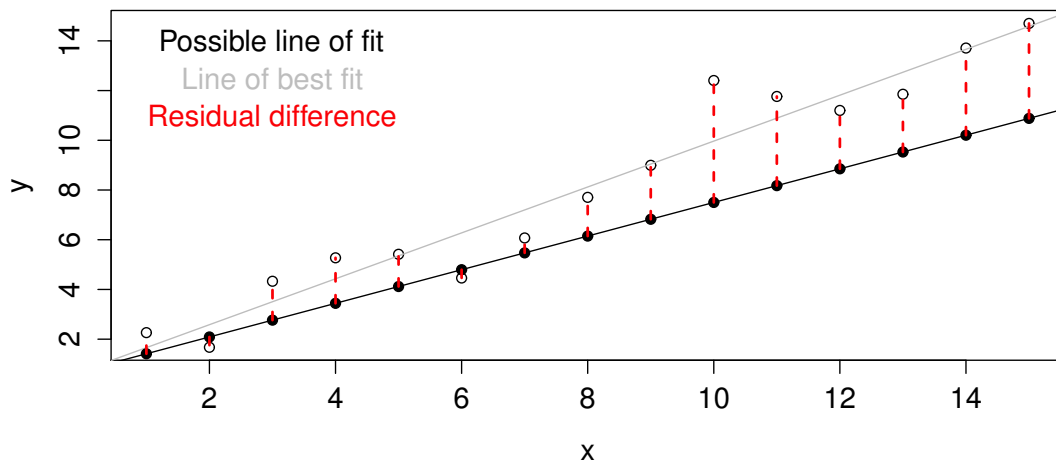
Although regression is actually a bit more than drawing a line to connect the dots, the basic idea behind it is that for two variables there exists an equation with two basic constants—a slope and an intercept—that allows us to predict the value of  $Y$  if we already know  $X$ .

But, although we know that equation exists, how do we figure it out? To do that we'll turn to something called **least squares estimators**.

### Least Squares Fitting

If you think back to Equation 6.1, our two coefficients were  $\beta_0$  and  $\beta_1$ . Those refer to the population coefficients: that is, if we were able to measure every single instance of  $X$  and  $Y$  in all of the universe, those would be the coefficients that we could derive from the population. In most cases, however, that isn't possible and so we use least squares estimates of  $\beta_0$  and  $\beta_1$  which we refer to as  $\hat{\beta}_0$  and  $\hat{\beta}_1$  (pronounced “beta-null hat” and “beta-one hat”).

We can visualize this process in Figure 6.1. Here, we see a number of data points as white circles. There is also a solid black line that represents a possible line of best fit. However, are the predicted data points (solid black circles) as close to the observed data as possible? Looking at the red dashed line between our predictions and observations, it looks like we could do a better job reducing that difference.



**Figure 6.1** Data with a least-squares line of best fit (grey), one potential line of best fit (black), and the residual difference between the predicted and observed values (red). Least squares fitting seeks to minimize the overall distance between every observed value and predicted value, resulting in a single line of best fit.

Minimizing that difference between our predicted and observed values is exactly what least squares estimation seeks to do. In fact, we can see the least squares line of best fit as the light grey line in Figure 6.1. The residual difference in this line is mathematically as small as it possibly can be.

How we actually do this involves a bit of calculus so we aren't going to go into too much detail here. However, from that calculus we can derive these two estimates of our coefficients:

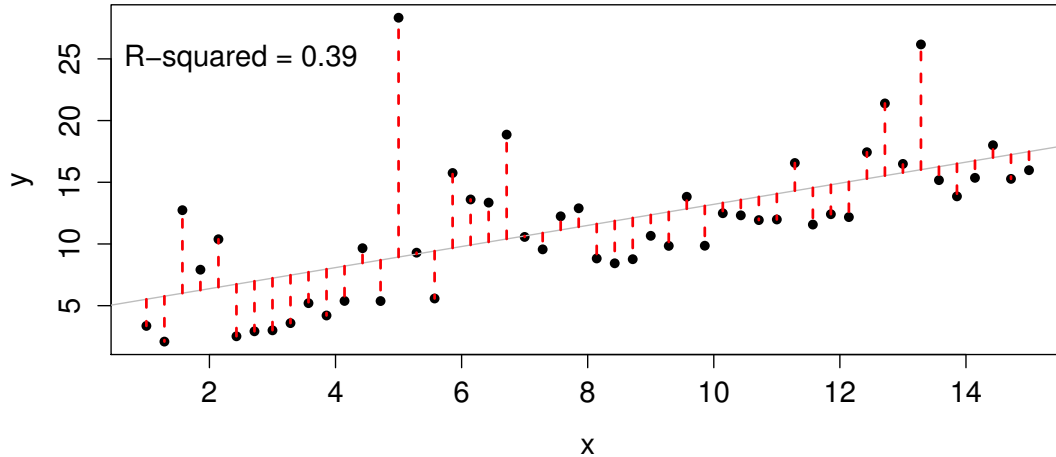
$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \cdot \bar{X}\end{aligned}$$

Thankfully, these aren't equations that you're liable to ever have to understand or use (unless you just really want to): least squares fitting is automated by every modern statistical package. Even your TI-84 calculator is happy to do it for you. The point, rather, is that given a set of data, there are these mathematical tools that we can use to attempt to model the relationship between our variables, even though this is mostly a behind-the-scenes operation.

### Coefficient of Determination

The whole of a simple linear regression boils down to just a couple numbers, one of which is the **coefficient of determination**, denoted  $R^2$  (pronounced “R-squared”). This statistic represents the proportion of the variance seen in the data that is explained by the model.

If you think back to the chapter *Measuring Uncertainty*, variance is the squared average distance of a data point from the mean. Let's take Figure 6.1 as an example. Here, we see that there is a regression equation fitted to the data with an  $R^2 = 0.39$ , indicating that about 39% of the variance observed within the data is explained by the model.



**Figure 6.2** Regression model with approximately 39% of the observed variance explained.

Now, it's important to recognize that the line of best fit that we've plotted doesn't describe our data perfectly. This is because there is an inherent bit of error—random variability—in any dataset (and thereby any equation that we use to model it). Given this, a better definition of Equation 6.1 is:

$$\hat{y} = \beta_0 + \beta_1 \cdot x + \epsilon \quad (6.2)$$

where  $\epsilon$  represents that error. In our example, we can see that each point does not vary above or below the regression line by an equivalent amount: if there were no error, then the distance of all of the points from the regression line, above and below, should sum out to 0. (That is, points above the line are equally far from it as are points below the line.)

The equation that we use to model our data will never be able to perfectly explain all of that random variation which is why we use the  $R^2$  statistic to quantify the proportion of it that **can be** explained. Correspondingly, a larger value of  $R^2$  corresponds to a model that better fits the data.

As with Pearson's  $r$  in correlational designs,  $R^2$  varies from 0 to 1. (Note: Pearson's  $r$  can fall on the interval  $(-1, 1)$ ; however, because the coefficient of determination is squared, it can only fall on the interval  $(0, 1)$ . Any negative number multiplied by another negative number will always equal a positive number.)

## Residuals

The formal calculation of the proportion of the variance that is explained by the model is a bit more complicated than what we presented above. To understand how  $R^2$  is calculated, we first need to understand the concept of **residuals**. A residual is defined as the difference between an observed and predicted value. If we look back to Figure 6.1, the residuals are represented by the dashed red lines. This is equivalent to the error term ( $\epsilon$ ) that we mentioned in the previous section.

When calculating  $R^2$ , we use the formula:

$$R^2 \equiv 1 - \frac{SS_{residual}}{SS_{total}} \quad (6.3)$$

where  $SS_{residual}$  is the residual sum of squares and  $SS_{total}$  is the total sum of squares. To get an idea of what these two terms mean, let's look at Figure 6.1. On the left we see a representation of the total sum of squares for four data points. This is derived by finding the squared distance between each observed data point and the arithmetic mean of  $y$  and adding them all together. Mathematically, we represent this as:

$$SS_{total} = \sum_i (y_i - \bar{y})^2$$

The idea here is that we want to find out the total distance of our data from their mean: a larger value means that the data are farther dispersed from their mean; a smaller value indicates that they all cluster closer to their mean value. (Likewise, a value closer to 0 indicates a weaker relationship between the two variables: it shows that  $y$  is unaffected by the value of  $x$ .)

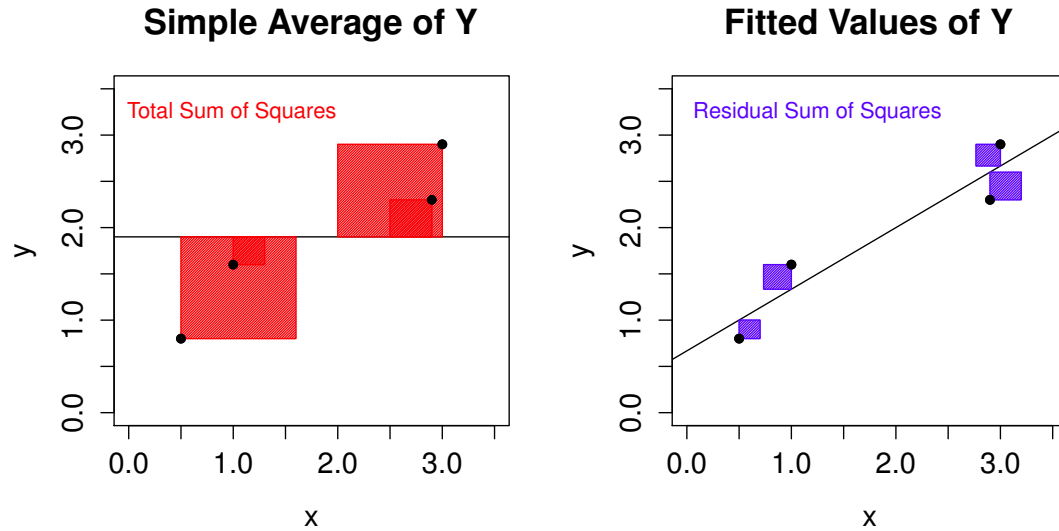
Next we have the residual sum of squares. This value represents the sum of the squared distances of each point from its predicted value. In other words, it's a measure of how accurately the data can be fitted to a line. Mathematically, we represent this as:

$$SS_{residual} = \sum_i (y_i - \hat{y}_i)^2$$

As with our total sum of squares, a higher value for our residual sum of squares indicates a larger discrepancy between our predicted and observed values of  $y$  whereas a smaller value indicates a better fit of the model to the data. By then comparing the ratio of these two values, we can assess what proportion of the total variance our model is able to explain, and thus how good of a fit it provides.

It may help if we visualize these, as in Figure 6.1. Here, the total sum of squares is represented in red on the left. We see that the horizontal line represents the average  $y$  value across our four data points—just below 2.0. To compute the total sum of squares, we measure how far apart the  $y$  value for each of our four data points is from that average and we square it. This is represented by each of the red boxes. To obtain our sum of squares, we just go ahead and add the areas of the four boxes.

You can also think of the total sum of squares as a measure of the relationship between our explanatory and response variables: if there were no relationship between them, then every single data point would be exactly on the line and our sum of squares would be 0.



**Figure 6.3**  $R^2$  is defined generally as  $1 - \frac{SS_{residual}}{SS_{total}}$ . Here, we see the total sum of squares (red) as the sum of the squared distance from each observation to the arithmetic mean of  $y$  ( $\bar{y}$ ). The residual sum of squares (blue) is computed as the sum of the squared distance from each observation to its predicted value.

Next we can look at the residual sum of squares, shown in blue on the right. Here we have our line of best fit along with the same four data points. To calculate this sum of squares, we follow the same procedure as before: we measure the distance on the  $y$ -axis between each data point and the line, we square it, and then we add together all those values. In this case, though, the difference represents the failure of our model to explain the relationship between our two variables. If it explained the relationship perfectly, then our residual sum of squares would be equal to 0.

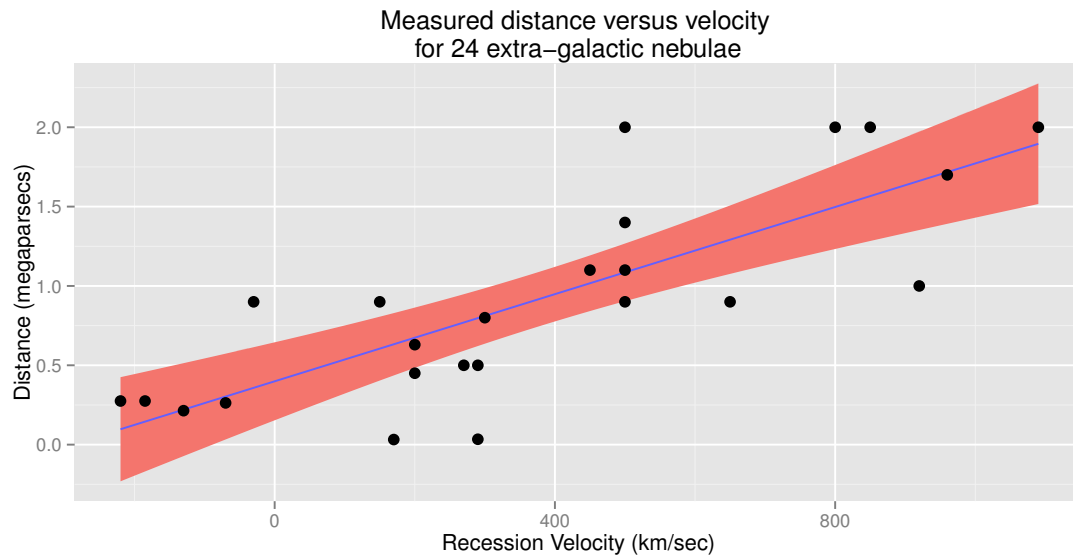
So, knowing that a larger value of our  $SS_{total}$  indicates a stronger relationship and that a smaller value of our  $SS_{residual}$  means a better model, we can see that larger values of  $R^2$  mean that a larger proportion of the variance seen in our data is explained by the model. (Remember, dividing a small number— $SS_{residual}$ —by a larger number— $SS_{total}$ —gives you an even smaller result. So if  $R^2$  is equal to 1 minus whatever that value is, its value will approach 1 as the model becomes better and 0 as the model becomes worse.

## 6.2 Cautions and Considerations

### Basic Assumptions

## Outliers and Influential Points

### 6.3 Case Study: [STUDY]



**Figure 6.4** Measured distance versus velocity for 24 extra-galactic nebulae. Data from Hubble, E. (1929). [A relation between distance and radial velocity among extra-galactic nebulae](#). *Proceedings of the National Academy of Sciences of the United States of America*, 15, 168-173. Line of best fit plotted in blue with a 95% confidence interval for the regression line in red.

## 6.4 Exercises

## 6.5 Additional resources

## CHAPTER 7

---

### MULTIPLE REGRESSION

---





## CHAPTER 8

---

## LOGISTIC REGRESSION

---



## CHAPTER 9

---

# MODEL REFINEMENT AND VARIABLE SELECTION

---



## PART III

---

# COMPARISONS AMONG MULTIPLE SAMPLES

---



## CHAPTER 10

---

### *T*-TOOLS

---





## CHAPTER 11

---

### ALTERNATIVES TO *T*-TOOLS

---



## CHAPTER 12

---

### MULTIFACTOR STUDIES

---



## CHAPTER 13

---

### MULTIFACTOR STUDIES

---



## CHAPTER 14

---

# MULTIFACTOR STUDIES WITH REPLICATION

---





## PART IV

---

# INFERENCES FROM UNUSUAL DATA STRUCTURES

---



## CHAPTER 15

---

### TIME-SERIES DATA AND SERIAL CORRELATION

---



## CHAPTER 16

---

### RANK-ORDERED DATA

---



## CHAPTER 17

---

### TABLES OF COUNTS, PROPORTIONS, AND ODDS

---





## CHAPTER 18

---

# FACTOR ANALYSIS AND PRINCIPLE COMPONENT ANALYSIS

---



## PART V

---

## SUPPLEMENTAL MATERIALS

---



## CHAPTER 19

---

### CHOOSING THE RIGHT ANALYSIS: A CHEAT SHEET

---



## CHAPTER 20

---

### GETTING OUTSIDE HELP: HOW TO SHARE DATA

---

