# APPLIED STATISTICS

## An Introduction to Statistical Analysis

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

CC0 1.0 Universal

**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

## PART I INTRODUCTORY MATERIALS TO THE BOOK

**1 Preface 3**

1.1 Introduction 3
1.2 To the Student 4

**2 Introduction to R 7**

2.1 Getting Started 7
2.2 Basic R Syntax 9
2.3 Exercises 12
2.4 Additional Resources 13

**3 Measuring Uncertainty 15**

3.1 Measures of Central Tendency and Spread 15
3.2 Standard Deviation 18
3.3 Sampling Distributions 21
3.4 Visualizing Data 23
3.5 Exercises 26

3.6 Additonal Resources 26

4 Research Design 27

4.1 Relating Design and Analysis 27
4.2 Designing Sound Experiments 29

5 Introduction to Statistical Inference 33

5.1 What Is Statistical Inference 33
5.2 Some Considerations about Statistical Inference 35
5.3 Recap: The Components of a Statistical Test 36

PART II COMPARISONS AMONG MULTIPLE SAMPLES

6 *t*-Tools 41

6.1 Overview of *t*-Tools 41
6.2 Cautions and Considerations 44
6.3 Implementation in R 45
6.4 Case Study: Medicare Spending on Patients 47
6.5 Exercises 49
6.6 Additional Resources 49

7 Alternatives to *t*-Tools 51

7.1 What Are Nonparametric Statistics? 51
7.2 Overview of Nonparametric Tools for Two Samples 52
7.3 Cautions and Considerations 53
7.4 Implementation in R 54
7.5 Case Study: Heavy Metal Accumulation 55
7.6 Exercises 56
7.7 Additional Resources 56

8 One-Way ANOVA 57

8.1 Overview of One-Way ANOVAs 57
8.2 Cautions and Considerations 60
8.3 Implementation in R 61
8.4 Case Study: [STUDY] 61
8.5 Exercises 61
8.6 Additional Resources 61

9 Multifactor Studies 63

9.1 Overview of Multiple ANOVAs 63
9.2 Cautions and Considerations 63

9.3 Implementation in R 63
9.4 Case Study: [STUDY] 63
9.5 Exercises 63
9.6 Additional Resources 63

10 Multifactor Studies with Replication 65

10.1 Overview of Repeated Measures Designs 65
10.2 Cautions and Considerations 65
10.3 Implementation in R 65
10.4 Case Study: [STUDY] 65
10.5 Exercises 65
10.6 Additional Resources 65

## PART III DESCRIBING RELATIONSHIPS AND PREDICTING VALUES

11 Correlation 69

11.1 Overview of Correlation 69
11.2 Cautions and Considerations 73
11.3 Implementation in R 74
11.4 Case Study: National Education Trends 76
11.5 Exercises 77
11.6 Additional Resources 78

12 Simple Linear Regression 79

12.1 Overview of Simple Linear Regression 79
12.2 Cautions and Considerations 85
12.3 Implementation in R 85
12.4 Case Study: Evidence for the Big Bang 87
12.5 Exercises 88
12.6 Additional resources 88

13 Multiple Regression 91

13.1 Overview of Multiple Regression 91
13.2 Cautions and Considerations 91
13.3 Implementation in R 91
13.4 Case Study: [STUDY] 91
13.5 Exercises 91
13.6 Additional Resources 91

14 Logistic Regression 93

14.1 Overview of Logistic Regression 93

14.2 Cautions and Considerations 93
14.3 Implementation in R 93
14.4 Case Study: [STUDY] 93
14.5 Exercises 93
14.6 Additional Resources 93

15 Model Refinement and Variable Selection 95

15.1 General Strategies for Model Refinement 95
15.2 Cautions and Considerations 95
15.3 Implementation in R 95
15.4 Case Study: [STUDY] 95
15.5 Exercises 95
15.6 Additional Resources 95

PART IV INFERENCES FROM UNUSUAL DATA STRUCTURES

16 Time-Series Data and Serial Correlation 99

16.1 Overview of Time-Series Tools 99
16.2 Cautions and Considerations 99
16.3 Implementation in R 99
16.4 Case Study: [STUDY] 99
16.5 Exercises 99
16.6 Additional Resources 99

17 Factor Analysis and Principle Component Analysis 101

17.1 Overview of Factor Analyses 101
17.2 Cautions and Considerations 101
17.3 Implementation in R 101
17.4 Case Study: [STUDY] 101
17.5 Exercises 101
17.6 Additional Resources 101

18 Rank-Ordered Data 103

18.1 Overview of Tools for Ranked Data 103
18.2 Cautions and Considerations 103
18.3 Implementation in R 103
18.4 Case Study: [STUDY] 103
18.5 Exercises 103

18.6 Additional Resources 103

19 Tables of Counts, Proportions, and Odds 105

19.1 Overview of Chi-Squared Tests 105
19.2 Cautions and Considerations 105
19.3 Implementation in R 105
19.4 Case Study: [STUDY] 105
19.5 Exercises 105
19.6 Additional Resources 105

PART V SUPPLEMENTAL MATERIALS

20 Choosing the Right Analysis: A Cheat Sheet 109

21 Getting Outside Help: How to Share Data 111

# PART I

# INTRODUCTORY MATERIALS TO THE BOOK

# CHAPTER 1

# PREFACE

## 1.1 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.

## 1.2 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

| | |
|---|---|
| **Statistic** | Any quantity that can be calculated from a set of data. |
| **Population** | 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." |
| **Sample** | 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." |
| **Parameter** | A statistic referring to a population. This is usually denoted by a Greek letter. |
| **Estimate** | 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. |
| **Mean** | 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"). |
| **Standard Deviation** | A measure of how far dispersed observations are from their mean. Denoted $\sigma$ when referring to the parameter or $s$ for the estimate. |
| **Standard Error** | 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.

## CHAPTER 2

# INTRODUCTION TO R

## 2.1 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 .xslx; 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.)

## 2.2 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
```

## 2.3 Exercises

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

```
download.file(
    "https://raw.githubusercontent.com/faulconbridge/appliedStats/
master/LaTeX/part01/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].`

**2.2** 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.

**2.3** 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.

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

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

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

**2.7** 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`.

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

## 2.4 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 3

# MEASURING UNCERTAINTY

## 3.1 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, $x_1, x_2, \ldots, x_n$, is denoted by $\bar{x}$ (pronounced “x-bar”). For a sample of $n$ observations, it is calculated:

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

Alternately, it may be represented as:

$$\bar{x} = \frac{\sum_i^n x_i}{n} \tag{3.1}$$

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 3.1.

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, out 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 3.2.

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

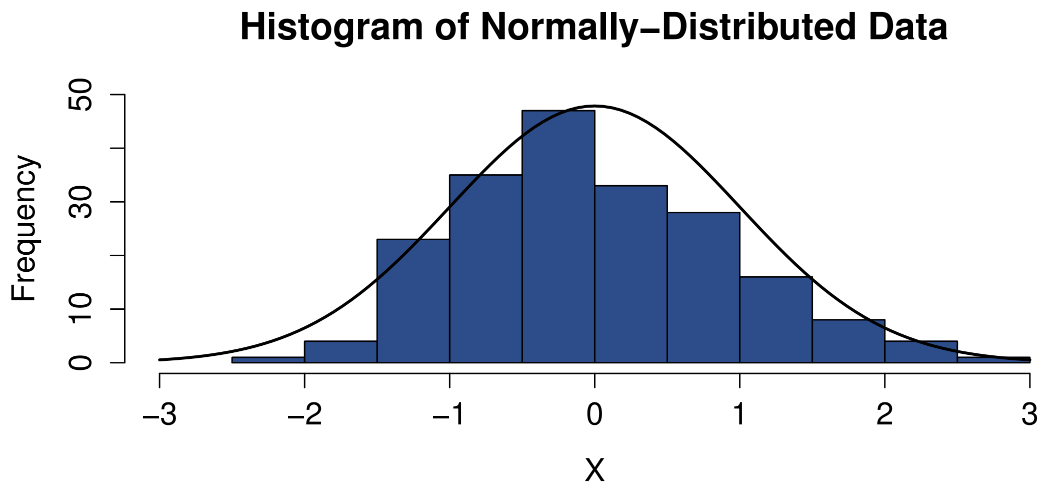


**Figure 3.1** Histogram of approximately normal data with a slight rightward skew

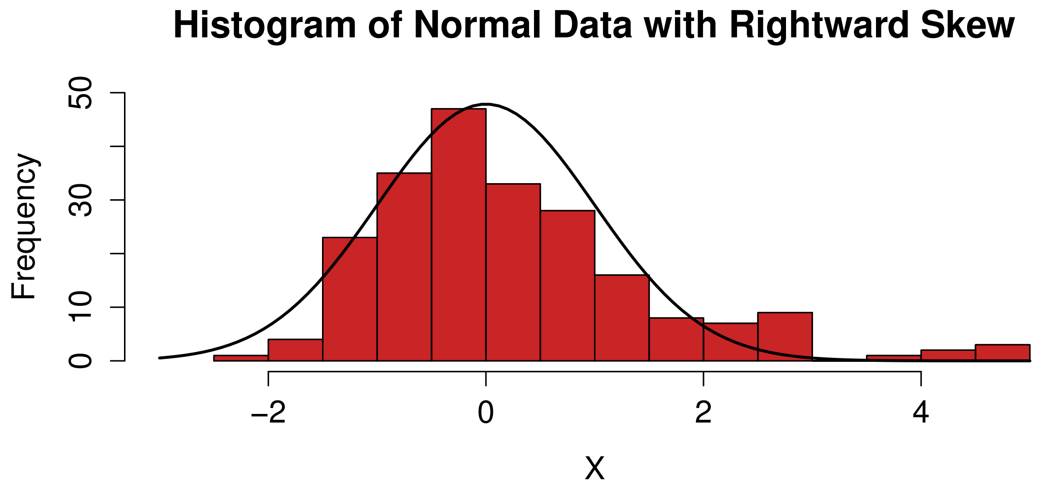


**Figure 3.2** Histogram of data heavily skewed right

```
# 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.

## 3.2 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 3.3.

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:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}, \text{ where } \mu = \frac{1}{n}\sum_{i=1}^{n}(x_i) \tag{3.2}$$

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.

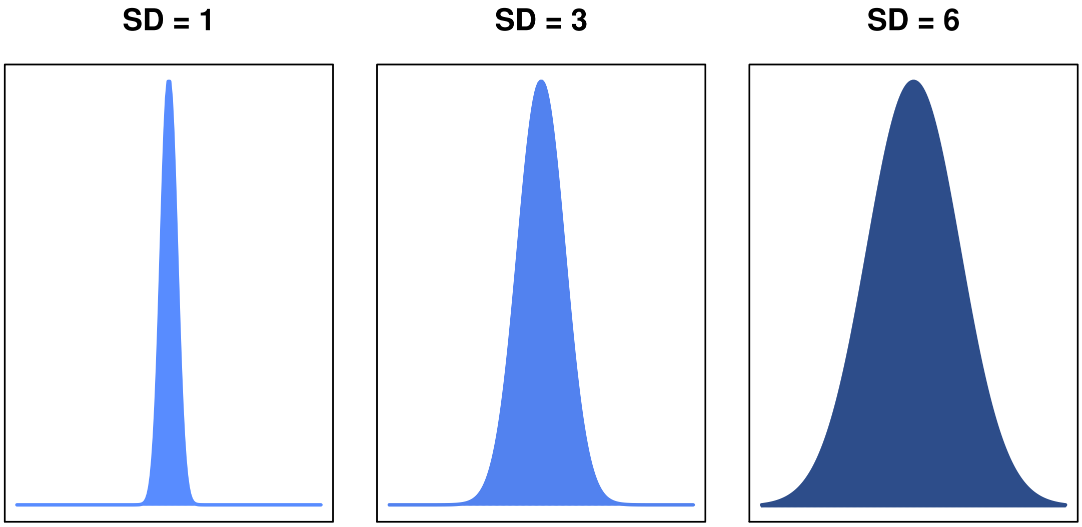


**Figure 3.3** Three samples of data with mean $\mu = 0$ and standard deviations of 1, 3, and 6, respectively.

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:

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

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 65-97-99.5% rule.

This rule tells us that 65% 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 3.4:

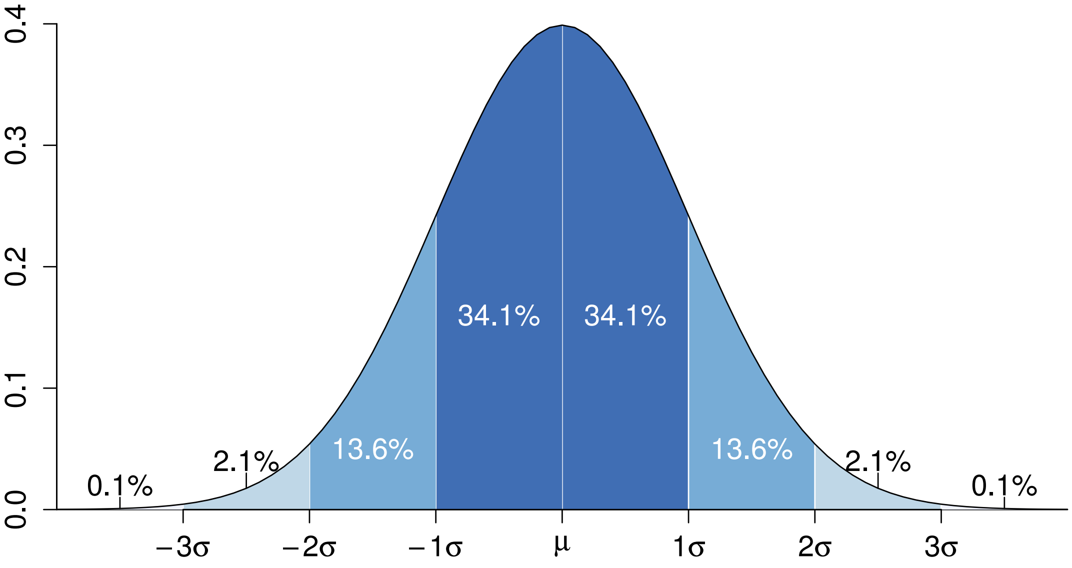


**Figure 3.4** The 68-97-99.5% rule states that nearly all observed data points will be within 3 standard deviations of the mean. This is often useful for flagging outliers or inappropriate distributions to fit your data to. Source: Wikipedia

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$^2$.

### Interquartile Range

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

| | |
|---|---|
| **Minimum** | The smallest value in your data set |
| **Q1** | The value in your data set that is larger than 25% of all other values |
| **Median** | The median value in your data set |

| | |
|---|---|
| **Q3** | The value in your data set that is larger than 75% of all other values |
| **Maximum** | 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.)

## 3.3 Sampling Distributions

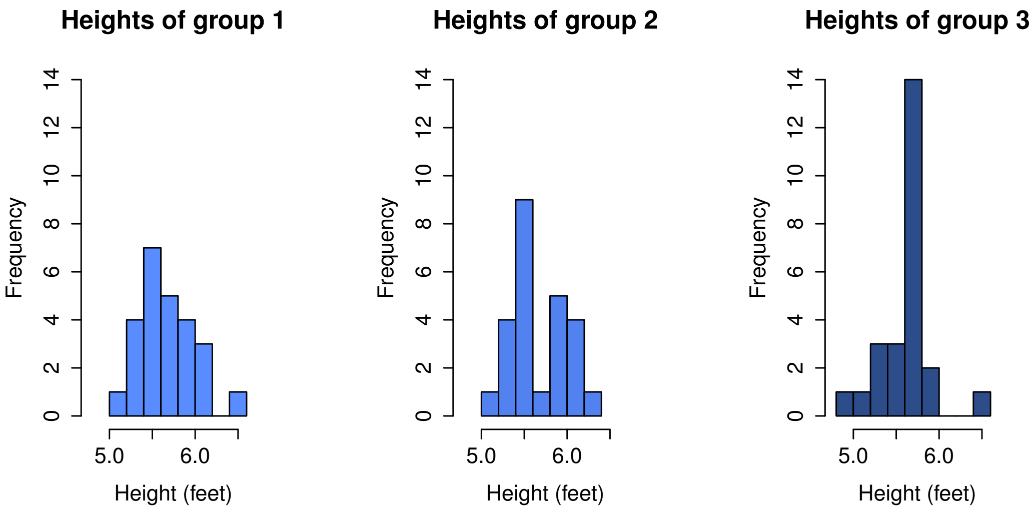
### Central Limit Theorem and Sampling Distributions

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 3.5.

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 3.6.


Heights of group 1

Heights of group 2

Heights of group 3

**Figure 3.5** Heights of three randomly-selected groups of men

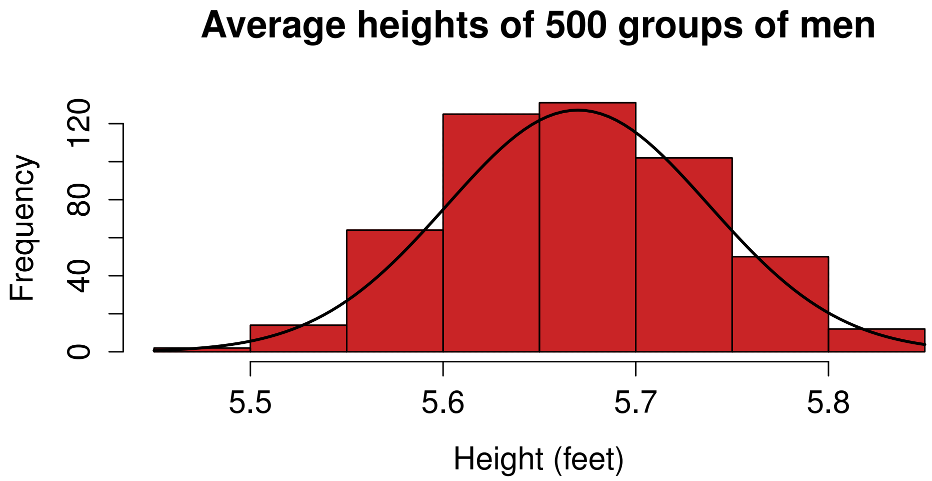

Average heights of 500 groups of men

**Figure 3.6** 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:

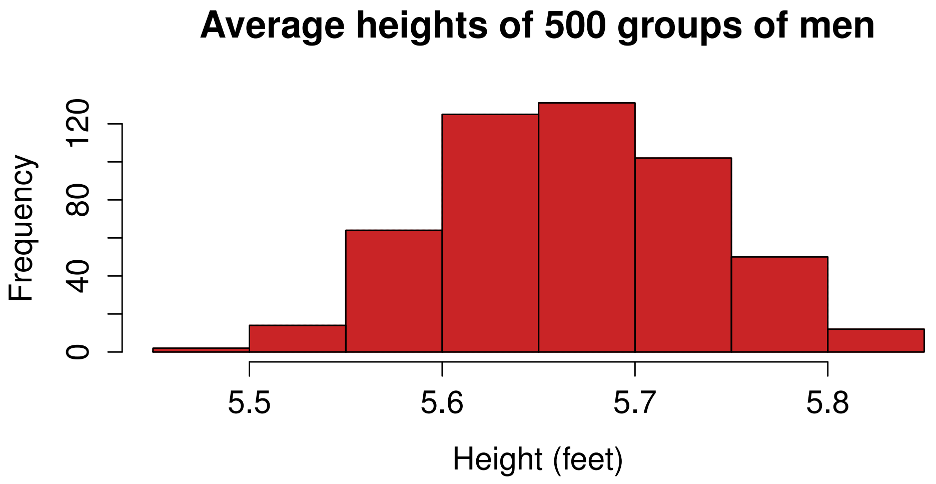
$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3.3}$$

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.

## 3.4 Visualizing Data

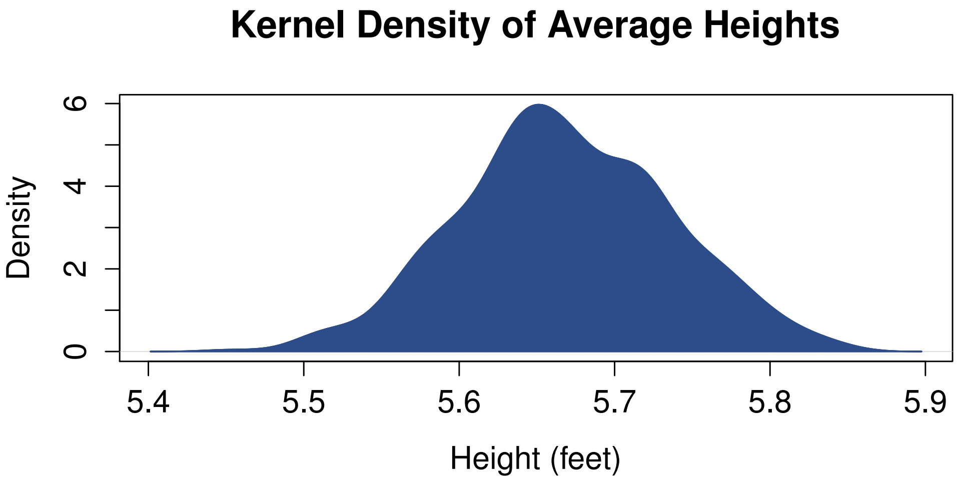
### Histograms

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:

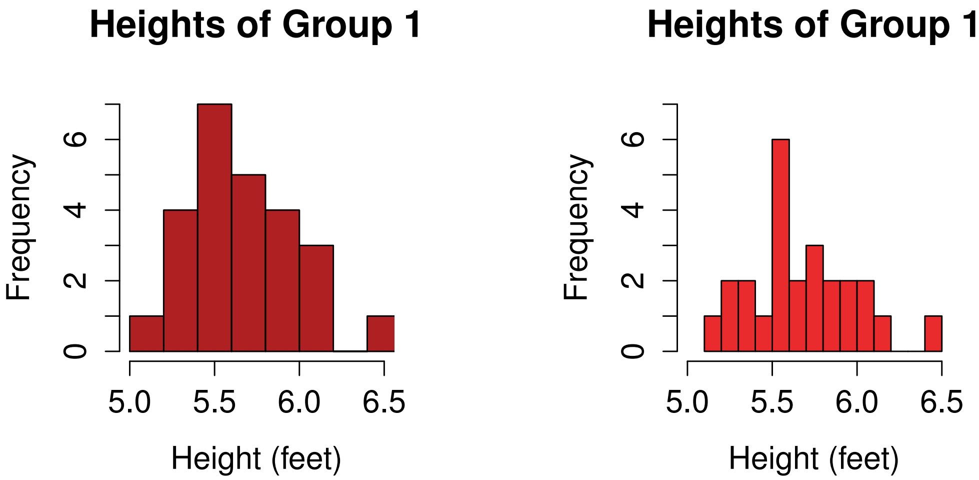


### Kernel Density Plots

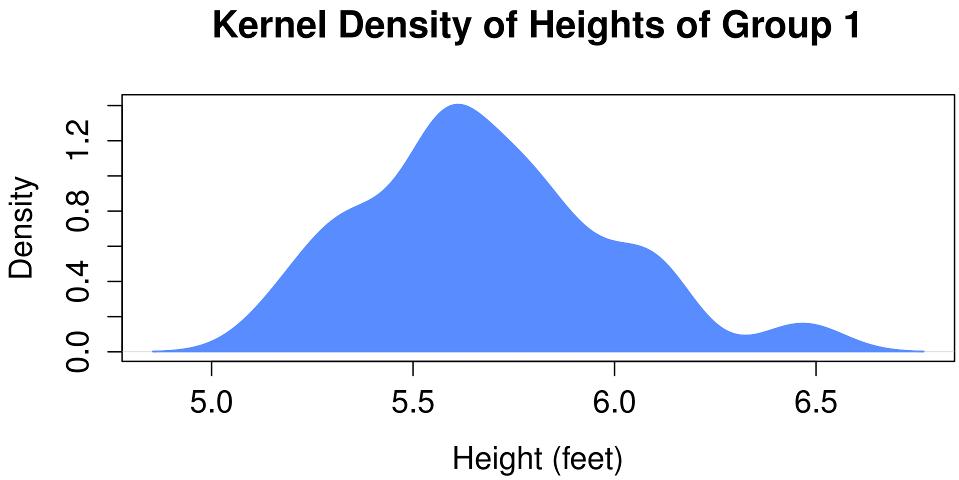
A kernel density plot is similar to a histogram in that it shows us about what proportion of our sample is likely to be at any certain value. The big difference between kernel density plots and histograms is that density plots are continuous: whereas histograms take all of your data and lump them together 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:



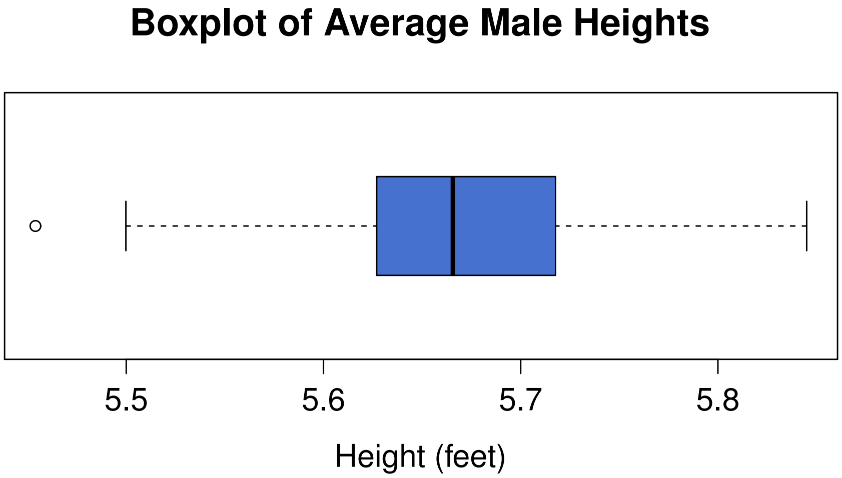
As the range of each bin varies, we can go from a nice-looking normal distribution to a much worse looking non-normal distribution. But with a kernel density plot, you remove those issues in interpretation:



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:



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.

## 3.5 Exercises

**3.1** 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))
}
```

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

**3.3** 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?

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

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

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

## 3.6 Additonal Resources

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

# CHAPTER 4

# RESEARCH DESIGN

## 4.1 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) lets 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).

### 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 4.1).

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!

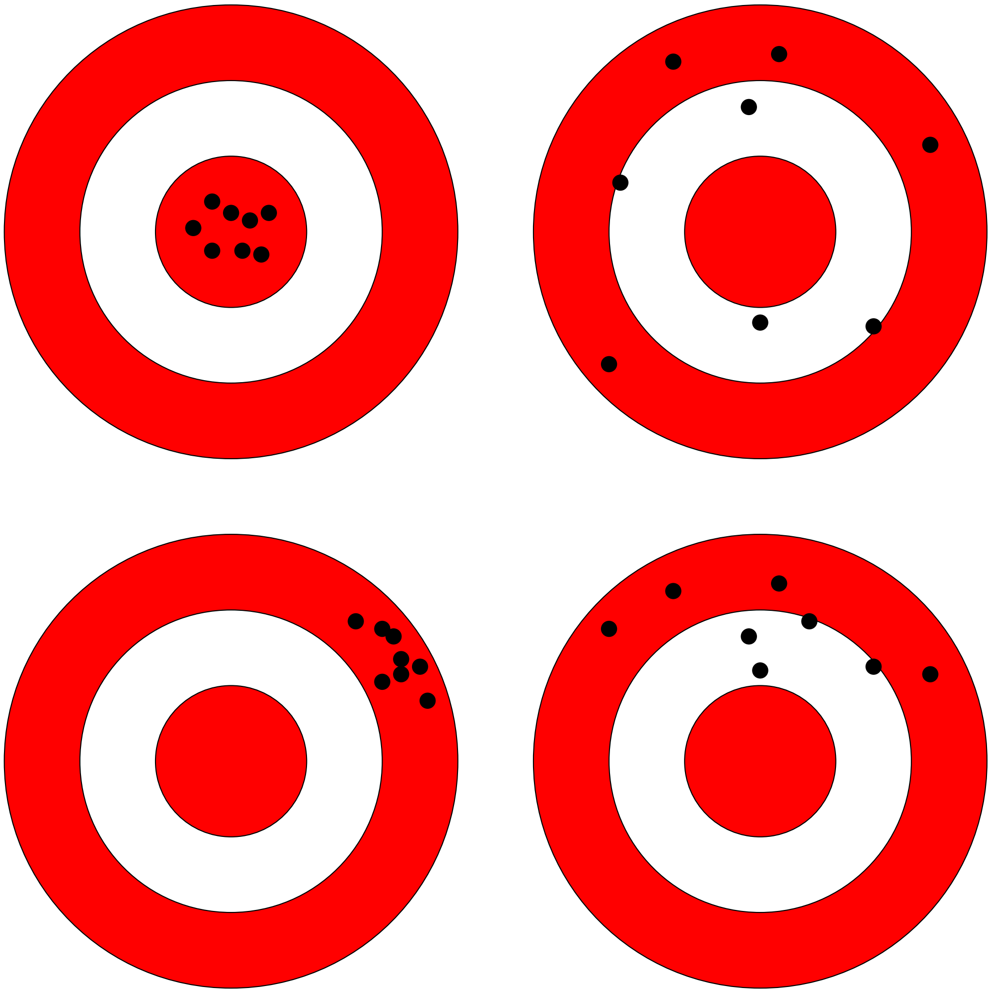


**Figure 4.1** Accuracy and precision in a game of darts. High accuracy and high precision (upper left); high accuracy and low precision (upper right); low accuracy and high precision (lower left); low accuracy and low precision (lower right). NOTE: The upper right circle is "accurate" because each of the darts **averages out** to near the bullseye.

## 4.2 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 (Figure 4.2). And pretty highly.

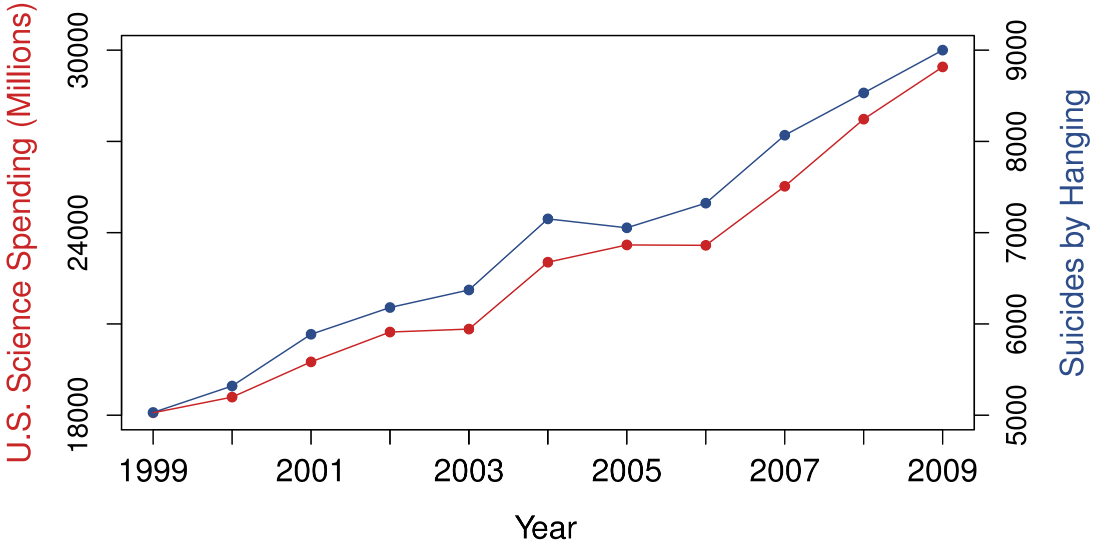


**Figure 4.2** Correlation between U.S. spending on science and suicides by hanging

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; $p$-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 $30^{th}$ 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 must 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 5

# INTRODUCTION TO STATISTICAL INFERENCE

## 5.1 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

***Standardized Test Statistics*** 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.

***Degrees of Freedom*** Degrees of freedom is one of those concepts that (at least in the social sciences) is mentioned briefly and then promptly forgotten for the rest of your career. And that's darned unfortunate, because degrees of freedom is actually an important concept and fairly nuanced to explain to a non-mathematical audience (and frankly deserving a better treatment than we give it here).

A basic definition of the concept is the "number of observations minus the number of necessary relations among these observations" (via Walker, H.. In other words it's the total number of observations in your sample minus the number of parameters being estimated.

The thinking behind this goes that your sample data can be used to estimate one of two things: the variance of the population or a parameter (e.g., a piece of data can be used to estimate the population mean or standard deviation, but not both). Now, generally, each parameter being estimated takes up only one degree of freedom with the rest left to estimate variance. That doesn't always hold true, such as when using Welch's *t*-test, but it's close enough for our purposes.

Again, we can't emphasize enough how much better an explanation this concept deserves and we strongly recommend that you take an afternoon to read up on it. Gerard Dallal's Degrees of Freedom might be a good place to start!

***Probability Values*** Accompanying the 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.

Important to keep in mind (really! This is important) is that a *p*-value is **not** a posterior probability of the truth of the null hypothesis; rather, it **is** a probability conditional on the null. What does that mean? Well, it's the difference between saying "Given our data, there is a 5% chance that the null hypothesis is true" and "Given that the null hypothesis were true, there is only a 5% chance that our data would look the way they do." See the difference there? (Not to belabor the grammar, but the second statement is careful to use the subjunctive, the idea being "if we assume this to be true although it might not actually be".)

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.

All told, *p*-values are picky and noisy statistics that are hard to compare across studies. If you're interested in reading more about cautions in their interpretation, try this article by Gelman for starters.

***Confidence Intervals*** 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.

## 5.2 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., $p$-value $< 0.05$) and rejects the null hypothesis when there is actually no evidence for doing so. If you recall from above, $p$-value $= 0.05$ 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, $\beta$.

## 5.3 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<br>$H_A$ : boy height $\neq$ girl height |
| Fundamental Statistics | Average boy height = 4'10"<br>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 $p$-value $= 0.05$, 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 **is** 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

# COMPARISONS AMONG MULTIPLE SAMPLES

# CHAPTER 6

# *T*-TOOLS

## 6.1 Overview of *t*-Tools

### One-Sample *t*-Test

The one-sample $t$-test is one of the simplest entry points to inferential statistics there is. Simply put, it seeks to determine if the average value of a set of data differs significantly from some theoretical quantity. Say, for instance, we are working in a lumber mill and it looks like a shipment of Douglas fir trees is a bit smaller than usual. We know that a typical Douglas fir from this region is about 225 feet tall and has a radius of 7.5 feet, giving us a total volume of about 39000 cubic feet.

So we take the measurements for each of the trees in this shipment and find that the average volume is $36500$ cubic feet with a standard deviation around $\sigma = 2000$ cubic feet. Do we have sufficient proof that these trees are actually smaller than our typical shipment?

Well, unfortunately, the human brain isn't great at intuitively figuring out if differences like that are meaningful or within the expected margin of error: that's where inferential statistics come into play. So let's first state our hypotheses as:

$$\begin{aligned} H_0 : &\quad \bar{x} = \mu \\ H_A : &\quad \bar{x} \neq \mu \end{aligned}$$

where $\bar{x}$ is the mean volume of our shipment of trees and $\mu$ is the mean volume of adult Douglas firs. Now, we can conduct our statistical test.

A *t*-test, speaking generally, takes the form:

$$t = \frac{\text{Estimate} - \text{Parameter}}{SE(\text{Estimate})} \tag{6.1}$$

The specifics of what you plug in for the estimate and the parameter will vary a bit depending on which type of *t*-test you use; however, they all are based off of this.

In our case, the one-sample *t*-test will look like:

$$t = \frac{\bar{x} - \mu}{s_x} \tag{6.2}$$

where each of the variables is what we defined in our hypotheses above. Plugging in our data we get:

$$\begin{aligned} t(124) &= \frac{36500 - 39000}{2000/\sqrt{125}} \\ &= -13.98 \end{aligned}$$

Although we have our standardized *t*-statistic, we still don't know if it reaches a level of significance or not. historically, we would have to go to a table of $p$-values and look up the one that was closest to our *t*-statistic given the degrees of freedom (here, 124, or $n - 1$). Thankfully, we now have software that is able to compute this value for us much more precisely and quickly. In this case, we end up with a $p$-value $< 0.001$, showing that the result is highly significant.

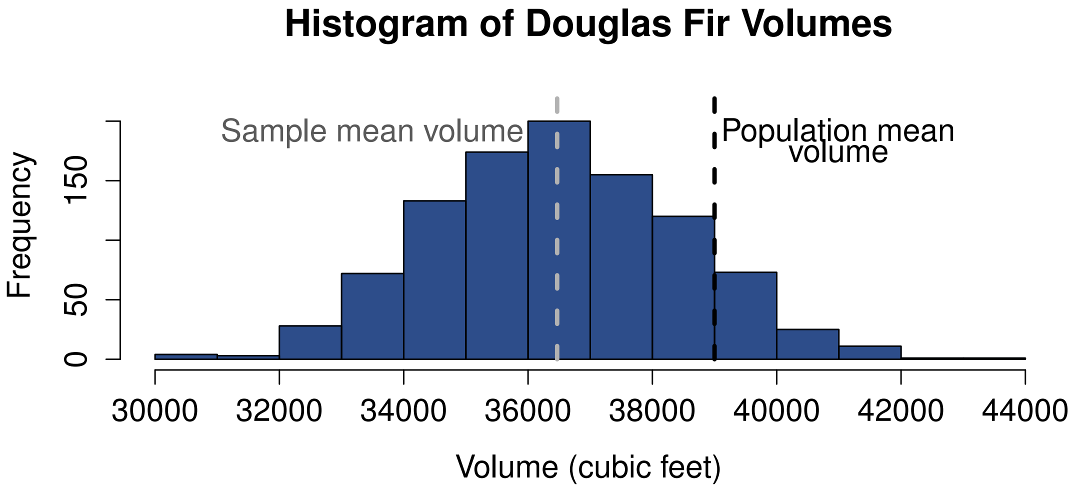


**Figure 6.1** Histogram of the volumes of a shipment of Douglas firs with mean volume 36500 cubic feet and standard deviation $\sigma = 2000$ cubic feet. A one-sample *t*-test shows that this shipment of firs is significantly smaller than the typical shipment (mean 39000 cubic feet), $t = -13.98$, $p$-value $< 0.001$.

Moreover, we can infer the direction of this difference. Remember that the *t*-statistic is computed by subtracting our estimated mean from our population mean: given this, a negative *t*-statistic means that the sample mean is less than the population mean whereas a positive *t*-statistic shows that our sample mean is greater than that of the population.

So, we can ultimately conclude that this shipment of Douglas firs is statistically significantly below the average size of a shipment, given $t(124) = -13.98$ with a $p$-value $< 0.001$. We can summarize this difference in Figure 6.1.

### Paired-Samples *t*-Test

So now that we're familiar with a one-sample *t*-test, the next two tests in the suite become really easy to understand. In a one-sample *t*-test, we're testing one sample of data against a mean that we already know; with a paired-samples test, we extend that to two samples of data that are usually matched along some criteria. For instance, we might be looking at the difference in grieving patterns between mothers and fathers; or personality differences between heterozygous twins; or blood pressure of clinic patients before and after a given drug treatment.

In every case, our two variables are somehow *related* to one another and for every data point in column A there is a corresponding data point in column B. For instance, let's say that we work at a large health clinic and we're testing a new drug, Procardia, that's meant to reduce hypertension. We find $1000$ individuals with a high systolic blood pressure ($M = 145$mmHg, $SD = 9$), we give them Procardia for a month, and then measure their blood pressure again. We find that the mean systolic blood pressure has decreased to $138$mmHg with a standard deviation $8$. We can visualize this difference in Figure 6.2.

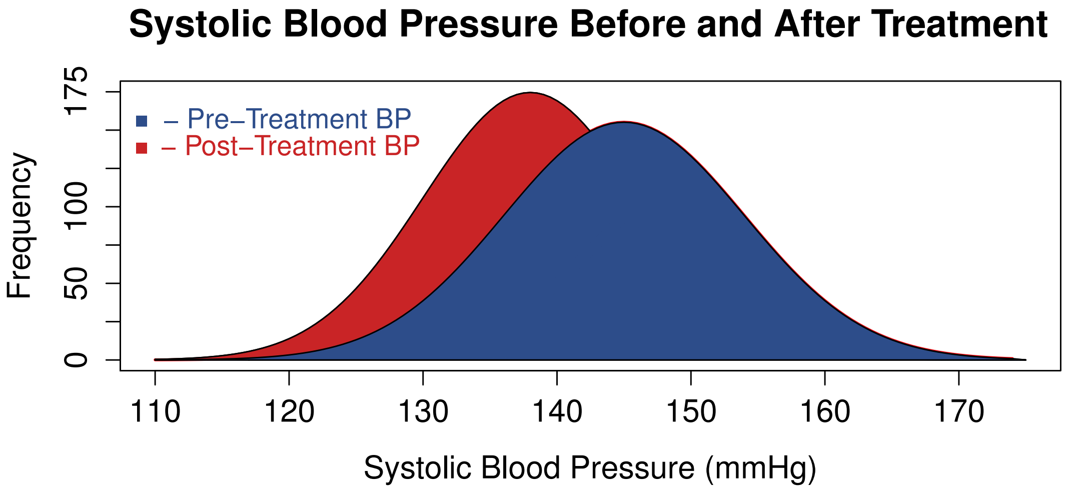


**Figure 6.2** Systolic blood pressures before and after treatment with an antihypertensive drug.

So, we know that these data are paired: we used the same individuals for both pre-treatment and post-treatment and there are $1000$ observations for each column of data. We know that the pre-treatment and post-treatment mean systolic blood pressures are different, but we want to know if this difference is statistically significant. In this case, our *t*-test will take the form:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{x_1 - x_2}}, \text{ where } s_{x_1 - x_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.3}$$

Plugging in our data, we end up with $t(999) = 19.75$, $p$-value $< 0.001$, indicating that (yay!) our experimental drug does appear to actually reduce hypertension in individuals!

### Independent-Samples *t*-Test

Finally is the independent-samples *t*-test. Functionally, this is nearly identical to the paired-samples test: the only major difference is that it is used when the two samples are not paired or somehow related to one another. We would use this test if we were interested in determining, for example, whether there was a difference between depression scores of outpatients at two different clinics; or whether New Yorkers and Clevelanders spend different amounts of money each month eating out; or do two strains of a bacteria respond differently to an antimicrobial gel?

In each of these examples, there is no pairing: New Yorkers are an entirely different population from Clevelanders; the two strains of bacteria have been isolated from one another in a lab for generations. There aren't any obvious commonalities between the two samples.

Now, with this test, choosing the formula to use gets a little complicated. In fact, there are three that we can choose among. If both samples have an equal number of observations and equal variances, we use:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{x_1x_2} \cdot \sqrt{\frac{2}{n}}}, \text{ where } s_{x_1x_2} = \sqrt{\frac{1}{2}(s^2_{x_1} + s^2_{x_2})} \tag{6.4}$$

Alternately, if our samples have an unequal number of observations, but equal variances, we will use:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{x_1x_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } s_{x_1x_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{6.5}$$

Or, finally, we can use Welch's *t*-test when the population variances are not assumed to be equal:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}, \text{ where } s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.6}$$

Woah, that's a lot of choice. So, how do we figure out which to use, exactly? Well, it's not strictly speaking wrong to always default to using Welch's test assuming unequal variances. It's maybe not something that you necessarily *should* always do, but it's better than using Equation 6.4 or 6.5 when the variances aren't actually equal. That said, we'll detail how exactly to determine which of these three tests to use in both the Equality of Variances and Implementation in R sections.

## 6.2 Cautions and Considerations

### Assumptions

The suite of *t*-tools assumes that:

1. the test variable is normally distributed;
2. the variances of the test variable are equal between the two populations;
3. the cases represent a simple random sample of the population; and
4. the scores on the test variable are independent of one another.

### Equality of Variances

As we mentioned above, the suite of *t*-tools largely assumes that the variances of the two samples are equal. However, when this is not actually the case, unless using Welch's test, your results can be a bit misleading. This difference arises from how each of the different tests computes the standard deviation for the two samples. Assuming unequal variances, the **pooled standard deviation** is calculated.

By this estimate, the standard deviations of the two samples are combined (or pooled). However, these are first **weighted** by sample size. The idea behind this is that a standard deviation for a smaller sample doesn't carry as much importance as a standard deviation from a large sample. (The standard deviation of the large sample should be a better reflection of the population standard deviation.) Accordingly, each sample is weighted by its degrees of freedom.

Yet, if the two samples don't actually have the same standard deviation, this statistic quickly becomes misleading: it's like combining elephants and elephant mice together and calling them the same thing. Pretty obviously, they aren't. And once you start trying to make inferences about them, things are going to get wonky pretty quickly.

So before deciding which type of independent-samples test to run, we will first conduct Levene's test for equality of variances. This operates under the hypotheses:

$$\begin{aligned} H_0 &: s_1 = s_2 \\ H_A &: s_1 \neq s_2 \end{aligned}$$

Without getting into the math behind it, the idea is that if it reaches a level of statistical significance, we can conclude that the two samples are not likely to actually have equal variances and we know to use Welch's test.

### Robustness of *t*-Tools

When we talk about the robustness of a test, we're talking about how well it holds up when we violate the assumptions that it relies on. Generally speaking, a more robust test is, of course, preferable: it allows us to still reach largely valid conclusions when some or all of our assumptions are not met.

When working with *t*-tools, normality is generally not the biggest concern. As long as the sizes of your two samples are sufficiently large the test will remain valid, regardless of departures from normality. However, if both of the samples have the same standard deviation, these tests are moderately affected by long-tailedness (e.g., Figure 6.3).

If our assumption of independence, however, is violated, the *t*-tool is likely to give misleading results. This assumption may be violated when there are cluster effects (samples come from the same cluster) or serial effects(samples are close together in time or space).

### Resistance to Outliers

We mentioned before (back when we introduced uncertainty, comparing the mean and median as measures of central tendency) the idea of resistance to outliers. The idea here is that a statistic or statistical procedure is resistant if the presence of one or a few outliers doesn't have an undue influence on the end result. In the case of means, outliers can significantly skew the result.

Unfortunately, because *t*-tools are based on sample means, they are not resistant to outliers. If outliers are present in your data, it may be more appropriate to either remove them from the dataset or to use a comparable test that does not compare sample means. (These tests are usually called nonparametric tests: we detail some of them in the following chapter.)

## 6.3 Implementation in R

Please note: all data used in these explanations are simulated.

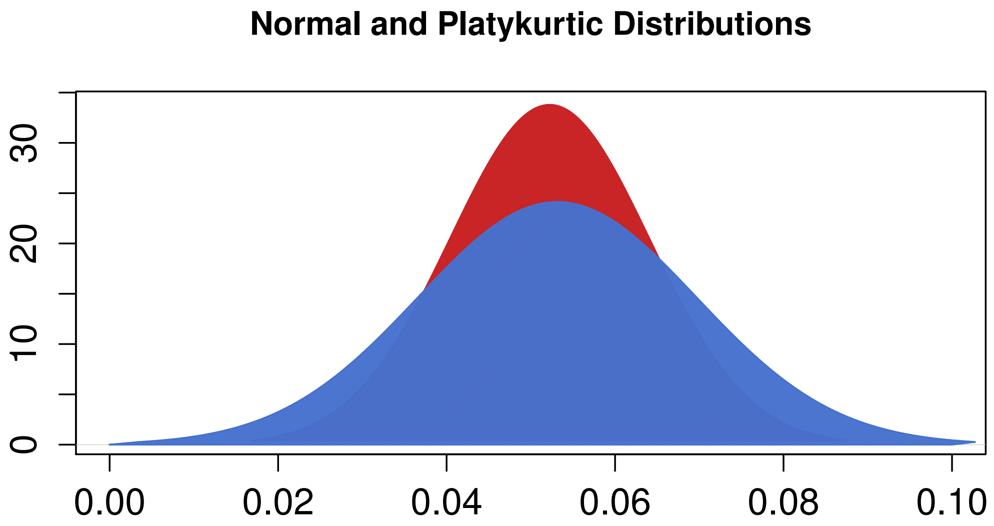


**Figure 6.3** Two samples with the same standard deviation and approximately the same shape. One sample (blue) has longer tails. This moderately impacts the validity of *t*-tools.

### One-Sample

To conduct a one-sample *t*-test in R, we use the syntax `t.test(y, mu = 0)` where `x` is the name of our variable of interest and `mu` is set equal to the mean specified by the null hypothesis.

So, for example, if we wanted to test whether the volume of a shipment of lumber was less than usual ($\mu_0 = 39000$ cubic feet), we would run:

```
# One-sample t-test
set.seed(0)
treeVolume <- c(rnorm(75, mean = 36500, sd = 2000))
t.test(treeVolume, mu = 39000) # Ho: mu = 39000
```

### Paired-Samples

To conduct a paired-samples test, we need either two vectors of data, $y_1$ and $y_2$, or we need one vector of data with a second that serves as a binary grouping variable. The test is then run using the syntax `t.test(y1, y2, paired=TRUE)`

Going back to our example above using pre-treatment and post-treatment systolic blood pressures, our test would take the form:

```
# Paired-samples t-test
# where y1 & y2 are numeric
set.seed(2820)
preTreat <- c(rnorm(1000, mean = 145, sd = 9))
postTreat <- c(rnorm(1000, mean = 138, sd = 8))
t.test(preTreat, postTreat, paired=TRUE)
```

### Independent Samples

The independent-samples test can take one of three forms, depending on the structure of your data and the equality of their variances. The general form of the test is `t.test(y1, y2, paired=FALSE)`. By default, R assumes that the variances of `y1` and `y2` are unequal, thus defaulting to Welch's test. To toggle this, we use the flag `var.equal=TRUE`.

In the three examples shown here we test the hypothesis that Clevelanders and New Yorkers spend different amounts monthly eating out. The first example assumes that we have two numeric vectors: one with Clevelanders' spending and one with New Yorkers' spending. The second example uses a binary grouping variable with a single column of spending data. (That is, there is only one column of spending data; however, for each dollar amount, the next column specifies whether it is for a New Yorker or a Clevelander.) Finally, the third example assumes that the variances of the two samples are unequal and uses Welch's test.

```
# Independent-samples t-test
# where y1 and y2 are numeric
set.seed(0)
ClevelandSpending <- rnorm(50, mean = 250, sd = 75)
NYSpending <- rnorm(50, mean = 300, sd = 80)
t.test(ClevelandSpending, NYSpending, var.equal=TRUE)

# [...] where y1 is numeric and y2 is binary
spending <- c(ClevelandSpending, NYSpending)
city <- c(rep("Cleveland", 50), rep("New York", 50))
t.test(spending˜city, var.equal=TRUE)

# with equal variances not assumed
t.test(ClevelandSpending, NYSpending, var.equal=FALSE)
```

## 6.4 Case Study: Medicare Spending on Patients

The data used in this example were taken from *Medicare hospital spending per patient* dataset, published by the Centers for Medicare and Medicaid Services, an agency of the Department of Health and Human Services.

Let's say that we're moving from New York to California and are interested in seeing how average Medicare spending per patient varies between the two states. We will first state our null and alternative hypotheses:

$$\begin{aligned} H_0 &: \mu_{\text{CA}} - \mu_{\text{NY}} = 0 \\ H_A &: \mu_{\text{CA}} - \mu_{\text{NY}} \neq 0 \end{aligned}$$

We're assuming that there is actually no difference in the mean Medicare spending per patient between the two states and are testing the hypothesis that a difference *does* exist. We aren't specifying the direction of the difference (i.e., we aren't specifying which state we think will have a higher mean spending rate).

We can also visualize the spending patterns of the two states in Figure 6.4. Here, we see the Medicare spending per patient in each of the two states as a proportion of the national average spending per patient. (E.g., 1 represents the national average spending; 0.9 represents 90% the national average spending, etc.)

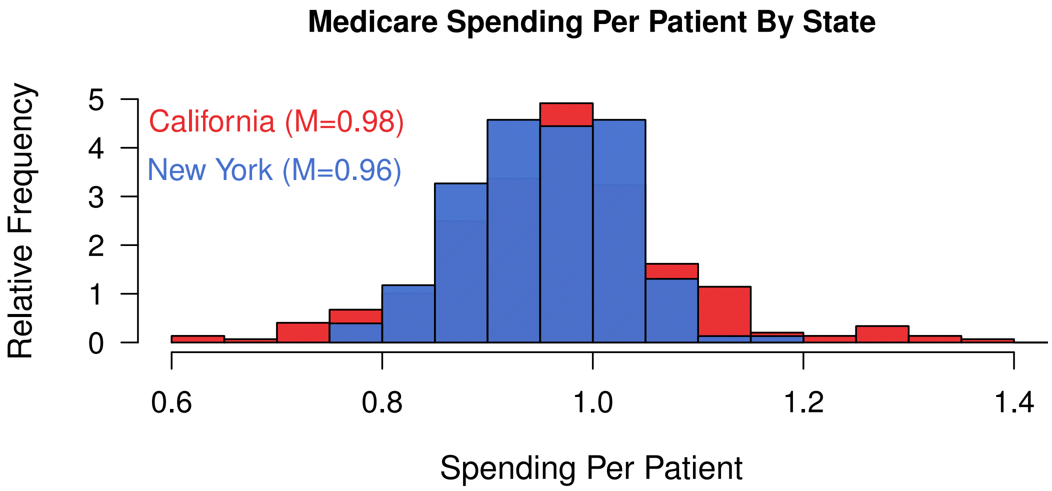


**Figure 6.4** Medicare spending per patient in New York and California. Rates of spending are reported as a proportion of the national average Medicare spending per patient.

Next, we conduct Levene's test to see if we can assume equal variances or not:

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   1  11.703 0.0006811 ***
      448
```

We see that there is significant evidence, $F = 11.70$, $p$-value $< 0.001$, that the variances of the two samples differ. Given this, we will go ahead and use Welch's test to evaluate our hypotheses. Doing so gives us the result:

```

        Welch Two Sample t-test

t = 1.8277, df = 430.923, p-value = 0.06828
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
 -0.001252522  0.034494749

sample estimates:
mean in group CA  |  mean in group NY
       0.9762290  |         0.9596078
```

So, here we have an interesting result: it looks like the two states *might* have different mean spending levels per patient, $t = 1.828$, $p$-value $= 0.0683$; however, this doesn't quite reach our level of statistical significance. Moreover, if we look at the 95% confidence interval provided, we see that the difference between the two population means is likely between -0.001 and 0.034, meaning that if there even is a difference between the two states, at most it's a difference of 3% in per-patient spending relative to the national average.

Given this, we can conclude that although there may be a difference between the average per-patient Medicare spending of California and New York, that difference does not quite rise to the level of statistical significance. Further, even if there were a significant difference, there would not be a practical difference, as the two states are not likely to vary in spending by more than 3% of one another.

## 6.5 Exercises

## 6.6 Additional Resources

# CHAPTER 7

# ALTERNATIVES TO *T*-TOOLS

## 7.1 What Are Nonparametric Statistics?

Nonparametric statistics are those that are not based on paramaterized probability distributions. Did that lose you? That's okay if you're scratching your head.

The idea is that most test rely on parameters such as mean, variance, etc. (Thus, this family of procedures is known as parametric statistics.) When we're talking about nonparametric statistics, we mean that we aren't making assumptions about the data belonging to any particular distribution.

Let's take *t*-tools from the previous chapter: those rely on sample means taken from two populations that we assume to be normally distributed. If those populations don't follow this distribution, our test can fall apart and give us misleading results.

With nonparametric tests, our data don't need to follow any particular distribution: our statistical tests are agnostic to this. We are able to achieve this break from dependence on probability distribution in one of two ways: through distribution-free statistical methods and through nonparametric statistics.

In either case, nonparametric tests will usually operate on ordinal data: that is, data that are numeric, but that don't have any true 0 or necessarily equal spacing between items. For instance, if we were administering a health survey and asked respondents, ”How would you rate your current health?”

Although the options might be “excellent,” “good,” “average,” “poor,” or “one foot in the grave,” for research purposes, they would be coded 1 through 5, respectively. Although we know that 1 means the respondent thinks he or she is in excellent health, the number itself doesn't signify anything meaningful: there is no meaningful 0 on this scale, and we don't know if the difference between a 1 and a 2 is the same as the difference between a 2 and a 3.

## 7.2 Overview of Nonparametric Tools for Two Samples

Although there are many types of nonparametric tests, we will only go over one or two right now. If it doesn't look like these are quite right for your needs, check out our chapter later on in the book on working with rank-ordered data: there we give a more general overview of nonparametric tests beyond only the context of a two-sample comparison.

### Man-Whitney U Test

First up is the Mann-Whitney U test, also variously known as the Mann-Whitney-Wilcoxon test, the Wilcoxon rank-sum test, and the Wilcoxon-Mann-Whitney test. Whew, that's a mouthful.

Thankfully, though, the test itself is a bit simpler than deciding what to call it! The basic idea here is that we are going to take the data from our two samples and rand-order them. For example, let's say that we're working in a call center and we launch a campaign to decrease customers' wait time before they are connected to a representative. We measure the wait times both before and after this campaign and want to determine whether there was an impact of the initiative.

To test this, we will (1) combine the data from the two samples; (2) rank all of the values in ascending order; (3) calculate the ranks for ties (by averaging the rank of each tied item); and (4) calculate the sum of the ranks for the first sample (in this example, the wait times before our intervention). Our process might look something like this:

| Condition | Wait time (mins) | Rank | Tied rank |
|---|---|---|---|
| Before | 0.1 | 1 | 1.5 |
| After | 0.1 | 2 | 1.5 |
| After | 0.1 | 3 | 1.5 |
| Before | 0.2 | 4 | 4 |
| Before | 0.3 | 5 | 5.5 |
| Before | 0.3 | 6 | 5.5 |
| After | 0.4 | 7 | 7 |
| Before | 0.5 | 8 | 8 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| After | 17.2 | 152 | 152 |
| After | 17.8 | 153 | 153 |
| Before | 17.9 | 154 | 154 |
| After | 18.0 | 155 | 155.5 |
| Before | 18.0 | 156 | 155.5 |
| Before | 18.1 | 157 | 157 |
| After | 18.2 | 158 | 158 |
| Before | 18.4 | 159 | 159 |
| Sum of ranks for Before Intervention | | | 9743.5 |

So, here we have our indicator variable in the first column, our raw data in the second column, ordered from smallest to largest, and our ranked values in the third and fourth columns. Again, whenever we have a tie, we take the average rank of all tied items at that measurement. Finally, we take all of the ranks for our "Before Intervention" and add them up to get our W statistic.

And, at last, we're ready to go and conduct our fun little statistical test. To do this, we'll use the formula (it's a doozy!):

$$Z_W = \frac{\left|W - \frac{n(m+n+1}{2}\right| - 0.5}{\sqrt{\frac{mn(m+n+1)}{12}}} \tag{7.1}$$

where $W$ is the Mann-Whitney test statistic that we calculated above, $n$ is the size of the first sample, and $m$ is the size of the second sample. Of course, this is only unless there are ties, in which case the equation gets even funkier. But it's all good! Modern statistical packages have you covered: no chance we're about to make you do this all by hand.

The point is, this ends up spitting out a handy $Z_W$ statistic and $p$-value, real similar to a *t*-test like the ones we just ran.

### Wilcoxon Signed Rank Test

Despite the similarity in names, this is actually an entirely separate test from the one that we just ran (although if you didn't already guess it, we're going to be doing some more ranking!).

In this test, we compute what's called a signed rank (so it's a signed rank test! Such creative names. . . ). It's a pretty similar procedure to the previous test, but differs in that we are working with paired values whereas the previous test is more analogous to an independent-samples *t*-test. Given this, we compute a difference score between the pre- and post-test values.

Let's take the example of a high school math test: a teacher wants to find out if a review session helps her students' understanding of the concepts they've covered in the last chapter. To test this, she administers a pretest and computes each student's percentage score, gives the lesson, and then conducts a post-test and computes students' percentage scores again.

For the purposes of this test, we will rank the scores based on the magnitude of their difference. For instance:

| Student | Pre | Post | Difference | Magnitude | Rank | Signed rank |
|---|---|---|---|---|---|---|
| 1 | 78 | 78 | 0 | 0 | — | — |
| 2 | 80 | 78 | 2 | 2 | 1 | 1 |
| 3 | 81 | 84 | -3 | 3 | 2 | -2 |
| 4 | 68 | 72 | -4 | 4 | 3 | -3.5 |
| 5 | 84 | 80 | 4 | 4 | 4 | 3.5 |
| 6 | 90 | 98 | -8 | 8 | 5 | -5 |
| ⋮ |  | ⋮ |  | ⋮ |  | ⋮ |

Here we (1) order the data by the magnitude of their differences; (2) rank those data; (3) compute the average rank for tied values; and (4) include the sign of the original difference (i.e., whether the difference is positive or negative). Then we add up all of the signed rank values to compute an overall $W$ statistic.

Following that, there's another formula that's just as fun as the last one—so fun in fact that we aren't even going to include it here! What's important is that once again you get a $Z$ statistic and $p$-value at the end of it all.

## 7.3 Cautions and Considerations

### Assumptions of the Mann-Whitney Test

This test assumes that:

1. all observations are independent of one another;
2. the data are measured on an ordinal scale;
3. the distributions of both groups are equal under the null hypothesis;
4. the responses are continuous; and
5. the alternative hypothesis is restricted to a shift in location (i.e., $F_1(x) = F_2(x + \delta)$).

### Assumptions of the Wilcoxon Signed-Rank Test

This test assumes:

1. data are paired and come from the same population;
2. each pair is chosen randomly;
3. each pair is independent of every other pair; and
4. the data are measured on an ordinal scale.

## 7.4 Implementation in R

Please note: all data used in these explanations are simulated. For both tests, the same function, `wilcox.test(x,y)` is called. One important tidbit to remember is that the default function does not handle ties gracefully. If you have any ties in your data, you should instead use `wilcox_test()` in the `coin` package. We detail the use of both below.

### Mann-Whitney Test

To specify that a Mann-Whitney test should be conducted, we set the `paired` flag to `FALSE`:

```
set.seed(0)
before <- c(rnorm(100, 4.9, 3.8))
after <- c(rnorm(100,3.8, 2.9))

wilcox.test(before, after, paired = FALSE)
```

### Wilcoxon Signed-Rank Test

This test can be conducted with either two vectors of paired data or a single vector of difference values (i.e., $x_1 - x_2$). For both tests, we indicate that a Wilcoxon signed-rank test should be conducted by setting the `paired` flag to `TRUE`.

```
set.seed(0)
pretest <- c(rnorm(100,70,10))
posttest <- c(rnorm(100,85,7))
difference <- pretest - posttest

# Paired-samples test
wilcox.test(pretest, posttest, paired = TRUE)

# Difference vector or one-sample test
wilcox.test(difference, paired = TRUE)
```

## 7.5 Case Study: Heavy Metal Accumulation

In 2004, Laureysensa, et al. looked at the ability of a number of poplars to resorb heavy metals from the soil in which they were planted. Specifically, they were interested in looking at seasonal variation in these resorption patterns: whether the poplars would more actively take up heavy metals from the soil during certain times of the year than others.

To this end, they looked at the resorption patterns of 13 different cloned poplars for aluminum, cadmium, iron, manganese, and zinc and measured their heavy metal content ($\mu g/g$) in both August and November of the same year. However, due to oddities of the data, a typical parametric test to look at seasonal differences might not be appropriate. Rather, we will use Wilcoxon's signed rank test.

For this study, let's focus only on one of the heavy metals measured: aluminum. We can start by visualizing the distributions of both samples (Figure 7.1).

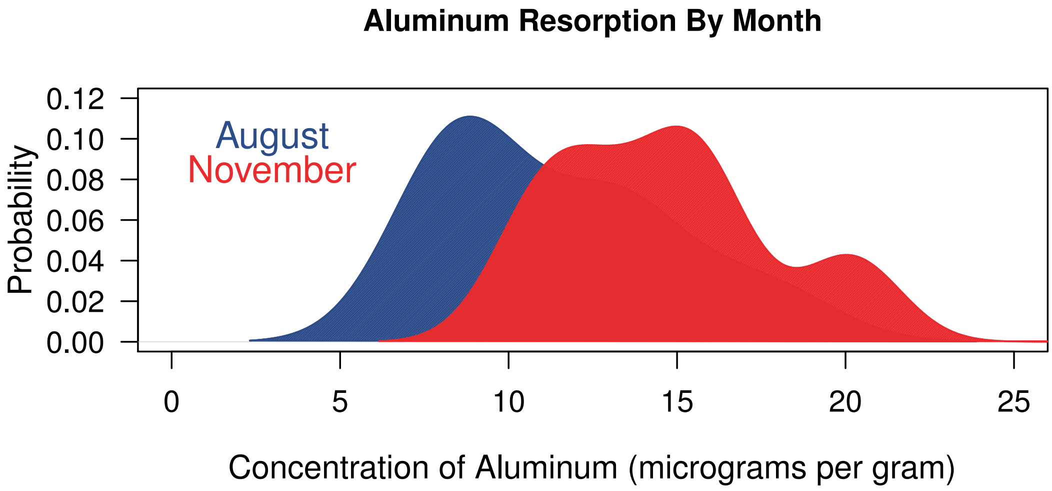


**Figure 7.1**

As we can see, their probability densities differ somewhat from one another and both depart from the normal distribution. Given this, we can assume that a *t*-test may not be most appropriate to judge the differences between these two samples.

Knowing that, we can construct our null and alternative hypotheses for these data:

$$\begin{aligned} H_0 : &\quad \tilde{\mu}_{\text{aug}} = \tilde{\mu}_{\text{nov}} \\ H_A : &\quad \tilde{\mu}_{\text{aug}} \neq \tilde{\mu}_{\text{nov}} \end{aligned}$$

where $\tilde{\mu}$ is the median for each sample. We then conduct a Wilcoxon signed rank test to evaluate these hypotheses:

```

        Wilcoxon signed rank test

data:  august and november
V = 16, p-value = 0.03979
alternative hypothesis: true location shift is not equal to 0

```

This test indicates that the the medians of the two samples do differ to a level of statistical significance, $V = 16$; $p$-value $= 0.040$, indicating that resorption of aluminum among poplars is more efficient during the winter months ($\tilde{x} = 15.3$) than during the fall ($\tilde{x} = 10.0$).

## 7.6 Exercises

## 7.7 Additional Resources

# CHAPTER 8

# ONE-WAY ANOVA

## 8.1 Overview of One-Way ANOVAs

To understand ANOVAs, let's think back once more to $t$-tests. There, we're looking to determine whether there is a difference between the means of two samples of data. But that if we want to look at more than two groups?

Well, using what we've learned up to now, we're stuck: there's nothing more we can do. But of course that's not *really* the case, is it? Otherwise we wouldn't be writing this chapter!

The idea with a one-way ANOVA (**AN**alysis **O**f **VA**riance) is that we have a variable of interest, but three or more levels of our grouping variable. For instance, we might be looking at student test scores as a function of classroom type: regular, honors, AP (advanced placement), or IB (international baccalaureate). That's way beyond the scope of any $t$-test that we've come across. Thankfully, though, the ANOVA is happy to step in to fill the gap.

Before we can actually run the test, however, there is some background information that we need to understand first.

### Fixed and Random Effects

For starters, we should be familiar with fixed, random, and mixed effects models. Broadly speaking, these refer to how the treatment factors are defined within the context of the experiment.

One caveat before we get started, though: there is no single agreed-upon definition for fixed and random effects. Rather, the precise definition is going to depend on the exact context of your experiment and your program of research. For a quick overview of these different definitions (or at least a spattering of them), check out Andrew Gelman's write-up, linked here.

***Fixed Effects*** In a fixed-effects model, there is a basic assumption that the individual-specific effect is correlated with the independent variable(s). The idea here is that person-to-person, there is little variance beyond that caused by the treatment effect.

***Random Effects*** Random effects models assume that there is some inherent amount of individual variance not accounted for by the treatment. So if we were talking about a regression equation with a random intercept $a_i$ and fixed slope $b$, each individual would have his or hew own regression line, parallel to every other individual but with a differing intercept (cf. Figure 8.1).

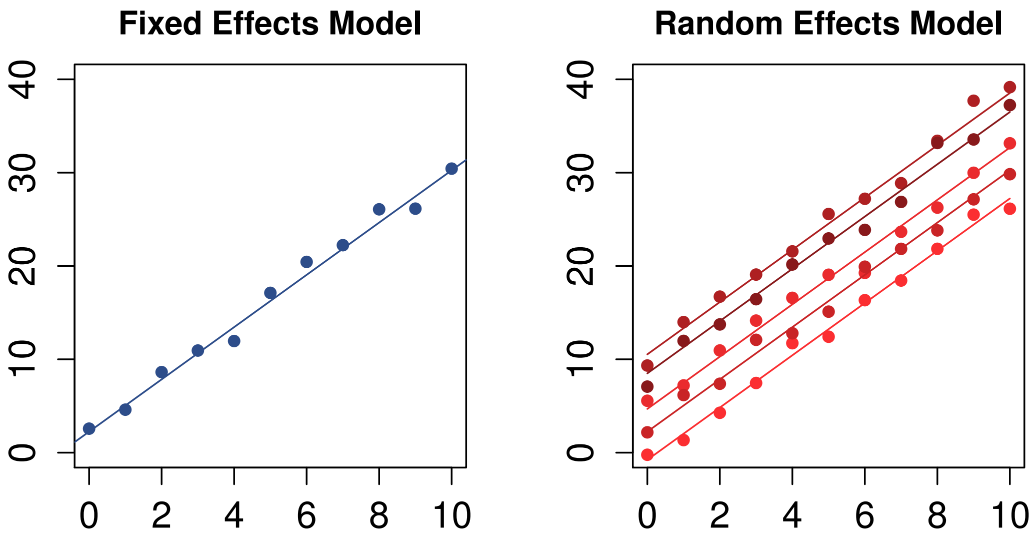


**Figure 8.1** A regression equation assuming fixed-effects factors for all individuals (left) and a random-effects intercept for 5 individuals (right).

### Sum of Squares

Residual, treatment, and total sums of squares are discussed more fully in our chapter on simple linear regression. The basic idea here is that there are two big sources of error: the treatment and the residual. Every time we have an experiment and are taking measurements, there is going to be some amount of background noise inherent in that: the point of statistical modeling is to reduce the amount of noise that we are unable to explain.

As such, there is noise from the treatment (individuals responding differently to the stimuli) and noise from the residual (those random fluctuations that our model is unable to explain). These combine to form a total measure of the error. In the context of an ANOVA, the residual error (termed the residual sum of squares) is equivalent to the variance within each group; the treatment sum of squares equivalent to the variance between groups. Take for instance Figures 8.2 and 8.3.

In Figure 8.2, we are looking at the difference in locations of the two treatment groups relative to one another. A greater distance between the two treatment groups indicates that there exists a more substantial effect of the treatment on patient outcomes. Understandably, larger between-groups variances make it

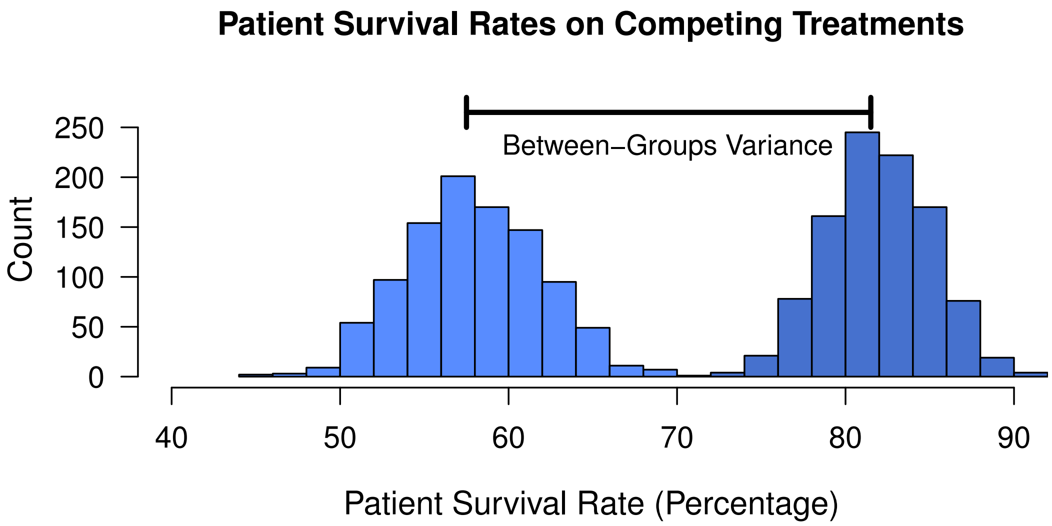


**Figure 8.2** Patient survival rates using two competing treatment regimens. Here, between-groups variance is illustrated, showing the distance between the means of the two treatment groups.

easier for our statistical test to establish the significance of a difference: the less overlap between groups, the more clear the difference becomes.

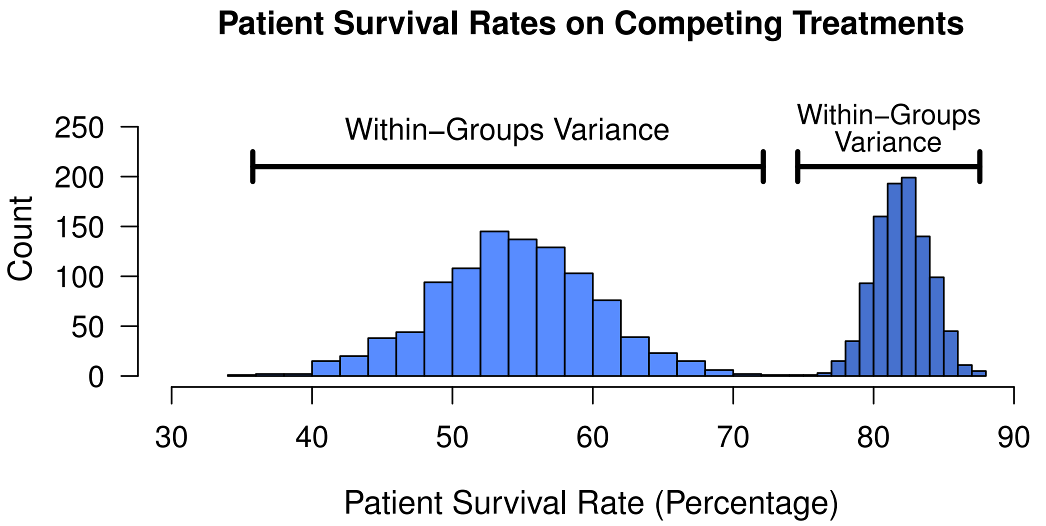


**Figure 8.3** Patient survival rates using two competing treatment regimens. Here, within-groups variance is illustrated, showing the spread of the measurements within each treatment group.

Likewise, in Figure 8.3, we see the variance within groups. The treatment group in light blue has a larger variance than does the treatment group on the right in dark blue. Unlike between-groups variance, here we rather see smaller within-group variances. The reason for this is that a larger within-groups variance means that we need an even **bigger** between-groups variance to compensate (or a much larger sample size): this type of variance obscures the effect of our treatment by introducing additional noise to our model.

Formally, we define these measurements as sums of squares, calculated:

$$\begin{aligned}
SS_{residual} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
SS_{treatment} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2 \\
SS_{total} &= SS_{residual} + SS_{treatment} \\
&= \sum_{i=1}^{n} (y_i - \bar{y}_i)^2
\end{aligned}$$

### The $F$-Test

The $F$-test is used to take these observed variances and give a standardized test statistic allowing us to quantify the magnitude of the difference observed between our treatment groups. It is calculated by:

$$F = \frac{MS_{treatment}}{MS_{residual}} \tag{8.1}$$

The mean square ($MS$) statistics are calculated by:

$$F = \frac{\text{Between-groups}}{\text{Within-groups}} = \frac{MS_{treatment}}{MS_{residual}} = \frac{SS_{treatment}/(I-1)}{SS_{residual}/(n-1)}$$

where $I$ is the number of treatments and $n$ is the total number of observations.

Values of the $F$-statistic above 1 indicate that the observed results are further and further inconsistent with the null hypothesis. However, the exact value of $F$ needed for statistical significance will vary as a function of your degrees of freedom.

### Follow-Up Analyses

#### *Post-Hoc Tests*

#### *Planned Comparisons and Linear Contrasts*

## 8.2 Cautions and Considerations

### Assumptions

This test assumes that:

1. the samples are normally-distributed;
2. the samples are independent;
3. the cases represent a simple random sample of the population; and
4. the variances of the treatment groups are equal.

### Robustness of the Test

## 8.3 Implementation in R

### aov() vs. lme() vs. lmer()

## 8.4 Case Study: [STUDY]

## 8.5 Exercises

## 8.6 Additional Resources

## CHAPTER 9

# MULTIFACTOR STUDIES

9.1 Overview of Multiple ANOVAs

9.2 Cautions and Considerations

9.3 Implementation in R

9.4 Case Study: [STUDY]

9.5 Exercises

9.6 Additional Resources

# CHAPTER 10

# MULTIFACTOR STUDIES WITH REPLICATION

**10.1 Overview of Repeated Measures Designs**

**10.2 Cautions and Considerations**

**10.3 Implementation in R**

**10.4 Case Study: [STUDY]**

**10.5 Exercises**

**10.6 Additional Resources**

# PART III

# DESCRIBING RELATIONSHIPS AND PREDICTING VALUES

# CHAPTER 11

# CORRELATION

## 11.1 Overview of Correlation

### 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 11.1).

***Scatterplots with Categorical Variables*** Additionally, we can use scatterplots to visualize data with categorical indicators. For example, if we were interested in the depression and anxiety scores of patients with mild, moderate, and severe bipolar diagnoses (Figure 11.2).

### 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

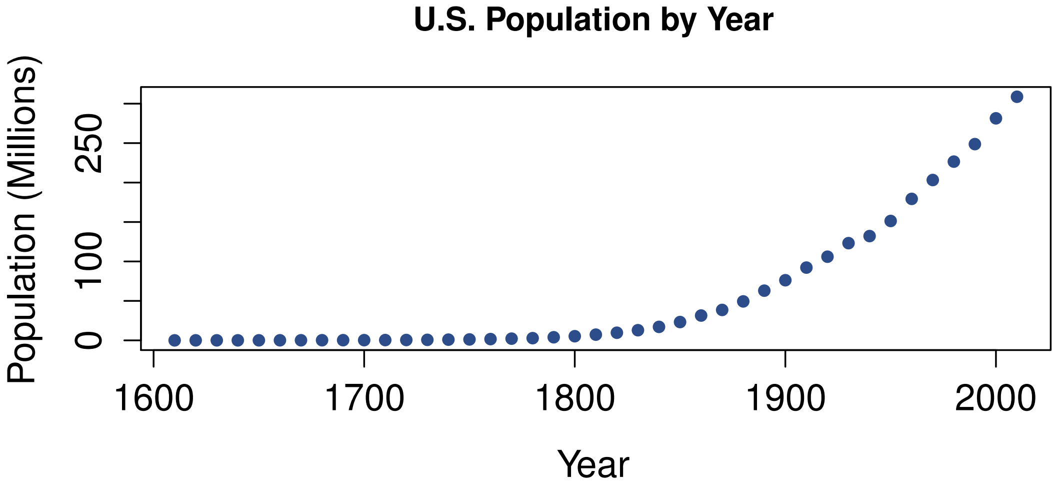


**Figure 11.1** United States population by decade: 1600 through 2010

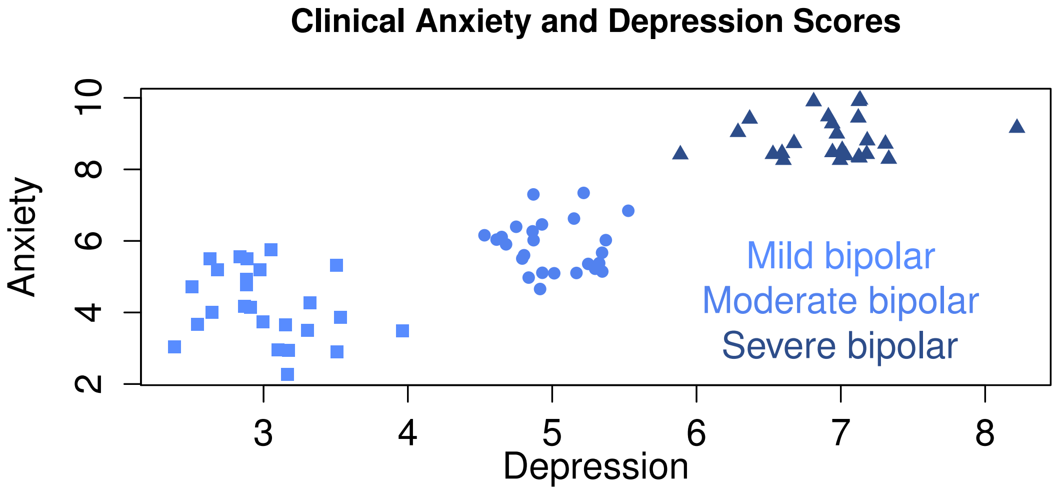


**Figure 11.2** Depression and anxiety scores of patients with bipolar depression

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:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{for a population} \tag{11.1}$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \quad \text{for a sample} \tag{11.2}$$

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:

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

### Correlation Matrices and 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 11.3):

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

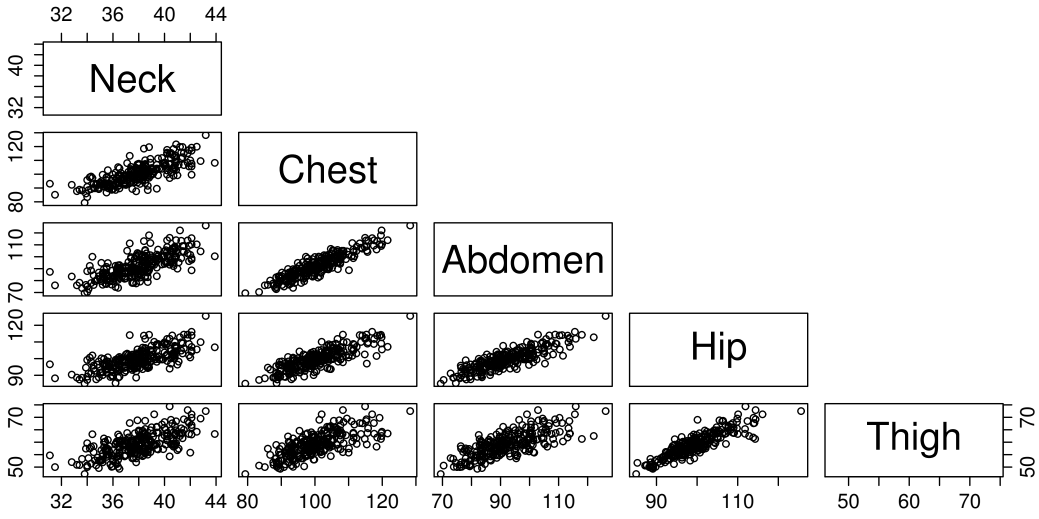


**Figure 11.3** Lower-triangular scatterplot matrix for multiple correlations

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.)

### 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. This approach is generally preferred if there are other known sources of covariance that you also measured.

## 11.2 Cautions and Considerations

### 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 by 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 11.4.

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.

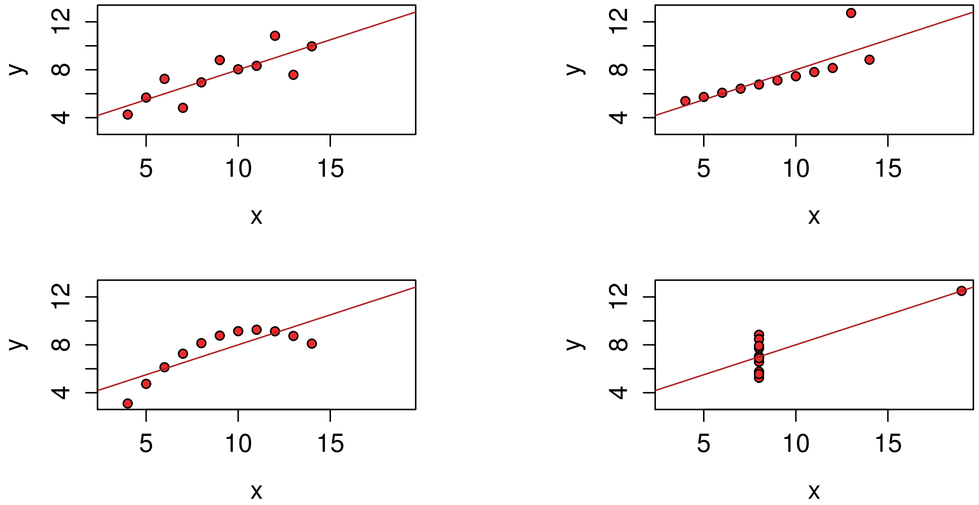


**Figure 11.4** Four separate data sets, each with the same mean and correlation. Source: Wikipedia

### 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.

Take, for instance, Figure 11.5. 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.

## 11.3 Implementation in R

### Correlation

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:

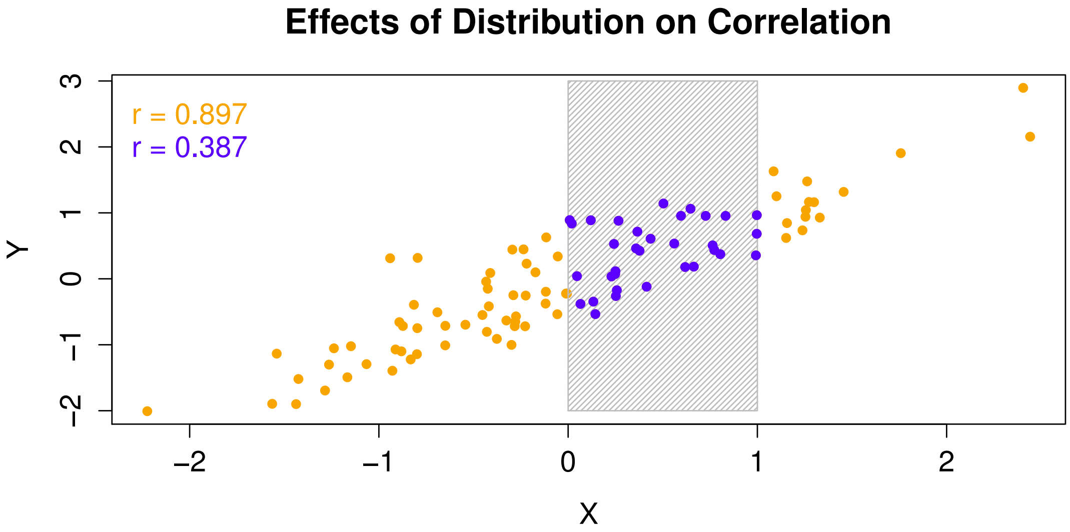


**Figure 11.5** Pearson correlation coefficients between X and Y are shown when the two variables' ranges are unrestricted (orange) and when the range of X is restricted to the interval (0,1) (blue)

```
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, $t = 18.78$; $p$-value $< 0.001$. Because our $p$-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.

## Partial Correlation

To run a partial correlation in R, we will execute something like:

```
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.

## 11.4 Case Study: National Education Trends

These data are taken from the Census' American Community Survey. Here, 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 to have completed high school or higher. So let's start by constructing a scatterplot of the two variables, seen in Figure 11.6:

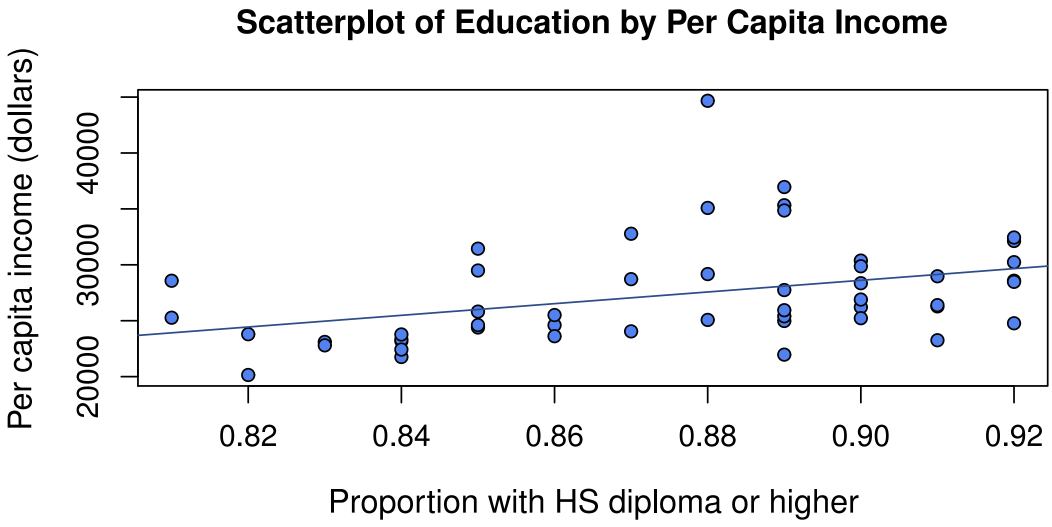


**Figure 11.6** Scatterplot of perCapitaIncome and HighSchoolOrHigher

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, giving us the results:

```
  Pearson's product-moment correlation

data:  perCapitaIncome and 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.

## 11.5 Exercises

**11.1** 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.

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

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

**11.4** 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.

**11.5** 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.

**11.6** **Test yourself:** Choose two new variables and performa 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?

**11.7** 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.

**11.8** 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)?

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

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

**11.11** **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.

## 11.6 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 12

# SIMPLE LINEAR REGRESSION

## 12.1 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:

$$\hat{y} = \beta_0 + \beta_1 \cdot x \tag{12.1}$$

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 12.1, 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 $\hat{\beta}_0$ (pronounced "beta-null hat" and "beta-one hat").

We can visualize this process in Figure 12.1. 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.

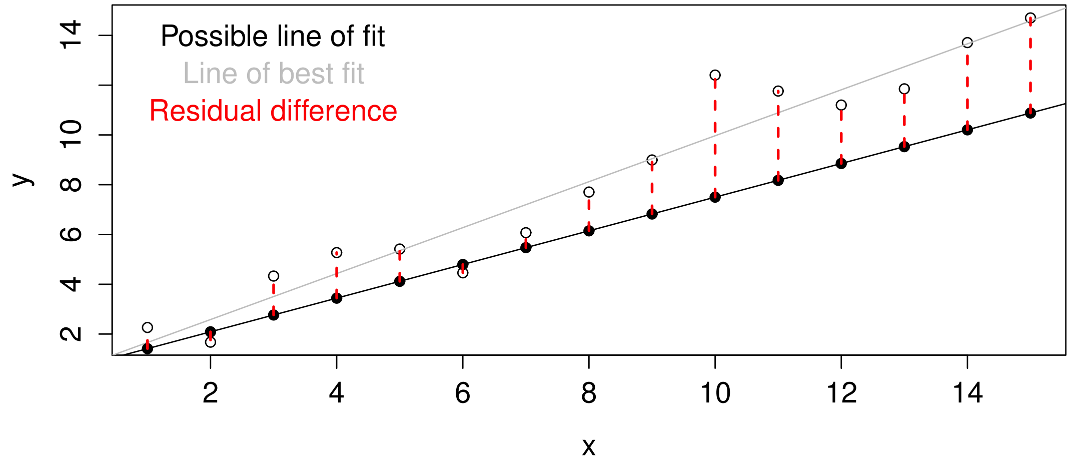


**Figure 12.1** Data with a least-squares line of best fit (grey), one potential line of best fit (black), and the residual difference between the predicted and observed values (red). Lease squares fitting seeks to minimize the overall distance between every observed value and predicted value, resulting 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 12.1. 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:

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

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 12.2 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.

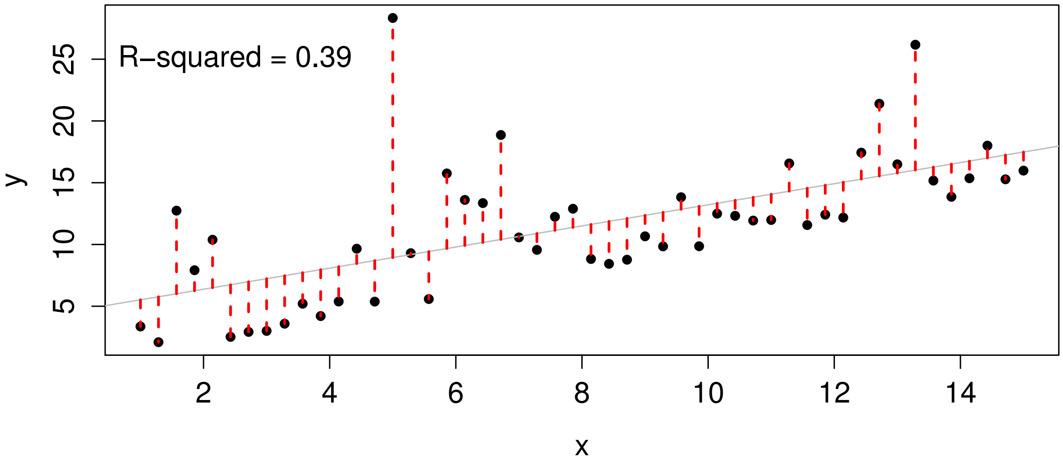


**Figure 12.2** Regression model with approximately 39% of the observed variance explained.

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 12.1 is:

$$\hat{y} = \beta_0 + \beta_1 \cdot x + \epsilon \tag{12.2}$$

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 12.2, the residuals are represented by the dashed red lines. Alternately, we can see a little more explicitly what a residual is by looking at Figure 12.3. This is equivalent to the error term ($\epsilon$) that we mentioned in the previous section.

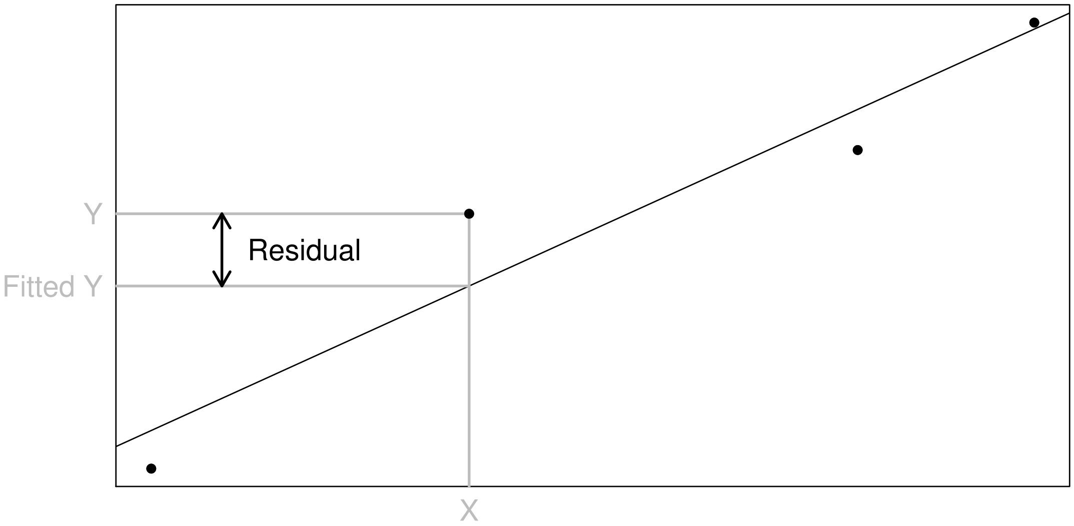


**Figure 12.3** A residual is the difference between an observed data point and its predicted value at a given point $X$. This represents the error that the model is unable to explain.

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

$$R^2 \equiv 1 - \frac{SS_{residual}}{SS_{total}} \tag{12.3}$$

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 12.4. 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:

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

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 ration 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 12.4. 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.

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.

***Residuals in the Context of a Linear Model*** When we construct a regression model, looking at a plot of the fitted values by the residuals can be a good diagnostic tool. (In R we achieve this by calling `plot(resid(fit), fitted(fit)` where `fit` is the linear model that we have created. See our Implementation in R section for details of how to create a linear model.) Figure 12.5 illustrates some common patterns among residual plots.

Here, what we really want to see is something like the graph in the top left—unbiased and homoscedasatic. When this is the case, our residuals will form roughly a rectangle on our plot. It indicates that the model fits our data fairly well. When we start seeing any of the other patterns illustrated, however, that may indicate either that the data need to be transformed or that a different model should be fit differently (e.g., a different type of regression is appropriate).

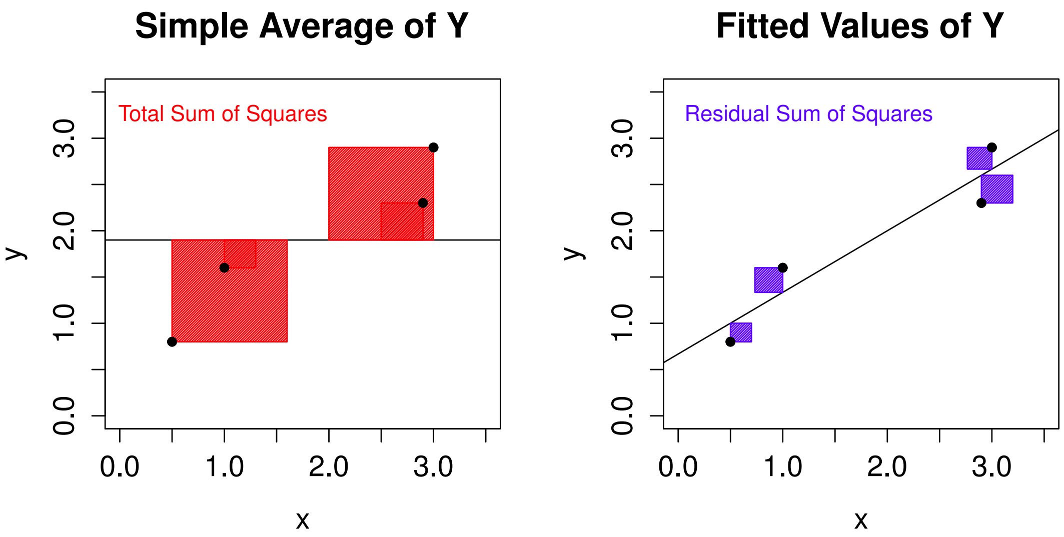


**Figure 12.4** $R^2$ is defined generally as $1 - \frac{SS_{residual}}{SS_{total}}$. Here, we see the total sum of squares (red) as the sum of the squared distance from each observation to the arithmetic mean of $y$ ($\bar{y}$). The residual sum of squares (blue) is computed as the sum of the squared distance from each observation to its predicted value.

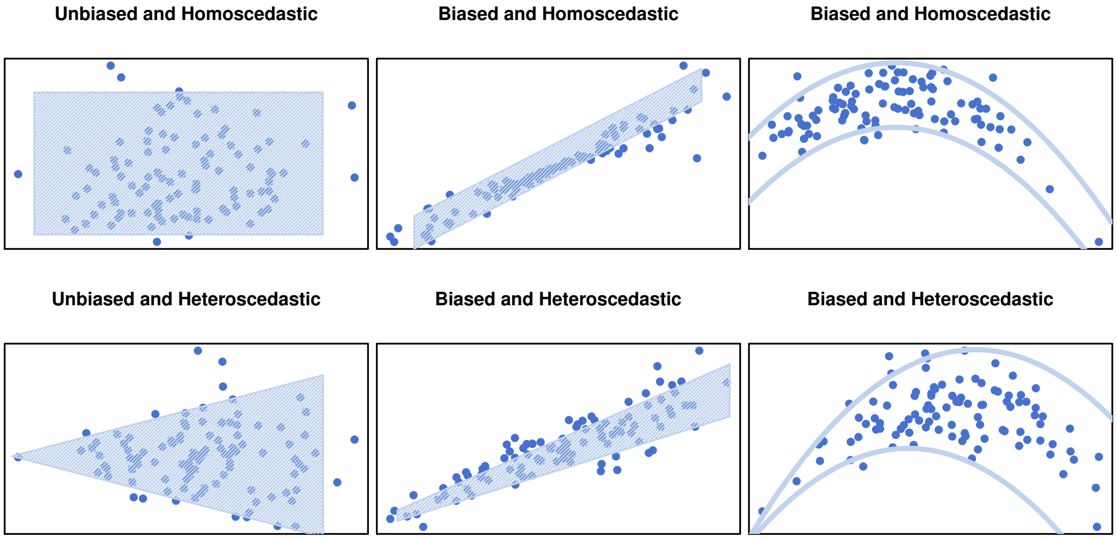


**Figure 12.5** Typical patterns for biased and unbiased, heteroscedastic and homoscedastic residuals. Heteroscedasticity means that not all observations in a sample or population have the same variance and usually indicates that the data should be transformed. A residual plot is biased if the fitted values at any given value of the residual do not average out to be 0.

Unfortunately, the examples given here are all highly stylized: in practice, it's often a much murkier picture that you're left to interpret and make a best guess on. There are some tools to help you that we outline in our chapter on Data Transformations, but by and large this is something that you just need to get a feel for through practice and repeated exposure. Typically, however, bias among your residuals indicates that the model that you chose is not appropriate for your data. Alternately, heteroscedastic residuals usually indicate the need for a data transformation.

## 12.2 Cautions and Considerations

### Basic Assumptions

When running a regression, there are certain conditions that the model assumes are met. These include:

1. The data represent a normally-distributed sample of
2. 
3. 
4.

### Outliers and Influential Points

## 12.3 Implementation in R

The basic command to implement a linear regression in R is `lm(y ˜ x, data = dataset)`. We can display the results of the regression using the `summary()` function. If, for instance, we were working with a dataset titled `demographics` and we wanted to use `yearsEducation` to predict `annualIncome`, we would run:

```
model <- lm(annualIncome ∼ yearsEducation)
summary(model)

Residuals:
    Min      1Q  Median      3Q     Max
-128.41  -35.08   -8.67   34.34  194.76

Coefficients:
                Estimate   Std. Error   t-value   Pr(>|t|)
(Intercept)    917.8839        9.6425    95.191    < 2e-16 ***
yearsEducation   0.4180        0.1166     3.585   0.000691 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.76 on 58 degrees of freedom
Multiple R-squared:  0.1814,        Adjusted R-squared:  0.1673
F-statistic: 12.86 on 1 and 58 DF,  p-value: 0.0006908
```

We can interpret this as a model with an overall significance of $F(1, 58) = 12.86$; $p$-value $= 0.0006$. Further, this one predictor is able to explain 18.1% of the variance observed in the data ($R^2 = 0.181$).

If we wish to assess goodness of fit of the model through a plot of its residuals against its fitted values, we run `plot(fitted(modelName), resid(modelName))`. In this example:

```
plot(fitted(model), resid(model),
     # Specify the labels for the x- and y-axes
     xlab = "Predicted Income",
     ylab = "Residuals",

     # Specify the plot title
     main = "Residuals by Fitted Values",

     # Specify the orientation of the axis values,
     # the type of point to use on the graph,
     # and the color of the points
     las = 1, pch = 16, col = "dodgerblue")
```

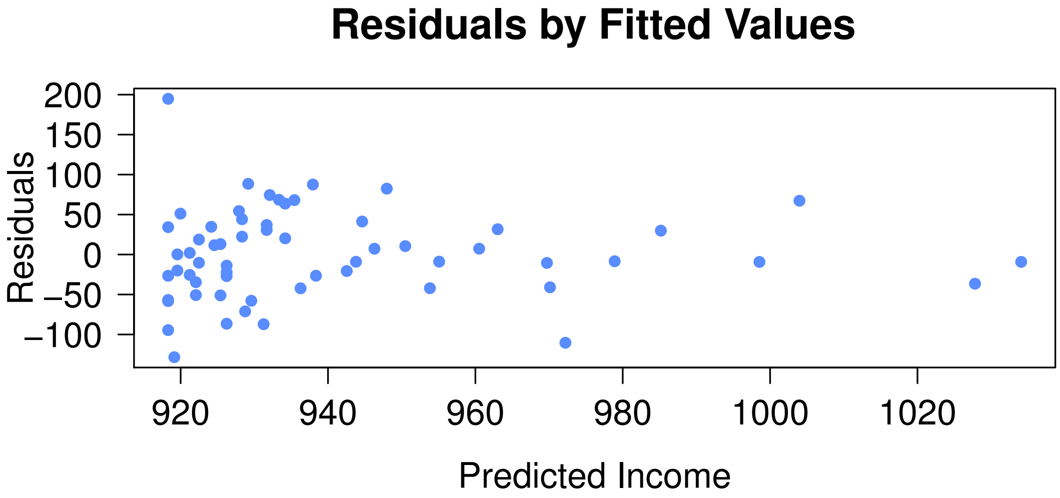


**Figure 12.6** Residual plot for the sample model

As we can see from Figure 12.6, however, the residual plot in this example a little problematic: many of the points are concentrated to the far left of the x-axis and are sparse to the larger x values. This may indicate that some transformation of our original data are necessary.

## 12.4 Case Study: Evidence for the Big Bang

Let's, for throwback Thursday's sake, head back to the late 1920s when physicists were hotly debating models of the universe, whether it was expanding, whether it was an eternal steady state universe, and so forth. Let's say for a moment we're working with Edwin Hubble and we're trying to demonstrate conclusively that in fact the Big Bang is the best cosmological model for the development of the early universe.

But this is a crazy idea! It's heretical! How can we hope to demonstrate its validity? Well, simplifying the story a bit, Hubble realized if this model were true, we could predict the distance of an extra-galactic nebula from Earth using its measured recession velocity. But, you ask, why does this matter? Well, let's take a look at Figure 12.7

**Figure 12.7** Movement of galaxies from a central point and all that fun stuff!

The idea is that, if the Big Bang were the correct model, then it would make sense for us to see galaxies further away from the earth to be receding at a faster speed than those closer to us. If a galaxy were moving in approximately the same direction as ours, then either (1) it would be moving faster than ours; (2) ours would be moving faster than it; or (3) they would be moving at approximately the same speeds. In either case 1 or 2, this would result in the two galaxies moving further apart. In the third case, as long as the two are not moving in exactly the same direction (i.e., $\theta \neq 0$ where $\theta$ is the interior angle between the trajectories of the two galaxies), we will also continue to move farther apart (but at a slower rate). And of course if the galaxies are moving in opposite directions, their recession velocity will be greater still.

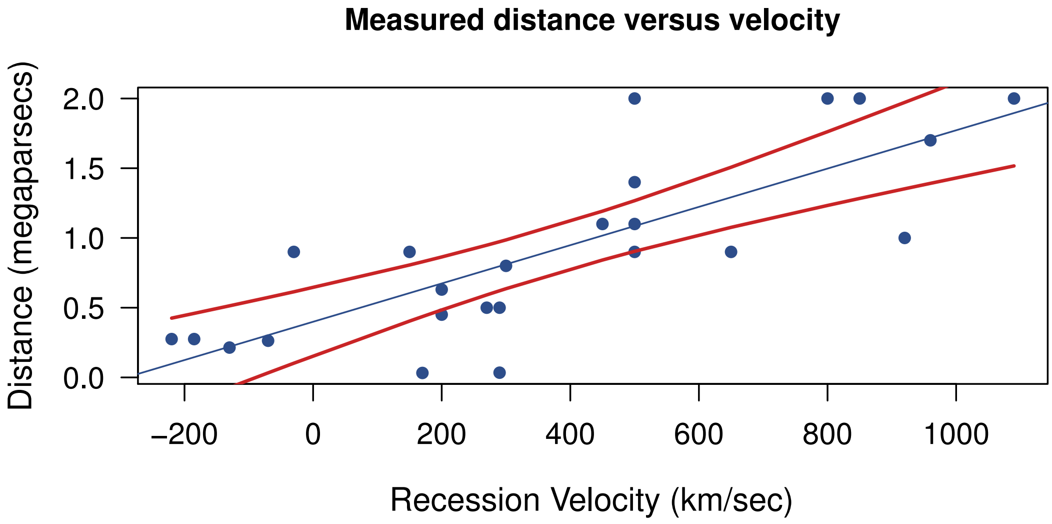


**Figure 12.8** Measured distance versus velocity for 24 extra-galactic nebulae. Data from Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences of the United States of America, 15*, 168-173. Line of best fit plotted in blue with a 95% confidence interval for the regression line in red.

So, knowing this, we can plot the relationship between recession velocity and distance for the 24 extra-galactic nebulae that Hubble observed (seen in Figure 12.8). Doing so, we see that there appears to be a positive relationship between the two. Indeed, if we construct a simple linear model, we see:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3990982  0.1184697   3.369  0.00277 **
Velocity    0.0013729  0.0002274   6.036 4.48e-06 ***

Residual standard error: 0.405 on 22 degrees of freedom
Multiple R-squared:  0.6235,        Adjusted R-squared:  0.6064
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06
```

We can see that we have obtained a significant overall model, $F(1, 22) = 36.44$; $p$-value $< 0.001$, with velocity contributing significantly to that model, $t = 6.04$; $p$-value $< 0.001$. This in turn gives us the ability to predict a galaxy's distance from Earth given its recession velocity, via the formula:

$$\text{Distance} = 0.399 + 0.00137 \cdot \text{Velocity}$$

Moreover, looking at a plot of our fitted valuer by our residuals (Figure 12.9), we see that it appears approximately unbiased and homoscedastic, indicating that we likely have a well-fitted model and that our data don't need to be transformed in any way. Given all this, we can fairly conclude that there is a relationship between the distance of an extra-galactic nebula from Earth and its recession velocity, implying that, indeed, all bodies in the universe are moving apart from one another, having originated from a central point at one time in the universe's history. That is, we find significant support for the Big Bang Theory.

## 12.5 Exercises

## 12.6 Additional resources

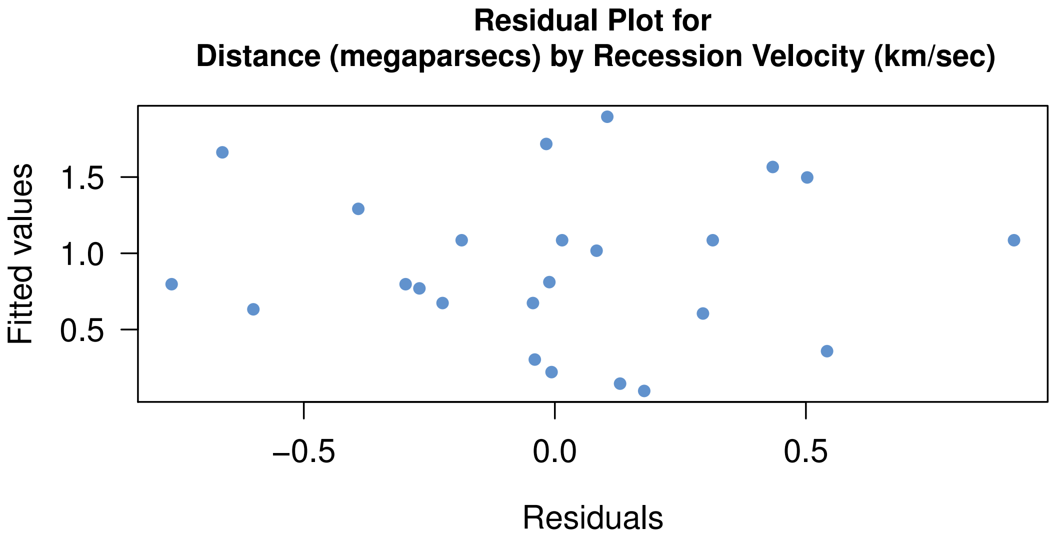


**Figure 12.9** Regression plot for our linear model predicting extra-galactic nebula distance by recession velocity

## CHAPTER 13

# MULTIPLE REGRESSION

13.1 Overview of Multiple Regression

13.2 Cautions and Considerations

13.3 Implementation in R

13.4 Case Study: [STUDY]

13.5 Exercises

13.6 Additional Resources

## CHAPTER 14

# LOGISTIC REGRESSION

**14.1 Overview of Logistic Regression**

**14.2 Cautions and Considerations**

**14.3 Implementation in R**

**14.4 Case Study: [STUDY]**

**14.5 Exercises**

**14.6 Additional Resources**

## CHAPTER 15

# MODEL REFINEMENT AND VARIABLE SELECTION

**15.1 General Strategies for Model Refinement**

**15.2 Cautions and Considerations**

**15.3 Implementation in R**

**15.4 Case Study: [STUDY]**

**15.5 Exercises**

**15.6 Additional Resources**

## PART IV

# INFERENCES FROM UNUSUAL DATA STRUCTURES

# CHAPTER 16

# TIME-SERIES DATA AND SERIAL CORRELATION

16.1 Overview of Time-Series Tools

16.2 Cautions and Considerations

16.3 Implementation in R

16.4 Case Study: [STUDY]

16.5 Exercises

16.6 Additional Resources

# CHAPTER 17

# FACTOR ANALYSIS AND PRINCIPLE COMPONENT ANALYSIS

17.1 Overview of Factor Analyses

17.2 Cautions and Considerations

17.3 Implementation in R

17.4 Case Study: [STUDY]

17.5 Exercises

17.6 Additional Resources

## CHAPTER 18

# RANK-ORDERED DATA

18.1 Overview of Tools for Ranked Data

18.2 Cautions and Considerations

18.3 Implementation in R

18.4 Case Study: [STUDY]

18.5 Exercises

18.6 Additional Resources

## CHAPTER 19

# TABLES OF COUNTS, PROPORTIONS, AND ODDS

**19.1 Overview of Chi-Squared Tests**

**19.2 Cautions and Considerations**

**19.3 Implementation in R**

**19.4 Case Study: [STUDY]**

**19.5 Exercises**

**19.6 Additional Resources**

## PART V

# SUPPLEMENTAL MATERIALS

## CHAPTER 20

# CHOOSING THE RIGHT ANALYSIS: A CHEAT SHEET

# CHAPTER 21

# GETTING OUTSIDE HELP: HOW TO SHARE DATA