

Roy Sabo · Edward Boone

Statistical Research Methods

A Guide for Non-Statisticians



Springer

Statistical Research Methods

Roy Sabo • Edward Boone

Statistical Research Methods

A Guide for Non-Statisticians

 Springer

Roy Sabo
Department of Biostatistics
Virginia Commonwealth University
Richmond, Virginia, USA

Edward Boone
Department of Statistical Sciences
and OR
Virginia Commonwealth University
Richmond, Virginia, USA

ISBN 978-1-4614-8707-4 ISBN 978-1-4614-8708-1 (eBook)
DOI 10.1007/978-1-4614-8708-1
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013949863

© Springer Science+Business Media New York 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

1	Introduction	1
1.1	Statistical Methods as a Part of the Research Process	1
1.1.1	Populations and Samples	1
1.1.2	Parameters and Statistics	2
1.2	The Statistical Method	3
1.2.1	Research Question and Hypothesis Generation	3
1.2.2	Statistical Assumptions	4
1.2.3	Statistical Method	5
1.3	Writing in the IMRaD Format	6
1.4	The R Statistical Software Package	7
1.4.1	Getting Started	10
1.4.2	Loading Data	10
1.4.3	Working with Data	11
2	One-Sample Proportions	13
2.1	Introduction: Qualitative Data	13
2.2	Establishing Hypotheses	14
2.3	Summarizing Categorical Data (with R Code)	15
2.4	Assessing Assumptions	16
2.5	Hypothesis Test for Comparing a Population Proportion to a Hypothesized Value	17
2.5.1	Behavior of the Sample Proportion	17
2.5.2	Decision Making	18
2.5.3	Standard Normal Distribution	20
2.6	Performing the Test and Decision Making (with R Code)	21
2.6.1	Test Statistic	21
2.7	Formal Decision Making	24
2.7.1	Critical Value Method	24
2.7.2	p -value Method	26
2.7.3	Conclusion	27
2.7.4	Confidence Intervals	28
2.8	Contingency Methods (with R Code)	29
2.9	Communicating the Results (IMRaD Write-Up)	32

2.10	Process	34
2.11	Exercises	34
3	Two-Sample Proportions	37
3.1	Summarizing Categorical Data with Contingency Tables (with R Code)	37
3.2	Hypothesis Test for Comparing Two Population Proportions	39
3.2.1	Generating Hypotheses About Two Proportions	39
3.2.2	Statistical Assumptions	40
3.3	Performing the Test and Decision Making (with R Code)	42
3.3.1	Critical Value Method	43
3.3.2	p -Value Method	44
3.3.3	Confidence Intervals	45
3.3.4	Chi-Square Test	47
3.4	Contingency Methods (with R Code)	51
3.5	Odds Ratio (with R Code)	51
3.6	Communicating the Results (IMRaD Write-Up)	53
3.7	Process	55
3.8	Exercises	56
4	Multi-category Data	59
4.1	Introduction: Types of Multi-categorical Data	59
4.2	Summarizing Categorical Data (with R Code)	60
4.3	Establishing Hypotheses: Difference Between Comparisons and Association	63
4.4	Assessing Assumptions (with R Code)	65
4.5	Performing the Test and Decision Making (with R Code)	67
4.5.1	Critical Value Method	68
4.5.2	p -Value Method	70
4.5.3	Interpretation of Results	70
4.6	Contingency Methods (with R Code)	72
4.7	Communicating the Results (IMRaD Write-Up)	73
4.8	Process	75
4.9	Exercises	76
5	Summarizing Continuous Data	79
5.1	Representative Values (with R Code)	80
5.1.1	Mean	80
5.1.2	Median	80
5.1.3	Other Measures	81
5.2	Measures of Variability (with R Code)	82
5.2.1	Standard Deviation	83

5.2.2	Range Measures	84
5.2.3	Empirical Rule	84
5.3	Assessing Normality (with R Code)	85
5.3.1	Histogram	86
5.3.2	Box Plot	88
5.3.3	QQ Plot	90
5.3.4	Outliers	94
5.4	Rounding and Reporting Conventions	95
5.4.1	Rounding	95
5.4.2	Reporting Based on Distribution	96
5.4.3	Standard Error	96
5.5	Exercises	97
6	One-Sample Means	101
6.1	Behavior of the Sample Mean	101
6.2	Establishing Hypotheses	106
6.3	Assessing Assumptions (with R Code)	107
6.4	Summarizing Data (with R Code)	108
6.5	Performing the Test and Decision Making (with R Code)	109
6.5.1	Critical Value Method	112
6.5.2	p -Value Method	113
6.5.3	Confidence Intervals	114
6.6	Contingency Methods (with R Code)	115
6.7	Communicating the Results	116
6.8	Process	118
6.9	Exercises	118
7	Two-Sample Means	121
7.1	Introduction: Independent Groups or Paired Measurements	121
7.2	Independent Groups	122
7.2.1	Establishing Hypotheses: Independent Groups	122
7.2.2	Assessing Assumptions (with R Code)	123
7.2.3	Summarizing Data (with R Code)	124
7.2.4	Performing the Test and Decision Making (with R Code)	128
7.2.5	Contingency Methods (with R Code)	132
7.2.6	Communicating the Results	133
7.2.7	Process for Two-Sample t -Test	135
7.3	Paired Measurements	136
7.3.1	Establishing Hypotheses: Independent Groups	136
7.3.2	Assessing Assumptions (with R Code)	137
7.3.3	Summarizing Data (with R Code)	139
7.3.4	Performing the Test and Decision Making (with R Code)	140

7.3.5	Communicating the Results	141
7.3.6	Process for Paired t -Test	143
7.3.7	Exercises	143
8	Analysis of Variance	147
8.1	Establishing Hypotheses	147
8.2	Assessing Assumptions (with R Code)	148
8.3	Summarizing Data (with R Code)	149
8.4	Performing the Test and Decision Making (with R Code)	152
8.4.1	Post-hoc Multiple Comparisons (with R Code)	154
8.5	Contingency Methods (with R Code)	160
8.6	Communicating the Results	161
8.7	Process	162
8.8	Exercises	163
9	Power	167
9.1	Making Mistakes with Statistical Tests	167
9.2	Determinants of Sample Size	169
9.3	Categorical Outcomes	170
9.3.1	One-Sample Case	170
9.3.2	Two-Sample Case (with R Code)	171
9.4	Continuous Outcomes	173
9.4.1	One-Sample Case (with R Code)	173
9.4.2	Two-Sample Case (with R Code)	173
9.4.3	Multi-sample Case (with R Code)	174
9.5	Post-hoc Power Analysis	175
9.6	Exercises	179
10	Association and Regression	181
10.1	Introduction	181
10.1.1	Association Between Measurements	181
10.1.2	Scatter Plots (with R Code)	182
10.2	Correlation Coefficients	186
10.2.1	Establishing Hypotheses	186
10.2.2	Assessing Assumptions (with R Code)	187
10.2.3	Summarizing Data	188
10.2.4	Estimating Correlation, Performing the Test, and Decision Making (with R Code)	188
10.2.5	Contingency Methods (with R Code)	190
10.2.6	Communicating the Results (with IMRaD Write-Up)	190
10.2.7	Process for Estimating Correlation	191
10.3	Simple Linear Regression	193
10.3.1	Establishing Hypotheses	195
10.3.2	Assessing Assumptions (with R Code)	195

10.3.3 Summarizing Data	195
10.3.4 Estimating the Regression, Performing the Test, and Decision Making (with R Code)	195
10.3.5 Establishing the Worth of the Regression	197
10.3.6 Communicating the Results (with IMRaD Write-Up)	205
10.3.7 Process for Simple Linear Regression	206
10.4 Exercises	207
Bibliography	211

Chapter 1

Introduction

1.1 Statistical Methods as a Part of the Research Process

1.1.1 Populations and Samples

The impetus for conducting research which utilizes statistical analyses is the desire to better understand some population of interest. A *population* is defined as the totality of any group of subjects sharing some characteristic(s). The characteristics that such groups of subjects share can be generally defined (e.g. nationality, ethnicity, gender) or specifically defined (e.g. low-income patients under 40 years of age with type II diabetes). Researchers typically study such populations because some feature of that group is unknown or under question (e.g. what is the success rate of patients surviving a particular treatment for a particular disease).

Though the research focus is on the population level, the use of an entire population is impractical for many reasons, each of which can be summarized in one word: resources. The resources needed to measure or examine the members of a population include money for research materials (drugs, laboratory space, recruitment, etc.) and the time needed to conduct the study. If a population is too large, then a great deal of money is needed to examine every subject within that population. Likewise, if members of a population are spread over a large area (e.g. the contiguous U.S.), the money and time required to reach them all will again be great. Importantly, the resources available to conduct research are usually constrained by factors external to the research. For instance, federal or industrial agencies sponsoring such research only have so much funding to offer, so population-level studies are usually out of the question. In other cases, such as in drug development, it would be unwise to test new treatments on large populations of subjects, especially when the risks of such treatments are severe or unknown.

Due to constrained resources, studies focus on subgroups of populations, which we refer to as *samples*. Samples are – by definition – smaller than the populations from which they are drawn and are thus more manageable, both from a resource-expenditure point of view as well as a conduct-of-research point of view. In sampling from a population we hope to capture the characteristics of the entire population in the smaller sample. For instance, if 57% of all undergraduate college students throughout the U.S. are female (and 43% male), then we would hope that a smaller sample drawn from this population would maintain a similar gender breakdown.

But therein lies one of the underlying facets of statistical theory and its applications: how do we know that a sample resembles the population from which it is drawn? The short answer is that we usually never know how closely a sample represents its parent population, or – in other words – how “good” the sample is. But we can take measures to help ensure that our samples are of the highest possible quality, none more important than our sampling technique. In order to obtain a sample, it must be taken from the population, meaning that certain subjects – but not all – from the population must be identified as also belonging to the sample. How these subjects are identified is essential to sound statistical practice. If we “draw” certain subjects from a population as opposed to others simply because it is easy for us to do so (e.g. we take those closest to us; we take those willing to participate without compensation; etc.), then we will have drawn what’s called a biased sample (these particular examples is also called a convenience sample), meaning that the reason we selected certain subjects has caused our sample to somehow not reflect the parent population. Except in certain cases (clinical trials for example), we try to avoid these convenience samples.

Collecting a *simple random sample* is the surest way we have of capturing the important characteristics from a population. A simple random sample is a process or quality more so than a noun, and it means that the process used to identify subjects ensured that every subject (or most subjects) in a population had an equal chance of being selected into the sample. The phrase “equal chance” implies that we are probabilistically selecting patients into the sample, and there are many ways of doing this (e.g. flipping a coin, picking numbers randomly from a phone book) that we won’t get into. If we know or if we can *reasonably assume* that this type of process was followed, and provided the sample is itself not too small, then we can typically expect our sample to be a microcosm of the population. If that is the case, then an analysis of our sample should mimic an analysis of our population, and the results we would get from both cases should be similar.

1.1.2 Parameters and Statistics

The populations in which we are interested often consist of many subjects (consider the number of citizens in the United States, or the number of diabetes sufferers worldwide), each consisting of many individual characteristics.

Naturally, a comprehensive understanding of all facets of the entire population is typically unattainable – aside from the fact that the population itself is usually unattainable. Thus, we focus on *parameters* that adequately summarize certain mathematical characteristics of the population. These parameters often take the form of proportions or means, and in most cases reflect the characteristic we would expect to observe in a typical subject from that population.

However, since we cannot collect populations, we must focus on the properties of the samples to which we have access. Any property of a sample that we measure (such as proportion or mean) is called a *statistic*, and is thus distinguished from its population counterpart, the *parameter*. As we will see in subsequent chapters, we can use a few statistics to summarize our entire sample (this is especially helpful if samples are large), and we can also use them to test hypotheses about the population in question. For instance, if we want to know something about a population parameter (say the success rate for a certain type of experimental cancer treatment), then we can use the sample statistic (say the success rate for 20 patients who underwent that treatment) to provide information on that population parameter. The most popular statistical method that turns a sample statistic into *inference* on a population parameter is called *hypothesis testing*, which will be the main focus of our foray into biostatistical methodology.

1.2 The Statistical Method

1.2.1 Research Question and Hypothesis Generation

A hypothesis test (note it's “a hypothesis”, not “an hypothesis”; you've been warned) is the process of using sample data to provide evidence toward some statement about a population parameter. Such a statement originally occurs in the form of a *research question*, where we boldly and unequivocally state what we feel or think about some parameter. As statisticians and biostatisticians (or the hopeful users of statistics), our first job is to translate this research question into a parametric or symbolic form that lends itself to being measured. For example, stating that you want to make cancer-victims better doesn't make good science, but saying you want to increase the median survival time of cancer-victims by 10 months through a particular treatment works well.

Once we have determined what population parameter we are interested in, we need to then turn the research question into a set of testable hypotheses. These competing hypotheses must be such that only one can be true at a time. Given that we have defined such hypotheses, we can then use our sample data to provide evidence for or against those hypotheses; naturally, the hypothesis that the evidence more closely supports becomes the “winner”. It is this process of using sample data to support a set of hypotheses about a population parameter that we are referring to in hypothesis testing.

1.2.2 Statistical Assumptions

In order to conduct a hypothesis test, several characteristics of our sample must be in order for us to place any stock in the worth of such a test. The characteristics we require of a sample are: that it is *representative*, that the subjects within that sample are *independently measured*, and that our sample is *large enough* for the planned statistical method to work correctly.

Representative Samples: A sample is representative of the population from which it is drawn if the sample is somehow a microcosm of that population, in that it maintains the important characteristics (e.g. gender or race proportions, disease susceptibility) of the population even though it only contains a fraction – often a small fraction – of its members. This is an important characteristic, the utility of which is easily observed through the unfortunate instance of an unrepresentative sample – if a sample is not representative of the population from which it was drawn, then what good is it? The idea is that if a sample is representative of a population, the numeric or mathematical characteristics of that population will be present in the sample. This attribute will ensure that statistical analysis of the sample would yield similar results to a (hypothetical) statistical analysis of the population.

Independent Measurements: The concepts of dependence and independence are somewhat difficult to explain without some basic foundation in statistical language, so we will save some of this discussion for later. However, it should suffice to say that we would not want a sample where the measurements or values we observe for some subjects are influenced by – or *depend* upon – the measurements or values for other subjects. This may at first seem like a weird phenomenon – in simple random samples this *rarely* happens – but examples are easy to imagine. For instance, if we are conducting a study where we are measuring the presence or absence of a certain gene, and we unknowingly sampled measurements from members of the same family, then the outcomes for those subjects within the same family will be related due to genetic inheritability. This is bad because measurements that are related – or dependent – make the sample measurements seem closer together than they actually may be in the grand population (this is called *variability* and will be discussed later). Regardless, we would like our sample measurements or values to be independent of one another, and if we responsibly sample from the parent population (i.e. create a simple random sample), then we can usually assume that this is the case.

Adequate Sample Size: Ask any statistician their biggest pet peeve, and one of the most popular responses will be analyzing samples that are *too small*. This happens for many reasons, such as small or esoteric populations, limited resources, etc., but it happens most often due to poor planning. The reason why this is a problem is that small samples can in no way represent the population from which they were drawn, and thus any statistical methodology dependent upon the sample's representativeness will break down (i.e. not work). Thus, we need our samples large enough to adequately reflect the

populations from which they are drawn (if you ask a statistician, no sample is large enough), yet manageable enough to be cost effective. For many of the procedures we will discuss throughout this text we will have rules for determining how large a sample we need. We will also focus – in Chapter 9 – on performing a sample size or power analysis, which helps us determine the sample size we need to collect before we conduct the study.

1.2.3 Statistical Method

We will follow a formal method for conducting statistical analyses that consists of several parts: statement of the research question, determining what method to use, assessing our statistical assumptions, summarizing the data, performing the test, and interpreting the results. These parts are designed for several reasons: so that we can be sure we are taking the correct steps for the analysis, so that we can easily communicate our methods and results, and so that our methods can be easily reproduced.

Statement of Research Question: Before we know what statistical procedure we're going to use (the statistical method provides the answer we're looking for), we have to know what question we are asking. We do this by taking our research question – which must be explicitly stated – and turning it into a set of testable hypotheses. We will spend a lot of time doing this throughout the text. At the end of the day, you cannot provide an answer if you don't know what the question is.

Determination of Statistical Method: Once we know our question, we can figure out how best to answer it. The remaining chapters in this text are arranged to provide different types of methods we can use to answer different types of questions we could potentially face. We determine what statistical method to use by looking at how measurements were observed, and the types of measurements we can come across vary considerably.

Assess Statistical Assumptions: Once we've identified the type of measurement we have, and what kind of statistical method we would like to use to analyze those measurements, we need to determine whether or not it is *appropriate* to use that method. In general, this is done by assessing whether our sample is *representative*, whether our measurements are *independent*, and whether we have a *large enough sample*, though on occasion there will be other considerations.

Summarize the Data: Provided our assumptions are met, we can then summarize the sample data with statistics that represent all of the important details of that sample. We will focus a great deal on how to appropriately summarize a sample, given the type of measurements we have and what assumptions are met.

Perform the Test: Once our data are appropriately summarized, we can then perform the statistical hypothesis test or use the desired statistical technique. Again, we will learn various methods throughout the semester, with each chapter presenting a new class of methods.

State the Result: Once we have conducted the statistical analysis, we will need to make sense of our results. As mentioned earlier, we do this by stating which hypothesis the evidence supports. Recall that though we are trying to learn some characteristic about some population, that *characteristic exists or is true*; we simply don't know what it is. So when we state our result, we can make one of two decisions: the evidence supports the first hypothesis, or the evidence supports the second hypothesis. Since the conditions stated in one of the two hypotheses we've created *must be true*, we can make two types of mistakes called Type I and Type II errors. We will focus on errors of the first type throughout this text, and we will cover errors of the second type in Chapter 9. In practice, if we have *set the table* correctly by following sound scientific methods in our data collection and sampling methodology, these types of errors are of little concern and we can put a great deal of faith in our statistical conclusions. Of course, the key aspect of any statistical result lies in translating it into a meaningful statement that can be understood by curious and critical readers.

1.3 Writing in the IMRaD Format

While we will spend a great deal of time performing statistical analyses, we must also learn how to communicate these results to the scientific community. We will spend a lot of time focusing on the “write-up” of our methods and results. This is not because we don't like you, or because we take peculiar pleasure in torturing students, but rather because the results from statistical analysis – regardless of how fancy or sophisticated – are useless unless they can be understood by those not involved in the study. This is not only true of statistical methods and results, but of science in general. If a scientific or statistical method is unclear, then readers of your research will not understand what you have done and will ultimately reject your work via the following, well-established assumption: “if I can't understand what the author is saying, then it must not be any good, for I am smarter than the author”.

A standard write-up that describes the key points of research that any sophisticated reader would need to know to make an informed decision – *are the results from this research reliable enough for me to use or believe?* – would then be a necessity for translating and communicating our results. The IMRaD style goes a long way toward providing such a standard format, and has been accepted by virtually all credible scientific research journals (in a strange irony, methodological research in statistics generally does not adhere to IMRaD, though all of the pieces are still there).

The IMRaD format consists of four main parts: the Introduction, the Methods section, reporting of Results, and a Discussion. Each part serves its own purpose – briefly described below – and contains specific information that matches up perfectly with the process we will follow for conducting statistical analyses. Sophisticated readers become accustomed to this format

and its placement of material, so much that they often skip to the parts they are interested in to glean information quickly. We will spend a great deal of effort understanding this method and its pieces, as well as practicing how they apply to specific statistical methods.

Introduction: Here we provide details on the scientific problem in which we are interested, and then describe the populations of interest. The treatment or intervention specific to the current study is introduced, and the scientific research question is un-categorically introduced (i.e. in the form from which you will create your testable hypotheses).

Methods: In an actual publication, this is the section where you would describe the setting of your sample, including such details as where and when subjects were observed. A thorough description of what was measured and the process under which those measurements were taken would then be provided. Any technological processes specific to the particular science and measurements in question would be described here. Generally, a description of the statistical methods used to analyze the sample measurements in light of the research questions and hypotheses would be placed in the last sub-section of the Methods Section (and often in small font to indicate its importance). Here you will state how you summarized your data, how you analyzed the data, and how you will make your decisions based on that analysis. You must also specify any details that aid in that process, such as the statistical software used for analysis.

Results: The details of the statistical analyses are presented here, starting with a summary of the sample data (including any tabular or graphical representations), continuing with the results from the analysis of the primary research question, and ending with any secondary or sub-analyses not specified in the primary research question. An unequivocal answer to the primary research question must be provided in this section.

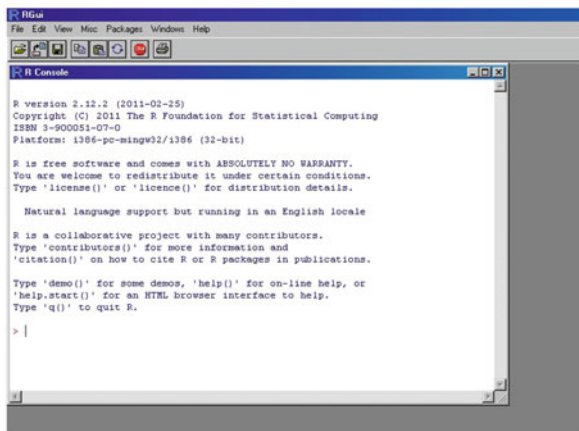
Discussion: A brief summary of the results is provided at the beginning of this section, where here the results are described in words (no statistics). The scientific or clinical implications of these results are then expounded upon, specifically with regards to how these results compare to those from previous research studies. Any study limitations – there will always be something – must be identified and described in this section, as should a justification as to how they do or do not affect your results. This section often ends with a prognostication of what these results mean for future research or what steps need to be taken to continue this research.

1.4 The R Statistical Software Package

While some common statistical procedures are simple enough to compute by hand, many are computationally intensive enough that we would not wish to do so. Further, modern data sets can be so large that calculation by hand is typically prohibited. Fortunately, there are many statistical software

Figure 1.1: The initial screen in R showing the R Console.

Windows Console



```

R Console
-----
R version 2.12.2 (2011-02-25)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-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.

Natural language support but running in an English locale

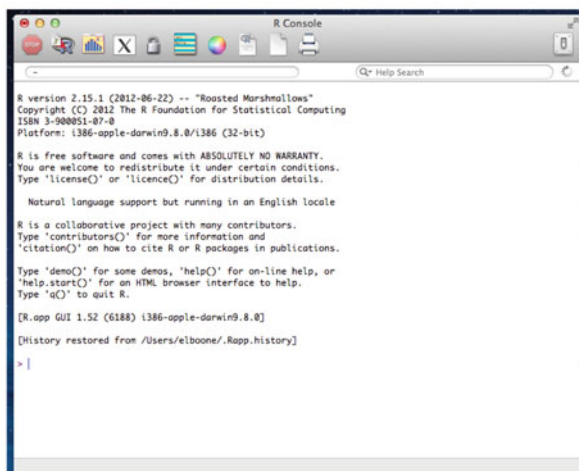
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |

```

Mac Console



```

R Console
-----
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-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.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.52 (6188) i386-apple-darwin9.8.0]

[History restored from /Users/elboone/.Rapp.history]

> |

```

packages available to perform these computations; some popular packages are SPSS, SAS, Minitab, Stata, JMP, etc. In this text we will focus on using the statistical package R, which is an open-source (read: free) software program that is continually updated with new and improved packages by its users. R can be downloaded at no cost from: cran.r-project.org. Once on that page simply select your operating system (Windows, MacOS X or Linux) and download the base package.

Once R is installed, you will have an extremely powerful piece of statistical software at your fingertips. The main drawback to using R is that *it is a language*, which means you will need to program the analyses yourself

Figure 1.2: Screenshot of the R window with the R Console (left window) and the New Script (right window) for Windows.

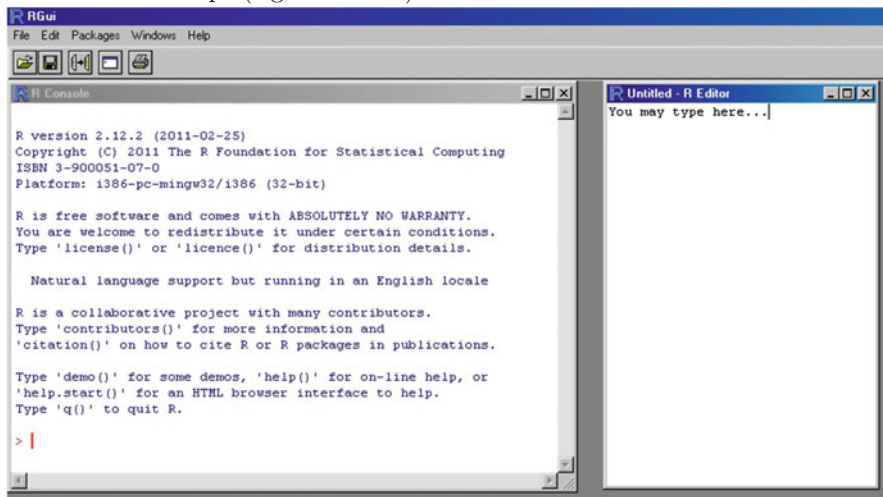
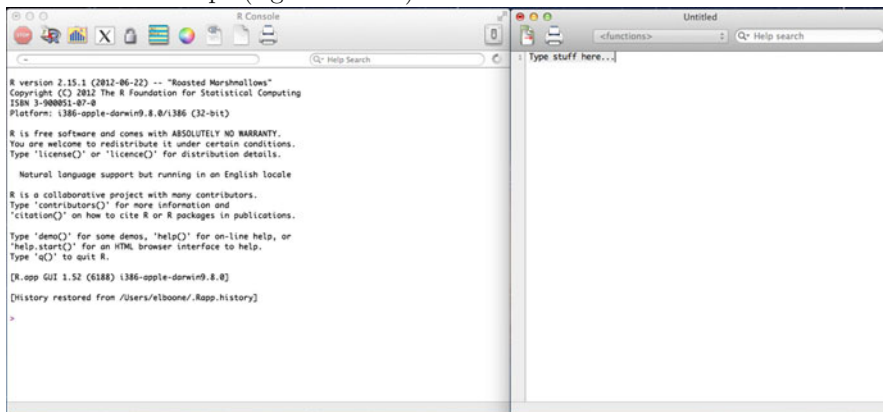


Figure 1.3: Screenshot of the R window with the R Console (left window) and the New Script (right window) for Macs.



(as opposed to the point-and-click functionality of SPSS or JMP). While this can be frustrating at first, the code is surprisingly simple (both to learn and to use) and refreshingly short. Further, by programming your own code in R, you will automatically save a record of your work, which is not easily achieved using a point-and-click program. Throughout this text we will show you how to conduct each of the analyses using the R statistical software.

1.4.1 Getting Started

Once you have downloaded and installed R you can open the application, which gives the screen shown in Figure 1.1 for both Windows and Mac users. The R Console is where written program commands can be entered into R and executed.

While you can type directly into the console, the experience will be enhanced if you instead type commands into an R script, which can then be submitted to the console. To open a blank script, simply go to the File menu and select New Script. This will open the R text editor window in which you may type your commands. Figure 1.2 shows the R program with the R Console (left window) and the R Editor (right window) in a Windows platform (Figure 1.3 shows the R Editor for Macs). After entering your commands into the R Editor, you can ask R to execute the commands by highlighting the desired code, hitting the right-mouse button, and then selecting Run; once you're more proficient, you may use the following sequence of hot-buttons: select-all ("control A"), then run ("control R").

1.4.2 Loading Data

For the purposes of working with R, the data file you wish to import should be in a Comma Separated Values (CSV) format, which is a compact format for your data. To read in a CSV into a dataframe you use the following code:

```
Data1 <- read.csv(file.choose(), header = TRUE)
```

The statement above can be broken down into several pieces. The data file will be read into a dataframe called `Data1`. The `<-` symbol is the assignment operator which puts the data file *into* the dataframe named `Data1`. The `read.csv()` function tells R that you wish to input a data file in CSV format. The `file.choose()` statement tells R that you wish to find the file on your computer using the windows environment. The `header = TRUE` statement tells R that the first row in the data file are labels for the columns, i.e. the first row is a header. If the first row does not contain column headings then this statement should be set to `FALSE`.

Example: Suppose we wish to load the data file `Chapter1data.csv`. The first thing to do is to download `Chapter1data.csv` to a location that you can easily find on your computer. Use the code in Program 1 to load the data. Remember to select the code, right mouse click and select "Run line or Selection".

From the output in Program 1 we can see that R prints the commands supplied (code) in the R Console. By simply including a line with only the dataframe name, R will print the data frame to the console. Notice that the `Chapter1data` dataframe has four columns (`Subject`, `Weight`, `Height` and `BP.Sys`) and has seven rows.

1.4.3 Working with Data

Once we have the data file imported into a dataframe we may want to work with certain parts of the data. To access a specific column in the dataframe we can use the following command format:

```
dataframe$column
```

For example in the `Chapter1data` dataframe we may want to work with the Height measurements. To do this we would need to refer to the data as:

```
Chapter1data$Height
```

In addition to accessing a specific column we may wish to access specific rows of data. There are several ways to do this, though the easiest is to specify the row you wish. In order to do this we need to understand how R organizes data. In a typical `dataframe` we can access any item by specifying the row and column of the item, which is done in R using the `data[row,column]` notation, often called *bracket* notation. For example, if in `Chapter1data` we wish to obtain the item in the third row and second column we would use the following syntax:

```
Chapter1data[3,2]
```

Here, the order is important: the first item in the bracket specifies the row while the second item specifies the column. If we leave one of these items blank then we will get the entire row or column for the item left non-blank. For example,

Program 1 Program to import the `Chapter1data.csv` file into R and display the results.

Code:

```
Chapter1data <- read.csv(file.choose(),header=TRUE)
Chapter1data
```

Output:

```
> Chapter1data <- read.csv(file.choose(), header=TRUE)
> Chapter1data
  Subject Weight Height BP.Sys
1       1  117.9   65.9    124
2       2  137.5   69.0    129
3       3  147.1   65.8    127
4       4  146.4   61.5    129
5       5  125.4   62.3    123
6       6  139.8   65.2    124
7       7  143.9   55.5    117
>
```

```
Chapter1data[3,]
```

will give us the entire third row of the dataset as no column was specified. We could similarly obtain just the second column by writing

```
Chapter1data[,2].
```

Using the bracket notation can be cumbersome if you don't know which row or column number for what you want. You may wish to use some logic to get the information you want. R can use the bracket notation in combination with logic operators to make subsets of your data. The logic operators are given in Table 1.1.

Table 1.1: Basic Logic Operators in R.

equal to	==
not equal to	!=
less than	<
greater than	>
and	&
or	

For example, suppose in the `Chapter1data` we want all the information for subject 4. Then we can use the bracket notation and the logic operators to obtain this information using the following syntax:

```
Chapter1data[ Chapter1data$Subject == 4, ]
```

Notice that inside the brackets we specified the column *and* dataframe that we need to check. This syntax can seem redundant but is ultimately flexible for adding more complicated logic arguments (or combinations thereof). We will not cover all possibilities here. However as we progress through the text you will see many of these logic arguments and data subsetting in use. Of course, we will explain what we are doing when we use the syntax.

Proficiency in R is a valuable skill, as it can satisfy most statistical needs and is freely available. What we have presented here is just enough to enable you to get you started in R so that you can proceed through the remainder of this book; additional concepts and programs will be provided in each chapter as the needs arise. There are plenty of reference texts to consult should you need additional assistance (e.g. see [Ekstrom 2012](#)).

Chapter 2

One-Sample Proportions

2.1 Introduction: Qualitative Data

The simplest types of measurements are qualitative in nature, meaning that they are non-numeric – or at least numeric manipulation of them is meaningless – and include names, labels and group membership. Examples of qualitative data are ubiquitous, but are best exemplified by dichotomous categorical data consisting of only two possible values, such as a patient's gender (male or female), diagnosis of a certain disease (positive or negative), or the result from a health intervention (success or failure).

Dichotomous categorical data are typically described in terms of the proportion (p) of some population with one of the two possible characteristics. This value is defined as the total number of subjects in some population *exhibiting* that specific characteristic *divided* by the number of total subjects (N) in that population. For instance, if 105 out of 200 physicians in a given hospital are female, then the proportion of these physicians who are female is $p_f = 105/200 = 0.525$. It should stand to reason that a proportion can only take values between 0 and 1.0, as you cannot have fewer than zero subjects with a given characteristic (reflecting $p = 0/N = 0$), just as you cannot have more than the total number subjects with a given characteristic (reflecting $p = N/N = 1.0$).

The Complement Rule: The outcomes for dichotomous data must also be mutually exclusive, in the sense that any given subject may assume only one of the two potential outcomes at one time. For instance, a subject cannot simultaneously test positive and negative for a disease. In general we will ignore instances where the outcomes are not mutually exclusive; in practice it is best to avoid these scenarios all together. One benefit of this characteristic is that we only need to know the proportion of one of the two outcomes to know the proportion for both. Returning to our previous example, if there are 105 female physicians, then there *must be* $200 - 105 = 95$ male physicians, meaning the proportion of male physicians is $p_m = 95/200 = 0.475$. Note here

that there are $105 + 95 = 200$ physicians who are either male or female, meaning that the proportion of physicians who are either male or female is $p_e = 200/200 = 1.0$. Further, note that $p_f + p_m = 0.525 + 0.475 = 1.0 = p_e$. This will *always be the case* for dichotomous categorical data. So if we know that $p_f = 0.525$, then we can use what is called the *complement rule* to find $p_m = 1 - p_f = 1 - 0.525 = 0.475$.

As a final note on proportions, there is a one-to-one relationship between proportions and percentages, meaning that for every proportion between 0 and 1.0, there is a corresponding percentage between 0 and 100%. This means that we should be able to transform proportions into percentages – and percentages into proportions – with ease. Without getting into the mathematical rationale, the algorithm is simple: to turn a proportion into a percentage, move the decimal two places to the right and add a percent sign (%). For example, if we have the proportion $p = 0.525$, we turn it into the percentage 52.5%. Likewise, if we have a percentage (say 47.5%), we turn it into a proportion by moving the decimal two places to the left and removing the percent sign (0.475).

2.2 Establishing Hypotheses

The key problem here is that we generally do not know the exact value of a population proportion, and at times we might not even know the total number of subjects comprising that population. This is problematic for those who may want to base their decisions or actions on such a proportion. For instance, in deciding between two different treatments to administer to a patient, a physician might want to know the success rates – read: *proportions* – of those two treatments before choosing between them. These population values are *rarely* known, but certainly the physician – or others in a similar situation – must make a decision, so something else must be done.

Thus enters the statistical method and the formation of a hypothesis. When a population proportion is unknown, we must formulate competing and mutually exclusive hypotheses about that proportion, collect data representative of the desired population, evaluate that data, and determine which hypothesis the evidence supports. The first step in this process is to set up competing hypotheses to test. There is generally some hypothesized value (p_0 – pronounced “*p-naught*”) in which we are interested; for instance, maybe it is commonly accepted that the success-rate of a given treatment is 0.5 (treatment is successful for half of all patients and unsuccessful for the other half). This value then becomes the central crux around which we form our hypotheses.

As mentioned in Chapter 1, we will create two mutually exclusive hypotheses, such that only one can be true at a time. We name these hypotheses the null and alternative hypotheses, and we have a formal process for determining which hypothesis gets which name (the naming procedure is actually important). The null hypothesis (H_0) is that hypothesis which states

our parameter is equal to some value (p_0), while the alternative hypothesis (H_A) indicates that the parameter is somehow different from p_0 . Depending upon our research question, there are three possible ways in which the parameter – proportion in this case – can differ from p_0 : it can be *less than* p_0 (represented by the symbol $<$), it can be *greater than* p_0 (represented by the symbol $>$), or it can be *not equal to* p_0 (represented by the symbol \neq). To choose between these three options we begin by translating our research question into symbolic form, which will include one of the following options: $<$, \leq , $>$, \geq , $=$ or \neq . As an example, suppose our research question is that the proportion of subjects with adverse toxic reactions to a particular drug is less than 0.3. In order to turn this into a symbolic statement, we must identify the operative phrase “is less than”, which is stating that $p < 0.3$.

The second step is to find the functional opposite of the statement from our research question. Based on the symbolic form from our research question, we create the functional opposite by pairing the following symbols: ($<$ and \geq), ($>$ and \leq) or ($=$ and \neq). Note that each of these pairs comprise all possibilities for a given situation (e.g. you are either strictly less than some value or greater than or equal to some value; either greater than some value or less than or equal to some value; either equal to some value or not equal to some value). Returning to our example, the functional opposite of the symbolic form of our research question ($p < 0.3$) is $p \geq 0.3$.

The third step is to identify which of our two symbolic forms is the null hypothesis and which is the alternative hypothesis, which is easier to do than to explain. Of the two symbolic forms, the form with some equality (meaning the $=$, \leq or \geq signs) becomes the null hypothesis, while the symbolic form without any equality (meaning the \neq , $<$ or $>$ signs) becomes the alternative hypothesis. Further, regardless of the symbol in the statement that belongs to the null hypothesis, we use the $=$ sign. (We do this for practical reasons, as we’re going to assume the null hypothesis is true, and doing so is much easier if H_0 contains only one value rather than a range of values. Keep in mind, however, that this practical reason is not the same as theoretical justification, which will be given elsewhere.) For our example, the statement $p \geq 0.3$ contains equality, while the statement $p < 0.3$ does not. So our alternative hypothesis becomes $H_A : p < 0.3$, while the null hypothesis becomes $H_0 : p = 0.3$. This process can be followed for most research statements concerning one population proportion, and Table 2.1 lists the possible hypotheses as well as key words to help in guiding you to the appropriate pair.

2.3 Summarizing Categorical Data (with R Code)

Sample Proportion: Given a set of hypotheses about a population proportion, the next step is to collect evidence that will (hopefully) support one of the two hypotheses. When we are interested in a population proportion, the

Table 2.1: Possible Sets of Hypotheses for a Population Proportion Based Upon Key Phrases in a Research Question.

Hypothesis	Key Phrases		
	“less than” “greater than or equal to” “at least”	“greater than” “less than or equal to” “at most”	“equal to” “not equal to”
Null	$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
Alternative	$H_A : p < p_0$	$H_A : p > p_0$	$H_A : p \neq p_0$

logical step would be to calculate a *sample proportion* from a representative sample drawn from the population of interest. The sample proportion \hat{p} serves as an *estimate* of the population proportion, and is calculated in a similar manner as its population analogue, being the number of subjects x in a sample exhibiting a particular characteristic (often referred to as the *frequency*) divided by the total number of subjects n in the sample (referred to as the *sample size*). So if we have a random sample of 25 physicians, 13 of whom are female, then the proportion of female physicians in this sample is $\hat{p}_f = 13/25 = 0.52$. Due to the dichotomous nature of this type of measurement, we can use the complement rule to find the sample proportion of male physicians ($\hat{p}_m = 1 - 0.52 = 0.48$).

Rounding: Depending on the sample size, you will have different rules for rounding proportions. For sample sizes greater than 100, round \hat{p} to at most three decimal places (e.g. $\hat{p} = 0.452$). For sample sizes less than 100 but greater than 20, round \hat{p} to two decimal places (e.g. $\hat{p} = 0.45$). For small sample sizes less than ~ 20 , the ratio of frequency to sample size should be reported as a fraction (x/n) (e.g. $5/11$) and the sample proportion should not be calculated (note that there is no universally agreed upon value for this last rule, and 20 was selected for presentation).

2.4 Assessing Assumptions

If our random sample is adequately representative of the parent population from which it is drawn, then our sample estimate \hat{p} should be close to the population value p . Unless we have clear evidence or information to the contrary, we will assume: (i) that the sample used to calculate \hat{p} is representative, and (ii) that the subjects in the sample from whom measurements were taken are independent of one another. To determine if the sample is of sufficient size, we need to check the following conditions: based on the condition set forth in the null hypothesis H_0 , we need to expect there to be *at least five* subjects taking either value of the categorical variable. This expectation is determined by noting that if the proportion of subjects taking the first category is $p_0 = 0.3$ according to H_0 (meaning the proportion taking the second

category is $q_0 = 1 - p_0 = 1 - 0.3 = 0.7$) and if $n = 30$, then we would expect there to be $p_0n = (0.3)(30) = 9$ subjects in the first category, and $(1 - p_0)n = q_0n = (0.7)(30) = 21$ subjects in the second category. So in this case we would have adequate sample size. However, if H_0 instead specified that $p_0 = 0.1$, then we would expect $p_0n = (0.1)(30) = 3$ subjects in the first category and $q_0n = (0.9)(30) = 27$ subjects in the second. Since we expect less than five subjects in the first category, we would not have adequate sample size to perform the desired hypothesis test. In cases of inadequate sample size we would report that we cannot perform the desired test (meaning we stop the entire process and either figure out how much more data we need or perform a different test). Note that we use p_0 to determine if we have adequate sample size rather than \hat{p} .

2.5 Hypothesis Test for Comparing a Population Proportion to a Hypothesized Value

At this point we have already translated our research question into testable hypotheses, verified our assumptions, and summarized our data. It is now time to combine the two pieces into a statistical test that will eventually support either the null hypothesis or alternative hypothesis. Since we have a sample proportion \hat{p} that should resemble the population proportion p upon which we are trying to make inference, it makes sense to base our test around the sample estimate. However, before we develop a formal test, we should study further the behavior of \hat{p} in order to better understand from where such a test might arise.

2.5.1 Behavior of the Sample Proportion

Consider a random and representative sample of 200 patients undergoing treatment to alleviate side-effects from a rigorous drug regimen at a particular hospital, where 33 patients experienced reduced or no side-effects. For this particular sample, we know that the sample proportion of patients who experienced little or no side-effects was $\hat{p} = 33/200 = 0.165$. So one could presume – based on the evidence from this sample – that between 16 and 17% of all patients would experience reduced side-effects when using this treatment regimen. But is this a reasonable presumption? What if we had collected a different sample of patients from this hospital (or from a different hospital, for that matter)? Would the sample proportion change? If so, how much would it change?

While we cannot answer these questions based on our particular sample, we can conduct studies that will allow us to see what we could expect to find if we could *repeatedly sample* from a population with a *known population proportion*. It is possible for us to conduct a *simulation study*, or a study in which we repeatedly simulate sets of data that reflect known population parameters (such as p_0), where summary statistics (such as frequencies or \hat{p}) are calculated for each of those samples and then summarized themselves. We can then determine the likelihood of the observed sample data (or sample estimate) compared to the results from the simulation study (more on this topic later).

Returning to our example of 200 hospital patients, maybe the historical rate of patients with little or no side-effects is 10.0%, and we want to determine if this new treatment increases that rate. (Imagine a bag filled with a large number – say many thousands – of chips, 10% of which are red and the rest blue, and we draw out 200 of those chips and count the number of red chips; this is not what we really do, but the idea is the same).

The results from a simulation study that generated 1,000 such samples are provided below in Table 2.2, which shows the number of samples for which we observed specific success counts (ranging from 5 to 36) out of 200. Note that there are many more success frequencies in this study other than the frequency observed in our sample (which was 33), which reflects the variability we might observe if we were to repeatedly sample from this population. Variability can mean many things, but here we are taking it to mean how our sample proportion could change in value if we were to resample.

Note that a frequency of 20 occurs most often ($\sim 12\%$ of the time) and represents the case when $\hat{p} = 20/200 = 0.100$, which is the value (p_0) assumed in this study. Also note that most of the simulated samples yielded frequencies slightly below or slightly above 20 (e.g. 16–19, 21–24), while relatively fewer studies yielded frequencies greatly below or greatly above 20 (e.g. 5–11, 29–36), which makes sense since 20 was the value we assumed was the *true population parameter*. Based on the results from this simulation study, we would then conclude that if $p = 0.10$ was indeed true, we would expect sample proportions close to 0.10 rather than far away from it.

2.5.2 Decision Making

So how likely is our sample value of 33? Based on our simulated data, 33 occurred once out of 1,000 total simulations; specifically, 33 successes appeared at a rate of $1/1,000 = 0.001$, which is not often. More generally, any value *greater than or equal to* 33 (which includes 33, 34, 35 and 36) occurred only 4 out of 1,000 times, or $4/1,000 = 0.004$, which is still not often. In either case, both the observed event (33) or any event *at least as extreme* as our observed event (> 33) seems unlikely if the true population success rate is $p = 0.10$.

Table 2.2: Results From Simulation Study of Samples with 200 Dichotomous Observations with a Known Success Rate of 0.10.

# of Successes Frequency Proportion out of 200			# of Successes Frequency Proportion out of 200		
5	1	0.001	21	75	0.075
6	0	0.000	22	78	0.078
7	1	0.001	23	71	0.071
8	0	0.000	24	54	0.054
9	2	0.002	25	33	0.033
10	8	0.008	26	25	0.025
11	9	0.009	27	27	0.027
12	20	0.020	28	11	0.011
13	25	0.025	29	8	0.008
14	35	0.035	30	11	0.011
15	48	0.048	31	4	0.004
16	65	0.065	32	4	0.004
17	69	0.069	33	1	0.001
18	99	0.099	34	1	0.001
19	94	0.094	35	1	0.001
20	119	0.119	36	1	0.001

This last statement brings us to the crux of statistical decision making: based on our assumption of $p = 0.10$, the observed success rate $\hat{p} = 0.165$ does not seem likely (if we are to believe the simulation study, which we should). So what do we conclude? There are two likely outcomes: (i) our assumption of the population proportion was correct and our sample data are wrong (or at best unlikely), or (ii) our sample data is more reflective of the “truth” and our assumption was wrong.

Since our sample is the only information we have that reflects any property of the population from which it was drawn, and since it was randomly selected from and is representative of that population, *we must base our conclusions on what the data and its summaries tell us*. This is one of the most important ideas in this entire textbook: if the data do not support a given assumption, then that assumption is most likely not true. On the other hand, if our data *did* support our assumption, then we would conclude that the assumption is likely to be true (or at least more likely than some alternative).

Returning to our example, since a frequency of 33 (or greater) did not occur often in our simulation study, we would logically presume that we are not likely to observe frequencies that high (or higher) in samples drawn from a population with a success rate of 0.10. Thus, we conclude that, based on our sample proportion of 0.165, the true population proportion of patients experiencing reduced symptoms or side-effects under this treatment is probably greater than 0.10.

2.5.3 Standard Normal Distribution

While the previously conducted simulation study was helpful in discussing the behavior of a sample proportion under some hypothesized value, it is important to note that we do not usually conduct simulation studies every time we want to conduct a hypothesis test (indeed, it is often difficult or impractical to do so). Rather, statisticians from centuries past have successfully characterized the behavior of a sample proportion in such a manner that the results we would like to obtain are readily available without the need for sophisticated computing power.

Consider the histogram in Figure 2.1, which shows in graphical form the results from our simulation study. In its entirety, this histogram represents the *distribution* of sample proportions assuming $p = 0.10$. Based on this distribution, we see that it is largest in the middle (corresponding to likely values based on our assumption of $p = 0.10$), and then slowly gets smaller as we move away from 0.10 (in both directions), so that eventually we have infrequent or non-occurring outcomes. These regions are called *the tails* and represent values that are unlikely to occur if our assumption of $p = 0.10$ is true.

As mentioned earlier, we *do not* want to rely upon simulation studies or the distributions they create, though we would like a distribution that resembles that created by the simulation study. Thus, we use what is called the *standard normal distribution*, whose properties are well known and easy to use. A random variable Z has a standard normal distribution if the probability that it is between two numbers a and b is given by the following integral (given in Equation 2.1).

Figure 2.1: Histogram Summarizing Results from a Simulation Study of 1,000 Samples of 200 Dichotomous Outcomes with an Assumed Success Rate of $p = 0.10$.

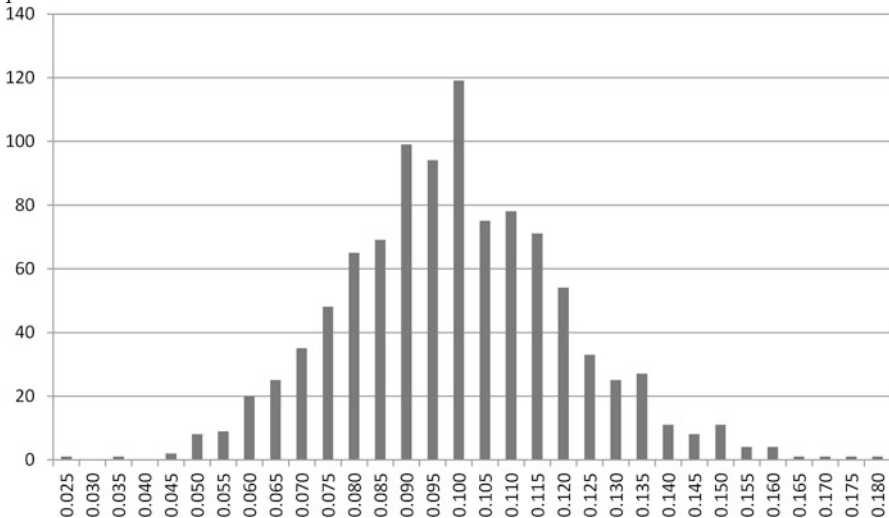
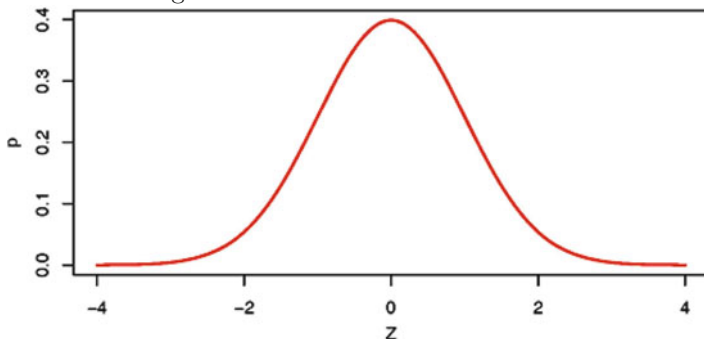


Figure 2.2: Standard Normal Curve.



$$P(a < Z < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dx \quad (2.1)$$

The standard normal distribution is centered at zero with a variance of one (we will formally define center and variance in later chapters). If the center of the distribution is any value other than zero or if the variance is not one, then we simply have a *normal distribution*. The standard normal distribution is graphically presented in Figure 2.2 below. Here we clearly see the *bulge* centered at zero, the gradual decline as we move away from zero, and the tails for unlikely large positive and large negative values far from the center.

To show how the normal distribution works, we have reproduced the histogram in Figure 2.3 and now overlaid a normal curve (like the one in Figure 2.2, but with a mean and variance matching those from the distribution in Figure 2.1). Notice how well the simulated data and the theoretical normal curve align. This generally happens if our simulation study is conducted adequately enough, and this result is actually supported by a statistical law (known as the *central limit theorem*, which we will discuss later). Thus, if our assumptions are met, we should feel comfortable using the normal distribution to represent the distribution of our sample estimate.

2.6 Performing the Test and Decision Making (with R Code)

2.6.1 Test Statistic

So while we can use the standard normal distribution to answer probabilistic statements about our data and hypotheses about the population parameter, a quick glance at Figure 2.1 will show you that the distribution of \hat{p} is *not*

Figure 2.3: Histogram from Simulation Study with Overlaid Normal Curve.

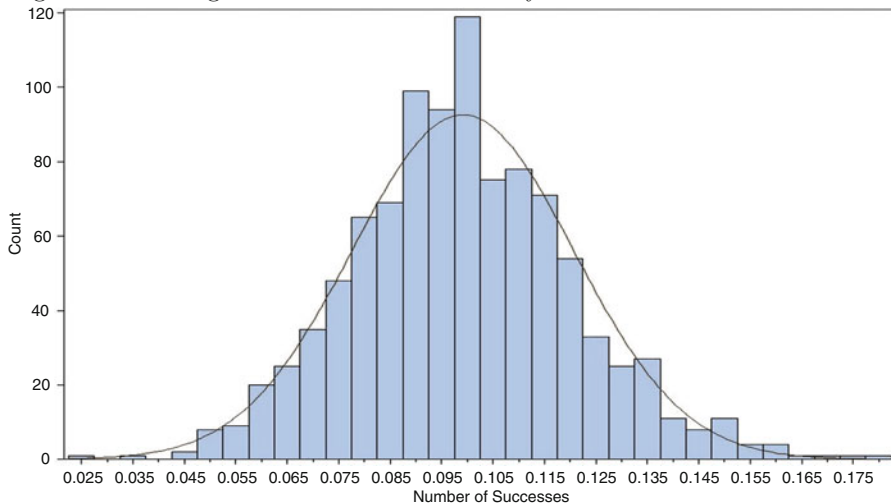
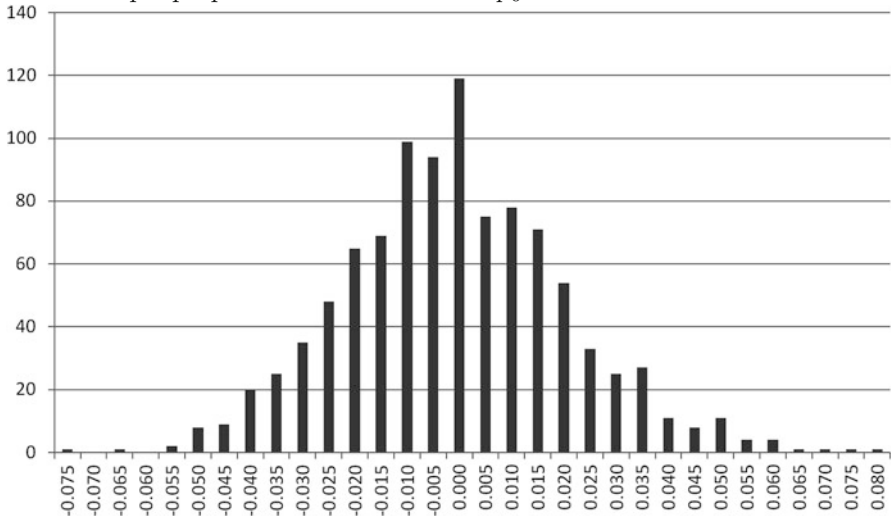


Figure 2.4: Histogram Summarizing Results from Simulation Study of 1,000 Samples of 200 Dichotomous Outcomes with Assumed Success Rate $p = 0.10$, where sample proportions are centered at $p_0 = 0.10$.



centered at zero. Since there is no other standard normal distribution for us to use, we will need to manipulate our sample estimate \hat{p} so that *it will* have a standard normal distribution.

To do this, note that the distribution of \hat{p} in Figure 2.1 is centered near the hypothesized value $p_0 = 0.10$. Thus, to get this distribution centered at zero, we can subtract the hypothesized value from our sample proportion: $\hat{p} - p_0$. Figure 2.4 shows the (nonsensical) adjusted distribution of \hat{p} (nonsensical since it contains negative proportions), which is centered at zero.

While the distribution is now centered correctly, it still requires a variance of one. For reasons that are easy to mathematically justify, but are not easy to explain, the variability of a sample proportion that was drawn from a population with known proportion p is given by the standard error: $\sqrt{\frac{p(1-p)}{n}}$ (note that we use the population proportion and not the sample proportion). Thus, to transform our sample proportion into a random variable that has a standard normal distribution, we center at zero by subtracting p_0 and scale to a variance of one by dividing by the standard error to get Equation 2.2 below.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (2.2)$$

Since the statistic z – known as a *test statistic* – has a standard normal distribution, we can use it to calculate probabilistic statements regarding our hypotheses and ultimately answer our question as to which hypothesis (the null or alternative) the data supports.

Returning to our example, recall that $x = 33$, $n = 200$, $\hat{p} = 0.165$, and $p_0 = 0.10$. Thus, our test statistics is

$$z = \frac{0.165 - 0.10}{\sqrt{\frac{0.10(1-0.10)}{200}}} = \frac{0.065}{0.0212} = 3.064$$

which means that the sample proportion 0.165 is slightly more than three standard deviations above the hypothesized population proportion 0.10 (standard deviation is another measure of variability, which we will define later). Three standard deviations is a lot, and means that in view of our sample data, the hypothesized value is unlikely. We can now use the test statistic z to make a formal decision. Note that we report test statistics – of the form z – to two decimal places, meaning we would report $z = 3.06$.

Unfortunately, R does not compute the test statistic just provided. However, the R function `prop.test()` does provide an equivalent (though cosmetically different) test for proportions, the syntax for which is as follows:

```
prop.test( x, n, p=p_0 )
```

The R code for our example is given in Program 2 below.

Notice that R provides a considerable amount of output (some of which is not needed or has not yet been defined). The `data:` line states what R is testing, which corresponds to the x , n and p_0 . The next line gives the value of the test statistic `X-squared = 9.3889` and states that `p-value =`

Program 2 Program to conduct a hypothesis test on a single proportion.

Code:

```
prop.test(x=33, n=200, p=0.1, correct=FALSE)
```

Output:

```
1-sample proportions test without continuity correction

data: 33 out of 200, null probability 0.1
X-squared = 9.3889, df = 1, p-value = 0.0021
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
 0.1199686 0.2226578
sample estimates:
      p
0.165
```

0.003216; this will be defined below. Hopefully, you are aware that the stated test statistic (9.3889) is different from the 3.064 we calculated by hand. However, note $\sqrt{9.3889} = 3.064$, which is the value we obtained. This relationship will be explained further in the next Chapter.

2.7 Formal Decision Making

2.7.1 Critical Value Method

The most traditional method of making a decision in a hypothesis test is to use critical values. A *critical value* is literally the boundary – in this case from the standard normal distribution – between values of our test statistic that seem likely and values that seem unlikely *if we were to assume that the null hypothesis was true*. Identification of such a critical value (or values) is helpful in the sense that we would only have to calculate our test statistic z and compare it to the critical value to make our decision.

Finding the critical value depends upon our alternative hypothesis and what is called the significance level. The *significance level* – denoted by the Greek letter α – is formally defined as the probability that we reject the null hypothesis (i.e. we don't believe it is true) when the situation it describes is actually true (i.e. rejecting H_0 was a mistake). Informally, α represents the lack of evidence we would need to observe in order for us doubt the veracity of the null hypothesis. Generally, we set the significance level at $\alpha = 0.05$, meaning that we would need to observe a statistic (or a statistic of a *more extreme value*) that we would expect to occur less than 5% of the time in order for us to reject the null hypothesis in favor of the alternative.

Given a specified significance level (and $\alpha = 0.05$ is generally used), the critical value then depends upon our alternative hypothesis. If the alternative specified that the proportion is less than some specified value ($H_A : p < p_0$) then we would expect small sample proportions (or negative values of our tests statistic z) to be rare if we assume $H_0 : p = p_0$ is true, and thus our critical value should be negative. For a similar reason, we have a positive critical value if our alternative hypothesis is $H_A : p > p_0$. In cases of a *two-sided alternative hypothesis* (or $H_A : p \neq p_0$), we need two critical values, since sample values much greater or much lower than our hypothesized value would lead us to reject the null hypothesis. All possible cases and decisions are presented in Table 2.3 for $\alpha = 0.05$ and $\alpha = 0.01$, which are the most commonly used significance levels. Thus, rather than having to determine critical values for each hypothesis test we wish to perform, we can consult Table 2.3 to obtain: (i) the critical value specific to the desired significance level and alternative hypothesis, and (ii) the criterion under which we would select the null or alternative hypothesis.

Table 2.3: Critical Values and Rejection (Acceptance) Regions for Hypothesis Test of a Proportion for Given Significance Levels (α) and Alternative Hypotheses.

Alternative Hypothesis	Critical Value	$\alpha = 0.05$	$\alpha = 0.01$
		Select Hypothesis	Select Hypothesis
$H_A : p < p_0$ (Left-Tailed Test)	-1.645	H_0 if $z \geq -1.645$	H_0 if $z \geq -2.33$
		H_A if $z < -1.645$	H_A if $z < -2.33$
$H_A : p > p_0$ (Right-Tailed Test)	1.645	H_0 if $z \leq 1.645$	H_0 if $z \leq 2.33$
		H_A if $z > 1.645$	H_A if $z > 2.33$
$H_A : p \neq p_0$ (Left-Tailed Test)	-1.96, 1.96	H_0 if $-1.96 \leq z \leq 1.96$	H_0 if $-2.575 \leq z \leq 2.575$
		H_A if $z < -1.96$ or $z > 1.96$	H_A if $z < -2.575$ or $z > 2.575$

In our example, say our original research statement was: *the proportion of subjects who experience reduced side-effects from this treatment is greater than 0.10*. This means our null hypothesis is $H_0 : p = 0.10$ and our alternative hypothesis is $H_A : p > 0.10$, and thus our critical value is 1.645 (if we have $\alpha = 0.05$), meaning that we will reject H_0 in favor of H_A if the test statistic is greater than 1.645, and we will not reject H_0 if the test statistic is less than or equal to 1.645. Earlier, we calculated our test statistic as $z = 3.06$,

which falls in the rejection region, so we reject H_0 in favor of H_A . We will discuss what this means and how we react later.

2.7.2 p -value Method

As an alternative to the critical value method, we can calculate what is called a p -value, which is defined as the probability of observing a test statistic *at least as extreme* as the one we actually observed, given that the null hypothesis is true. This is a tricky definition that has three distinct pieces. First, we must assume that the null hypothesis is true, otherwise we have no bearing to gauge how likely or unlikely the observed data are. Second, the meaning of *at least as extreme* depends upon the alternative hypothesis. If we have a left-tailed test (i.e. $H_A : p < p_0$), then at least as extreme means less than or equal to our observed test statistic. If we have a right-tailed test (i.e. $H_A : p > p_0$), then at least as extreme means greater than or equal to our observed test statistic. If we have a two-tailed test (i.e. $H_A : p \neq p_0$), then at least as extreme means both greater than or equal to the absolute value of our observed test statistic ($|z|$) and less than or equal to the negative of the absolute value of our observed test statistic ($-|z|$).

The third part is calculating the desired probability, which of course depends upon our observed test statistic (z , which itself depends upon the null hypothesis) and the meaning of “at least as extreme” (which is particularly dependent upon the alternative hypothesis). The standard normal distribution is used to calculate p -values, and we generally rely upon statistical software for their computation. Z -tables are used in many elementary Statistics courses, but we will not consult them. P -values can be calculated in Microsoft Excel, and are routinely provided by most statistical software packages (including R; see Program 2 above).

Regardless of the method of computation, the probability being calculated will be the same. If we have a left-tailed test, we calculate the probability that a standard normal random variable Z is less than our observed test statistic z given that the null hypothesis is true (or $P(Z < z|H_0)$). If we have a right-tailed test, we calculate the probability that Z is greater than z (or $P(Z > z|H_0)$). If we have a two-tailed test, then we calculate *two-times* the probability that Z is greater than $|z|$ (or $2P(Z > |z| |H_0)$), or we calculate *two-times* the probability that Z is less than $-|z|$ (or $2P(Z < -|z| |H_0)$). Admittedly, these definitions are complicated, but the good news is that you will not have to calculate them by hand.

To put the p -value into practice, we must compare it to the stated significance level α . In order to reject the null hypothesis, we would need an outcome (or something more extreme) that is less likely than our significance level. Thus, we reject the null hypothesis if our p -value is less than the significance level ($p\text{-value} < \alpha$), and we fail to reject the null hypothesis when our p -value is greater than or equal to the significance level ($p\text{-value} \geq \alpha$).

For our example, our observed test statistic ($z = 3.06$) and the right-tailed hypothesis test means that our p -value is 0.001092. This is less than the significance level $\alpha = 0.05$, so we reject the null hypothesis in favor of the alternative hypothesis. Note that you will make the same decision with the p -value method as you will using the critical value method (meaning, if you come to different conclusions, at least one of them is wrong). We also round p -values to at most four decimal places, so we should report p -value = 0.0011.

2.7.3 Conclusion

Whether we used the critical value or p -value method, we report our results in the same manner. First, we firmly declare whether we rejected or failed to reject the null hypothesis, the former case in favor of the alternative. We then state *in words* what this statistical decision means; as mentioned earlier, statistical methods – such as hypothesis testing – are only useful if we can phrase the results in ways that clinical or non-statistical researchers can understand and interpret.

In our example, our test statistic fell in the rejection region (the p -value was also smaller than the significance level), so we rejected the null hypothesis ($H_0 : p = 0.10$) in favor of the alternative ($H_A : p > 0.10$). So we would declare that the evidence suggests the success rate of this treatment at reducing side-effects is *likely* greater than 0.10. Notice that we did not claim that the success rate is greater than 0.10. This is because we only have statistical evidence, which is not the same as definitive proof.

The R software conducts the two-sided test ($H_A : p \neq p_0$) by default, though we can easily modify the code to conduct either of the one-sided tests. By adding the `alternative` statement to the R function `prop.test()`, R performs the test corresponding to the specified hypothesis. The specific syntax of the `alternative` statement for each type of hypothesis test is given below. Note that if you do not specify the `alternative` statement, R will default to the "two.sided" case and will perform the two-sided test.

$H_A : p \neq p_0 :$

```
prop.test( x, n, p=p0, alternative="two.sided", correct=FALSE)
```

$H_A : p > p_0 :$

```
prop.test( x, n, p=p0, alternative="greater", correct=FALSE)
```

$H_A : p < p_0 :$

```
prop.test( x, n, p=p0, alternative="less", correct=FALSE)
```

For the right-tailed hypothesis in our example ($H_A : p > 0.10$) we would use the following R code (`prop.test(x=33, n=200 p=0.10, alternative = "greater", correct=FALSE)`), to produce the correct right tailed test; note that the p -value you get with this code (0.001092; try it yourself) matches what we reported for the z -test results.

2.7.4 Confidence Intervals

The sample proportion \hat{p} is dependent upon the sample we collect and the particular subjects observed within that sample. In other words, \hat{p} may change if we collect a different sample consisting of different subjects. This is a source of variability that is not expressed if we focus solely upon the current sample estimate. Thus, we often accompany each sample estimate with a *confidence interval* that takes into account sampling variability.

A *confidence interval* is straight forward to calculate, though somewhat tricky to define. What is definitive is what a confidence interval is not. A confidence interval has a stated level of confidence (defined as the complement of the stated significance level, or $1 - \alpha$). For instance, if our significance level is 0.05, then our confidence level is $1 - 0.05 = 0.95$, and we would then construct a 95% confidence interval. This level of confidence is often taken as the quantification of our belief that the true population parameter resides within the estimated confidence interval; this is false. Once calculated, a population parameter is either in a confidence interval or it is not. Rather, the confidence level reflects our belief *in the process of constructing confidence intervals*, so that we believe that 95% of our estimated confidence intervals would contain the true population parameter, if we could repeatedly sample from the same population. This is an important distinction that underlies what classical statistical methods and inference can and cannot state (i.e. we don't know anything about the population parameter, only our sample data).

To calculate a $(1 - \alpha)\%$ confidence interval (or CI) we need three things: a point estimate, a measure of variability of that point estimate, and a probabilistic measure that distinguishes between likely and unlikely values of our point estimate. With these three pieces, our CI would take the form:

$$(\text{Point Estimate} \pm \text{Measure of Variability} \times \text{Probabilistic Measure})$$

The \pm sign indicates that by adding and subtracting the second part from the first we will obtain the upper and lower bounds, respectively, of our confidence interval. For a point estimate we use \hat{p} ; in our example, this value is 0.165. As a measure of variability, we will use the square root of $\hat{p}(1 - \hat{p})/n$, which is similar to the standard error used in hypothesis testing, except here we use \hat{p} instead of p_0 since we don't necessarily want to use a null hypothesis to summarize our data (e.g. sometimes we may only want the CI and not the hypothesis test). Based on our sample data, this value would be $SE = \sqrt{0.165(1 - 0.165)/200} = 0.026$. As a probabilistic measure, we use the positive critical value from a two-tailed test for the given confidence level. For instance, if we want 95% confidence, then we would have $\alpha = 0.05$, and a two-tailed test would yield critical values ± 1.96 , of which we take the positive value 1.96. Putting these values from our example together, our 95% confidence interval is

$$(0.165 - 1.96 \times 0.026, 0.165 + 1.96 \times 0.026) = (0.114, 0.216)$$

To interpret this interval, we would say “a 95% confidence interval of the population proportion of subjects who experienced reduced side-effects with this treatment is (0.114, 0.216)”. In general, we round the confidence interval to the same degree of precision as our point estimate, in this case the sample proportion. Note that some researchers use confidence intervals to conduct hypothesis tests, where they estimate a confidence interval and determine whether some hypothesized value is within the interval (if not, reject H_0 ; if so, fail to reject H_0). While the confidence interval approach is similar in many ways to hypothesis testing, they are not the same and may not produce the same inference. For this and other reasons, we will use confidence intervals only as a form of data summarization, and will not use them for inference. For the record, we do not recommend or condone the use of confidence intervals for making statistical decisions or inference, and strongly encourage you to refrain from this practice.

Note that we could have used R to produce this confidence interval, but it will not *immediately* be the same, since R calculates confidence intervals using what is called the “continuity correction”. This adjustment and the resulting type of interval is an equally valid but all together different type of confidence interval than the method described above; note that what we learned is by far the most commonly accepted form of calculating confidence intervals for dichotomous data. Moving forward, you can choose to calculate 95% CIs on a proportion using the method outlined in this chapter (which requires you to calculate the interval by hand), or you may use the two methods provided by the R software. To get the 95% CI in R, we make use of the `prop.test()` function with the following specifications (`x=33`, `n = 200`, `p = 0.1`, `alternative="two.sided"`), which produces a 95% CI of (0.118, 0.225). Note that this method uses what’s called the continuity correction, which we can turn off by specifying “`correct=False`” in the `prop.test()` function, which gives a 95% CI of (0.120, 0.223). Both of these intervals are similar to but not equal to the interval provided above (0.114, 0.216); ultimately, we would have to create our own code in R (which is not too difficult) to obtain the confidence interval we obtained by hand.

2.8 Contingency Methods (with R Code)

Occasionally we will experience the situation where we wish to compare the proportion to some hypothesized value, but (at least) one of our expected frequencies is less than 5, meaning we do not have a large enough sample size to perform the z -test. In that case, we must instead use the *Binomial Test*, which is a test that compares the proportion to a hypothesized standard and *is valid for any sample size*. This test works for any sample size because it is based on the concept of *enumeration*, or counting all possible outcomes that *could be observed* within one group of categorical data. In this instance enumerating all possible outcomes is not difficult, and can even be done by hand when the sample sizes are small enough.

For instance, imagine the case where someone gives you a cup and tells you it is either filled with Pepsi or Coke (let's say it is actually Pepsi). If you were asked to taste the soda and guess which soda was in which cup, there are only two possible outcomes: you guess correctly or incorrectly. This scenario is numerically represented in Table 2.4. If we had no way to discern between the unique tastes of Pepsi and Coke (i.e. we were simply guessing), then we would assume that either outcome (we guess correctly or incorrectly) would have the same probability (0.5). Given this assumption (which is our null hypothesis), we can calculate the p -value of having as many or more correct guesses than what we observed. Based on the one-cup experiment (Table 2.4), if we were to guess 0 correct, then the p -value = 1.0, because we are certain of getting an outcome at least as extreme as the one we got (i.e. 0 or more correct) the next time we do this experiment. If we guessed correctly, then the p -value = 0.5, meaning there is an equal likelihood of getting a 1 or 0 the next time we do this experiment (the "1" being at least as extreme). Both p -values are much larger than 0.05, so even if we selected correctly, this is not enough evidence for us to reject the null hypothesis.

Table 2.4: Enumeration of Outcomes from One-Cup Experiment (Y: Correct Guess, N: Incorrect Guess).

Actual Soda in Cup		# Correct	Frequency	Proportion	p -value
Pepsi	N	0	1	0.5	$0.5+0.5 = 1.0$
	Y	1	1	0.5	0.5

Now let's assume that we have two cups, where the first is filled with Pepsi and the other with Coke. Of course, we do not know which sodas are actually in each cup, so we could guess that they are both filled with Pepsi, they are both filled with Coke, or they are filled with one soda each (and there are two ways in which this can happen: Pepsi in the first and Coke in the second, or Coke in the first and Pepsi in the second). Thus there are four ways in which we can guess, one resulting in no correct guesses, two resulting in one correct guess (and one incorrect guess), and one resulting in two correct guesses. These outcomes are summarized in Table 2.5. Since the four outcomes are equally probable if we are only guessing (assuming the null hypothesis is true), then each particular outcome has a 0.25 chance of occurring. So in this case, even if we guess the contents of both cups correctly, our p -value (0.25) would still not lead us to reject the null hypothesis.

If we have three cups (filled with Pepsi, Coke and Coke, respectively), there are now eight possible ways in which we can guess, which lead to 0, 1, 2 or 3 correct guesses. The possibilities are listed in Table 2.6. Here we see that even if we were to guess correctly the contents of each cup, the evidence that we actually know what we are doing is still low (p -value = 0.125). So in

Table 2.5: Enumeration of Outcomes from Two-Cup Experiment (Y: Correct, N: Incorrect).

Actual Soda in Cups						
Pepsi	Coke	# Correct	Frequency	Proportion	p -value	
N	N	0	1	0.25	$0.25 + 0.50 + 0.25 = 1.0$	
N	Y	1	2	$2 \times 0.25 = 0.50$	$0.50 + 0.25 = 0.75$	
Y	N	1				
Y	Y	2	1	0.25	0.25	

Table 2.6: Enumeration of Outcomes from Three-Cup Experiment (Y: Correct, N: Incorrect).

Actual Soda in Cups						
Pepsi	Coke	Coke	# Correct	Frequency	Proportion	p -value
N	N	N	0	1	0.125	1.0
N	N	Y	1	3	$3 \times 0.125 = 0.375$	$0.375 + 0.375 + 0.125 = 0.875$
N	Y	N	1			
Y	N	N	1			
N	Y	Y	2	3	$3 \times 0.125 = 0.375$	$0.375 + 0.125 = 0.500$
Y	N	Y	2			
Y	Y	N	2			
Y	Y	Y	3	1	0.125	0.125

this case guessing all of the cups correctly would still lead us to not reject the null hypothesis.

While we will not enumerate the outcomes, Table 2.7 presents the outcomes from both four-cup and five-cup experiments. Here there is still only one way of getting them all correct, but the number of ways in which we can get 0, 1, 2, 3 (or 4) correct answers is quite larger than previously seen. Note that if we select all of the cups correctly in the five-cup experiment, we get a p -value of 0.03125 (in the four-cup case, we still get a high p -value for guessing all cups correctly: p -value = 0.0625). So in this case, the only way we could convince someone that we know how to discern between the tastes of Pepsi and Coke is if we guessed the contents of 5 cups correctly, since there is a small likelihood that we could guess our way to 5 correct cups if we didn't know what we were doing.

In each of these cases, the number of correct guesses follows a *binomial distribution*, where the probability of a given number of correct guesses depends upon both the probability of any given event and the number of different ways in which that outcome can occur. Using this method, we can also calculate the probability (in the form of a p -value) of observing 33 or more successes out of 200 trials assuming the actual rate is 0.10. This value comes out to 0.002916, which is sufficiently small compared to our significance level of $\alpha = 0.05$, and thus we reject the null hypothesis that the population success rate is 0.10 in favor of the alternative that the population success rate is larger.

Table 2.7: Outcomes from Four- and Five-Cup Experiments.

# Correct	Four-Cups			# Correct	Five-Cups		
	Frequency	Proportion	<i>p</i> -value		Frequency	Proportion	<i>p</i> -value
0	1	0.0625	1.0000	0	1	0.03125	1.00000
1	4	0.2500	0.9375	1	5	0.15625	0.96875
2	6	0.3750	0.6875	2	10	0.31250	0.81250
3	4	0.2500	0.3125	3	10	0.31250	0.50000
4	1	0.0625	0.0625	4	5	0.15625	0.18750
				5	1	0.03125	0.03125

We use the `binom.test()` function to calculate the *exact binomial test* in R. The general syntax is similar to the test on proportions using the `prop.test()` function and is given by:

```
binom.test(x, n, p=p0, alternative = c("two.sided", "less",
                                     "greater"), conf.level = 0.95)
```

where `x` is the number of successes, `n` is the number of trials, `p0` is the hypothesized value, `alternative` corresponds to which type of alternative hypothesis you have (with options: `"two.sided"`, `"less"` and `"greater"`) and `"conf.level"` is the desired confidence level (thus, the significance level is *one minus* the confidence level). The statements `x`, `n`, and `p0` are required for a hypothesis test. The default values are 0.5 for `p0`, `"two.sided"` for `alternative`, and 0.95 for `conf.level`. For the example where we have 33 successes in 200 trials we can calculate the exact *p*-value as given in Program 3.

Note in Program 3 that the output is similar to that of `prop.test()`. Here we see the *exact p*-value is given by 0.002916. Using the $\alpha = 0.05$ significance level we would conclude that it is *likely* that the population success rate is greater than 0.10.

2.9 Communicating the Results (IMRaD Write-Up)

The following write-up is an example of the material that we need to communicate if we had actually conducted the example study used for the majority of concepts included in this section. This will form a template of the material that you should include in the write-ups for actual data analyses, though you must note that the specific material that we include in any given IMRaD write-up will depend upon the types of statistical methods we use as well as the specific research question at hand (which will itself call for additional material than what is provided here).

Introduction: Treatments designed to treat certain diseases or conditions often have adverse side-effects that can complicate a patient's reaction to

Program 3 Program to conduct an *exact* hypothesis test on a single proportion.

Code:

```
binom.test(x=33, n=200, p=0.1, alternative="greater")
```

Output:

```

Exact binomial test

data: 33 and 200
number of successes = 33, number of trials = 200, p-value =
0.002916
alternative hypothesis: true probability of success is
greater than 0.1
95 percent confidence interval:
 0.1232791 1.0000000
sample estimates:
probability of success
                0.165

```

the treatment, and can ultimately result in a worse disease or condition prognosis. Clinicians and practitioners are then interested in treatments that have minimal to no side-effects. It was of interest to determine whether the proportion of patients undergoing a particular treatment who experienced little to no adverse side-effects was greater than 0.10.

Methods: The frequency of subjects reporting reduced side-effects from treatment out of 200 subjects is reported, and the proportion of subjects reporting reduced side-effects is summarized with a sample proportion and a 95% confidence interval. We test the null hypothesis of a 0.10 success rate ($H_0 : p = 0.10$) against a one-sided alternative hypothesis that the success rate is greater than 0.10 ($H_A : p > 0.10$) by using a chi-square-test with significance level $\alpha = 0.05$. We will reject the null hypothesis in favor of the alternative hypothesis if the p -value is less than α ; otherwise we will not reject the null hypothesis. The R statistical software was used for all analyses.

Results: Assuming that the sample was representative and subjects were independent, the sample was large enough to conduct the statistical analysis. Out of a sample of 200 total patients, 33 reported reduced symptoms ($\hat{p} = 0.165, 95\%CI : 0.120, 0.223$). Using this data, a chi-square test yielded a p -value of 0.0011, which is less than the stated significance level. Thus, we reject the null hypothesis in favor of the alternative hypothesis.

Discussion: The sample data suggest that the proportion of patients who reported reduced side-effects using this treatment is greater than 0.10. Thus, clinicians and practitioners interested in treating patients with reduced treatment-related side-effects may wish to consider this treatment.

2.10 Process

1. State research question in form of testable hypothesis.
2. Determine whether assumptions are met.
 - (a) Representative sample.
 - (b) Independent measurements.
 - (c) Sample size: calculate expected frequencies
 - i. If $np_0 > 5$ and $n(1 - p_0) > 5$, then use z -test or chi-square test (in R).
 - ii. If $np_0 < 5$ or $n(1 - p_0) < 5$, then use binomial test.
3. Summarize data.
 - (a) If $np_0 > 5$ and $n(1 - p_0) > 5$, then report: frequency, sample size, sample proportion and CI.
 - (b) If $np_0 < 5$ or $n(1 - p_0) < 5$, then report: frequency and sample size.
4. Calculate test statistic.
5. Compare test statistic to critical value or calculate p-value.
6. Make decision (reject H_0 or fail to reject H_0).
7. Summarize with IMRaD write-up.

2.11 Exercises

1. A researcher is interested in the proportion of active duty police officers who pass the standard end of training fitness test. They take a random sample of 607 officers from a major metropolitan police force and administer the fitness test to the officers. They find that 476 of the officers were able to successfully pass the fitness test. Create 96% confidence interval for the proportion of all active duty police officers on the police force that can pass the fitness test.
2. Occupational health researchers are interested in the health effects of sedentary occupations such as call center workers. Specifically they are interested in lower back pain. They conduct a survey of 439 call center workers and record whether or not the worker has back pain at the end of their shift. The surveys show that 219 workers reported back pain. In the general population there are reports that 25% of workers have back pain. Conduct a hypothesis test to determine if call center workers have a higher rate of back pain than the general population of workers.

3. In December 2012 Gallup Poll conducted a survey of 1,015 Americans to determine if they had delayed seeking healthcare treatment due to the associate costs. Of the participants, 325 reported delaying seeking treatment due to costs. Create a 95% confidence interval for the proportion of all Americans who have delayed seeking healthcare treatment due to costs.
4. [Norman et al. \(2013\)](#) Consider the preference for walkability of neighborhoods for obese men. They studied 240 obese men and asked them their preference for walking behavior in neighborhoods. They found that 63 responded as they walked for transportation. Create a 99% confidence interval for the proportion of obese men who walk for transportation.
5. [Barrison et al. \(2001\)](#) are interested in the reasons that proton pump inhibitors were prescribed. Of the 182 gastroenterologists 122 of them prescribed proton pump inhibitors to patients. Create a 98% confidence interval for the proportion of all gastroenterologists who prescribe proton pump inhibitors.
6. [Barrison et al. \(2001\)](#) were interested in the proportion of physicians who deemed that proton pump inhibitors (PPI) should be sold over the counter. Of the 391 physicians surveyed 59 responded that PPIs should be sold over the counter. Create a 92% confidence interval for the proportion of physicians who think that PPIs should be sold over the counter.
7. [Keightley et al. \(2011\)](#) are concerned with obese peoples self perceptions. They hypothesize that a majority of obese people can identify their own body type. They conduct a study with 87 obese people and find that 7 can correctly identify their body type. Conduct a hypothesis test to determine whether or not their hypothesis is warranted.
8. [Salerno et al. \(2013\)](#) is interested in determining the current infection rate of Chlamydia and Gonorrhea infections. They obtained a sample of 508 high school students who consented to a urine test to screen for the two diseases. Of the participants 46 tested positive for at least one of the diseases. Create a 99% confidence interval for the proportion of all high school students who have one of the two diseases.

Chapter 3

Two-Sample Proportions

In the last Chapter we focused on estimating and conducting a hypothesis test on a proportion from a single group. In practice, we are often interested in comparing proportions from two separate groups, and as such we would perform a hypothesis test comparing the proportions from those two different samples. The process for the two-sample case is similar to that for the one-sample case, in that we will go through the same general steps, though the details of those steps will be different. Further, there are additional statistical techniques that we perform, depending upon the status of our assumptions.

3.1 Summarizing Categorical Data with Contingency Tables (with R Code)

In this chapter we still focus on the case where our outcome measure – in both samples – is dichotomous, meaning that those sample measures are best represented by sample proportions. As always for dichotomous data, we want to report the frequency, sample size and sample proportion for each of our two groups; note that you want to calculate (but not report) the overall sample proportion (i.e. pool the data from both groups) for reasons that will become clear later.

For the two-sample case, the data summaries can be efficiently presented in what is called a *contingency table*. In its simplest form, a 2×2 (read: “two-by-two”) contingency table lists the frequencies of each outcome for each treatment in tabular form, where the rows represent *group membership* and the columns represent *outcome membership*. An example is presented in Table 3.1 below. In this case, the two groups consist of n_1 and n_2 subjects, respectively, for a total of $n_1 + n_2 = n$ total subjects. Note that in the first group, a of the n_1 subjects take the outcome “Yes”, while b subjects take the outcome “No”, and that $a + b = n_1$. Likewise, in the second group there

are c “Yes” outcomes and d “No” outcomes, such that $c + d = n_2$. The total number of subjects taking a particular outcome across group membership is generally not of interest.

Table 3.1: Example 2×2 Contingency Table.

Group	Outcome		Group Sample Size
	Yes	No	
1	a	b	$n_1 = a + b$
2	c	d	$n_2 = c + d$
Total	$a + c$	$b + d$	$n = n_1 + n_2$

In practice, we will expand upon the simple 2×2 contingency table to also include the sample proportions and 95 % CI confidence intervals for the proportions in each group. Returning to the example used in Chapter 2, where 33 out of 200 patients reported little or no side-effects associated with a particular treatment, let us now assume that 8 of 100 patients on a control treatment reported little or no side-effects. Thus, there is a 0.165 success rate for the treatment group, and a $8/100 = 0.080$ success rate in the control group. To compare these two groups we look at the data summary presented in the contingency table found in Table 3.2.

Table 3.2: Contingency Table for Symptom-Relief Example.

Group	Outcome			Proportion Reporting Little or No Symptoms	
	Yes	No	Sample Size	Observed	95 % CI
Treatment	33	167	200	0.165	0.114, 0.216
Control	8	92	100	0.080	0.027, 0.133
Total	41	259	300		

Note that the confidence intervals for the proportion of “Yes” outcomes in each treatment group were calculated using the methods provided in Chapter 2. We could have used R to produce confidence intervals, but – as mentioned in Chapter 2 – they are not going to be the same, since R calculates these intervals using an alternative method. Using the `prop.test()` function in R, we obtain a (0.120, 0.223) 95 % CI for the treatment group and a (0.041, 0.150) 95 % CI in the control group. Again, note that these intervals are similar but not equal to the confidence intervals provided in Table 3.2.

Before continuing with the hypothesis testing process, we must take time to discuss what can and cannot be compared in this table. Let us consider the frequency of “Yes” outcomes in the two groups. One might be tempted to conclude – since there are 33 “Yes” outcomes in the Treatment group and only 8 “Yes” outcomes in the Control group – that the treatment works

better than the control at reducing side-effects. However, *this is not a fair comparison* since the sample sizes for the two groups are not equal. Rather, we look to the sample proportions for comparisons between the two groups, since the sample proportion adjusts the sample frequency by the sample size to give an *average response rate*. In our example, we can see that the rate of those reporting little or no side effects is higher (0.165) in the treatment group than the corresponding rate (0.080) in the control group. This is indeed a fair comparison to make, though we came to the same conclusion reached by looking at the sample frequencies. So consider another example, where 20 out of 200 subjects in the first group responded “Yes”, and 10 out of 100 subjects in the second group responded “Yes”. Here, we would erroneously claim that the first group has a higher rate of “Yes” respondents than the second group (even though it actually does have more “Yes” responses: 20 compared to 10), since once we adjust for sample size, both sample proportions are 0.10. In summary, do not compare frequencies, even if the sample sizes are the same. Rather, look to the sample proportions and conduct a hypothesis test.

To create a contingency table in R, we need to organize the data in “matrix” form. This is done by entering the number of successes and failures for each group into the `c()` (see below). By also specifying the desired number of rows and columns (using the `nrow` and `ncol` commands, we can create the table using the `matrix()` function in the following coding.

```
Table1 <- matrix(c( 33, 8, 167, 92 ),
                 nrow=2, ncol=2)
```

This R code above creates our table, which we have named `Table1`. After the `matrix()` statement we list values for the table using the `c()` function, and then the numbers of rows and columns. Note that if the number of items in the list (i.e. the `c()` function) do not match the specified number of elements in the matrix (i.e. the product of the numbers of rows and columns), then R will not produce the table. Of course, it is often a simple process to create a contingency table using word processing software (e.g. Microsoft Word), and we recommend doing so, as it encourages you to check your work.

3.2 Hypothesis Test for Comparing Two Population Proportions

3.2.1 Generating Hypotheses About Two Proportions

Like we did in Chapter 2, we need to turn a research question into a set of testable hypotheses (a null and alternative) that will allow for statistical application. Unlike the last Chapter, we now have *two proportions* in which we are interested, so hypotheses we wish to test must arise from a *single statement* of both proportions. We now are concerned with population proportion

p_1 for group 1 and population proportion p_2 for group 2, and to make matters simple we will take their difference $p_1 - p_2$. Focusing on the difference between the two group proportions will allow us to phrase our research statements in the same way as we had previously, and even use the same symbols. For instance, if our research statement is that the proportion in group 1 is larger than in group 2 (or $p_1 > p_2$), we get the following statement by focusing on the difference ($p_1 - p_2 > 0$). Likewise, if our research statement is that the group 1 proportion is smaller than the group 2 proportion (or $p_1 < p_2$), then we get ($p_1 - p_2 < 0$) by focusing on the difference, and if we state that the proportions are equal (or $p_1 = p_2$), we could equivalently focus on ($p_1 - p_2 = 0$). This logic follows for statements that include \leq , \geq or \neq . The process for turning a research statement into a set of hypotheses is similar to that covered in Chapter 2, and Table 3.3 contains the three different null and alternative hypotheses that can arise from certain types of research questions for comparing two proportions.

Table 3.3: Possible Sets of Hypotheses for Comparing Two Population Proportions Based Upon Key Phrases in Research Question.

Hypothesis	Key Phrases of p_1 relative to p_2				
	"less than"		"greater than"		"equal to"
	"greater than or equal to"	"at least"	"less than or equal to"	"at most"	"not equal to"
Null	$H_0 : p_1 - p_2 = 0$		$H_0 : p_1 - p_2 = 0$		$H_0 : p_1 - p_2 = 0$
Alternative	$H_A : p_1 - p_2 < 0$		$H_A : p_1 - p_2 > 0$		$H_A : p_1 - p_2 \neq 0$

3.2.2 Statistical Assumptions

Before proceeding with the hypothesis test, we need to ascertain whether the assumptions necessary to conduct that test are met. Since there are two samples under consideration in this case, we need to be certain that *both samples* are representative of the populations from which they are drawn. In this textbook we generally assume that the samples we work with are representative, though in practice this depends upon the sampling methods used by the researchers who collected the data. The subjects within these two samples (and between) also need to be independent of one another, in the sense that the value one subject takes for a particular outcome cannot depend upon the value that any other subject takes. We generally assume that subjects (and thus the samples) are independent if we know that they were collected randomly. In the two-sample case, it may also be necessary for subjects to be allocated randomly into one of the two groups (such as in a clinical trial with competing treatments), though this would not be necessary if the groups are based on *fixed concepts* (such as gender or ethnicity).

Determining adequate sample size is a little more complicated than it was in the one-sample case. Again, we need to *expect* at least five subjects to take both values of the dichotomous outcome, though now we need to expect this for both groups of subjects, and we will again base our expectations on the null hypothesis. Regardless of our research question and alternative hypothesis, the null states that p_1 and p_2 are equal. Our best guess of what the population proportion would look like if there was *actually no difference* between the two groups is to pool the data from those two groups together to form a *grand proportion* $\bar{p} = (a + c)/(n_1 + n_2)$. If this value \bar{p} was truly the population proportion, then we can determine what we would expect to observe in each group by multiplying \bar{p} by the sample size in each group. These resulting values are our *expected frequencies of successes*, and the *expected frequencies of failures* can be calculated by using the compliment rule (i.e. subtracting) for each group.

The expected frequencies for our example are listed in Table 3.4. The grand proportion obtained by pooling the outcomes from the treatment and control groups is $\bar{p} = 0.137$, which – as one might expect – is somewhere between the Treatment group proportion of 0.165 and the Control group proportion of 0.080 (it is closer to 0.165 because the treatment group has more subjects than the control group). Based on this grand proportion, we see that the expected frequencies of “Yes” outcomes for both groups (27.3 and 13.7, respectively) are greater than 5, as are the expected frequencies of “No” outcomes (172.7 and 86.3, respectively). Thus, we have adequate sample size to conduct the hypothesis test. Note, however, that we would claim that *we do not* have adequate sample size if *any* of our expected frequencies (for both outcomes in either group) were less than 5. We will develop a more general rule in Chapter 4.

Table 3.4: Observed and Expected Frequencies for Two-Sample Symptom Relief Example.

Group	Observed			Expected	
	Yes	No	Sample Size	Yes	No
Treatment	33	167	200	27.3	172.7
Control	8	92	100	13.7	86.3
Total	41	259	300	$\bar{p} = 41/300 = 0.137$	

If we do not wish to calculate these values by hand, we can ask R to do the calculations for us. After entering the cell frequencies in matrix form (as `table1`) in Program 4 below), we need to call the `chisq.test()` function, which under normal circumstances produces results from the chi-square test (more on this below). For our purposes, we name this function (here we chose `expval1`), and then ask for the expected values using the `expval1$expected` line; note here that the key is adding the function `$expected` to our named output `expval1`. We then see that the resulting output matches what we calculated by hand.

3.3 Performing the Test and Decision Making (with R Code)

Since our hypotheses are in terms of the difference $p_1 - p_2$, our sample estimate of that difference ($\hat{p}_1 - \hat{p}_2$) will be the focus of our test statistic. We also need a statistic to measure the variability of our sample estimate *under the condition that the null hypothesis is true*. Since we do not have specific hypothesized values for p_1 or p_2 , we again make use of the grand proportion \bar{p} in the calculation of our standard error. In the case of comparing two sample proportions, we use the following test statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}. \quad (3.1)$$

Program 4 Program to generate expected values for symptom relief example.

Code:

```
# Create the table
table1 <- matrix(c(33,8,167,92),nrow=5,ncol=2)

# Name and invoke the chi-square test
expval1<-chisq.test(table1)

# Ask for expected values
expval1$expected
```

Output:

```
      [,1]      [,2]
[1,] 27.33333 172.66667
[2,] 13.66667  86.33333
```

Note that under the null hypothesis the difference $p_1 - p_2$ is equal to zero. Using the data from our current example, we get the following value for our test statistic

$$z = \frac{0.165 - 0.080}{\sqrt{\frac{0.137(1-0.137)}{200} + \frac{0.137(1-0.137)}{100}}} = \frac{0.085}{0.043} = 2.02, \quad (3.2)$$

which implies that the difference in the sample proportions (0.085) is slightly more than two standard deviations above the hypothesized difference of zero. Whether or not two standard deviations is a lot is unclear at this point, so we need to look at the critical value and p -value methods to make a more informed comparison of these two proportions. As always, we round the test statistic z to two decimal places.

3.3.1 Critical Value Method

For reasons similar to those covered in the single-proportion case of Chapter 2, the test statistic in the 2-sample case will also follow a standard normal distribution if our assumptions are met (especially that of sample size). Note that we could show that the test statistic has a standard normal distribution by conducting another simulation study, but at this point it should suffice to simply state that the result holds. This means that the critical values used for choosing between our competing hypotheses are the same as they were before and depend only upon our choice of significance level and the direction of the inequality found in the alternative hypothesis. These values are presented for the two-sample case in Table 3.5 below. Note that the only differences from this table and Table 2.3 are the parameters found in the alternative hypothesis; everything else, including our decision making process, is the same.

Table 3.5: Critical Values and Rejection (Acceptance) Regions for Hypothesis Test of Two Proportions for Given Significance Levels (α) and Alternative Hypotheses.

Alternative Hypothesis	Critical Value	$\alpha = 0.05$		$\alpha = 0.01$	
		Select Hypothesis	Critical Value	Select Hypothesis	Critical Value
$H_A : p_1 - p_2 < 0$ (Left-Tailed Test)	-1.645	H_0 if $z \geq -1.645$	-2.33	H_0 if $z \geq -2.33$	
		H_A if $z < -1.645$		H_A if $z < -2.33$	
$H_A : p_1 - p_2 > 0$ (Right-Tailed Test)	1.645	H_0 if $z \leq 1.645$	2.33	H_0 if $z \leq 2.33$	
		H_A if $z > 1.645$		H_A if $z > 2.33$	
$H_A : p_1 - p_2 \neq 0$ (Left-Tailed Test)	-1.96, 1.96	H_0 if $-1.96 \leq z \leq 1.96$	-2.575, 2.575	H_0 if $-2.575 \leq z \leq 2.575$	
		H_A if $z < -1.96$ or $z > 1.96$		H_A if $z < -2.575$ or $z > 2.575$	

For our example, let's say we had the following research statement: the success rate for the treatment group is greater than the success rate for the control group. Based on the phrasing of this statement (notably the words "greater than"), our null and alternative hypotheses are $H_0 : p_1 - p_2 = 0$ and $H_A : p_1 - p_2 > 0$, and our critical value is 1.645. This means that we will reject H_0 if our test statistic is greater than 1.645, and we will fail to reject H_0 if our test statistic is less than or equal to 1.645. We have previously seen that our test statistic is 2.02, which is greater than our critical value

of 1.645, so we reject the null hypothesis in favor of the alternative. Thus, the data suggest that the difference in success rates ($p_1 - p_2$) is greater than zero, which is another way of saying that the data suggests that p_1 is greater than p_2 .

3.3.2 p -Value Method

Like the critical value method just covered, the p -value method is mostly the same in both the one- and two-sample proportion cases. The calculation of a p -value again depends upon the magnitude (numbers) and direction (+ or - sign) of the test statistic, as well as the alternative hypothesis. If we have a left-tailed test, we calculate the probability that a standard normal random variable Z is less than our observed test statistic z given that the null hypothesis is true (or $P(Z < z|H_0)$). If we have a right-tailed test, we calculate the probability that Z is greater than z (or $P(Z > z|H_0)$). If we have a two-tailed test, then we can either calculate the two-times the probability that Z is less than $-|z|$ (or $2P(Z < -|z||H_0, z < 0)$), or two-times the probability that Z is greater than $|z|$ (or $2P(Z > |z||H_0, z > 0)$). As always, we reject the null hypothesis if the p -value is less than the significance level, and we fail to reject the null hypothesis if the p -value is greater than or equal to the significance level. Note that the p -value and critical value methods will always give the same conclusion for the two-sample proportion case.

Returning to our example, note that we have a right-tailed alternative hypothesis $H_A : p_1 - p_2 > 0$, so we calculate our p -value as the probability of having a test-statistic greater than or equal to our observed test statistic, assuming that the null hypothesis is true (equal proportions). Thus, our p -value is 0.0217 (rounded to four decimal places), which is less than $\alpha = 0.05$, so we reject H_0 in favor of H_A , and conclude that the data suggests that the success rate in the treatment group is larger than the success rate in the control group. Using the `prop.test()` function in R – learned in Chapter 2 – provides us with the p -value for this test (but unfortunately not the critical value). Here we need to organize the information by the number of successes and the sample size for each group. Since there are 33 successes in the treatment group (out of $n_1 = 200$) and 8 successes in the control group (out of $n_2 = 100$), we will create two “vectors” of this information using the `c()` function, which in our example would be `c(33, 8)` for the successes and `c(200, 100)` for the sample sizes. Program 5 shows the code and the resulting output for our test.

We specified the appropriate right-tailed alternative hypothesis by including the `alternative="greater"` command; we also could have selected from `"two.sided"` or `"less"` which would have provided the two-tailed and left-tailed tests, respectively. Here we see that the p -value we obtained is 0.02167, which matches what we obtained by hand – once we round appropriately. However, notice that this test does not provide the observed test

Program 5 Program to generate hypothesis test for comparing two proportions.

Code:

```
# Call prop.test()
prop.test(c( 33, 8 ), c( 200, 100 ), alternative="greater",
correct=FALSE)
```

Output:

```
2-sample test for equality of proportions with continuity
correction
```

```
data:  c(33, 8) out of c(200, 100)
X-squared = 4.0823, df = 1, p-value = 0.02167
alternative hypothesis: greater
95 percent confidence interval:
 0.02291086 1.00000000
sample estimates:
prop 1 prop 2
 0.165  0.080
```

statistic for the z -test and instead provides a value called **X-squared**; we will talk about the source of this value momentarily. Also notice in the first line of the output it says "...without continuity correction". The continuity correction is an alternative approach that adjusts the specified rates (successes and failures) to better match the normal distribution.

3.3.3 Confidence Intervals

In an effort to further summarize our data, we can produce a confidence interval on the observed difference between the two sample proportions ($\hat{p}_1 - \hat{p}_2$). As explained in Chapter 2, the confidence interval combines our estimator (in this case the difference in sample proportions) with a measure of the variability of that estimator (the standard error) and a probabilistic measure indicating reasonable likelihood (critical value). As in Chapter 2, the standard error takes the form of the denominator of the test statistic used for hypothesis testing, where we now use the actual sample proportions rather than a hypothesized value or the grand proportion, and takes the form

$$SE = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}. \quad (3.3)$$

The probabilistic measure is the same as the critical value we would use for a two-sided alternative hypothesis, and depends upon the significance level. If $\alpha = 0.05$, then we use 1.96; if $\alpha = 0.01$, then we use 2.575.

For our example, based on the observed sample proportions and sample sizes, we get the following 95 % confidence interval for the difference in sample proportions

$$((\hat{p}_1 - \hat{p}_2) \pm 1.96 \times SE) = ((0.165 - 0.08) \pm 1.96 \times 0.038) = (0.011, 0.159). \quad (3.4)$$

So the 95 % CI for the difference in success rates between the treatment and control groups is (0.011, 0.159). While we use this CI for data summary purposes – as we would with the 95 % CIs for the treatment group (0.114, 0.216) and the control group (0.027, 0.133) – note that it corroborates our result from the hypothesis test since 0 is not contained within the interval. Again, since our standard error expression is different for CIs than it is for hypothesis testing, these results are not necessarily the same over all examples, and as such we will not use confidence intervals to make inference on the two population proportions. To calculate the confidence interval on the difference between two proportions in R, we again make use of the `prop.test()` function. Being sure to specify the `alternative="two.sided"` option, Program 6 shows the code and the resulting output for our test. Here we see that Program 6 produces both the hypothesis test and the confidence interval on the difference. In this case the 95 % confidence interval is (0.01101623, 0.15898377), which we round to (0.011, 0.159) and exactly matches what we obtained by hand. If we were to incorporate the continuity correction, we would obtain (0.004, 0.166), which is slightly wider than our original confidence interval.

Program 6 Program to generate a confidence interval for a difference in proportions.

Code:

```
# Call prop.test()
prop.test(c( 33, 8 ), c( 200, 100 ), alternative="two.sided",
correct=FALSE)
```

Output:

```
2-sample test for equality of proportions without continuity
correction
```

```
data:  c(33, 8) out of c(200, 100)
X-squared = 4.0823, df = 1, p-value = 0.04333
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01101623 0.15898377
sample estimates:
prop 1 prop 2
 0.165  0.080
```

3.3.4 Chi-Square Test

If our assumptions are met (especially a large enough sample size), then we can use an alternative method for conducting a two-tailed hypothesis test (i.e. $H_A : p_1 - p_2 \neq 0$). This test is based upon the *chi-square* probability distribution, and is actually a special case of the more general test we will learn in Chapter 4. To conduct this test, we must revisit the contingency table, where here we focus on the number of success and failures in both groups. This revisited contingency table is found in Table 3.6. Here we note

Table 3.6: Contingency Table for Chi-Square Test.

Group	Observed		Total
	Yes	No	
1	a	b	$a + b$
2	c	d	$c + d$
Total	$a + c$	$b + d$	$n = a + b + c + d$

that the marginal column and row totals (i.e. the sums we obtain by adding all values down one column or across one row) are just as important as the observed values, as noted in the following test statistic

$$\chi^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}. \quad (3.5)$$

We will expect this test statistic to take small values if ad is close in value to bc , which happens when the number of successes in Group 1 is close (relative to sample size) to the number of successes in Group 2. If the relative numbers of successes in the two groups are not close, then the test statistic χ^2 will take larger values.

From our example, we get

$$\chi^2 = \frac{300((33)(92) - (167)(8))^2}{(41)(259)(200)(100)} = 4.08. \quad (3.6)$$

Note that the square root of this value is our test statistic z from earlier (i.e. $\sqrt{\chi^2} = \sqrt{4.0823} = 2.02047 = z$). This result will always hold for the two-sample/two-outcome case, and implies that both the z -test and the chi-square test will always give the same result. Note that this test statistic is automatically produced in R (see Program 6 above).

We now need to determine how to make a decision for the chi-square test based on the observed test statistic χ^2 . As stated earlier, this test statistic will follow the chi-square probability distribution, which is a probability distribution for certain random variables that only take positive values. Since we can view the chi-square test statistic as the square of the standard normal test statistic z , we know that χ^2 will always take positive values and the

use of the chi-square probability distribution makes sense. However, because of this relationship, we can only use the chi-square test when we have two-sided alternative hypotheses, since the test would not be able to distinguish between positive and negative test statistics for one-sided alternatives.

The chi-square distribution is parametrically dependent upon the so-called *degrees of freedom*, which can be thought of as the number of independent pieces of information we have available in our contingency table. To illustrate the concept of degrees of freedom, imagine we had a contingency table (Figure 3.1) where we knew the marginal totals of the columns and rows, but not the particular numbers within the table. Based on only the marginal column and row totals, we do not have enough information to fill in the rest of the table. However, observe what happens when we fill in any one of the four interior parts: the rest of the table *must take certain values due to the complement rule*. So by knowing just one of those four pieces (along with the marginal totals), we can figure out the remaining three. By no coincidence, our chi-square test has 1 degree of freedom.

As mentioned earlier, the chi-square distribution is characterized solely by the degrees of freedom, which take positive values (usually integers). Various plots of this distribution for degrees of freedom 1 through 5 are provided in Figure 3.2. For small degrees of freedom (less than 2), the curve is highest at 0 and slowly tapers off (indicating that values slightly above 0 are more likely

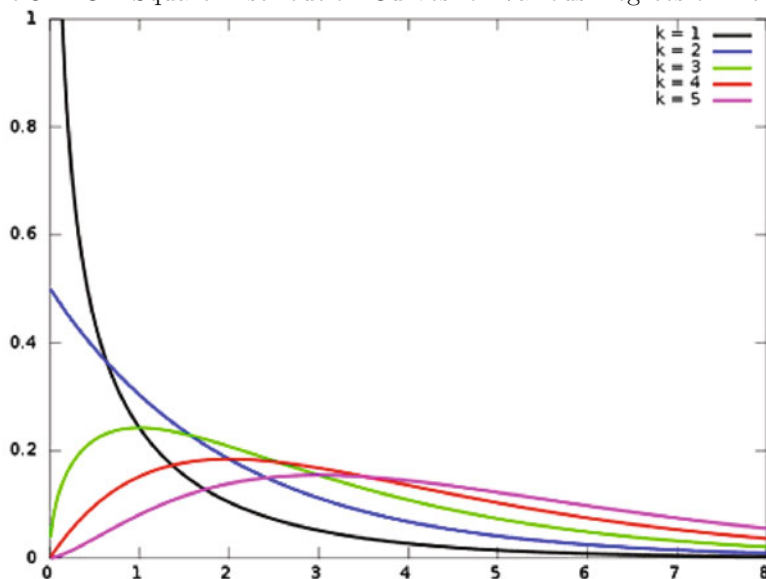
Figure 3.1: Illustration of Degrees of Freedom to Complete a Contingency Table.

Outcome			
Group	Yes	No	Total
Treatment	?	?	200
Control	?	?	100
Total	41	259	300

Outcome			
Group	Yes	No	Total
Treatment	33	?	200
Control	?	?	100
Total	41	259	300

Outcome			
Group	Yes	No	Total
Treatment	33	167	200
Control	8	?	100
Total	41	259	300

Figure 3.2: Chi-Square Distribution Curves for Various Degrees of Freedom.



than larger values). For larger degrees of freedom (3 and above), the center of the curve is removed from zero, implying that values under that center are more likely than smaller or larger values (from that center). Of special importance is noting what happens to the curve as the degrees of freedom increases: as the degrees of freedom changes from 3 to 4 to 5, the central mound of the curve (indicating the mode, or most likely value) is shifting to the right, and the curve more closely resembles a symmetrical curve. In fact, for large enough degrees of freedom (k), the chi-square curve is indistinguishable from a normal curve with mean $k - 2$ and variance $2k$.

For the chi-square test, we can obtain a critical value or use the p -value method, though we will rely upon statistical software for the calculations. When we have one-degree of freedom and significance level $\alpha = 0.05$, the critical value for a chi-square distribution is 3.841 (note that $(1.96)^2 = 3.841$). From our example, since our test statistic $\chi_1^2 = 4.08$ is greater than 3.84, we reject the null hypothesis in favor of the alternative (the two proportions are most likely not equal). To obtain the p -value, we would calculate the probability that a chi-square random variable with 1 degree of freedom takes a value greater than or equal to our test statistic. For our example, we get a p -value = $P(\chi^2 > 4.08) = 0.0433$, which is less than our significance level, so we again reject the null hypothesis in favor of the alternative. Note that for the chi-square test we most often use the p -value method, and report the test statistic, degrees of freedom and p -value in the following manner: $\chi_1^2 = 4.08$, p -value = 0.0433.

Note also that this is the same p -value we would get from a two-sample z -test with a two-sided hypothesis, and it is also 2-times the p -value we would obtain from a right-tailed z -test. Thus, if we wanted to turn the p -value from a chi-square test into the p -value from either the left-tailed or right-tailed z -test (but not both), we simply divide by 2. However it is not always clear to which test (the left- or right-tailed z -test) this halved p -value will apply. If we have a right-tailed alternative and our z -test statistic is positive, then the chi-square p -value will be two-times the z -test p -value. Likewise, if we have a left-tailed alternative and our z -test statistic is negative, then the chi-square p -value will again be two-times the z -test p -value. However, if we have a right-tailed alternative and a negative z -test statistic, or a left-tailed alternative and a positive z -test statistic, then the two p -values do not coincide.

In our previous use of the `prop.test` function (which earlier provided us with the chi-square test statistic, degrees of freedom, and p -value), we can also use the `chisq.test()` function to obtain the relevant information. Using the same definitions used in creating the contingency table (i.e. `table1`), we simply place `table1` in the `chisq.test()` function, as shown in Program 7 below.

Program 7 Program to conduct a chi-square test on a contingency table.

Code:

```
# Create the table
table1 <- matrix(
  c( 33, 8, 167, 92 ),
  nrow=2,
  ncol=2
)

# Run the test
chisq.test(table1,correct=FALSE)
```

Output:

Pearson's Chi-squared test

```
data: table1
X-squared = 4.0823, df = 1, p-value = 0.04333
```

The output for the test conducted in Program 7 gives the basic information for the chi-square test: the test statistics **X-squared**; the degrees of freedom **df** associated with the test; and the p -value. Notice that the result from the p -value method (reject H_0 since $0.04333 < 0.05$) matches what we

observed earlier using the critical value method, and also indicates a significant result. Note that – as was the case with the `prop.test` function – we need to “turn off” the continuity correction by specifying the `correct=FALSE` option. Had we failed to do so, R would have provided a slightly different result.

3.4 Contingency Methods (with R Code)

Occasionally we will experience the situation where we wish to compare the proportions from two groups of subjects, but (at least) one of our expected frequencies is less than 5, meaning we do not have a large enough sample size to use either the z -test or the χ^2 -test. In that case, we must instead use *Fisher’s exact test*, which is a test that compares the two proportions and is *valid for any sample size*. Fisher’s exact test works for any sample size because – like the binomial test from Chapter 2 – it is based on the concept of counting all possible outcomes that could be observed between two groups of categorical data. In this instance enumerating all possible outcomes is not difficult, and can be done by hand when the sample sizes are small enough; unlike in Chapter 2, we will not show how Fisher’s exact test works.

In practice, computer software will do these types of enumerations for us. To calculate Fisher’s exact test in R we use the `fisher.test()` function by specifying the contingency table (e.g. `table1` above). The output for the two-tailed test is provided below in Output 8, where the two-tailed hypothesis p -value is 0.049886, which we round to 0.0499. We could obtain p -values for the left-tailed or right-tailed alternative hypotheses by specifying `alternative="less"` or `alternative="greater"`, which would have provided p -value = 0.9887 for the left-tailed hypothesis and p -value = 0.02945 for the right-tailed hypothesis. If we stick with our original right-tailed alternative hypothesis, since the p -value = 0.0295 is less than our significance level $\alpha = 0.05$, we reject the null hypothesis in favor of the alternative hypothesis that the treatment success rate is most likely larger than the control success rate. Note that since this method counts all possible outcomes, it may take a considerable amount of time for enumeration when sample sizes are large, and may require a computer with a sufficient amount of RAM in order to complete all enumerations.

3.5 Odds Ratio (with R Code)

An alternative measure used to compare the relative success of some measurement between two groups is the *odds ratio*. This measure – while ubiquitously used in the health sciences – can be somewhat challenging to fully understand, as it is based on the probabilistic concept of odds. Most of us use the idea of odds qualitatively (e.g. the odds of a team winning a game are high), but we may be less familiar with how to use them quantitatively. Probabilistically, the odds of some event are defined as the ratio of the frequency (a)

Program 8 Program for Fisher's Exact Test.

Code

```
fisher.test(table1, alternative="two.sided")
```

Output:

Fisher's Exact Test for Count Data

```

data:  table1
p-value=0.04986
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9744107 5.9231475
sample estimates:
odds ratio
 2.266913

```

with which some event did occur to the frequency (b) with which some event did not occur, such that the odds are listed as $a : b$ or a/b (read “ a -to- b ”). For example, if we are to flip a coin and are interested in the likelihood of having a “head” landing facing up, then there are two possible outcomes we could observe (i.e. heads or tails). Of these, heads is the outcome where our event occurs, and tails is the outcome where our event does not occur. Thus, the odds for a “head” are 1:1 or 1/1 (read “1-to-1”). These are even odds, meaning that a heads is equally likely to occur or not occur (we know this because the numbers on either side of the colon “:” are equal). If the number to the left of the colon is larger than the number to the right, then the event is more likely to occur than to not occur, and if the number to the right of the colon is larger than the number to the left, then the event is more likely to not occur than to occur. Note that the odds for some event ($a : b$) are directly tied to the probability of that event ($a/(a+b)$). In our coin example, the 1:1 odds for a head translates into a $1/(1+1) = 1/2$ probability of having a head land up.

Turning to our example, the number of successes in the treatment group was 33 (out of 200). Thus, the odds of success in the treatment group are 33:167; the odds of success in the control group are 8:92. These odds imply that a success is less probable than a failure in both groups. While unfortunate, these values do not answer our question of whether the treatment reduces symptoms compared to the control. This is where the odds ratio becomes useful. Evaluating the fractions implied in each of the odds ($33/167 = 0.1976$; $8/92 = 0.0870$) and – as the name implies – taking their ratio gives the odds ratio ($OR = 0.1976/0.0870 = 2.27246$), which we round to at most two decimal places ($OR = 2.27$).

Note that if the odds of success in each group were equal, the odds ratio would be 1; conversely, an odds ratio of 1 implies that the odds of some event are equal between two groups. If the odds ratio is less than one, then the odds of the event are greater in the second group than in the first, and an odds ratio greater than one implies that the odds of the event are greater in the first group than in the second. For our example, $OR = 2.27$ implies that the odds of reduced symptoms are greater in the treatment group than in the control group. Specifically, we can state that the odds of having reduced symptoms in the treatment group are 2.27 times the odds of having reduced symptoms in the control group (try it: $2.27246 * 0.0870 = 0.1976$). Or for those with more confidence in their quantitative skills, we could say that the odds of reduced symptoms are 127 % larger in the treatment group than in the control group. To calculate this difference, turn the odds ratio into a percentage by moving the decimal two places to the right, add a percent sign, and then subtract 100 %. For example, if the $OR = 2.0$, then 2.0 turns into 200 %, and after subtracting 100 % we are left with 100 % (i.e. the odds in the first group are 100 % larger than the odds in the second group, or twice as large). In our example, since the $OR = 2.27$, then 2.27 turns into 227 %, and after subtracting 100 % gives us 127 % (i.e. the odds in the treatment group are 127 % larger than the odds in the control group). This process can also be used when an odds ratio is less than one to determine by what percentage that the odds in one group are *smaller* than the odds in another group.

There are several methods for generating the confidence interval of an odds ratio, each of which is somewhat involved. We will thus report the confidence interval without explaining its derivation. Recall that R provided the odds ratio in the output for Fisher's Exact Test in Program 8. This value was listed as 2.266913, which – after rounding to 2.27 – we see is nearly identical to what we calculated by hand. Rather than rely upon this output, we will use the `Oddsratio` function, as shown in Program 9 below. In this output we see the estimated odds ratio is 2.272455 (which we round to 2.27), and the 95 % confidence interval for the odds ratio is (1.007657, 5.124812) (which we round to (1.01, 5.12)). Note that this interval lies completely above 1, which implies that the odds of success in the treatment group are larger than that in the control group. Note that we specify `method="Wald"` in the `oddsratio` function to use the same method of calculating the odds ratio as when done by hand, though others can be specified.

3.6 Communicating the Results (IMRaD Write-Up)

The following is an example of the IMRaD write-up for our two-sample example.

Introduction: Treatments designed to treat certain diseases or conditions often have adverse side-effects that can complicate a patient's reaction to

the treatment, and can ultimately affect the disease or condition prognosis. Clinicians and practitioners are interested in treatments that have no or minimal side-effects. It was of interest to determine whether the proportion of patients reporting reduced side-effects from a particular treatment was greater than the proportion of patients reporting reduced side-effects from a control treatment.

Program 9 Program for Odds Ratio.

Code

```
oddsratio(table1, method="Wald", correction=FALSE)
```

Output:

Outcome

Predictor	Disease1	Disease2	Total
Exposed1	33	167	200
Exposed2	8	92	100
Total	41	259	300

\$measure

odds ratio with 95% C.I.

Predictor	estimate	lower	upper
Exposed1	1.000000	NA	NA
Exposed2	2.272455	1.007657	5.124812

Methods: The frequency of subjects reporting reduced side-effects as well as the total sample size are reported for both the treatment ($n = 200$) and control ($n = 100$) groups, and the proportions of subjects reporting reduced side-effects in both groups are summarized with sample proportions and 95% confidence intervals. The difference in sample proportions is also presented, as is a 95% confidence interval on the difference between the two group proportions. (*If z-test is used:*) We test the null hypothesis of no difference in success rates ($H_0 : p_1 - p_2 = 0$) against a one-sided alternative hypothesis that the difference in success rates is greater than 0 ($H_A : p_1 - p_2 > 0$) by using a two-sample z-test with significance level $\alpha = 0.05$. (*If chi-square test is used:*) We test the null hypothesis of no difference in success rates ($H_0 : p_1 - p_2 = 0$) against a two-sided alternative hypothesis that the difference in success rates differs from 0 ($H_A : p_1 - p_2 \neq 0$) by using a chi-square test with significance level $\alpha = 0.05$. We will reject the null hypothesis in favor of the alternative hypothesis if the p -value is less than α ; otherwise we will not reject the null hypothesis. The R statistical software was used for all statistical analyses.

Results: The data are summarized in Table 3.7 below. Assuming that the two samples are representative and subjects are independent, the two samples are large enough to conduct the statistical analysis. The observed success rate in the treatment group (0.165, 95 %CI : 0.114, 0.216) is significantly larger than that in the control group (0.080, 95 %CI : 0.027, 0.133), with an observed difference of 0.085 (95 %CI : 0.011, 0.159). (*If z-test is used:*) The two-sample z-test yielded p -value = 0.0217, so we thus reject the null hypothesis in favor of the alternative hypothesis. (*If chi-square test is used:*) The chi-square test ($\chi_1^2 = 4.1$, $df = 1$, p -value = 0.0433) yielded a small p -value, so we thus reject the null hypothesis in favor of the alternative hypothesis

Table 3.7: Data Summary.

Group	Outcome		Sample Size	Proportion Reporting Little or No Symptoms	
	Yes	No		Observed	95 % CI
Treatment	33	167	200	0.165	0.114, 0.216
Control	8	92	100	0.080	0.027, 0.133
			Diff	0.085	0.011, 0.159

Discussion: The sample data suggest that the proportion of patients who reported reduced side-effects using the treatment is greater than the proportion who reported reduced side-effects using the control. Thus, clinicians and practitioners interested in treating patients with reduced side-effects due to the treatment may wish to consider this treatment.

3.7 Process

1. State research question in form of testable hypotheses.
2. Determine whether assumptions are met.
 - (a) Representative
 - (b) Independence
 - (c) Sample size: calculate grand proportion and expected frequencies
3. Summarize data with contingency table.
 - (a) If sample size is adequate: summarize groups with frequencies, sample sizes, proportions and CIs, and report difference in sample proportions and CI for difference.
 - (b) If sample size is inadequate: report frequencies and samples sizes for each group.

4. Perform Test.
 - (a) If sample size is adequate: calculate z-test or chi-square test statistic.
 - (b) If sample size is inadequate: perform Fisher's Exact test.
5. Compare test statistic to critical value or calculate p -value.
6. Make decision (reject H_0 or fail to reject H_0).
7. Summarize with IMRaD write-up.

3.8 Exercises

1. Police officer fitness is important for the ability for the police force to complete its mission. A researcher is interested in determining if differences exist between the fitness levels of female and male officers. He collects a sample of 212 female officers and 316 male officers. For each of the officers a fitness test is given and it is recorded whether or not the officer passed the test. The results of the tests were as follows: 162 females passed the test and 222 males passed the test. Determine if there is difference in the proportion who pass the fitness test across gender.
2. In October 2012 Gallup Poll conducted a survey comparing rates of exercise between Britons and Germans. The survey consisted of 7,786 Germans and 7,941 Britons aged 18 or older. The participants were asked if they exercised at least 30 min three times a week or more. This showed that 4,288 Britons and 5,840 Germans reported that they exercise at least 30 min three or more times per week. Conduct a test to determine if Germans exercise more than Britons.
3. [Justesen et al. \(2003\)](#) conducted a retrospective pharmacokinetic study to determine the long-term efficacy in HIV patients of a combination of indinavir and zidovudine. Of partial interest was the number of patients who remained in the treatment regimen for the entire 120 weeks, as per the study design. Compare the rate of patients who remained on treatment for the entire duration between patients who had or who had not previously experienced protease inhibitors. The data are provided in the following table.

Previous protease inhibitor experience	Remained in Regimen 120 weeks	
	Yes	No
Yes	8	4
No	2	7

4. In a study by [Engs and Hanson \(1988\)](#), college students were asked whether they had ever driven an automobile after having consumed alcoholic beverages. One goal of this study was to determine if the percentage of students responding “yes” had changed after a law (students were originally assessed in 1983, and the law passed in 1987) raised the minimum age permitting the purchase of alcohol. Using the data provided in the following table, compare the rates of students who stated that they did not drive after consuming alcohol.

Drove after drinking	Year	
	1983	1987
Yes	1,250	991
No	1,387	1,666

5. [Flynn and Allen \(2004\)](#) are interested in the reporting deficiencies in documentation from operating rooms. When a surgeon performs an operation a comprehensive operative note should be generated to document the procedure, give indication for why the procedure was needed and to have a record for billing and reporting purposes. Certified professional coders reviewed 550 operative notes from a multi-specialty academic practice to determine the proportion of reporting deficiencies. Of the 550 records reviewed 213 were dictated by a faculty member and 337 were dictated by residents. Faculty member reports contained 107 deficiencies and resident reports contained 201 deficiencies. Determine if there is a difference in the proportion of operative note deficiencies between faculty members and residents.
6. [Salerno et al. \(2013\)](#) is interested in determining the current infection rate of Chlamydia and Gonorrhea infections. They obtained a sample of 508 high school students (226 males and 282 females) who consented to a urine test for the two diseases. Of the participants 14 males and 32 females tested positive at the screening for at least one of the diseases. Based on this information can we say that the infection rate differs across genders?

Chapter 4

Multi-category Data

4.1 Introduction: Types of Multi-categorical Data

The discrete data we considered in the last two chapters were of the dichotomous nature, where each subject could take one of only two values for some measurement (Yes or No, Success or Failure, etc.). We now generalize to the case where one or more of our variables under consideration has more than two possible values. These types of polytomous or multi-category data are more complicated than their dichotomous brethren, and as such must be handled differently.

Before discussing the hypothesis testing process, we must first designate between two different types of multi-category data: nominal and ordinal. Ordinal data are a type of multi-category data where the various levels that a subject can assume have a distinct and informative ordering. For instance, the letter grade a student can receive in a course could be one of five distinct grades (F, D, C, B and A). These grades are in fact categories (they most certainly aren't numbers), and the distinct values have a natural ordering to them: a D is *better or higher* than an F, a C is better than both a D or an F, a B is better than a F, D and C, and an A is better than all other grades. This order to the categories exists *without* us having any subjects to measure, and as such is a characteristic of the multi-category data itself. There are many other examples of ordinal data that we will experience, such as age groups (20–29, 30–39, 40+) and subject preference (dislike, indifferent, like). In contrast, nominal data are a type of multi-category data where the various levels that a subject can assume have no natural ordering, where one particular ordering is just as informative as any other ordering. For instance, patient ethnicity can generally assume one of several values (African American, Caucasian, Hispanic, etc.), but there is no meaningful way to order or rank them (AA-C-H is no more informative than H-AA-C or C-H-AA).

For nominal data, there is no information other than what is provided by each subject (i.e. how we present the categories – we have to do it somehow – means nothing).

While the distinction between nominal and ordinal data is important, we will not focus too much more on their differences. In fact, the method in which we analyze multi-category data (in this Chapter at least) is the same regardless of whether that data are nominal or ordinal. Rather, we will focus on the differences between two different cases of multi-category measurements. In the first case we have one multi-category measurement (which may be nominal or ordinal) and one clearly defined dichotomous outcome (e.g. success or failure). This can be viewed as the generalization of the two-sample proportion comparison case to the multi-sample proportion comparison case, and as such the hypothesis for this case is called a *test of homogeneity of proportions*. In the second case we have two multi-category measurements (either may be nominal or ordinal), where we may not have a clearly defined outcome. Since there may not be a natural set of proportions suitable for comparison in this case, we instead *test for associations* between such multi-category variables. The distinctions between testing for homogeneity and testing for association will affect how we summarize and report our results, though the statistical methods used for hypothesis testing are the same for both cases.

4.2 Summarizing Categorical Data (with R Code)

When one of our categorical measurements is a dichotomous outcome and the other is a multi-categorical measurement, we will be interested in comparing the proportions for one of the two outcome levels across each level of the multi-level variable. As such, our contingency table for summarizing data in this instance should, along with reporting all cell frequencies and sample sizes for each level of both categorical variables, present the proportion of interest for each level of the descriptive categorical variable. For example, Table 4.1 presents a contingency table of two categorical variables measuring 283 student grades from an undergraduate statistics course, along with an indicator of whether or not those students got into the graduate program of their choice. From the observed counts, we can see that the numbers of students accepted to their preferred program increases with their undergraduate statistics course grade, and the counts for not getting into their preferred program generally decrease as the course grades increase.

Analyzing the trends across grades can be more appropriately accomplished by calculating proportions, but we must *decide which proportions to analyze*. Note that R can provide us with three different proportions: column, row and total percentages. Except in the rarest of circumstances, we will not report the *total percentages*, which present the cell percentage with respect

Table 4.1: Contingency Table of Student Grades in an Undergraduate Statistics Course and Whether or Not the Student was Accepted to Graduate Program of Choice.

Grade	Accepted to program of choice		Total
	Yes	No	
A	52	7	59
B	48	13	61
C	41	35	76
D	11	33	44
F	4	39	43
Total	156	127	283

to the total across all cells. *Column percentages* present the cell percentage with respect to the total within a particular column, while *row percentages* present the cell percentage with respect to the total within a particular row. Our choice between column and row percentages boils down to the placement of the outcome variable. If our levels of the outcome variable are presented as the columns of the contingency table, then we would want to compare the success proportion/percentage (generically) for each level of the other multi-category variable (or the rows of the contingency table), so we would select row percentages. However, if the levels of the outcome variable are presented as the rows of a contingency table, then we would want to compare the success proportions across the levels of the other multi-category variable (in this case the columns), so we would select column percentages. So as a general rule, if we are comparing proportions for one outcome level across the rows of a contingency table, then we select row percentages, and if we are comparing proportions for one outcome level across the columns of a contingency table, then we select column percentages.

Returning to our example, we see in Table 4.1 that the outcome levels are presented as the columns. Since we are interested in comparing the proportion of “Yes” outcomes across the various grade levels, we want to present the row proportions (which are produced by selecting the row percentages). Specifically, we will present only the proportions for the value “Yes”, since we will be comparing those proportions across the levels of the Grade variable. The updated contingency table is presented in Table 4.2. Note, we only have to compare the value of one level of the outcome across the values of the other variable; since the outcome is here dichotomous, comparing the “Yes” proportions is the same as comparing the “No” proportions. We can obtain the “Yes” proportions in R (as well as the “No” proportions) using the `prop.table` function. After entering our frequencies in matrix form `table1<-matrix(c(52,48,41,11,4,7,13,35,33,39),nrow=5,ncol=2)`, we get the row proportions for both outcomes by specifying `prop.table(table1,1)`, where the number “1” indicates “row percentages”; we would specify `prop.table(table1,2)` if we had wanted column percentages.

We also have data from a hospital survey of patient satisfaction of treatment for their medical needs based on the severity of their condition. Satisfaction is ordinally measured in three levels (Low, Medium and High), while condition severity is also ordinally measured in three levels (Minimal, Moderate and Severe). The data from this study are presented in Table 4.3. Note that in this case that a clear pattern is difficult to immediately discern. Further, even though patient satisfaction can be thought of as a clear outcome here (the level of satisfaction must be determined after treatment of the condition was given), a comparison of proportions does not seem to make sense, and if it did, it may not be clear as to which proportions we should compare. Regardless, we need to look at *some proportions* so as not to possibly be misled by the cell frequencies and differing sample sizes across the different levels of both variables.

Table 4.2: Contingency Table (with Proportions of Interest) of Student Grades in an Undergraduate Statistics Course and Whether or Not the Student was Accepted to the Graduate Program of Choice.

Grade	Accepted to program of choice		Total	Proportion “Yes”
	Yes	No		
A	52	7	59	0.88
B	48	13	61	0.79
C	41	35	76	0.54
D	11	33	44	0.25
F	4	39	43	0.09
Total	156	127	283	

Table 4.3: Contingency Table of Patient Satisfaction and Severity of Condition.

Disease Severity	Patient satisfaction			Total
	Low	Medium	High	
Minimal	7	3	18	28
Moderate	7	10	8	25
Severe	19	4	11	34
Total	33	17	37	87

Recalling that *total proportions* are non-sensical for most purposes, we need to decide between reporting *row* or *column proportions*. If neither of our two variables were outcomes, then we could choose either *row* or *column proportions*, or whichever best exemplified any relationship between the two variables (remember: you want to place your variables so that you would

look at row proportions). Since the disease severity variable can be seen as a grouping variable (since patient satisfaction can be seen as an outcome), it makes more sense to report *row proportions* for this example, which are seen in the complete Table 4.4 below. This enables us compare the distribution of patient satisfaction counts for each level of disease severity. For instance, we can see that the percentage of patients reporting “Low” satisfaction increases with disease severity, while the percentage of patients reporting “High” satisfaction decreases as disease severity increases. The percentage of patients reporting “Medium” satisfaction is highest for moderately severe diseases. Provided our frequencies are entered in matrix form (say `table2`) we can obtain the percentages in R using the following code: `prop.table(table2,1)`.

4.3 Establishing Hypotheses: Difference Between Comparisons and Association

When we have two multi-category variables, we have to choose between testing for homogeneity (equivalent proportions) and testing for association (subjects taking combinations of values between the two variables in a non-random manner). This choice can be determined by the research question, but is often determined by the types of multi-category variables we have. Generally speaking, when we have a dichotomous outcome, or if we are explicitly interested in comparing the proportion of one level of the multi-category outcome across the levels of the other variable, then we will perform a test of homogeneity. The null hypothesis for this test is that the proportions for one specific level of the outcome (e.g. the proportion of “Yes” responses: p_Y) are the same across all “A” levels of the other multi-category variable, or $H_0 : p_{Y1} = p_{Y2} = \dots = p_{YA}$. The alternative is a little more complicated than what we have seen previously. One would be tempted to state that the alternative is that all of the pairwise proportion combinations differ, or

Table 4.4: Contingency Table of Patient Satisfaction and Severity of Condition with Counts and Row Percentages.

Disease Severity	Patient satisfaction			Total
	Low	Medium	High	
Minimal	7	3	18	28
	25 %	11 %	64 %	
Moderate	7	10	8	25
	28 %	40 %	32 %	
Severe	19	4	11	34
	56 %	12 %	32 %	
Total	33	17	37	87

$H_A : p_{Yi} \neq p_{Yj}$, for all i and j . However, this is only one way in which the null hypothesis can be false. Recall that the null and alternative hypotheses must be mutually exclusive: either one or the other must be true. Another requirement is that the null and alternative hypotheses should together account for all possible outcomes. Since the null accounts for only one scenario (all proportions are equal), the alternative hypothesis must account for all of the possibilities that *could occur* if H_0 is not true. Another possibility of a non-null outcome is if all pairs involving the first proportion p_{Y1} were not the same, but all other pairings were the same. If we follow this type of example to its extreme, we can achieve a catch-all phrase that reflects all possibilities that are non-null if we state $H_A : p_{Yi} \neq p_{Yj}$, for at least one combination i and j . Thus, if there was only one true difference between any of the success proportions, then H_0 would be violated and the alternative hypothesis (H_A) would apply. Likewise, if several or all of the proportions were different, then H_0 would be false and H_A would be true. Only in the event that all of the proportions were equal (or not significantly different) would we believe that H_A is unlikely and H_0 was more likely.

When we have more than two values for our outcome, or if there is no natural outcome between our two variables, we are likely to be interested in the relationship between them and will test for association rather than homogeneity. The null hypothesis for this test cannot be easily expressed in symbolic notation, so we would state that under the null hypothesis the two variables are not associated. If the null hypothesis was true and the two multi-category variables are not associated, then the distribution of proportions for one variable level across all values of the other variable would approximate the proportion distributions for all other levels of the first variable. Statistically, this will include cases when any differences in the various proportions are too small to notice. The alternative hypothesis states that there is a relationship between the two variables, but what that relationship could be is not explicitly stated. While there are countless examples of what this relationship could be, generally it means that the distribution of proportions for specific values of one variable changes depending upon the level of the other variable. An easily explained instance of this is with ordinal data, when the proportions of subjects taking higher values of one variable increase with as the level of the other variable increases. For an alternative hypothesis that is mutually exclusive of the null, we simply state that there is a relationship or association between the two variables. What that relationship could be is described after viewing the results.

Returning to our running our examples, it should be clear in the case of the graduate admissions example that we should perform a test of homogeneity, since we have a dichotomous outcome and we are explicitly interested in comparing success proportions across the various grade levels. For this example, our null hypothesis of equal proportions is $H_0 : p_{YA} = p_{YB} = p_{YC} = p_{YD} = p_{YF}$ against the alternative $H_A : p_{Yi} \neq p_{Yj}$ for at least one pair of $i \neq j = A, B, C, D$ or F . In the patient satisfaction example, there are too many levels for both variables for us to conduct a test of homogeneity,

so we will instead test for association. Our null hypothesis in this instance is that disease severity is unrelated to patient satisfaction with treatment, while our alternative hypothesis is that the level of patient satisfaction depends upon disease severity.

4.4 Assessing Assumptions (with R Code)

As always, we require a representative sample and independent subjects, which we will assume if our data were collected through random sampling. Checking to see if we have the required sample size is complicated, but is done by calculating *expected frequencies*, as was done in Chapter 3. This is done by again calculating the overall or grand proportion \bar{p} for each level of one of the variables. By essentially ignoring the levels of the other variable, we are effectively calculating the proportions of subjects who take each level of the specified variable if the other variable did not have an effect on it. For the graduate admissions example, irrespective of grade, there are 156 students who were accepted into their graduate program of choice and 127 who were not, meaning that the grand proportion of students accepted into their program of choice is $\bar{p}_Y = 156/283 = 0.551$, and the grand proportion of students not accepted is $\bar{p}_N = 0.449$. Using these grand proportions and the actual sample sizes for each level of the heretofore ignored grade variable, we can calculate expected frequencies by multiplying the grand proportions by those sample sizes. The expected frequencies are rounded to the first decimal place, and are presented in Table 4.5 below. In this most general of cases for categorical data, we are looking to see that at most 20% of the expected frequencies are less than 5. So if there are 10 cells, no more than 2 of them can be 5 or less. If there are 20 cells, no more than 4 of them can be 5 or less. In cases where 20% of the total number of cells is not an integer (e.g. if there are 8 cells, 20% of which is 1.6), then we round down, so that in this case at most 1 cell can be 5 or less. In Table 4.5, we see that none of the cells have expected frequencies lower than 5 (none are lower than 19.0), so we have adequate sample size to perform this test.

Fortunately for us, these expected values can be generated in R. After having placed our cell frequencies in matrix form (say `table1` that we created earlier), we must invoke the `chiq.test(table1)` function as if we wished to conduct a chi-square test. However, as seen in Program 10 below, we must label the test so that we can call it again later. In this case, we name the test `expval1<-chisq.test(table1)` so that we can call it using the line `expval1$expected`. This latter command produces the expected values in Program 10, which we see matches the expected values we calculated by hand in Table 4.5.

We can assess the sample size for the patient satisfaction example in a similar way, though in this case we would have to calculate three grand

Table 4.5: Expected Frequencies for Graduate Admissions Example

Grade	Accepted to program of choice			Grand	Expected		
	Yes	No	Total	Proportions	Yes	No	Yes
A	52	7	59	0.551	0.449	32.5	26.5
B	48	13	61	0.551	0.449	33.6	27.4
C	41	35	76	0.551	0.449	41.9	34.1
D	11	33	44	0.551	0.449	24.3	19.7
F	4	39	43	0.551	0.449	23.7	19.3
Total	156	127	283	0.551	0.449		

Program 10 Program to generate expected values for graduate admissions contingency table.

Code:

```
# Create the table
table1 <- matrix(c(52, 48, 41, 11, 4,
                  7, 13, 35, 33, 39),
                nrow=5, ncol=2)

# Name and invoke the chi-square test
expval1<-chisq.test(table1)

# Ask for expected values
expval1$expected
```

Output:

```
      [,1]      [,2]
[1,] 32.52297 26.47703
[2,] 33.62544 27.37456
[3,] 41.89399 34.10601
[4,] 24.25442 19.74558
[5,] 23.70318 19.29682
```

proportions either for each level of patient satisfaction (0.397, 0.195 and 0.425) or for each level of disease severity (0.322, 0.287 and 0.391). Proceeding the same way as we did for the previous example, we get the expected counts in Table 4.6 below. In this table, the total column proportions (for patient satisfaction) were used to calculate the expected counts, which were found by multiplying the three total proportions by the total sample size in each row. In this case, note that the cell for “Moderate” disease severity and “Medium”

patient satisfaction is less than 5. However, this is only one cell out of a total of nine, which is less than 20 % ($1/9 = 0.11 < 0.20$), so we have an adequate sample size to perform the hypothesis test for this problem. These values can be generated in R in the same manner as in the graduate admissions example by first labeling the chi-square test (`expval2<-chisq.test(table2)`) and then asking for the expected values (`expval2$expected`). You may verify these values on your own.

4.5 Performing the Test and Decision Making (with R Code)

Regardless of the type of hypothesis we have (either homogeneity or association) the test statistic is the same. There is a closed form equation for this test, but it is not as simple to use as the tests from Chapters 2 and 3. The statistic takes the following form:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (4.1)$$

Here O_{ij} is the observed frequency for cell ij (the i th level of the row variable and the j th level of the column variable), E_{ij} is the expected frequency for cell ij , and the summation (Σ) is over all cells. For each cell in the contingency table, we take the difference between the observed and expected frequency, square that difference, divide that squared-difference by the expected frequency, and then sum those values across all cells. Since the differences are squared, this test statistic cannot take negative values, and as such will follow a chi-square distribution (assuming the null hypothesis of equal proportions or no association is true).

Based on the observed and expected frequencies found in Table 4.5 for the graduate admissions example, we get the differences found in the first part of Table 4.7. Note that the differences for each row are mirror images of each other (e.g. 19.5 and -19.5), which results from the null conditions used in calculating the expected frequencies. Further, while the values for the squared-differences in each row are the same, once we divide by the expected cell frequencies the resulting values become unique. For this example, the test statistic takes the value 92.36423, which we round to the first decimal place as 92.4.

To obtain this test statistic in R, we use the `matrix()` function to enter our cell frequencies (as previously mentioned). Program 11 shows how to code the graduate admissions example in R and provides the corresponding output, where we find the χ^2 statistic (as well as the degrees of freedom and p -value),

Table 4.6: Expected Frequencies for Patient Satisfaction Example.

Severity	Patient satisfaction				Grand proportions			Expected frequencies		
	Low	Med	High	Total	Low	Med	High	Low	Med	High
Min	7	3	18	28	0.379	0.195	0.425	10.6	5.5	11.9
Mod	7	10	8	25	0.379	0.195	0.425	9.5	4.9	10.6
Severe	19	4	11	34	0.379	0.195	0.425	12.9	6.6	14.5
Total	33	17	37	87	0.379	0.195	0.425			

Table 4.7: Process for Calculating Test Statistic for Graduate Admissions Example.

Grade	Observed–expected		(Observed–expected) ²		(Observed–expected) ² / Expected	
	Yes	No	Yes	No	Yes	No
A	19.5	–19.5	380.3	380.3	11.7	14.3
B	14.4	–14.4	207.4	207.4	6.2	7.6
C	–0.9	0.9	0.8	0.8	0.0	0.0
D	–13.3	13.3	176.9	176.9	7.3	9.0
F	–19.7	19.7	388.1	388.1	16.4	20.1
					Sum = 92.36423	

which matches the result we obtained earlier by hand. We should remember to examine the degrees of freedom to ensure that the correct analysis was performed.

The process for calculating the test statistic in the patient satisfaction example is listed below in Table 4.8. Note that the symmetry in the observed differences (the “mirror image” effect in the graduate admissions example) is slightly more complicated in this case, but is ultimately meaningless and disappears if we don’t round the values. After squaring the differences and dividing by the expected cell counts, we get a test statistic of 16.89274, which we round to 16.9. We can obtain these results in R using Program 11 with the following matrix: $matrix(c(7, 7, 19, 3, 10, 4, 18, 8, 11), nrow = 3, ncol = 3)$.

4.5.1 Critical Value Method

Large values of this test statistic indicate that the observed frequencies differ from the expected frequencies for at least some cells. Since the expected cell frequencies are calculated assuming the null hypothesis is true, these large differences would indicate that the conditions of the null hypothesis are not likely. Small values of the test statistic indicate that the observed cell frequencies are close in value to the expected cell frequencies, which would indicate that the conditions of the null hypothesis are more likely.

Program 11 Program to conduct a chi-square test on a multi-category contingency table.

Code:

```
# Create the table
table1 <- matrix(c(52, 48, 41, 11, 4,
                  7, 13, 35, 33, 39),
                nrow=5, ncol=2)
```

```
# Run the test
chisq.test( table1 )
```

Output:

Pearson's Chi-squared test

```
data: table1
X-squared = 92.3642, df = 4, p-value = 2.2e-16
```

Table 4.8: Process for Calculating Test Statistic for Patient Satisfaction Example.

Severity	Observed-expected			(Observed-expected) ²			(Observed-expected) ² /Expected		
	Low	Med	High	Low	Med	High	Low	Med	High
Min	-3.6	-2.5	6.1	13.1	3.1	37.1	1.2	1.1	3.1
Mod	-2.5	5.1	-2.6	6.2	26.2	6.9	0.7	5.4	0.7
Severe	6.1	-2.6	-3.5	37.3	7.0	12.0	2.9	1.1	0.8
							Sum = 16.89274		

To determine what differentiates a large value from a small value, we can find a critical value from the chi-square distribution. To find the degrees of freedom we can follow the same process we used in Chapter 3, which was to determine the number of “free cells” in a contingency table in which we can enter a given frequency before all of the other cells are known (assuming the marginal row and column totals are known and fixed). Though this process is more complicated for multi-category variables than it was for dichotomous variables, there exists a mathematical expression for this number, which is the product $(a - 1)(b - 1)$, where a is the number of levels for the first variable and b is the number of levels of the second variable. This value is then the degrees of freedom we use to calculate the critical value, which is the 95th percentile (generally the $100 \times (1 - \alpha)$ th percentile) from the chi-square distribution with the stated degrees of freedom.

In the graduate admissions example, there are $a = 5$ grade levels and $b = 2$ outcome levels, so that there are $4 \times 1 = 4$ degrees of freedom. Thus the critical value from a chi-square distribution with 4 degrees of freedom is 9.488. Since our test statistic is $\chi^2 = 92.36 > 9.488$, we reject the null hypothesis in favor of the alternative and declare that the success proportions are not homogeneous. For the patient satisfaction example, there are $a = 3$ disease severity levels and $b = 3$ patient satisfaction levels, so there are $2 \times 2 = 4$ degrees of freedom. The critical value from the chi-square distribution with 4 degrees of freedom is 9.488 (the same as the previous example), and since our test statistic $\chi^2 = 16.9 > 9.488$, we reject the null hypothesis in favor of the alternative hypothesis and declare that disease severity and patient satisfaction are related.

4.5.2 p -Value Method

Rather than find a critical value, we can calculate a p -value using the test statistic and the stated significance level α . As always, if the p -value is less than α , we reject the null hypothesis in favor of the alternative, and claim that either (i) the proportions of interest are not likely homogeneous, or (ii) the two variable are likely associated. If the p -value is greater than or equal to α , then we fail to reject the null hypothesis.

For the graduate admissions example, the test statistic was 92.36 with 4 degrees of freedom, resulting in a p -value < 0.0001 . Since this is less than $\alpha = 0.05$, we reject the null hypothesis in favor of the alternative and conclude that the success proportions are not likely homogeneous. For the patient satisfaction example, the test statistic is 16.89 and 4 degrees of freedom yields a p -value of 0.0020. Since this is less than $\alpha = 0.05$, we reject the null hypothesis in favor of the alternative and conclude that the two measures are likely related. Recall that R automatically presents the p -value when calling the `chisq.test()` function.

4.5.3 Interpretation of Results

If the chi-square test yields a significant result, then we need to interpret how the null hypothesis is violated. This can be done in several ways, but is best done by observing the patterns in the contingency table. At times this can be a simple process, especially when the number of levels for both variables is small, but it often requires us to ask more of the contingency table than it may at first appear to be worth. What we first need to observe is the “cell chi-square” values, which are the values $(Observed - Expected)^2 / Expected$ for each cell that contribute to the test statistic. When these values are close to zero, it means that the observed cell frequencies are close to the expected cell frequencies and – more importantly for our purposes – the observed values closely match what we would expect to observe if the null hypothesis was true. However, when these values are greater than zero – generally larger

than 2 – it means the observed cell frequencies are not close to the expected cell frequencies, and the observed values differ somewhat from what we would expect if the null hypothesis was true.

Once these *large cell chi-squares* have been identified, we need to determine how the cell frequencies *as a whole* differ from the null. Students are often misled into thinking that the cells that have large chi-square values are those that differ from the null hypothesis. This is a half truth: yes, the observed frequencies in those cells do differ from the expectation under the null hypothesis, but it is the pattern of differences across all cells that is important. So what we do at this point is compare the observed frequencies with the expected frequencies in each cell, and try to derive any patterns of those differences across all cells. While there are no general rules to follow (since there are countless ways these differences can present themselves), we generally look for regions where the observed counts are all greater/less than the expected counts, and columns or rows for which the differences change from positive to negative over the values in the corresponding rows or columns.

For instance, let's observe the cell chi-squares for the graduate admissions example, which are presented in the last two columns of Table 4.7. Here we note that all of the cell chi-squares are large except for those corresponding to the *C* grade level. Comparing the observed and expected cell frequencies, we see that the “Yes” frequencies are larger than expected for the *A* and *B* grade levels and are lower than expected for the *D* and *F* grade levels, while the “No” frequencies are lower than expected for the *A* and *B* grade levels and are higher than expected for the *D* and *F* grade levels. However, this is not the whole story. What we really see when we look *across the grade levels* is that the observed “Yes” cell frequencies are decreasing with respect to the observed cell frequencies as the grade level gets worse; likewise, the observed “No” cell frequencies increase with respect to the expected frequencies as the grade level gets worse. In this light, it is not so much that the low cell chi-squares at the *C* grade level mean that those students get accepted as expected, but rather, they are coincidentally the transition from students that exceed expectations to those who do not meet expectations. If we translate these results into proportions, we state that the proportion of students admitted into the graduate program of their choice decreases as those students' grade level decreases.

We can ask R to calculate the cell chi-square values in a manner similar to that used to calculate the expected cell frequencies. As seen in Program 12, we must first label the chi-square test (here entitled `cellchi1`), and then ask for the residuals from the test (done by the line `cellchi1$residuals`). However, this last step will only provide the standardized differences between the observed and expected frequencies (which leads to positive and negative values), so we square them using the code `cellchi1$residuals*cellchi1$residuals`. From the output given in Program 12, we see that these values match what we had calculated by hand.

For the patient satisfaction example, the cell chi-square values are listed in the last three columns of Table 4.8. According to these values, the high cell chi-squares correspond to subjects with minimal disease severity and

Program 12 Program to obtain cell chi-square values from chi-square test graduate admissions example.

Code:

```
# Create the table
table1 <- matrix(c(52, 48, 41, 11, 4,
                  7, 13, 35, 33, 39),
                nrow=5, ncol=2)

# Name and invoke the chi-square test
cellchi1<-chisq.test(table1)

# Ask for cell chi-square values
cellchi1$residuals*cellchi1$residuals
```

Output:

```
          [,1]      [,2]
[1,] 11.66421114 14.32769242
[2,]  6.14498772  7.54817390
[3,]  0.01907728  0.02343351
[4,]  7.24319901  8.89715784
[5,] 16.37819512 20.11809764
```

high satisfaction, moderate disease severity and medium satisfaction, and severe disease severity and low satisfaction. Comparing the observed and expected cell frequencies for these three cells, we see that there are more patients than expected for minimal severity/high satisfaction, for moderate severity/medium satisfaction, and for severe severity/low satisfaction. The remaining cells – those with low cell chi-squares – would then more closely resemble the conditions under the null hypothesis. Turning these results into a meaningful pattern, we see that in general patients with severe diseases report lower levels of satisfaction than do patients with minimal or moderate diseases, while those patients with minimal or moderate diseases tend to report higher levels of satisfaction than do patients with more severe diseases. These values can also be generated in R by calling the chi-square test (say `cellchi2<-chisq.test(table2)`) and then using the `cellchi2$residuals*cellchi2$residuals` command.

4.6 Contingency Methods (with R Code)

In the event that more than 20% of the cells have expected frequencies less than 5, one should not proceed with the hypothesis test. The chi-square test is dependent upon adequate sample size, since the chi-square distribution is an approximation of the actual distribution of the test statistic, an approximation

which improves as sample size increases. As a remedial action, one might consider combining entire rows or columns, provided that the resulting combinations are meaningful. Only as a last resort should deleting entire rows or columns be considered an option. Note that in some cases there are exact methods (akin to Fisher's Exact test) that could be used regardless of the sample size, but they are beyond the scope of this text.

4.7 Communicating the Results (IMRaD Write-Up)

The following is an example of the IMRaD write-up for the graduate admissions example.

Introduction: Educators are interested in metrics or predictors of performance in graduate academic programs. Of particular interest is the performance in an undergraduate statistics course, which is required for most undergraduate majors and graduate programs. We test the hypothesis that the proportion of students who are accepted into the graduate program of their choice is the same across all grade levels those students achieved in an undergraduate statistics course.

Methods: A total of 283 students participated in this study. We report the frequency of students who stated whether or not they were accepted into the graduate program of their choice (Yes or No) along with the total number of students and the proportion stating "Yes" for each possible grade level (A, B, C, D or F) those students achieved in an undergraduate statistics course. A test of homogeneity on the success proportions for each grade level is conducted using a chi-square test with four degrees of freedom. The null hypothesis is that the success proportions for each grade level are equal, while the alternative hypothesis is that at least two of those proportions differ. We will reject the null hypothesis if the resulting p-value is less than the significance level $\alpha = 0.05$, and we will fail to reject the null hypothesis otherwise. The R statistical software was used for all statistical analyses.

Results: The data are summarized in Table 4.9 below for the 283 students included in this study. Assuming that the data are representative and subjects are independent, the sample size is large enough to conduct statistical analysis since all expected cell frequencies are greater than 5. The test produced the following results ($\chi_4^2 = 92.4$, p -value < 0.0001), so we reject the null hypothesis in favor of the alternative and claim that there is a significant difference between at least two of the success proportions. By comparing the observed with expected cell frequencies, we see that the observed "Yes" cell frequencies are decreasing with respect to the observed cell frequencies as the grade level gets worse, while the observed "No" cell frequencies increase with respect to the expected frequencies as the grade level gets worse.

Discussion: The proportion of students admitted to the graduate program of their choice decreases as those students grade level decreases. Educators and students alike can use undergraduate performance in a statistics course as a part of the decision making process for admitting students into graduate programs.

The following is an example of the IMRaD write-up for the patient satisfaction example.

Table 4.9: Contingency Table of Student Grades in an Undergraduate Statistics Course and Whether or Not the Student was Accepted to the Graduate Program of Choice.

Grade	Accepted to program of choice		Total	Proportion “Yes”
	Yes	No		
A	52	7	59	0.88
B	48	13	61	0.79
C	41	35	76	0.54
D	11	33	44	0.25
F	4	39	43	0.09
Total	156	127	283	

Introduction: Clinicians, health provider administrators and supervisors have a vested interest in ensuring patients are satisfied with the treatment they receive. The level of patient satisfaction is likely tied to the alleviation of symptoms or the “curing” of disease, both of which are highly correlated with the severity of the patient’s disease. We test the hypothesis that the severity of disease and patient satisfaction of the clinical experience with treating that disease are related.

Methods: A total of 87 patients participated in this study. The frequency and proportion of patient level of satisfaction with their treatment (low, medium or high) as well as the total number of patients are reported for each level of disease severity (minimal, moderate or severe). A test of association between patient satisfaction and disease severity is conducted using a chi-square test with four degrees of freedom. The null hypothesis is that there is no relationship between patient satisfaction and disease severity, while the alternative hypothesis is that there is a relationship between the two measures. We will reject the null hypothesis if the resulting p-value is less than the significance level $\alpha = 0.05$, and we will fail to reject the null hypothesis otherwise. The R statistical software was used for all statistical analyses.

Results: The data are summarized in Table 4.10 below for the 87 patients included in this study. Assuming that the data are representative and subjects are independent, the sample size is large enough to conduct statistical analysis as fewer than 20% of all cells have expected frequencies less than 5

($1/9 = 11\%$). The test produced the following results ($\chi^2_4 = 16.9$, p -value = 0.002), so we reject the null hypothesis in favor of the alternative and claim that there is a relationship between patient satisfaction and disease severity. Comparing the observed and expected cell frequencies, we see that there are more patients than expected with minimal disease severity reporting high satisfaction, more patients than expected with moderate disease severity reporting medium satisfaction, and more patients than expected with severe disease severity reporting low satisfaction.

Table 4.10: Contingency Table of Patient Satisfaction and Severity of Condition.

Disease Severity	Patient satisfaction			Total
	Low	Medium	High	
Minimal	7 25 %	3 11 %	18 64 %	28
Moderate	7 28 %	10 40 %	8 32 %	25
Severe	19 56 %	4 12 %	11 32 %	34
Total	33	17	37	87

Discussion: In general, patients with severe diseases report lower levels of satisfaction than do patients with minimal or moderate diseases, while those latter patients tend to report higher levels of satisfaction than do patients with more severe diseases. Clinicians and health service administrators may want to consider these patterns when choosing between treatment options.

4.8 Process

1. State research question in form of testable hypothesis.
2. Determine whether assumptions are met.
 - (a) Representative
 - (b) Independence
 - (c) Sample size: calculate grand proportions and expected frequencies
3. Summarize data with contingency table.
 - (a) Summarize groups with frequencies, sample sizes, and proportions.
4. Perform Test.
 - (a) If sample size is adequate: calculate chi-square test statistic.
 - (b) If sample size is inadequate: do not perform test; collapse rows/columns.

5. Compare test statistic to critical value or calculate p-value.
6. Make decision (reject H_0 or fail to reject H_0).
7. Summarize with IMRaD write-up.

4.9 Exercises

1. In a larger study examining the associations between subject beliefs of body size and smoking, [Boles and Johnson \(2001\)](#) collected measurements (raw counts found in following table) on subject's perception of their own body size (overweight, appropriate or underweight), as well an indicator of whether or not the subject smoked (yes or no). Determine whether the proportions of subjects who smoke are different between the three weight status classifications.

Weight Status	Smoking status	
	Yes	No
Overweight	17	97
Appropriate	25	142
Underweight	96	86

2. [Schottenfeld et al. \(1982\)](#) conducted a study to compare the information provided on a death certificate with the cause of death determined after a formal autopsy. Three classifications were provided, which stated that the initially determined cause of death was accurate, inaccurate or incorrect. These assessments were made at two hospitals. Determine if there is an association between the type death certificate classification and the hospital from which the cause was determined.

Hospital	Certificate status		
	Accurate	Inaccurate	Incorrect
1	157	18	54
2	268	44	34

3. A researcher is interested in the PCB concentrations in popular sport fishing. To study this he has designed a study where he takes fish from three habitat types: Pool, Run and Riffle. After the fish is taken it is recorded on whether or not the fish has an unacceptably high level of PCB. They do this at different sites over different times and obtain the following data where each cell is the number of fish in each habitat/PCB category: Determine if there is an association with the habitat and excessive PCB.

Excessive PCB	Habitat		
	Pool	Run	Riffle
Y	121	57	86
N	713	343	527

4. Suppose a researcher is interested in the lifestyle choices for obese and non-obese people. Specifically which types of recreation activities they prefer which were then classified as one of the following: Sedentary, Outdoor, and Sports. A survey was conducted where participants were asked various questions and one question regarded whether or not they are obese. Another question regarded which type of recreation activities they typically engage in and which one is the most frequent activity. Below is a table of the data they obtained. It shows the self reported obesity category as well as the most frequently engaged in recreation activity.

Obese	Habitat		
	Sedentary	Outdoors	Sport
Y	135	108	92
N	168	262	245

Determine if there is an association with the obesity status and recreation activities.

5. Suppose a researcher is interested in the species distributions of fresh-water fish across several river basins in New South Wales. The species of interest are Gudgeon, Jollytail, Smelt, Bass and Other. The river basins to be considered are the: Richmond, Manning, Hunter and Brogo rivers. Using the data below determine if the data suggests there are differences in species distributions across these river basins:

Species	River basin			
	Richmond	Manning	Hunter	Brogo
Gudgeon	38	54	33	50
Jollytail	91	100	92	94
Smelt	99	95	94	80
Bass	119	131	115	104
Other	125	114	102	115

6. Suppose a researcher is interested in if various antihistamines are prescribed at similar rates across practice types. The antihistamines

considered are: Loratadine, Cetirizine, Desloratadine, Levocetirizine, Fexofenadine and other. The practice types of interest are: Pediatrics, Internal Medicine and Family Medicine. The researchers took a survey of various types of practices and had the practice head nurse give the most prescribed/recommended antihistamine for the practice. Using the data below determine if there are differences in prescription/recommendation rates of antihistamines by practice type.

7. A researcher is interested in the relative efficacy of penicillin and spectinomycin in the treatment of gonorrhea. Three treatments are looked at (1) penicillin (2) spectinomycin (low dose), spectinomycin (high dose). Three possible responses are recorded (1) positive smear, (2) negative smear, positive culture or (3) negative smear, negative culture. Using the data below determine whether or not there is any relationship between the type of treatment and the response.

Antihistamine	Practice type		
	Pediatrics	Internal	Family
Loratadine	131	110	116
Cetirizine	68	44	57
Desloratadine	135	154	134
Levocetirizine	113	97	91
Fexofenadine	121	114	124
Other	35	39	23

Treatment	+Smear	Response	
		+Culture	Smear -Culture
Penicillin	40	30	130
Spectinomycin L	10	20	70
Spectinomycin H	15	40	45

Chapter 5

Summarizing Continuous Data

Chapters 2–4 dealt with categorical data, where the measurements were qualitative in nature. Summarizing these data required frequencies, proportions and contingency tables. In this chapter we begin our study of *continuous data*, where the measurements we will deal with are quantitative in nature. Since this type of data consists of actual numbers, we will be able to arithmetically manipulate them (this isn't as malicious as it sounds). However, what type of manipulation we perform depends – of course – on the type of question we're attempting to answer, as well as the nature of the data itself.

An under-appreciated aspect of continuous data is that – by their very nature – measurements are likely to be unique. While categorical measurements can only take one of a few particular values, continuous measurements can most often take one of countless values, and as such calculating statistics like a proportion is not sensible. Rather, a summarization of continuous data must capture two unique aspects of the measurements: values that represent what a “typical” measurement look like, and a metric of how similar (or dissimilar) the measurements are. The “typical” value – the types of which are called *measures of center* – represents what value a typical or average subject could take. Values that reflect the degree of similarity – the metrics of which are called *measures of variability* – naturally represent how different we can expect any two values to be. Capturing these characteristics is necessary because we can use them to represent the data as a whole, especially in cases with large numbers of subjects.

5.1 Representative Values (with R Code)

As previously mentioned, measures of center capture what a “typical” value looks like for a given sample. For instance, if we were to sample one subject from some population, the measure of center would be our best guess as to the value of that subject’s continuous measurement. These measures are often used to represent the sample from which they were calculated (most notably, they aim to represent the population mean “ μ ”), and are generally involved in the testing of hypotheses or the construction of confidence intervals. There are several types of measures of centers, and which measure we use depends upon characteristics of the particular sample under consideration.

5.1.1 Mean

Generally, when a researcher mentions the mean of sample, they are referring to the arithmetic average of all the values. This is a bit awkward for statisticians, since the term mean has a precise meaning that depends on integral calculus and – depending upon the type of data – might refer to something other than an average. Without getting too specific with regards to the correct definition, we will begrudgingly refer to the mean in the commonly used fashion as the average of all measurements in a data set. This is generally not a problem if our continuous data follow a normal distribution (see below), and this happens often in practice. For a sample of n subjects taking measurements $x_i, i = 1, \dots, n$, the mean is calculated using the following equation:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}. \quad (5.1)$$

Simply put, we sum together the values for each subject and divide the resulting sum by the number of subjects. Provided that there is a center to some set of continuous measurements, the mean often has the best chance of representing it (when compared to other measures of center). This is because the mean includes information from every single piece of data that we have (unlike some other measures), where each measurement contributes equally to the final result.

One drawback of the mean is that it is not robust in the presence of severe outliers. If there are a few values that are much larger (or smaller) than most of the values in a sample, those values tend to “pull” the mean in their direction. This results from the mean placing equal weight ($1/n$, in fact) on each value.

5.1.2 Median

Percentiles are values within the data set (or are between values) that separate a given percentage of the (ranked) data from the rest of the data. For instance, the 10th percentile is that value that separates the smallest 10% of all

observations from the largest 90%. Likewise, the 47th separates the smallest 47% of all observations from the largest 53%. As we will see below with the Empirical Rule, we can use percentiles to tell us some details about symmetrically distributed data sets.

The median is the most popular percentile, and is used as an alternative measure of center to the mean. Passively defined, half of the values in a sample are less than the median, while the other half of the values in a sample are greater than the median. In other words, the median is the “middle-most” value in the sample, designated as the 50th percentile. Calculating the median is a two step process, the first step of which is for us to rank the data from smallest to largest. The second step depends upon how many subjects comprise the sample. If there is an odd number of subjects, then the median is the middle-most value. The “middle-most value” is identified by dividing the sample size (n) by 2 and rounding up to the next largest integer (say k); the median is then the value from the k th subject in the ranked data set. If there is an even number of subjects, then the median is the average of the two middle-most values, where here the “middle-most value” is identified as the average between the values in the $n/2$ and the $(n/2) + 1$ positions. For example, if we have the small sample 3, 7, 1, 9 and 8, we first rank the data points in ascending order: 1, 3, 7, 8 and 9. Since there are five values, we take as the median the middle-most value, which is the third value ($5/2 = 2.5 \rightarrow 3$), or 7. Alternatively, if we have the sample 3, 7, 1, 9, 8 and 2, we rank the data as 1, 2, 3, 7, 8 and 9, and take as the median the average of the third ($6/2 = 3$) and fourth ($6/2 + 1 = 4$) values ($(3 + 7)/2$), or 5.

5.1.3 Other Measures

The mean and median are the most commonly used measures of center in research. On occasion, however, other methods are used. The *mode* is defined as that value (or values) that occurs most frequently in a data set. The mode can be ascribed to both continuous and categorical data sets, though the definition of “most frequently” for continuous data is not entirely clear. For continuous measurements, the *idea* of the mode is much more informative than its actual estimate. The *geometric mean* is defined as the n th root of the product of all values (provided they are all positive and non-zero), or $(\sqrt[n]{x_1 x_2 \dots x_n})$. There are other measures, such as the *harmonic mean*, but they are even more rarely used than either the mode or geometric mean.

Most summary statistics are easily obtainable in R using the `summary` function. The `summary` function produces the mean and median of a sample (as well as other measures described below). Consider measurements from 30 female participants in the Fels Longitudinal Study (FLS). The FLS (Roche 1992) is a longitudinal study originated to study human growth and development, particularly with regards to obesity, metabolic function and cardiovascular health. In this particular sample, serum levels of cholesterol

Table 5.1: Cholesterol Data from 30 Female Fels Longitudinal Study Participants

261	160	259	223	169	127	221	190
224	228	229	294	204	177	199	212
186	207	192	241	162	249	206	210
200	213	185	171	189	159		

were measured, which are given in Table 5.1. The `summary` function is used on this data in Program 13, for which we also show the output for. Here we see the mean of this sample is 204.9, indicating the arithmetic average of these observations. We can also view percentiles, which includes the median, which for these women is given as 205.0, which is nearly identical to the mean. Note that there exist other functions for generating specific summary statistics, such as the `mean` and `median` functions, though these are admittedly less useful.

Program 13 Program to generate summary statistics for center of FLS data.

Input:

```
# Enter the data into a variable named FLS
FLS <- c(261, 160, 259, 223, 169, 127, 221, 190,
        224, 228, 229, 294, 204, 177, 199, 212,
        186, 207, 192, 241, 162, 249, 206, 210,
        200, 213, 185, 171, 189, 159)
```

```
# Run the summary function on FLS
summary(FLS)
```

Output:

```
> summary(FLS)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 127.0  185.2   205.0   204.9  223.8   294.0
```

5.2 Measures of Variability (with R Code)

Measures of variability aim to measure to what degree measurements in a given sample differ from one another. Measures that are small indicate the data are fairly homogenous or similar in value (though not exactly the same), while measures that are large indicate the data are fairly heterogeneous or dissimilar in value. The magnitude of the measure of variability in and of

itself is not a good or bad thing, meaning that a sample is not necessarily good if it has a small degree of variability, and a sample is not necessarily bad if it has a large degree of variability. Provided that the variability in a sample is not *caused by* the method in which the data were collected (errors in measurement, cheap or inefficient measurement devices, etc.), then the variability *is what it is*, and we accept it as a characteristic of the data. Unlike the case for measures of center, the most commonly used measures of variability do not exactly attempt to capture the same things, which can lead to confusion in practice.

5.2.1 Standard Deviation

The standard deviation is a purely algebraic characteristic of a sample (i.e. the equation precedes its lay definition) that is best described as the difference in value one would expect to observe between any two randomly selected subjects. It is not the average difference in the sense that the mean of a sample is the arithmetic average, but rather a typical or expected difference. Mathematically, the standard deviation is defined as the square root of the variance, and is expressed in the following equation

$$SD = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (5.2)$$

where n is the sample size, x_i is the i th value in the sample, and \bar{x} is the sample mean. Procedurally, we first calculate the mean, then subtract the mean from each observation, square those differences, sum the differences, and then divide by $n - 1$. Note that $n - 1$ is used so that the sample standard deviation is an unbiased estimator of the population standard deviation (as opposed to using n , which is why this is an expected value and not an average). The standard deviation is often written as the lower-case letter “ s ”, to differentiate it from the sample variance “ s^2 ”. The population variance is denoted using Greek lettering as “ σ^2 ”, while the population standard deviation is denoted as “ σ ” (note the convention among statisticians to use Greek letters for population values and Roman letters for the corresponding sample value).

The value of the standard deviation has little absolute meaning, or at least it is not often interpreted *in a vacuum*. The SD of a sample is often reported without comment (you will not often see the SD described as large or small). The main reason for this is that the magnitude of the SD must be interpreted *with respect to something*, for instance the mean of the sample, or the SD from another sample of similar measurements. Thus, the standard deviation is a *relative measurement*, and while we will report it, its utility will predominantly come through hypothesis testing.

5.2.2 Range Measures

Whereas the variance and standard deviation attempt to describe the variability between the data, range measures attempt to “bound” the data. The *range* of a measure is indicated by the interval between the smallest value (known as the minimum) and the largest value (known as the maximum). The range is found by simply identifying the smallest and largest values in a sample, and once the range is identified, this is the interval within which all values in a sample lie. This is a gross summary of the sample, and does not provide much information on its variability. The interquartile range (IQR) is another interval bounded by the 25th and 75th percentiles (also known as the first and third quartiles; the median is the second). The IQR requires a process for identification similar to that for median, where the data must be ranked before the 25th and 75th percentiles can be identified. Since 25% of the data are less in value than the 25th percentile, and since 25% of the data are greater in value than the 75th percentile, the IQR represents the interval between which 50% of the sample values reside. Unlike the range, the width of the IQR – with respect to the median – provides some indication of the variability of the sample values, where clearly wider intervals indicate greater variability, and smaller intervals indicate less variability.

In R, the `sd` function can be used to obtain the standard deviation, while the `summary` function (introduced earlier) can be used to obtain the range or interquartile range, as it produces percentiles such as the minimum, maximum, Q1, and Q3. As an example we return to the female cholesterol values from the FLS study. Using the `sd` function, the standard deviation is reported as 35.31958. What this value means is that for any two randomly selected subjects from this sample, we would expect their cholesterol values to differ by nearly 35 units (keep in mind that we would not state this in a write-up), though whether this typical difference is large or small cannot be determined without additional information. By looking at the output from the `summary` function we can obtain the range and IQR. For the range we take the `Min.` and `Max.`, which we report as (127.0, 294.0), meaning that all cholesterol values in this sample are between 127 and 294. For the IQR we take the values associated with `1st Qu.` and `3rd Qu.`, so that the IQR is (185.2, 223.8), which means that 50% of the female subjects have cholesterol values between 185 and 224 (note that in a write-up we just report the range or IQR, and do not explain them). This output is shown in Program 14.

5.2.3 Empirical Rule

One interesting application of the standard deviation is through the empirical rule, which gives us a broad idea of where particular data values are situated with respect to the mean. This rule states that for data with a nearly symmetrical distribution (see below): ~ 0.67 (or $2/3$) of all measurements fall within one standard deviation of the mean (or within the interval

Program 14 Program to generate summary statistics for variability of FLS data.

Input:

```
# Enter the data into a variable named cpk1
FLS <- c(261, 160, 259, 223, 169, 127, 221, 190,
        224, 228, 229, 294, 204, 177, 199, 212,
        186, 207, 192, 241, 162, 249, 206, 210,
        200, 213, 185, 171, 189, 159)

# Run the sd and summary functions on FLS
sd(FLS)
summary(FLS)
```

Output:

```
> sd(FLS)
35.31958

> summary(FLS)
  Min.  1st Qu.  Median    Mean  3rd Qu.   Max.
127.0  185.2    205.0   204.9   223.8   294.0
```

$(\bar{y} - s, \bar{y} + s)$); $\sim 95\%$ of all measurements fall within two standard deviations of the mean $(\bar{y} - 2s, \bar{y} + 2s)$; and $\sim 99\%$ of all measurements fall within three standard deviations of the mean $(\bar{y} - 3s, \bar{y} + 3s)$. The empirical rule is a special case of Chebychev's theorem, which states that for distributions of any shape, the proportion of subjects within k standard deviations of the mean is at most $1 - 1/k^2$. Returning to the cholesterol example, this would mean that $\sim 67\%$ of the cholesterol levels for female FLS participants would be between $204.9 - 35.31958 = 169.5804$ and $204.9 + 35.31958 = 240.2196$, $\sim 95\%$ of the levels would be between $204.9 - 2 \times 35.31958 = 134.2608$ and $204.9 + 2 \times 35.31958 = 275.5392$, and $\sim 99\%$ of the levels would be between $204.9 - 3 \times 35.31958 = 98.94127$ and $204.9 + 3 \times 35.31958 = 310.8587$.

5.3 Assessing Normality (with R Code)

If you recall from Chapters 2 and 3, we made the assumption that our test statistic z had a normal distribution. This assumption was checked by observing the expected cell frequencies in the contingency table, at least 80% of which had to be greater than 5. This was an ad hoc assessment, since we had only one sample, and thus only one test statistic, and we could not really determine whether or not the test statistic was normally distributed. Continuous data are another matter, where we now have the ability and tools

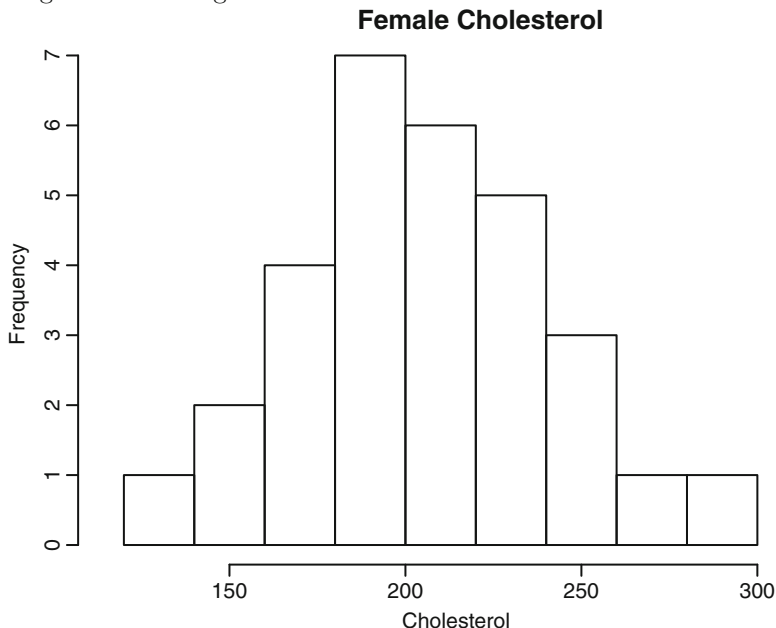
to determine whether *our sample* is normally distributed. To be certain, we can never state with certainty that a given sample is normally distributed, but we can provide graphical evidence that will allow us to safely assume one way or the other.

5.3.1 Histogram

One of the most powerful tools we have to visualize the distribution of a sample of continuous measurements is the histogram, which is a bar chart of frequencies, which in turn represent the number of subjects in the sample that fall within a series of intervals. A histogram for the female cholesterol values in the FLS database is provided in Figure 5.1. Here we see that the bars correspond to a series of nine intervals, each 20 units wide. The height of each bar corresponds to the frequency of females within the sample that take values within that interval. From Figure 5.1, we see that more women have cholesterol levels between 180 and 200 than any other interval, and this corroborates with our measures of center, which we know to be 204.9 according to the mean and 205 according to the median. This interval is also known as the mode for this histogram, since it contains more subjects than any other interval. In contrast, relatively fewer women have low cholesterol between 120 and 140 or high cholesterol between 260 and 300.

Program 15 shows the R code to generate Figure 5.1 for Female Cholesterol from FLS Database. To create the histogram we use the `hist` function

Figure 5.1: Histogram of Female Cholesterol from FLS Database.



Program 15 Program to generate Figure 5.1 for Female Cholesterol from FLS Database.

Code:

```
### Read in the dataset
FLS <- c(261, 160, 259, 223, 169, 127, 221, 190,
        224, 228, 229, 294, 204, 177, 199, 212,
        186, 207, 192, 241, 162, 249, 206, 210,
        200, 213, 185, 171, 189, 159)

### Create a histogram
hist(FLS,
     xlab="Cholesterol",
     main="Female Cholesterol"
    )
```

Output:

Figure 5.1 is the output.

on our data set FLS. We use the `xlab="Cholesterol"` statement to label the horizontal axes and the `main="Female Cholesterol"` statement to title the plot. There are many other options that can be used to modify the histogram, however we will not consider those here.

What we are really looking for in a histogram is its general shape. Normally or nearly normally distributed data have a symmetric distribution, where the shape to one side of the mode is similar to the shape on the other side of the mode. While the two sides do not need to be identical, they should both slightly decrease at about the same pace as we move away from the mode. The histogram in Figure 5.1 shows us a more or less symmetric histogram, even though there are differences between the two sides (or the tails, as they are commonly called). In all seriousness: try squinting your eyes when viewing a histogram so that small and meaningless differences do not capture your attention. If we superimpose a normal curve over the histogram (see Figure 5.2), we can see that they match up fairly well. This indicates that the female cholesterol levels are nearly normally distributed (at the very least, whatever distribution those values do have is not too distinguishable from normality).

Figure 5.3 presents a histogram for the triglyceride levels in male FLS subjects. The distribution we see here is in stark contrast to that observed in Figure 5.1, where now the mode decidedly occurs more toward one side of the distribution (between 50 and 100) than the other. This is an example of a skewed distribution, and in this case arises from the fact that triglycerides – like many biomarkers – cannot take values less than zero (they are said to be *truncated at zero*). Because of this truncation point, the data pile up

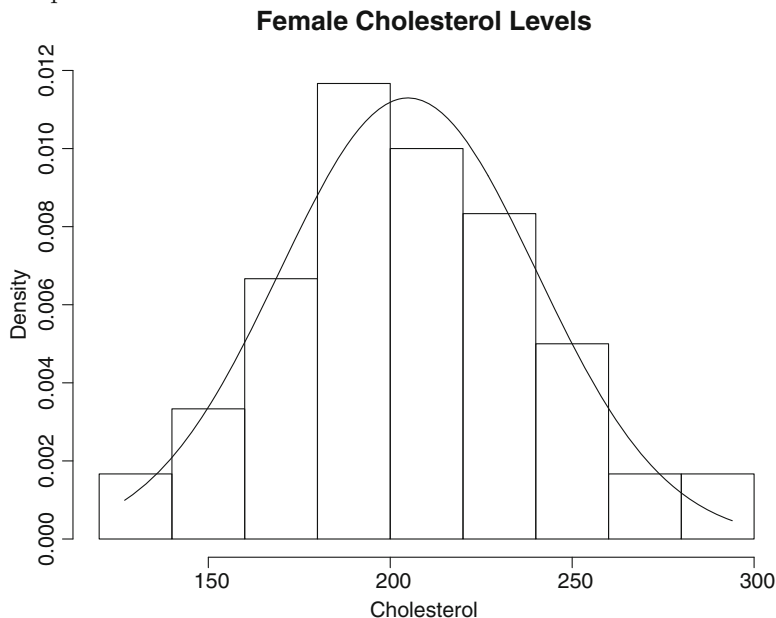
close to this point, and the values can only extend openly in one direction. This concept is called skewness, and since they extend to larger values here (toward the 600), this is called *right-tail skewness*. If the “tail” of the data were in the direction of smaller values (toward the 0), then this would be called *left-tail skewness*. Skewed data like these offer an example of data that are *not normally distributed* (recall that the most obvious characteristic of normally distributed data is symmetry about the mode). The superimposed normal curve in Figure 5.3 also shows how ill-fit the data are to the normal distribution.

Program 16 shows the R code to create Figure 5.3 for the Male Triglycerides from FLS Database. The first line reads the "Chp. 5 Male Trig.csv" file using the `read.csv` and writes it into the `tri1.m` dataset. We again use the `hist` function to create the base histogram similar to that in Program 15. However we wish to plot a distribution over the histogram so we will need to use the `freq=FALSE` statement to change the vertical axis scale to percentage instead of frequency. Once we have the histogram we will want to put a normal distribution over it. To do this we need to collect some summary statistics about the sample. We find the mean of our variable `tri1.m` using the `mean` function. The data has some missing values and hence we need to tell R what to do with these values. In this case we use the `na.rm=TRUE` statement that will remove the missing values from the dataset. We also need to find the standard deviation using the `sd` function, the minimum value using the `min` and the maximum value using the `max` function. We use the `na.rm=TRUE` option on all of these functions to remove the missing values. Once we have the summary statistics we need to create a set of values on which to evaluate the normal distribution. The `seq` function helps us create a sequence of values from the minimum (`tri1.min`) to the maximum value (`tri1.max`) with increments of 0.1 (using the `by=0.1` statement). We can then evaluate the normal distribution at these points using the `dnorm` function which requires the values, `x1`, the mean, `tri1.mean` and the standard deviation `tri1.sd1`. Now that we have the distribution evaluated at a set of points we can overlay the line on the histogram using the `lines` function. The `lines` function requires the evaluation points `x1` and the distribution points `tri1.density`. The same process can be used to overlay the normal curve to the female cholesterol data in Figure 5.2.

5.3.2 Box Plot

Another graphical display for assessing normality is the box plot. This plot is useful in that from it we can assess the general shape of the distribution, and it also identifies the position of several descriptive statistics. The IQR is indicated by the top and bottom of the box, with the top indicating the 75th percentile and the bottom indicating the 25th percentile. The horizontal line inside the box indicates the median. The dashed lines (called “whiskers”) above and below the box extend to 1.5 times above and below the IQR,

Figure 5.2: Histogram of Female Cholesterol from FLS Database with Superimposed Normal Curve.

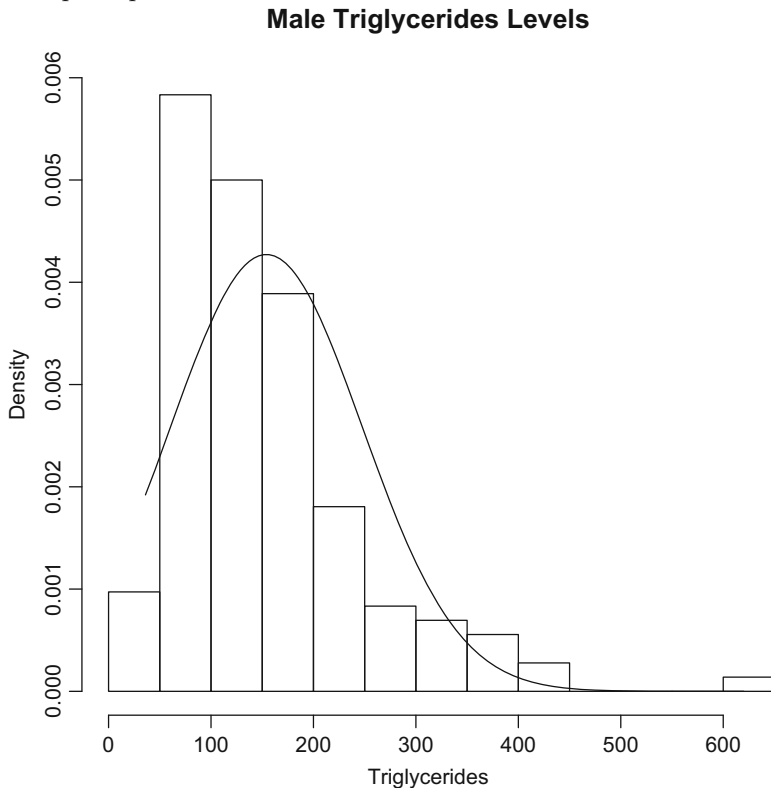


respectively. Any remaining dots outside the extent of the dashed lines are deemed outliers.

In the right-hand panel of Figure 5.4, we see the box plot corresponding to the female cholesterol values in the FLS database (note this is actually a larger sample of female FLS subjects than was used earlier). Note the median line is close to the center of the box, which is what we would expect in normally distributed data (recall that for symmetrically distributed data the mean and median are equal). Further, the distance of the 25th and 75th percentiles from the median are about equal, as are the distances of the dashed lines above and below the box. These are clues telling us that this data are nearly symmetric. Note that there are a few outliers in the high values of this plot, and while there are none in the lower values of the plot, there are not enough of them – nor are they extreme enough – for us to be alarmed.

In the left-hand panel of Figure 5.4, we see the box plot corresponding to the male triglyceride values in the FLS database. Here we note the distance from the median to the 75th percentile is larger than the distance from the median to the 25th percentile, and the “whisker” above the box are further away from the median than the “whisker” below the box. Further, there is at

Figure 5.3: Histogram of Male Triglycerides from FLS Database with Normal Curve Superimposed.



least one extreme outlier taking a value near 600. All of these characteristics indicate that this sample is most likely not normally distributed.

5.3.3 QQ Plot

The “quantile-quantile plot” – or “QQ plot” for short – is a fairly accurate tool in assessing normality. To construct this plot, we rank the data in our sample from smallest to largest, and match those values with the ordered percentiles taken from the standard normal distribution (if there are 100 subjects in our sample, then we take 100 percentiles; if $n = 257$, then we take 257 percentiles, etc.). If the data are *perfectly normally distributed*, then the plot of these matched values will form a 45° line. This *perfect case* is indicated by the solid red lines in Figures 5.5 and 5.6. The closer the sample is to normality, the closer the matched values – indicated by the circles – will adhere to the red line. Since we would expect any sample to differ somewhat from normality, the plot also provides the dashed red lines, which indicate the

Program 16 Program to generate Figure 5.3 for Male Triglycerides from FLS Database.

Code:

```
### Read in the dataset
tri1.m <- read.csv("Chp. 5 Male Trig.csv")

### Create histogram with normal density
hist(tri1.m$BCtrigly,
      freq=FALSE,
      xlab="Triglycerides",
      main="Male Triglycerides Levels")

### Find summary statistics to generate normal density
tri1.mean1 <- mean(tri1.m$BCtrigly, na.rm=TRUE)
tri1.sd1 <- sd(tri1.m$BCtrigly, na.rm=TRUE)
tri1.min <- min(tri1.m$BCtrigly, na.rm=TRUE)
tri1.max <- max(tri1.m$BCtrigly, na.rm=TRUE)

### Create a set of points to generate the density
x1 <- seq(tri1.min, tri1.max,
          by=0.1
          )
tri1.density1 <- dnorm(x1, tri1.mean1, tri1.sd1)

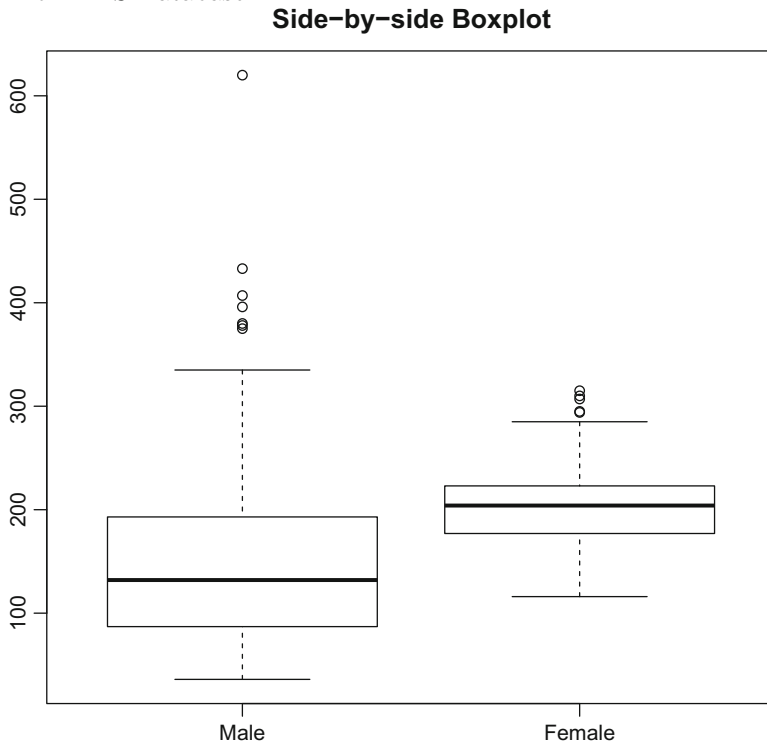
### Plot the density
lines(x1,tri1.density1)
```

Output:

Figure 5.3 is the output.

region of similarity within which a sample is more or less normally distributed. Thus, if all of the matched pairs (circles) fall within these bounds, as mostly seen in Figure 5.5 for the female cholesterol values, then we can assume that the data are normally distributed. However, if the circles – or sizeable portion of them – fall on or outside the boundaries, then it may be unsafe for us to assume normality. This is the case in Figure 5.6, which shows the QQ plot for male triglyceride values, where we see that in the lower tail of the plot (small triglyceride values), the black dots fall above the permissible range (indicating that there are more sample data in this region than we would expect under normality), and in the center of the plot the black dots fall slightly below the permissible range (indicating that there are fewer sample data in this region than we would expect under normality). These violations indicate that the triglyceride values are not normally distributed. A note of

Figure 5.4: Side-by-side boxplots for Male Triglycerides and Female Cholesterol from FLS Database.



caution: analyzing a QQ plot is not an exact science, and the assessment of normality is in some instances in the eye of the beholder. My advice is this: if you're uncertain whether a data set is normally distributed, it probably isn't.

Program 18 shows the R code to generate Figure 5.5 the QQ plot for Female Cholesterol from FLS Database. To get a nice QQ-plot we will want to use a contributed package named `car`. The first time we need to install the `car` package using the following code: `install.packages("car")`, which downloads the `car` package and installs it in R. To use the package we need to add a `library` statement to our R code. The line `library(car)` loads the functions associated with the `car` package, which allows us to use the `qqPlot()` function. As done before, we import the data using the `read.csv` function similar to Program 17. We first specify the data (`Chol1.f$BCholes`) in the `qqPlot` function, use the `ylab="Cholesterol"` statement to label the vertical axes, and use the `main="QQ-Plot of Female Cholesterol"` to title the plot.

Program 17 Program to generate Figure 5.4 for Male Triglycerides and Female Cholesterol from FLS Database

Code:

```
### Read in the dataset
tri1.m <- read.csv("Chp. 5 Male Trig.csv")
Chol1.f <- read.csv("Chp. 5 Female Chol.csv")

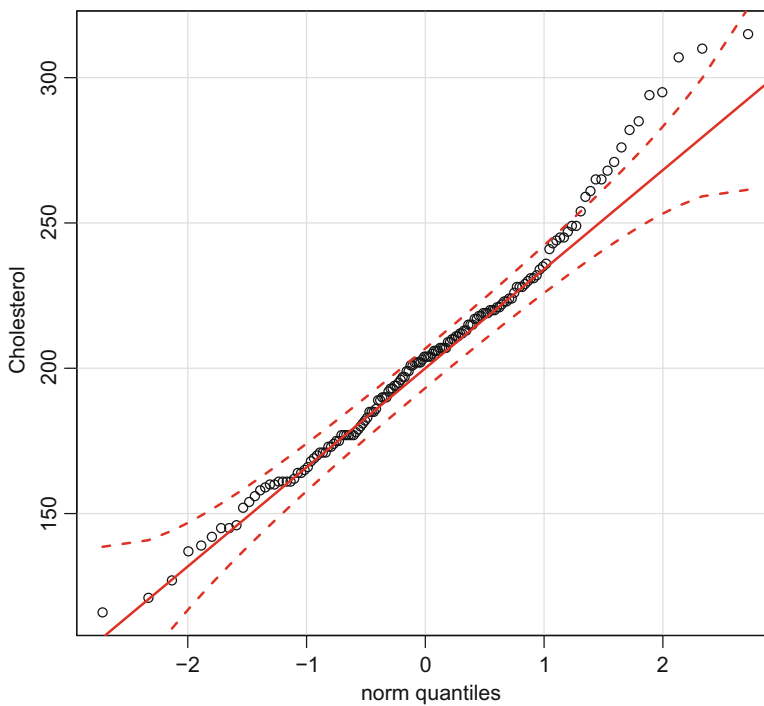
### Create a side-by-side boxplot
boxplot( tri1.m$BCtrigly, Chol1.f$BCcholes,
         names=c("Male", "Female"),
         main="Side-by-side Boxplot"
       )
```

Output:

Figure 5.4 is the output.

Figure 5.5: QQ Plot for Female Cholesterol from FLS Database.

QQ-Plot of Female Cholesterol



Program 18 Program to generate Figure 5.5 the QQ Plot for Female Cholesterol from FLS Database.

Code:

```
### Load the car package
library(car)
Chol1.f <- read.csv("Chp. 5 Female Chol.csv")

### Create the QQ Plot
qqPlot(Chol1.f$BCcholes,
       ylab="Cholesterol",
       main="QQ-Plot of Female Cholesterol")
```

Output:

Figure 5.5 is the output.

5.3.4 Outliers

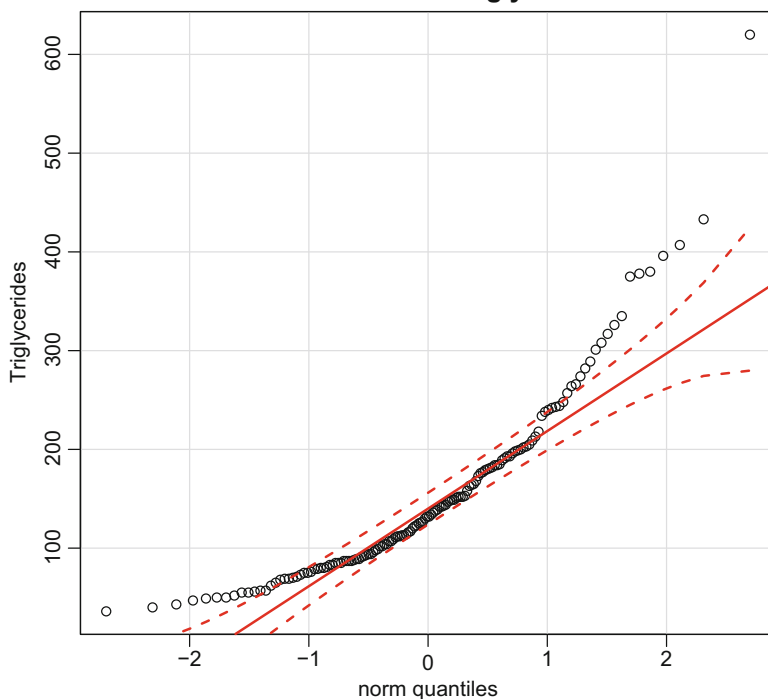
In Figure 5.4 we noted that there were outliers in the male triglyceride sample. Outliers are simply values that are somewhat removed from the center of the data, and they are not bad in and of themselves (usually). We mostly use them to inform us of the shape of our sample. In general, we would expect the same amount of outliers above and below the center of our distribution (in this case, the outliers would *cancel each other out* and have no effect on the measure of center), though if they are few in number we can relax this assumption. At times, one outlier or a small number of outliers will extend greatly from the center of the distribution, so much that they can interfere with our interpretation of the QQ plot. We do not remove these values, but rather investigate to determine if there are any causes as to why they exist (or at least why they are so extreme). If it turns out that an outlier was erroneously entered into the database, we may correct or delete it, but under no other circumstances can we delete or remove outliers in order to improve the distribution of our data. There are remedial actions you can take to “correct” for outliers, but we will not cover them here. Outliers have the greatest affect on the mean and standard deviation, since outliers with large enough values will “pull” these measures in their direction. Provided there are “few” outliers, the median and inter-quartile range will be unaffected by outliers, regardless of their magnitude. To see this characteristic for yourself (in a silly example), calculate and compare the means and medians in one data set (1, 3, 5) and another (1, 3, 1,000,000).

5.4 Rounding and Reporting Conventions

5.4.1 Rounding

The mean is rounded to no more than one decimal place beyond the data from which it is calculated. So if the data are measured in integers (e.g. the cholesterol for patient number 1 is 265), then the mean of such data can be expressed to the *first decimal place* (e.g. if the mean for female cholesterol is given as 203.76974, we round to 203.8). The standard deviation is rounded to no more than two decimal places beyond the level expressed in the data, though it is sometimes reported to the same degree of accuracy as the mean. So based on the female cholesterol levels, the standard deviation is most appropriately rounded to the second decimal place (meaning 37.985963 is rounded to 37.99). Note that unlike categorical data, the extent to which we round our data summaries is not directly dependent upon the sample size.

Figure 5.6: QQ Plot for Male Triglycerides from FLS Database.
QQ-Plot of Male Triglycerides



The median is generally rounded to the same decimal place expressed in the data, though the median can be expressed to additional decimal places if it is calculated as the average of two or more values. The same guidelines are followed for the range and inter-quartile ranges, which generally reflect

observed values and not estimates. So based on the female cholesterol levels, the median is 204, the range is (116, 315), and the inter-quartile range is (117, 223).

5.4.2 Reporting Based on Distribution

In cases of either normally or non-normally distributed data, we always report the sample size. When data are normally or (at least) symmetrically distributed, we report the mean and standard deviation to describe the center and variability in our data. The reason for this is that under the optimal case of normally distributed data, the mean is the most accurate measure of center while the standard deviation is the most accurate measure of variability. More technically, the mean and standard deviation are components of the test statistics used for hypothesis testing; since they are going to be used, they might as well be reported. However, when data are non-normal the mean does not accurately measure the center of the sample; likewise, the standard deviation will not accurately measure the variation in the sample. In cases of non-normally distributed data, we report the median as the measure of center and the interquartile range as the measure of variability. These measures are more robust to departures from normality than are the mean and standard deviation, and will suffice for all but the most severely skewed samples. Further, since the data in this case are skewed, these percentiles provide more information as to how the data are skewed. Note that we would prefer to report the mode, which indicates the most likely value in the sample directly under the “hump” of the distribution, but there are currently no suitable estimators of the mode.

In extreme cases, we may observe that the sample median and either (or both) of the components of the IQR are the same value, meaning that more than half of the data take the same value. We would still report the median and IQR, but we also state in words how much of the data take the same value, and we would clearly not even consider calculating the mean or standard deviation. In such cases we may want to consider alternative approaches – and consult a statistician – before continuing.

5.4.3 Standard Error

In many research papers, normally distributed data are summarized with the mean and *standard error* (or SE). This is a mistake, whether the authors of such studies realize it or not. The standard error is formally defined as the standard deviation divided by the square root of the sample size (s/\sqrt{n}), and represents the variability we would expect to observe in the mean if we could repeatedly sample from the same population, taking the mean each time. Since it is a property of the mean, it does not describe any characteristic of the *observed sample* (remember, we will only collect one), and since the SE is always smaller than SD, it will always *underestimate* the variability in a sample (since n is always greater than 1). For example, in the female

cholesterol measurements in the FLS database, the standard deviation is reported as 37.99, while the standard error is reported as 3.08. If we report the SE, one could mistakenly presume that subjects typically differ by around 3 units, which is much less than the 38 indicated by the standard deviation, and readers may think the variability is much less than it actually is. The SE actually indicates in this case our belief that if we were to repeatedly resample the cholesterol levels for groups of 152 women, the means of those groups would typically differ by about 3 units. This is not a property of our sample, but is actually a property of our sampling method. Further, the standard error will decrease as our sample size increases. This means that we can make the SE arbitrarily small by observing more subjects, whereas the standard deviation is relatively more stable and unrelated to the sample size. The bottom line is that if you are describing a sample, you report measures that capture characteristics of the sample, such as the mean and standard deviation.

5.5 Exercises

1. [Lansford et al. \(2010\)](#) is interested in the number of sexual partners for adolescents between age 16 and 22. They collected a sample of 526 people in this age group and asked them the number of partners they have engaged in which they have engaged in sexual activities. The following sample is consistent with their data: Completely summarize the above data using both numeric and visual summary tools.

2	7	3	1	2	0
0	1	3	2	0	1
0	4	3	0	6	7
1	2	1	8	1	3
0	2	6	1	6	0

2. [Rossi et al. \(2009\)](#) is interested in the waiting times for knee replacement surgery. They took a survey of 161 patients and calculated the days from the initial visit to surgery. The following sample is consistent with their data. Completely summarize the above data using both numeric and visual summary tools.

53	67	69	68	53
64	72	63	73	69
64	77	65	54	52
63	66	61	70	57
66	72	75	72	77

3. [Winer-Muriam et al. \(2002\)](#) is interested in determining the theoretical radiation/energy absorbed by a pregnant woman during different gestational periods. Information was gathered from eight patients during

the first trimester. The data collected is the theoretical dose/energy absorption to a set of points in each patient. Completely summarize the below data using both numeric and visual summary tools.

9.5	3.3	5.7	4.1
13.6	4.6	20.2	20

4. The Glasgow Coma Scale is (GCS) is used to measure the severity of a brain injury. The scale ranges from 3 to 15, with 3 indicating completely unconscious state (no response to any stimuli) to 15 indicating completely conscious state (normal response to all stimuli). There is some evidence that a patients initial GCS score may be correlated with recovery prognosis. Brain injury researchers generally report mean GCS scores of their patients. [Kreutzer et al. \(2009\)](#) give the data below for GCS scores for patients in their study. Completely describe the data using both numeric and appropriate visual summary tools.

3	15	7	3	8	3	3
3	3	6	9	3	13	12
4	3	15	3	15	3	15
3	14	8	6	15	3	6
7	4	5	7	7	3	3
14	15	15	3	3	3	3
15	4	13	3	10	15	3
15	8	14	15	13	12	2
7	12	10	3	12	15	15
5	9	3	6	4	15	12
5	9	3	15	3	10	10

5. We can consider example from [Green et al. \(2005\)](#) who is interested in estimating the amount of *diethylhexyl phthalate* (DEHP) that leach from IV tubing and bags into intravenous medications. Suppose they take 25 standard IV bags and standard tubing of length 1 m and put distilled water in the bag and let it sit for 8 h and then drain the bag through the tube into a container. From each of the containers they measure the DEHP in ng/mL and suppose they obtain the following data:

Completely describe the data using both numerical and visual tools.

53.0	40.4	39.1	39.6	52.9
32.8	51.7	42.9	55.0	43.8
51.1	44.2	38.3	44.3	47.7
43.7	44.2	40.0	60.1	42.9
27.0	50.8	37.0	47.5	69.6

6. [Yoshinaga et al. \(2004\)](#) are interested in the amount of radiation people working in a x-ray lab receive. In general, the typical person in the USA receives on average 3.6 mSv (milli Sievert) of radiation per year. Specifically they wish to know if x-ray labs technicians receive more than 0.01 mSv per day. They take a sample of 15 workers and places a device on each technician that records the amount of radiation they receive. Suppose this is the data they collected. Completely describe the data using both numerical and visual tools.

0.0023	0.0072	0.0054	0.0092	0.0114
0.0013	0.0017	0.0047	0.0069	0.0078
0.0082	0.0087	0.0044	0.0056	0.0087

7. As part of a study on an implantable medication system for insulin delivery, [Saudek et al. \(1989\)](#) measured the percentage above ideal body weight in eighteen patients (found below). Completely describe the data using both numerical and visual tools.

107	119	99	114	120	104	88	114	124
116	101	121	152	100	125	114	95	117

Chapter 6

One-Sample Means

In Chapter 5 we were introduced to continuous data, and in this Chapter we take our first steps into the realm of inference by focusing on hypothesis testing and estimating confidence intervals for one-sample continuous data. In many ways, the material we introduce here will resemble the material covered in Chapter 2, where we focused on the case of a one-sample proportion. While the specific details will clearly be different, most of the steps we take will be the same in both cases. For instance, in both cases we will summarize our data, generate hypotheses, evaluate the veracity of our assumptions, perform the statistical test, and make inference from the output of that test. Most importantly, since we are only dealing with one sample, we must again pay close attention to the hypothesized value upon which we base our test. Before we get to such a test, we must examine the behavior of the sample mean, much as we did for the sample proportion in Chapter 2.

6.1 Behavior of the Sample Mean

The sample mean, as described in Chapter 5, is the arithmetic sum of the values in a particular sample. For normally distributed sample data, the sample mean is the most consistent measure of the center of that data, due mainly to the fact that of all measures, only the mean captures data from every subject and thus includes the most information. For this reason – and others – we will use the mean as the basis for hypothesis testing, much as we centered the hypothesis testing process for categorical data around the sample proportion. What we then need is a way to determine how the sample mean behaves in practice, particularly with respect to its distribution.

The easiest way for us to accomplish this is to conduct a simulation study, in many ways similar to what we did for the sample proportion in Chapter 2. Rather than simulate values from a known distribution, here we will sample values directly from the larger sample of participants from the

Fels Longitudinal Study (FLS) used in Chapter 5. For instance, we had 152 non-missing female subjects who provided cholesterol values, whose values *in aggregate* were taken to be normally distributed. If we treat this sample as the population, we can take sub-samples of various sizes from this larger group, and by taking the means of these samples and repeating the process many times, we can obtain a large sample of sample means. This is the so-called *sampling distribution of the mean*, and we can analyze its characteristics just as we would any other sample by calculating means and standard deviations and also constructing histograms.

As a formal process, we will generate 1,000 random samples from the parent sample, with each sample of a specific size (n). For each of those samples, we will calculate the sample mean and standard deviation, and we will then take the average of those means and standard deviations over the 1,000 samples. We would like to see that the *mean of the means* is close to the “population” mean from the parent sample ($\mu = 203.8$) and also that the *mean of the standard deviations* is close to the population standard deviation from the parent sample ($\sigma = 37.86$). Note that the equation for a population standard deviation is different from that for the sample standard deviation, and thus the value used here is different from that presented in Chapter 5 ($s = 37.99$). In addition, we will find the standard deviation of the sample means, and also look at histograms for certain sample sizes.

The results from this simulation study are found in Table 6.1. Here we see for small samples ($n = 9$), the average mean and average standard deviations are reasonably close to the “population” values. This means that even for small samples of normally distributed data, the sample mean does a fairly good job of describing the center of the sample and the sample standard deviation does a good job of describing the variation of the sample. As the sample size increases from 9 to 100, we see that the means of both the sampling distribution of means and the sampling distribution of standard deviations continue to stay close to the respective population values.

However, while the mean of the means gives a good indication of the accuracy of the sample mean, the mean of standard deviations does not tell us anything about the accuracy of those means. In fact, the mean of the standard deviations tells us only that regardless of the sample size, the variability observed in the data is nearly the same, which we would expect. To analyze the variability of the sample mean, we must take the standard deviation of the 1,000 sample means, the so-called *standard error of the mean*. We see that this value (12.33) is lower than the average standard deviation (36.46) for small samples ($n = 9$), and gets smaller as the sample size increases. There is good reason for this, as the standard deviation measures the variability among the sample values, whereas the standard error measures the variability in the sample means. While the particular values in each of these samples vary in a manner reflected in the variability of the parent sample, the sample means vary to a lesser extent since they are capturing the center of the data set by averaging all values in those particular samples.

Table 6.1: Results from simulation studies for FLS Data. 1. Results based on 1,000 samples of size (n) from (A) the Female Cholesterol sample with mean $\mu = 203.8$ and standard deviation $\sigma = 37.86$., and (B) the Male Triglyceride sample with mean $\mu = 154.08$ and standard deviation $\sigma = 93.07$.

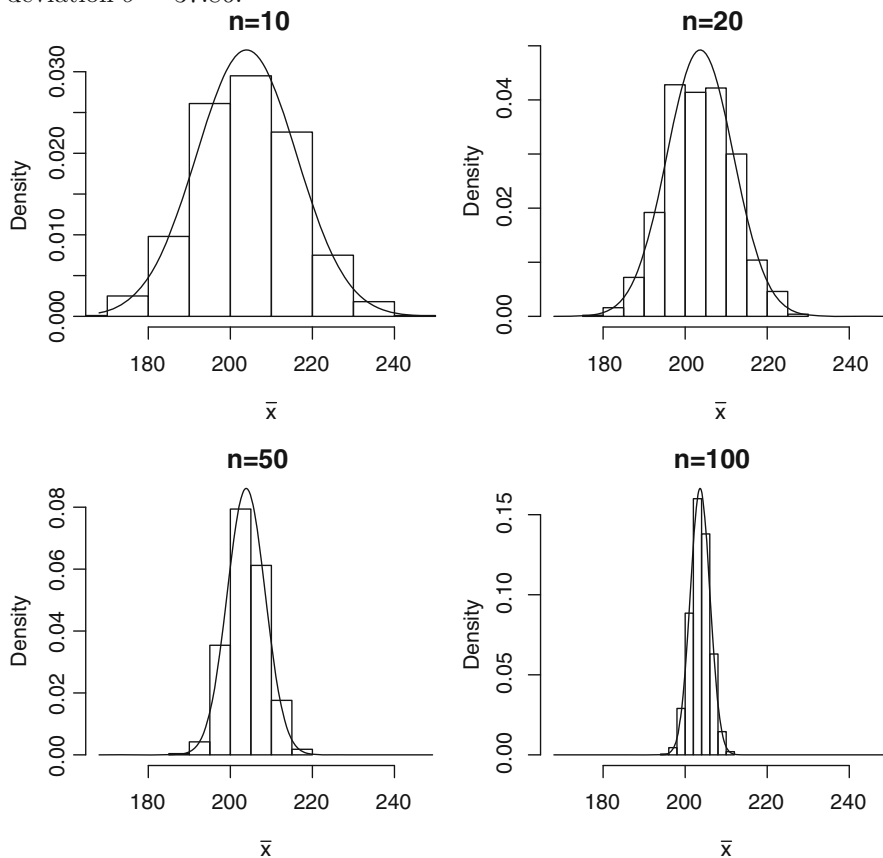
n	A. Female Cholesterol				B. Male Triglycerides			
	\bar{x}	SD	$SE_{\bar{x}}$	σ/\sqrt{n}	\bar{x}	SD	$SE_{\bar{x}}$	σ/\sqrt{n}
9	204.38	36.46	12.33	12.62	154.07	86.51	31.43	31.02
16	204.31	37.09	9.50	9.47	153.93	88.86	23.07	23.27
25	203.87	37.20	7.82	7.57	153.58	90.05	18.55	18.61
36	204.07	37.46	6.54	6.31	153.29	90.47	15.31	15.51
49	204.11	37.59	5.52	5.41	153.49	91.14	13.19	13.30
64	204.15	37.67	4.86	4.73	153.59	91.87	11.51	11.63
81	204.14	37.65	4.20	4.21	153.52	92.00	10.50	10.34
100	204.11	37.68	3.88	3.79	153.67	92.29	9.14	9.31
1,000	204.20	37.85	1.20	1.20	153.66	92.81	2.94	2.94

Further, as the number of values (i.e. sample size) increases, the information going into the sample mean increases, making its measure of center that much more accurate, and thus making its variability decrease. (As an aside, this is the reason that statisticians always tell their collaborators that they need large samples: the larger the sample size, the more precise the sample mean, the better the test.) Note also that the relationship between the standard error and the sample standard deviation is predictable, as seen by comparing the last two columns in the first part of Table 6.1. For each sample size, these values are nearly identical, which means that the standard deviation of the sample mean is equal to the standard deviation of the sample divided by the square root of n .

While the sample mean is centered around the population mean, and its variability is accurately measured by the standard error, we still do not know the distribution of the sample means. Histograms of the 1,000 sample means are found in Figure 6.1 for sample sizes $n = 10, 20, 50$ and 100 . We see that even for small samples, the histograms closely approximate the superimposed normal curves. In one way we might have expected this, since data that are sampled from a normal population will most likely follow a normal distribution themselves.

Table 6.1 also presents the results from a simulation study based upon the male triglyceride data from the FLS study. Even though the values in this sample were distinctly non-normal, we can see that the average of the sample means for each sample size is close to the population value. The average standard deviations are somewhat lower than the population value for small sample sizes, but that difference decreases as the sample size increases. As was the case in the female cholesterol data, the standard error of the

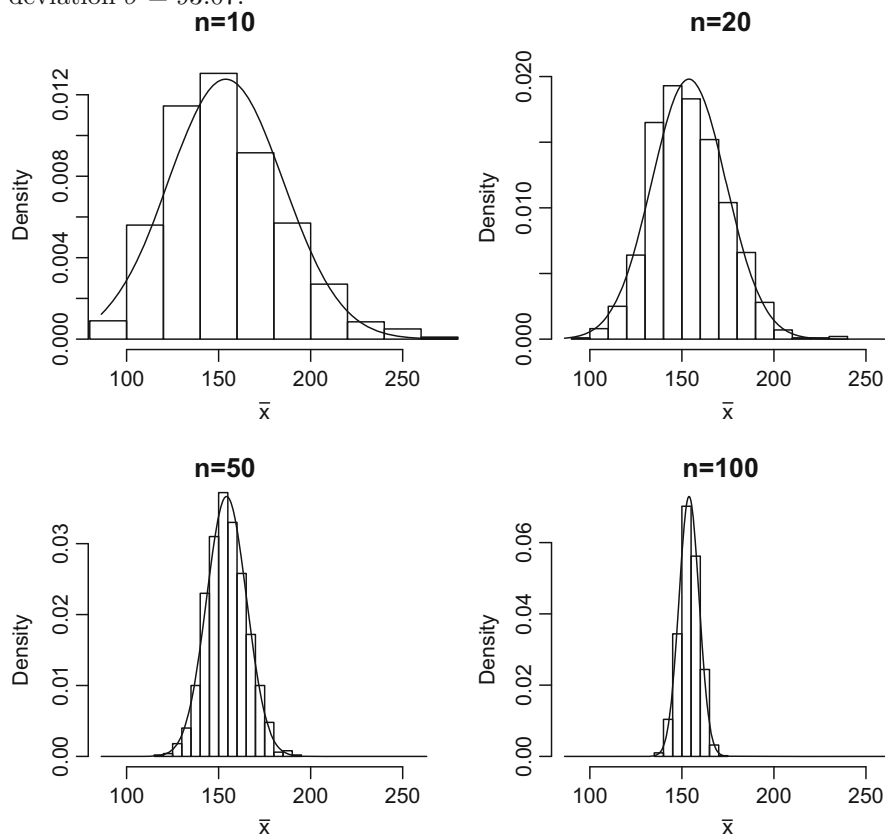
Figure 6.1: Histograms of 1,000 sample means of size $n = 10, 20, 50$ and 100 from the Female Cholesterol sample with mean $\mu = 203.8$ and standard deviation $\sigma = 37.86$.



mean decreases as the sample size increases and in all cases is close in value to that predicted from the population standard deviation. However, comparing the standard errors of the mean between the female cholesterol and male triglyceride studies shows that the mean for the triglyceride study has more variability, which is expected since the sample exhibits more variation. Histograms of the triglyceride means are found in Figure 6.2 for sample sizes 10, 20, 50 and 100. Here we see that, while not perfectly symmetrical, the sample means have an approximate normal distribution even for small sample sizes. Most importantly, the approximation to the normal distribution improves as the sample size increases.

The results from these two simulation studies represent two important characteristics of the sample mean. The first is that if the underlying population values are normally distributed, then samples obtained from that population are likely be normally distributed, and *most importantly*, the

Figure 6.2: Histograms of 1,000 sample means of size $n = 10, 20, 50$ and 100 from the Male Triglycerides sample with mean $\mu = 154.08$ and standard deviation $\sigma = 93.07$.



means of those samples are also likely to be normally distributed. The second characteristic is that even if the underlying population and samples are not normally distributed, the resulting sample means are likely to be normally distributed if the sample size is large enough. These results form the core of what is known as the Central Limit Theorem (CLT), which states – for our purposes – that for large enough sample sizes, the sample mean will be normally distributed, regardless of the underlying probability distribution. The implication of this result – the reason it is the *central* limit theorem – is that we can use the sample mean for inference in most cases, even if our sample is distinctly non-normal. In practice, we will implement the CLT using the following three rules:

1. If sample values are normally distributed, then the population from which they were drawn is likely to be normally distributed, and thus the resulting sample mean \bar{x} will be normally distributed.

2. If the sample size n is greater than or equal to 30, then the sample mean \bar{x} is normally distributed regardless of the distribution of the underlying sample data.
3. If the sample values are not normally distributed and the sample size is less than 30, then the central limit theorem does not apply and the sample mean does not necessarily have a normal distribution.

Note: it is important to distinguish between the distribution of the data, which may or may not be known, and the distribution of \bar{x} , which is described by the central limit theorem. We will make use of the CLT in making our decisions about which test we should use (see below).

6.2 Establishing Hypotheses

In the one-sample case for continuous data we were predominantly concerned with the population mean μ . Since we do not know the value of this parameter, we will use our sample data – most notably the sample mean \bar{x} – to construct and perform a hypothesis test on that population mean. Naturally, this requires us to assume some value for μ , which will arise from the research question we are interested in answering, and we ascribe this value the symbol μ_0 . In some cases, we will have a zero-valued population mean (where $\mu_0 = 0$), but we will often have a specific, non-zero value in mind.

Regardless of the assumed value of the population mean, we need to take this value in its context of the research question and form a set of hypotheses (H_0 and H_A). Though the parameter involved here (μ) is different from the parameter in Chapter 2 (p), the process of constructing null and alternative hypotheses is the same. We first translate the research question into a symbolic statement involving one of six possible symbols ($<$, $>$, \leq , \geq , $=$ and \neq). We then construct the symbolic statement that must be true if the original statement is false, recalling that the symbols ($<$, \geq), ($>$, \leq) and ($=$, \neq) are always paired with one another. We define the statement that contains some equality (\leq , \geq or $=$) as the null hypothesis (H_0 , being sure to change \leq or \geq to $=$), and we define the statement that does not contain equality ($<$, $>$ or \neq) as the alternative hypothesis (H_A). Table 6.2 below presents the possible pairs of null and alternative hypotheses we could create for a population mean μ with hypothesized value μ_0 , based upon certain key words from the research question.

The dataset “Female Chol Sample” contains cholesterol measurements from a subsample of 20 females from the larger FLS database. It was of interest to determine whether the average female had cholesterol levels greater than 200. Turning this research question into a symbolic statement, we isolate the phrase “greater than”, which means that our null hypothesis becomes $H_0 : \mu = 200$, while the alternative hypothesis becomes $H_A : > 200$.

Table 6.2: Possible Sets of Hypotheses for a Population Mean Based Upon Key Phrases from a Research Question.

Hypothesis	Key Phrases		
	“less than”, “greater than or equal to”, “at least”	“greater than” or equal to”, “at most”	“equal to”, “not equal to”
Null	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
Alternative	$H_A : \mu < \mu_0$	$H_A : \mu > \mu_0$	$H_A : \mu \neq \mu_0$

6.3 Assessing Assumptions (with R Code)

Before performing any test, we must determine whether the assumptions required for such a test are met. In any event, our sample must be representative of the population from which it’s drawn, and the subjects must be independent of one another. The best way to determine the veracity of these statements is to note whether or not the measurements were collected in a simple random sample (i.e. were subjects randomly sampled from the population). If we know this to be true, then we can proceed, otherwise we would either have to assume the sample was randomly collected (for our purposes) or not proceed with the test (in real life).

The determination of adequate sample size is based upon our assessment of the Central Limit Theorem (for those interested, we statisticians capitalize this theorem because it’s that important). Recall that if our sample size is greater than 30, then the CLT holds and our sample mean will be normally distributed, and we will thus know the distribution of our test statistic based upon that mean (what we “know” about the test statistic will be discussed below). Thus, regardless of the distribution of the data (normal, skewed, angry!), if our sample size is greater than or equal to 30, we can assume we have enough data. If we have fewer than 30 subjects, then one of two things will happen. In the first case, the sample data may still be close to normally distributed (as ascertained via histogram, box plot, or QQ plot). In this case, the CLT still holds, our sample mean is still normally distributed, and we can assume we have a large enough sample size. In the second case, when the sample data are not normally distributed, the CLT will not hold and we are unsure if the sample mean is normally distributed (it still *could be*, but we can’t be certain). Because of this uncertainty, we cannot use the parametric test (the so-called “t-test” that we will cover shortly) which relies upon the assumption of normality for the sample mean. Instead, in this case we will conduct a non-parametric test that is robust (i.e. is not adversely affected) to both small sample sizes and non-normally distributed data.

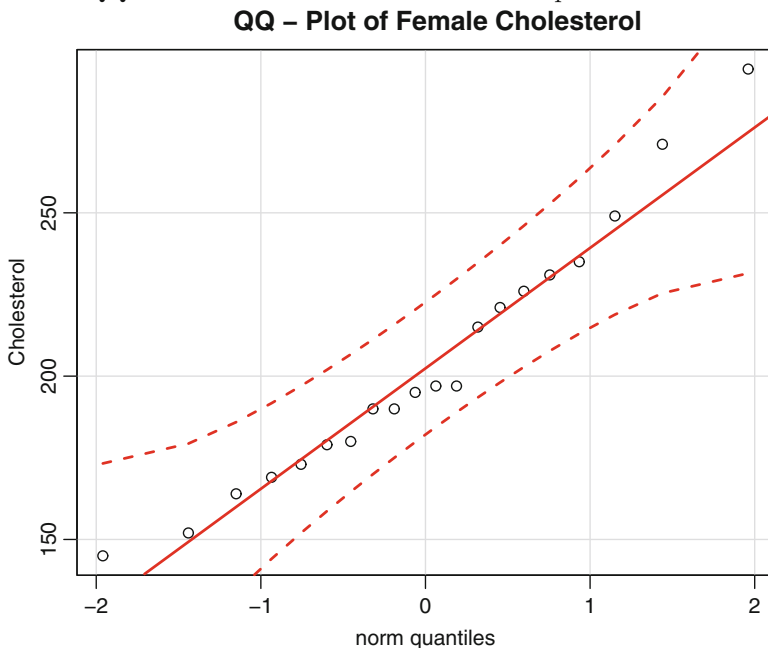
Based on the subsample from the FLS database, we first note that there are only 20 females who provided cholesterol values, which is less than the 30 required by the CLT. However, based upon the QQ plot shown in Figure 6.3, which was generated using the `qqplot()` function described in Chapter 5,

we can see that the cholesterol measurements appear normally distributed. Thus, we can conclude that the sample mean for this data set will be normally distributed as well.

6.4 Summarizing Data (with R Code)

For summarizing continuous data we will combine the conventions outlined in Chapter 5 with our assessment of the Central Limit Theorem (CLT). If the CLT holds (i.e. either our sample size is large OR there is evidence that the sample data are normally distributed, or both), then we will summarize our sample with the sample mean and standard deviation (not the standard error), along with a 95% confidence interval of the population mean (we will cover this later). However, if the CLT does not hold (i.e. the sample size is small AND there is evidence that the sample data are not normally distributed), then we will summarize our sample with the sample median and interquartile range; in this case you would not estimate a 95% confidence interval of the mean. Based on our assessment of the cholesterol subsample, we earlier concluded that the CLT holds. Thus, we summarize (using the `summary` and `sd` functions) the female cholesterol values with the mean and standard deviation, which for these data are $\bar{x} = 203.7$ and $SD = 39.03$ (recall your rounding specifications). We will calculate the confidence interval later.

Figure 6.3: QQ Plot for Female Cholesterol Subsample from FSL Database.



6.5 Performing the Test and Decision Making (with R Code)

As with most test statistics, we will take our summary measure of center (in this case the sample mean \bar{x}), subtract from it our hypothesized center of the data (μ_0), and divide that difference by the measure of variability for the population mean. If we (somehow) knew the population standard deviation was (say σ), then we know the standard error of the mean will be σ/\sqrt{n} , and our test statistic would become

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}. \quad (6.1)$$

The test statistic z would have a standard normal distribution, and we would be able to use that distribution to find critical values and p -values for the hypothesis test.

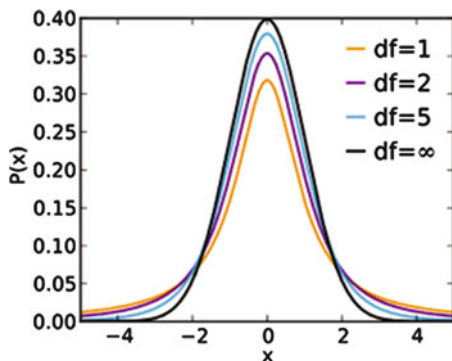
However, we do not (and will not) know the “true” population standard deviation σ . Intuitively, if we don’t know the population mean – which is *in the equation* to calculate standard deviation – there is no way for us to know the population standard deviation. Since we don’t know the population standard deviation, we replace it with the sample standard deviation s , the value we think will most closely resemble the population value. This yields the following test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}. \quad (6.2)$$

Note that the equation for test statistic t has the same general shape as that for test statistic z (Equation 6.1), with the only difference being the replacement of σ with s . However, while these test statistics have similar equations, their distributions are not the same. Notably, if we think of \bar{x} as a random variable – since it consists of n random variables – with a probability distribution, which we know to be normal under the CLT, then s must also be a random variable with a probability distribution of its own (the distribution of s^2 is actually proportional to a chi-square distribution). Since t is a ratio of two random variables, it actually has more variability than what is provided by the standard normal distribution attributed to z .

The distribution that most closely approximates the variability observed in test statistic t is called the Student’s t -distribution (or t -distribution, for short). This distribution is so-entitled, not because of its preponderant use by students in statistical courses, but because its creator published under the pseudonym “Student”. William Gosset derived the probability distribution while he worked for Guinness Brewery – the adult-beverage arm of the company who publishes the world-record books. At the time his employer would not allow him to publish under his actual name; hence the pseudonym.

Figure 6.4: Probability Curves for Student’s t -Distribution for Various Sample Sizes.



Compared to the standard normal distribution, the t -distribution is also centered at zero, but has a shorter “hump” in the center and thicker “tails” at both edges. This latter characteristic reflects the added variability in the test statistic due to it consisting of two random variables. Unlike the standard normal distribution, the shape of the t -distribution is dependent upon the observed sample size. Specifically, the relationship is through the *degrees of freedom* of the distribution, which in this case is $n - 1$. Note that this is the same value as the denominator of the standard deviation, which is actually the reason why the t -distribution has $n - 1$ degrees of freedom. As shown in Figure 6.4, the center of the t -distribution increases as the sample size (through the df) increases, while the tails get smaller. If we increase the sample size enough, there is practically no distinction between the t - and standard normal distributions, which is plotted as the case $df = \infty$.

Repeating the simulation example we conducted earlier for the female cholesterol data, for each of the simulated samples of size n we can also calculate the test statistic t , which will be based upon the mean, standard deviation and size for each sample. Part A of Table 6.3 shows the results from the Female Cholesterol parent sample for various sample sizes. Note that the test statistic was calculated in each case by using the “population” mean of 203.8. For each sample size we can see that the average test statistic value is close to zero, and as the sample size increases the standard error of the test statistic more closely approximates one. In fact the standard errors closely match with those produced by an actual t -distribution (SE_t).

Part B of Table 6.3 shows similar results for the Male Triglyceride example. While the average test statistics are a bit more removed from zero and the standard errors of the test statistic are a little larger than they were in the Female Cholesterol case, they are still quite close to zero and one, respectively, especially as the sample size increases. These two results show us that for adequately large sample sizes, the t -distribution is a good approximation of the distribution for test statistic t .

Table 6.3: Results from simulation studies for FLS Data. 1. Results based on 1,000 samples of size (n) from (A) the Female Cholesterol sample, and (B) the Male Triglyceride sample.

n	(A) Female Cholesterol			(B) Male Triglycerides		
	t	SE_z	SE_t	t	SE_z	SE_t
9	-0.02	1.14	1.15	-0.38	1.39	1.15
16	0.00	1.07	1.07	-0.24	1.20	1.07
25	-0.05	1.09	1.04	-0.22	1.15	1.04
36	0.00	1.09	1.03	-0.20	1.08	1.03
49	0.02	1.05	1.02	-0.18	1.07	1.02
64	0.04	1.04	1.02	-0.15	1.04	1.02
81	0.05	1.01	1.01	-0.16	1.03	1.01
100	0.06	1.00	1.01	-0.13	1.02	1.01

For the female cholesterol example, we have reported that the mean is 203.7, the standard deviation is 39.03, and the sample size is 20. Further, our null hypothesized value is stated as $\mu_0 = 200$. Thus, the test statistic for this problem is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{203.65 - 200}{39.027/\sqrt{20}} = 0.418. \quad (6.3)$$

While the t - and standard normal distributions are different, the interpretation of the test statistics are the same. In this case, the observed $t = 0.42$ implies that there is little evidence that the mean female cholesterol level is larger than 200.

Program 19 shows the R code to conduct the hypothesis test above concerning mean Female Cholesterol levels. We first need to read in the data using the `read.csv` and write it into the `Chol1.f` dataset. To conduct the t -test we use the `t.test` function in R. The first item in the `t.test` function is the data we want to test against, in this case `Chol1.f$BCcholes`. We also need to specify the assumed mean from the null hypothesis, which is `mu=200`. We must also state the alternative hypothesis using the `alternative` statement, where here we specify `"greater"` in order to test $H_0 : \mu > 200$.

The output from Program 19 provides the information we need for making inferences about the mean Female Cholesterol level, including the test statistic `t=0.4182`, the degrees of freedom, `df=19` and the `p-value=0.3402`. The output also informs us that we are conducting a `One Sample t-test`, the name of the data we are using `data: Chol1.f$BCcholes` and the alternative hypothesis. We can use this output to verify if R has analyzed the data correctly. The output also creates a confidence interval, which we will not use in this Chapter; we will return to the confidence intervals later.

Program 19 Program to generate hypothesis test for testing the mean Female Cholesterol level is greater than 200: $H_0 : \mu \leq 200$ versus $H_a : \mu > 200$.

Code:

```
# Read in the data
Chol1.f <- read.csv("Chp 6 Female Chol Sample.csv")

### Generate a test
t.test( Chol1.f$BCcholes,
        mu = 200,
        alternative = "greater"
      )
```

Output:

One Sample t-test

```
data: Chol1.f$BCcholes
t = 0.4183, df = 19, p-value = 0.3402
alternative hypothesis: true mean is greater than 200
95 percent confidence interval:
 188.5605      Inf
sample estimates:
mean of x
 203.65
```

6.5.1 Critical Value Method

Since the shape of the t -distribution is dependent upon the observed sample size, the critical value needed for hypothesis testing will also be sample size-dependent. The critical values for various sample sizes and significance levels are provided in Table 6.4. For small sample sizes, we can see that the critical values are larger than what would be used for the z -test based on the standard normal distribution (shown in the last row of Table 6.4). This makes sense, as for small samples we would expect the sample mean and standard error to differ more (on average) from the population values they are estimating than we would expect for larger sample sizes. As sample size (and thus degrees of freedom) increases, we can see that the critical values converge toward the critical value from the standard normal distribution, just as the shape of the t -distribution converges to standard normal for larger sample sizes. Note that if we are conducting a one-sided test, we would place all of our significance in that one tail, so we would use the stated α in Table 6.4. However, if we used a two-tailed test, we would divide the stated significance level in half and locate the critical value under that *halved critical value*. For example, if

our significance level was $\alpha = 0.05$, but we had a two tailed test with $n = 30$, then we would look under $\alpha = 0.025$ and select 2.045 (and -2.045) as our critical value(s).

Table 6.4: Critical Values from the t -Distribution for given Sample Size (n) and Significance Level (α).

	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$
n	Critical Value	Critical Value	Critical Value
5	3.747	2.776	2.132
10	2.821	2.232	1.833
15	2.624	2.145	1.761
20	2.539	2.093	1.729
25	2.492	2.064	1.711
30	2.462	2.045	1.699
50	2.405	2.010	1.677
100	2.366	1.984	1.660
∞	2.326	1.960	1.645

6.5.2 p -Value Method

Calculating p -values using the t -distribution is complicated and cannot be done with much fidelity through a table such as Table 6.4. We thus rely upon statistical software to calculate p -values. As always, we compare the correct p -value (the one appropriate for our set of hypotheses) to the stated significance level: if p -value $< \alpha$, then we reject the null hypothesis in favor of the alternative; otherwise we fail to reject the null hypothesis. For our example with the right-tailed alternative, the correct p -value is 0.3402, which is not smaller than the significance level $\alpha = 0.05$, so we fail to reject the null hypothesis that the mean female cholesterol level is equal to 200.

Our interpretation depends upon our decision. If we reject the null hypothesis, then we conclude that the population mean is most likely less than, greater than, or different from the hypothesized value. Alternatively, if we fail to reject the null hypothesis, then we state that there is no evidence to suggest that the population mean is somehow different from the hypothesized value. Remember that hypothesis tests make statements about the population parameters, not the sample statistics.

For either the critical value method or p -value method, we report the degrees of freedom and observed test statistic. We must then also report either the critical value or p -value (dependent upon which method we use) and whether we reject or fail to reject the null hypothesis. Note that due to the simplicity in reporting the results we generally prefer to report the results of the p -value method rather than those from the critical value method.

6.5.3 Confidence Intervals

As an additional aspect to the data summary, we generally provide a confidence interval of the population mean. This interval simultaneously provides an indication of both the center and variation in the data, since the interval is centered at the sample mean, and the width of the interval is dependent (mostly) upon the standard deviation. For one-sample continuous data – much like the case for one-sample proportions – the confidence interval consists of three things: a point estimate, a measure of variability, and a probabilistic measure. As the point estimate of the center of the data we will take the sample mean (\bar{x}), and as the measure of variability of the data we will take the sample standard error of the mean (s/\sqrt{n}). Since we used the t -distribution for hypothesis testing, we will again make use of that distribution for our probability measure. This measure will be dependent upon the sample size n through the degrees of freedom ($n - 1$), but the significance level we will use must be split into two pieces, one for each tail of the confidence interval. So if we want a 95% confidence interval, we will use $\alpha/2 = 0.05/2 = 0.025$. Likewise, if we want a 99% confidence interval, we will use $\alpha/2 = 0.01/2 = 0.005$. Together, we use the degrees of freedom and significance level to find the critical value $t_{n-1, 1-\alpha/2}$. Putting the three components together, the general formula for a $(1 - \alpha)100\%$ confidence interval of the mean is

$$(\bar{x} - t_{n-1, 1-\alpha/2}s/\sqrt{n}, \bar{x} + t_{n-1, 1-\alpha/2}s/\sqrt{n}). \quad (6.4)$$

Returning to the cholesterol example, recall that our sample $n = 20$ females had a mean cholesterol level of 203.65 and a standard deviation of 39.027. When $n = 20$ ($df = 19$) and with 95% confidence ($\alpha = 0.05$), the critical value from the t -distribution is 2.093. Thus, the 95% confidence interval for the mean cholesterol level in females is

$$\begin{aligned} (203.65 - (2.093)(39.027/\sqrt{20}), & \quad 203.65 + (2.093)(39.027/\sqrt{20})) \\ (185.38, & \quad 221.92) \end{aligned}$$

As always, we round the 95% CI to the same degree of specification as the mean, so we report (185.4, 221.9).

Program 20 shows the R code to generate a 95% confidence interval for the mean cholesterol levels in females. To generate the confidence interval we again use the `t.test` function, but here we omit the `mu=` and `alternative=` functions, and rather specify our desired confidence level using the `conf.level` statement. Here we wish to have a 95% confidence interval so we need to specify `conf.level=0.95` since `conf.level` must always be between 0 and 1. The `t.test` function provided in Program 20 gives as lower and upper confidence limits 185.3850 and 221.9150, respectively, which matches what we calculated by hand. Notice from the output in Program 20 that `t.test` produces much additional information (e.g. `t`, `df`, and the p -value) that we do not need. Since we are here only interested in estimating the confidence interval, this information should be ignored.

Program 20 Program to generate a 95% confidence interval for the mean Female Cholesterol level.

Code:

```
# Read in the data
Chol1.f <- read.csv("Chp 6 Female Chol Sample.csv")

### Generate the confidence intervals
t.test( Chol1.f$BCcholes,
        conf.level=0.95
      )
```

Output:

One Sample t-test

```
data: Chol1.f$BCcholes
t = 23.3366, df = 19, p-value = 1.895e-15
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 185.3850 221.9150
sample estimates:
mean of x
 203.65
```

6.6 Contingency Methods (with R Code)

In the event that the Central Limit Theorem does not apply (sample size too small and data not normally distributed) we cannot use the t -test method. This is because if the CLT doesn't hold, we have no idea if the mean has a normal distribution and whether our test statistic t will have a t -distribution. In this case, we can use an alternative method which does not require the CLT to hold: the Wilcoxon Signed-Rank test (or simply, the Sign test).

Unlike the t -test, the Wilcoxon test actually tests hypotheses about the median, rather than the mean. The underlying concept is this: if the median is truly μ_0 (or close to μ_0), then half of the observed values should be less than μ_0 and half of the observed values should be greater than μ_0 . If μ_0 differs somewhat from the true median, then the percentage of observations above and below the hypothesized median μ_0 will both differ from 50%. Of course, the greater the degree of these discrepancies, the more evidence there is of a difference between the hypothesized median μ_0 and the actual median.

While the calculations for the Wilcoxon test are not that complicated – they're actually kind of fun – we will not cover them. Instead, we will rely upon R to calculate the p -value for us (note that there is no meaningful test

statistic or degrees of freedom for this test). Thus, all we report is the p -value for our relevant hypothesis, which is formed in the same manner as for the t -test, except we replace the hypothesized mean with a hypothesized median (we usually assume these would be the same value). In R, you obtain the Wilcoxon Signed-Rank test by using the `wilcox.test` function.

Using a different example, we have a small sample of blood creatine phosphokinase (CPK) from 11 cross-country relay participants 12h into a 24h relay race (Zuliani et al. 1983). The data are listed below in Program 21. Since the sample size is small ($n=11$), and since the data are not necessarily normally distributed (see for yourself), we could arguably use the nonparametric test. We use the `wilcox.test` function to perform the test, specifying also the null-hypothesized value (here we specify $\mu=300$) and the alternative hypothesis (here we ask for a two-sided test). The p -value we obtain (0.0283) is less than our significance level, which means we reject the null hypothesis. While we should not have conducted a t -test, it is interesting to note that for this example we would have obtained the following results: $t = 2.10$, $df = 10$, $p\text{-value} = 0.06246$. So if we had erroneously performed the t -test, we would not have found a significant result.

Furthermore, note that the `wilcox.test` produced two warning messages, informing us that it could not compute an *exact* p -value; instead it produced an *approximate* p -value. This does not invalidate our results, but rather informs us that the stated p -value is an approximation. The only time we should be worried about using an approximate p -value is when the p -value is near our significance level α . For example, if we chose our significance level to be $\alpha = 0.01$ and our approximate p -value is 0.0098 then we should be careful about how strong of a statement we wish to make about rejecting H_0 . Since the p -value is approximate we are unsure if the *exact* p -value is actually greater than $\alpha = 0.01$ or possibly less than $\alpha = 0.01$.

6.7 Communicating the Results

The following is an example of the IMRaD write-up for the cholesterol example.

Introduction: Increased levels of cholesterol are an important co-morbidity to obesity, and are thought to be a strong predictor of cardiovascular disease. Using data from a large-scale epidemiological study, researchers tested the hypothesis that female cholesterol levels were greater than 200.

Methods: Cholesterol levels were obtained from 20 female participants in the Fels Longitudinal Study. These cholesterol levels were checked for normality using QQ plots, and were summarized with means, standard deviations and 95% confidence intervals. The alternative hypothesis that the mean cholesterol level was greater than 200 was assessed using a one-sample t -test against a null hypothesis that the mean cholesterol level was equal to 200. We will reject the null hypothesis if the resulting p -value is less than

Program 21 Program to generate a Wilcoxon Signed Rank hypothesis test for testing the center of the CPK distribution is different from 300: $H_0 : \mu = 300$ versus $H_a : \mu \neq 300$.

Code:

```
# Read in the data
cpk <- c(200, 300, 520, 490, 440, 380,
        1040, 1340, 260, 365, 400)

### Generate a test
wilcox.test(cpk,
            mu = 300,
            alternative = "two.sided"
            )
```

Output:

Wilcoxon signed rank test with continuity correction

```
data:  cpk
V = 49.5, p-value = 0.02831
alternative hypothesis: true location is greater than 300
```

Warning message:

```
1: In wilcox.test.default(cpk, mu = 300, alternative = "two
.sided") : cannot compute exact p-value with ties
2: In wilcox.test.default(cpk, mu = 300, alternative = "two.
sided") : cannot compute exact p-value with zeroes
```

the stated significance level, and we will fail to reject otherwise. All data summaries and tests were performed using the R statistical software, and all tests were conducted using a significance level of $\alpha = 0.05$.

Results: The sample is assumed representative, subjects are assumed independent, and inspection of the QQ plot shows evidence of normality, so that we can conclude a large enough sample size. The mean cholesterol level of the 20 female study participants was 203.7 ($SD = 39.03$, 95 % $CI : 185.4, 221.9$). The results from the one-sample t -test are $t_{19} = 0.42$, p -value = 0.3402. Since p -value = 0.3402 > 0.05, we fail to reject the null hypothesis that the mean cholesterol level in females is equal to 200.

Discussion: This study found little evidence that females exhibit high levels of cholesterol. Since sustained elevated cholesterol levels can be an early indicator of cardiovascular degeneration, this is good news for females.

6.8 Process

1. State research question in form of testable hypotheses.
2. Determine whether assumptions are met.
 - (a) Representative
 - (b) Independence
 - (c) Sample size: check distribution of data and sample size.
3. Summarize data.
 - (a) If sample size OR normality is adequate: summarize with sample size, mean, standard deviation and 95% CI.
 - (b) If sample size AND normality are inadequate: summarize with sample size, median and IQR.
4. Perform Test.
 - (a) If sample size is greater than 30 OR data are normally distributed, then use t -test.
 - (b) If sample size is less than 30 AND data are not normally distributed, then use Wilcoxon signed-rank test.
5. Compare test statistic to critical value or calculate p -value.
6. Make decision (reject H_0 or fail to reject H_0).
7. Summarize with IMRaD write-up.

6.9 Exercises

1. We can consider example from [Green et al. \(2005\)](#) who is interested in estimating the amount of *diethylhexyl phthalate* (DEHP) that leach from IV tubing and bags into intravenous medications. Suppose they take 25 standard IV bags and standard tubing of length 1 m and put distilled water in the bag and let it sit for 8 h and then drain the bag through the tube into a container. From each of the containers they measure the DEHP in ng/mL and suppose they obtain the following data:

53.0	40.4	39.1	39.6	52.9
32.8	51.7	42.9	55.0	43.8
51.1	44.2	38.3	44.3	47.7
43.7	44.2	40.0	60.1	42.9
27.0	50.8	37.0	47.5	69.6

The researchers want a 98% confidence interval for the mean DEHP level that leaches from the IV bag and tubing.

2. [Yoshinaga et al. \(2004\)](#) are interested in the amount of radiation people working in a x-ray lab receive. In general, the typical person in the USA receives on average 3.6 mSv (milli Sievert) of radiation per year. Specifically they wish to know if x-ray labs technicians receive more than 0.01 mSv per day. They take a sample of 15 workers and places a device on each technician that records the amount of radiation they receive. Suppose this is the data they collected. Answer their question.

0.0023	0.0072	0.0054	0.0092	0.0114
0.0013	0.0017	0.0047	0.0069	0.0078
0.0082	0.0087	0.0044	0.0056	0.0087

3. As part of a study on an implantable medication system for insulin delivery, [Saudek et al. \(1989\)](#) measured the percentage above ideal body weight in 18 patients (found below). Test the hypothesis that the mean percentage above ideal body weight is less than 100.

107	119	99	114	120	104	88	114	124
116	101	121	152	100	125	114	95	117

4. In there study on the effects of milk conception on hypervitaminosis of vitamin D, [Jacobus et al. \(1992\)](#) also collected measurements on serum creatinine. Estimate a 95% confidence interval for the mean creatinine level (in micromols/liter) using the data provided below. Perform the hypothesis test that the mean or median creatinine level is less than 115 mm/l, which is the largest value in the established normal range for creatinine.

159, 44, 80, 309, 80, 186, 433, 380

5. The Glasgow Coma Scale is (GCS) is used to measure the severity of a brain injury. The scale ranges from 3 to 15, with 3 indicating completely unconscious state (no response to any stimuli) to 15 indicating completely conscious state (normal response to all stimuli). There is some evidence that a patients initial GCS score may be correlated with recovery prognosis. Brain injury researchers generally report mean GCS scores of their patients. [Kreutzer et al. \(2009\)](#) give the data below for GCS scores for patients in their study. Create a 99% confidence interval for the mean GCS for brain injury patients.

3	15	7	3	8	3	3
3	3	6	9	3	13	12
4	3	15	3	15	3	15
3	14	8	6	15	3	6
7	4	5	7	7	3	3
14	15	15	3	3	3	3
15	4	13	3	10	15	3
15	8	14	15	13	12	2
7	12	10	3	12	15	15
5	9	3	6	4	15	12
5	9	3	15	3	10	10

6. [Lansford et al. \(2010\)](#) is interested in the number of sexual partners for adolescents between age 16 and 22. They collected a sample of 526 people in this age group and asked them the number of partners they have engaged in which they have engaged in sexual activities. The following sample is consistent with their data: Create a 95% confidence interval for the mean number of sexual partners for all adolescents between 16 and 22.

2	7	3	1	2	0
0	1	3	2	0	1
0	4	3	0	6	7
1	2	1	8	1	3
0	2	6	1	6	0

7. [Rossi et al. \(2009\)](#) is interested in the waiting times for knee replacement surgery. They took a survey of 161 patients and calculated the days from the initial visit to surgery. The following sample is consistent with their data. Create a 97% confidence interval for the mean waiting time for all knee replacement surgeries.

53	67	69	68	53
64	72	63	73	69
64	77	65	54	52
63	66	61	70	57
66	72	75	72	77

Chapter 7

Two-Sample Means

7.1 Introduction: Independent Groups or Paired Measurements

Building off the methods for the one-sample case covered in Chapter 6, it is natural to extend those ideas to cases where there are two samples. However, we run into the fact that there are two different cases where we are interested in comparing the means from two samples. The first is the case where we have two distinct groups of subjects, meaning that any subject providing a measurement in one group is prohibited from providing a measurement in the other group. These samples are then assumed to be independent of one another, in much the same way observations between different subjects are generally thought to be independent. This case will resemble the process covered in Chapter 6 in some ways, though the details will differ. The second case is where the two samples consist of *exactly the same subjects*, meaning that the same subjects were measured twice (sequentially or simultaneously). These samples are then assumed to be dependent, meaning that we have every reason to believe that the values taken by particular subjects in one sample are going to be related to the measurements from those same subjects in the second sample (i.e. the values the measurements take depend on one another). This case will also resemble the process for one-sample means, but in different ways and for different reasons than the first case. Though both cases involve two samples, they are handled in entirely different manners, and as such they will be covered separately in this Chapter.

7.2 Independent Groups

7.2.1 Establishing Hypotheses: Independent Groups

In the event that we have two independent samples of subjects, we are generally interested in comparing the means from those two samples. Examples occur often in practice, such as making comparisons between male and female subjects, or determining if there is an increased response of some biomarker at one dose level as compared to another. Thus, any hypothesis will involve direct comparisons of one population mean (μ_1) to another (μ_2). For instance, if we believe – before hand – that the mean of one group is larger than the mean of the other group, then we would write $\mu_1 > \mu_2$. As was explained in the two-sample proportion case in Chapter 3, it will be easier for us to work with the difference $\mu_1 - \mu_2 > 0$ and have us consider $\mu_1 - \mu_2$ as the parameter of interest. Once we focus on the difference between two population means, setting up the null and alternative hypotheses follows much in the same manner as we have previously established: translate the research question into a symbolic statement, find that statement’s logical complement, and then assign one as the null hypothesis and the other as the alternative hypothesis. Table 7.1 shows the possible sets of null and alternatives that we could use.

Table 7.1: Possible Sets of Hypotheses for a Difference in Two Independent Population Means Based Upon Key Phrases in the Research Question.

Hypothesis	Key Phrases of μ_1 relative to μ_2		
	“less than”, “or equal to”, “at least”	“greater than”, “or equal to”, “at most”	“equal to”, “not equal to”
Null	$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : \mu_1 - \mu_2 = 0$	$H_0 : \mu_1 - \mu_2 = 0$
Alternative	$H_A : \mu_1 - \mu_2 < 0$	$H_A : \mu_1 - \mu_2 > 0$	$H_A : \mu_1 - \mu_2 \neq 0$

As an example, a clinical trial was conducted to determine whether a certain treatment reduced blood pressure in hypertensive male adults. Participants in this trial were randomized into two groups: those randomized to the treatment group received the active treatment, while those randomized to the placebo group received a non-active placebo. The clinicians administering this study hypothesized that end-of-trial mean systolic blood pressure (SBP) for subjects in the treatment group would be lower than the mean SBP in the placebo group. If we let μ_T represent the treatment group population mean and μ_P represent the placebo group population mean, the symbolic representation of the research hypothesis is $\mu_T < \mu_P$, or $\mu_T - \mu_P < 0$. The opposite of this statement is $\mu_T - \mu_P \geq 0$, so we assign the first statement as the alternative hypothesis $H_A : \mu_T - \mu_P < 0$, and we assign the second statement (with only an equal sign) as the null hypothesis $H_0 : \mu_T - \mu_P = 0$.

7.2.2 Assessing Assumptions (with R Code)

As always, we must start with an assessment of the independence of subjects (all subjects irrespective of group membership) and the samples' representativeness of the population from which they were drawn. Since there are two groups here, one could make the argument that the values from subjects *within the same group* are closer together in value than are values from subjects *between groups*. However, this is not what we mean or require here for independence. All we need to know is that the measurement provided by one subject was not or could not be influenced by the measurement of any other subjects (regardless of group membership). The representativeness of the two samples is best determined by the randomization process for allocating subjects: if the study adequately randomized subjects, we can assume the samples are representative.

Since we have two samples, the determination of adequate sample size is somewhat more challenging than it was in Chapter 6. To be safe, one must ascertain that the Central Limit Theorem applies to both samples. Thus, for each sample, we first look at the shape of the data via histograms, box plots or QQ plots. If both samples appear normally distributed, then we can assume we have enough data. Otherwise, we must check to see if we have at least 30 observations in any sample that is not normally distributed. If the CLT applies to both samples (either through distribution or sample size; note that we could have a case where we have one small sample that is normal and one large, non-normal sample), then we can continue on to the parametric t -test (described below). However, if the CLT does not apply to one or both of the samples, then we cannot use the parametric t -test and must instead use non-parametric methods.

Aside from the three usual assumptions, the two independent samples case requires us to check a fourth assumption, that being equal variances between to the two samples. There are several ways to do this, ranging from the subjective to the objective. Subjectively, you can simply look at the standard deviations and determine whether they are close together (e.g. 4.52 and 4.54) or far apart (e.g. 4.52 and 25.67). In some cases this visual inspection will be sufficient, but it won't work for others (e.g. 4.52 and 7.12). This is an informal method, so we may need to rely upon a formal hypothesis test, of which 4 are commonly reported: "Levene's Test", "Bartlett's Test", the "F Test", and the "Brown-Forsythe Test". In R, we can invoke these tests using the `levene.test`, `bartlett.test`, and `var.test` functions. For our purposes, we will use the `var.test` function, which provides the results for the "F Test", where large p -values (greater than 0.05) indicate roughly similar variances, while small p -values (less than 0.05) indicate different variances. The decision to visually inspect the data or perform a formal test is not a formalized process, but when in doubt, you can always assume the variances are unequal (it's better to mistakenly assume the variances are unequal when they actually are, then to mistakenly assume that they are equal

when they aren't). We will explain the repercussions of unequal variances with regards to summarizing data, the test statistic and confidence intervals below.

The data from the blood pressure clinical trial are presented in Table 7.2. The sample sizes in each group (Treatment: 15; Placebo: 14) are too small for us to automatically invoke the CLT, so we will rely upon histograms and QQ plots to assess normality. From an inspection of the QQ plots (Figure 7.1) we can assume that the two samples are normally distributed. Thus, the CLT applies and we can use the parametric test. The standard deviation for the Treatment group in this example is 4.35, while the standard deviation for the Placebo group is 7.47. A quick comparison shows that the Placebo SD is almost twice as large as the Treatment SD, while, the F test (obtained using the `var.test()` command with data sets listed in Program 23 below) yields a p -value = 0.05452. While we technically wouldn't reject this null hypothesis of equal variances based on the p -value, it is close to 0.05. This might be the type of case where either hypothesis (equal or unequal variances) is debatable. For our purposes, we will "believe the test" and assume that the variances are equal (though you do not have to if you "believe your eyes and not the test").

Table 7.2: Systolic Blood Pressure Measurements from Subjects Enrolled in a Placebo-Controlled Clinical Trial.

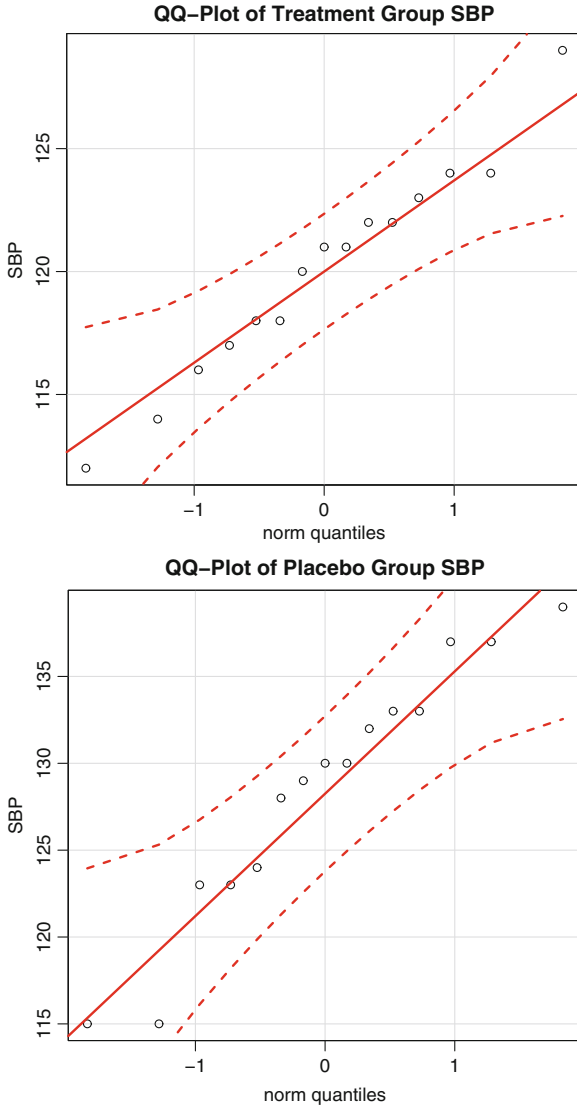
Treatment Group			Placebo Group		
121	112	114	137	130	115
129	121	117	128	133	129
118	122	124	132	130	137
118	120	123	124	115	139
124	116	122	133	123	

7.2.3 Summarizing Data (with R Code)

If the CLT applies to both of our samples, then we will summarize both samples with their respective sample size, mean, standard deviation and 95% confidence interval (see below). In addition, we will also summarize the difference between the sample means with the observed difference (e.g. $\bar{x}_1 - \bar{x}_2$), the standard error of that difference, and a 95% confidence interval. The standard error we report for the difference will depend upon our assessment of the assumption of equal variances. If we have assumed equal variances, then we will use the following measure for the standard error of the difference in means

$$(\text{Pooled})s = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (7.1)$$

Figure 7.1: QQ Plots for Systolic Blood Pressure Clinical Trial Data based on Group Status



where s_p^2 called the pooled estimator

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{7.2}$$

The estimator s_p effectively adds the variability between the two samples and approximates the overall standard deviation. On the other hand, if we

have assumed unequal variances, then we use the following measure for the standard error of the difference in means

$$(\text{Unpooled})s = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \quad (7.3)$$

It should be clear to see that the unpooled estimator is just a weighted sum of the two sample variances (weighted inversely by sample size). If we assume equal variances, we will report the difference in sample means, the pooled standard error based on s_p , and a 95 % confidence interval on the difference also calculated using the pooled standard error (see below). If we assume unequal variances, we will report the difference in sample means, the unpooled standard error, and a 95 % confidence interval on the difference also calculated using the unpooled standard error estimate (see below). Note here that if the sample sizes for each group are equal, then the two standard error estimates will be the same as well. If the CLT does not apply to either sample, then we will summarize each sample with the sample size, median and interquartile range, and we will not report the difference in either the means or medians.

The summary data for each group in the blood pressure clinical trial example are provided in Table 7.3. Based on our assessment of the data, we assumed that the variances are equal, so we report the pooled standard error for the difference (though not the treatment means). We see that the observed difference between sample means is the same for both the equal- and unequal-variance cases (as we would expect), while the standard errors for that difference are similar but not equivalent, indicating that there is some (small) difference in the group standard deviations. We will discuss the confidence intervals later in this chapter. In R, we may obtain the sample-specific data summaries using the `summary`, `sd`, and `t.test` functions discussed in Chapter 6. For the differences and standard errors in both the equal and unequal variance cases, we need to calculate the values “manually” (i.e. by coding equations directly in R to get the desired numbers), as shown in

Table 7.3: Summary Data for Systolic Blood Pressure Measurements Based on Group Status.

Group	n	Mean	SD	95 % CI
Treatment	15	120.1	4.35	(117.7, 122.5)
Placebo	14	128.9	7.47	(124.6, 133.2)
		Mean	SE	95 % CI
Difference (T-P, Equal-Var.)	–	–8.9	2.25	(–13.5, –4.2)
Difference (T-P, Unequal-Var.)	–	–8.9	2.29	(–13.6, –4.1)

Program 22. Even for those of us who are somewhat tepid in their computing skills, calculating this information (including Equations 7.1, 7.2, and 7.3) is rather simple: the difference in sample means is obtained by the `diff` equation; the pooled standard error (for the equal variance case) is obtained by the `sep` equation; and the unpooled standard error (for the unequal variance case) is obtained by the `seu` equation. We will discuss using R to calculate the confidence intervals on the differences below.

Program 22 Program to calculate difference in two sample means and both pooled and unpooled standard errors.

Code:

```
[!h]
# Input the data
treatment1 <- c( 121, 112, 114, 129, 121,
                117, 118, 122, 124, 118,
                120, 123, 124, 116, 122)

placebo1 <- c ( 137, 130, 115, 128, 133,
                129, 132, 130, 137, 124,
                115, 139, 133, 123)

diff<-mean(treatment1)-mean(placebo1)
nt<-15
np<-14
sdt<-sd(treatment1)
sdp<-sd(placebo1)
sp<-((nt-1)*sdt*sdt+(np-1)*sdp*sdp)/(nt+np-2)
sep<-sqrt(sp*(1/nt+1/np))
seu<-sqrt(sdt*sdt/nt+sdp*sdp/np)
diff
sep
seu
```

Output:

```
diff
-8.861905
sep
2.250064
seu
2.290133
```

7.2.4 Performing the Test and Decision Making (with R Code)

For both parametric cases (i.e. the CLT applies and we have either equal or unequal variances), the test statistic we use will be based on the t -distribution, much like the case described in Chapter 6. In this Chapter our point estimate of the population difference $\mu_1 - \mu_2$ will be the observed mean difference $\bar{x}_1 - \bar{x}_2$. We generally assume that the population difference is zero-valued, so that $\bar{x}_1 - \bar{x}_2$ is the numerator of our test statistic. In the case where we have assumed equal variances, the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \quad (7.4)$$

In the case of where we have assumed unequal variances, the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \quad (7.5)$$

For both the equal and unequal variance cases, the test statistics are similar (they differ only in the standard errors in their denominators), though their distributions are not the same. For the record, they both have t -distributions, but they each have different degrees of freedom. The equal variance test statistic has $n_1 + n_2 - 2$ degrees of freedom, while the unequal variance test has ν degrees of freedom. The symbol ν (pronounced “nu”) comes from what is called the “Saitterthwaite approximation”, which is used because the actual expression for the degrees of freedom is difficult to derive, and has the following equation.

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}. \quad (7.6)$$

While complicated, R will determine the value of ν so that it won’t have to be calculated by hand. This value often is reported as a decimal (no more than one decimal place (e.g. 34.2)), though you may report it as an integer. In any event, you must round ν down, as you cannot report greater degrees of freedom than you actually have.

Program 23 shows the R code and output for performing the equal-variance, two-sample t -test for comparing the systolic blood pressure means between the treatment and placebo groups. For this test we see that the observed test statistic is -3.93851 (which we would report as -3.94), while the reported degrees of freedom are 27. To obtain the unequal variance test statistic, we need to change the setting for the `var.equal` function to `FALSE`. This would provide an observed test statistic of -3.8696 (which we report as

-3.87), while the reported degrees of freedom are 20.619 (which we report as 20.6). Note that in both of these case, specifying the value for the null hypothesis ($\mu=0$) and alternative hypothesis (`alternative="less"`) are the same as they were in Chapter 6.

Program 23 Program to conduct an equal-variance, two-sample t -test on Systolic Blood Pressure Measurements P.

Code:

```
# Input the data
treatment1 <- c( 121, 112, 114, 129, 121,
                117, 118, 122, 124, 118,
                120, 123, 124, 116, 122)

placebo1 <- c ( 137, 130, 115, 128, 133,
                129, 132, 130, 137, 124,
                115, 139, 133, 123)

# Call t.test()
t.test(treatment1, placebo1, mu=0, alternative="less",
       var.equal=TRUE)
```

Output:

Two Sample t-test

```
data: treatment1 and placebo1
t = -3.9385, df = 27, p-value = 0.0002603
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -5.029398
sample estimates:
mean of x mean of y
 120.0667  128.9286
```

Critical Value Method

Since the test statistics for the equal- and unequal-variances cases have different t -distributions, it comes as no surprise that the critical values we obtain from those distributions will not be the same. If we assume equal variances, then we will draw our critical value from a t -distribution with $n_1 + n_2 - 2$ degrees of freedom. If we have a two-tailed alternative hypothesis ($H_A : \mu_1 - \mu_2 \neq 0$), we will take the $(1 - \alpha/2)100$ th percentile from the t -distribution, or $t_{n_1+n_2-2, 1-\alpha/2}$, and take the positive and negative values of this percentile to obtain our two critical values. If we have a right-tailed

alternative hypothesis ($H_A : \mu_1 - \mu_2 > 0$), we will take the $(1 - \alpha)$ 100th percentile from the t -distribution, or $t_{n_1+n_2-2, 1-\alpha}$, and for a left-tailed test we simply take the negative of the critical value from the right-tailed test (since the t -distribution is symmetric), or $-t_{n_1+n_2-2, 1-\alpha}$. For the unequal-variance t -test, the critical values are obtained in the same way, except we use ν degrees of freedom to obtain the relevant percentiles from the t -distribution. Since R does not automatically or immediately present the critical value, we will not cover this method.

p -Value Method

In the equal-variance case, we will use the observed test statistic to obtain our p -value from a t -distribution with $n_1 + n_2 - 2$ degrees of freedom. Likewise, in the unequal-variance case, we will obtain a p -value from a t -distribution with ν degrees of freedom. In either case we reject the null hypothesis if the p -value is less than the stated significance level α . Using our blood pressure example – and assuming equal variances – and recalling that our alternative hypothesis is that the treatment group has lower mean than the placebo group ($H_A : \mu_T - \mu_P < 0$), our left-tailed test means we have p -value = 0.0003, which we can obtain in R by using `t.test` function with the `var.equal=TRUE` specification. A similar set p -values are produced for the unequal-variance t -test, and selecting the appropriate value based on our alternative hypothesis we get p -value = 0.0005, which we can obtain in R by using `t.test` function with the `var.equal=FALSE` specification

Interpretation

For the equal-variance independent-sample t -test, our p -value (0.0003) was less than the significance level ($\alpha = 0.05$). This means that the statistical evidence suggests that mean systolic blood pressure for subjects on the active treatment is lower than the mean systolic blood pressure for subjects on the placebo. In other words, the evidence *seemingly suggests* that the active treatment lowers blood pressure compared to the control, however, we have to be careful with this kind of statement (were the SBP measurements already lower in the treatment group, or higher in the placebo group, before the treatment was administered?). Keeping the comments strictly on the means and less on how the drug works would be an advisable approach here.

Interestingly, if we conservatively assume that the two variances are unequal when they actually differ to some trivial extent, the critical values and p -values we get from the two methods will be close (when the two variances are close in value the unequal-variance test usually provides a slightly larger critical value and p -value than the equal-variance test). In fact, if the two variances are *exactly equal*, then the test statistics we obtain from both methods will be equal as well. Recall that the standard errors (and thus the test statistics) will be identical if the sample sizes are equal. These facts lend

credence to the rule that we should choose the unequal-variance test if we are unsure about the equality of the variances: even if we erroneously assume that the variances are unequal when they actually are equal, the tests will perform similarly (with only a modest loss in the ability to find a significant result). Note that this was the case in our blood pressure example, where the standard errors for the differences were similar (equal-var.: 2.25; unequal-var.: 2.29), as were the test statistics (equal-var.: -3.94 ; unequal-var.: -3.87) and p -values (equal-var.: 0.0003; unequal-var.: 0.0005).

Confidence Intervals

Along with the confidence intervals on the population means for each of the two groups, we must also provide a confidence interval on the difference between the two means. Naturally, the observed mean difference $\bar{x}_1 - \bar{x}_2$ is used as the point estimate for this interval, but the measure of variability and the probabilistic measure (as you might expect) depend upon our assumption of equal variances. If we have assumed that the two variances are equal, then we use the same standard error for the difference that was used for the equal-variance two-sample t -test, and the probability measure is taken from a t -distribution with $n_1 + n_2 - 2$ degrees of freedom ($t_{n_1+n_2-2, 1-\alpha/2}$). Then the $(1 - \alpha)100\%$ confidence interval for the difference between the means from two independent samples with equal variances is

$$\left((\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2, 1-\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2, 1-\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \quad (7.7)$$

where s_p is part of the pooled estimate of the standard error described previously. If we have assumed that the two variances are unequal, then we use the same standard error for the difference that was used for the unequal-variance two-sample t -test, and the probability measure is taken from a t -distribution with ν degrees of freedom (recall that ν was defined earlier). Then the $(1 - \alpha)100\%$ confidence interval for the difference between the means from two independent samples with equal variances is

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\nu, 1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{\nu, 1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right). \quad (7.8)$$

Fortunately, R will calculate these intervals for us using the `t.test` function with the `alternative="two.sided"` optionality included. For the equal-variance confidence interval using the pooled standard error, we specify the `var.equal=TRUE` option, while for the unequal-variance confidence interval using the unpooled standard error, we specify the `var.equal=FALSE` option. Program 24 shows the R-code for creating a 95% confidence interval for the Systolic blood pressure example in the equal-variance case. Notice that the reported 95% confidence interval on the difference ($-13.478654, -4.245156$) is the same as what we calculated by hand and reported in Table 7.3.

Program 24 Program to create a 95% confidence interval on the difference for the Systolic blood pressure example.

Code:

```
# Input the data
treatment1 <- c( 121, 112, 114, 129, 121,
               117, 118, 122, 124, 118,
               120, 123, 124, 116, 122)

placebo1 <- c ( 137, 130, 115, 128, 133,
               129, 132, 130, 137, 124, 115,
               139, 133, 123 )

# Call t.test()
t.test(treatment1, placebo1, alternative="two.sided",
       conf.level=0.95, var.equal=TRUE)
```

Output:

```
Two Sample t-test

data:  treatment1 and placebo1
t = -3.9385, df = 27, p-value = 0.0005206
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -13.478654  -4.245156
sample estimates:
mean of x mean of y
 120.0667  128.9286
```

7.2.5 Contingency Methods (with R Code)

In the event that the Central Limit Theorem does not apply – due to a small sample size and non-normal data in at least one of the two groups of subjects – and we cannot use the parametric t -test, we have one non-parametric alternative: the Wilcoxon Rank-Sum test (also known as the Mann-Whitney U test; for many reasons, non-parametric statistical methods usually have several names). This test is similar to the Wilcoxon Signed-Rank test from Chapter 6, and in fact similarly begins with us ranking the data irrespective of group membership from 1 to n (1 being the smallest; n being the largest). Once ranked, we then sum the ranks *separately* for each

group. If there is no difference between the “centers” of the two groups (i.e. if there is no mean or distributional difference) then we would expect the sum of the ranks to be nearly equal. However, if there is a difference between the “centers” of the two groups, then we would expect the summed ranks to differ (if the center of the first group is lower or less in value than the center of the second group, then we would expect the sum of the first group ranks to be lower than the summed ranks for the second group, and vice versa). So like the t -test, we can start from the assumption that the two distributions have the same center – as measured by the median – and determine if the observed data back that claim. In the two-sample case, the Wilcoxon Rank-Sum test will have one degree of freedom, and along with that we report the resulting p -value (note that we do not report the test statistic and degrees of freedom).

For our blood pressure data, let’s momentarily assume that the CLT did not apply. Then we would have summarized the blood pressure in the Treatment group with a median of 121 and an inter-quartile range of (117, 123), and we summarize the Placebo group with a median of 130 and IQR of (123.8, 134). Using the `wilcox.test` function in R (Program 25) to obtain the results of the Wilcoxon Rank-Sum test, we see that our p -value is 0.0008529. This value is smaller than the significance level $\alpha = 0.05$, so we reject the null hypothesis of *equal population medians*, and declare that the evidence suggests the medians of the Treatment and Placebo groups are not the same (particularly, the Placebo group median is larger than the Treatment group median).

We should note that the output also had a warning message telling us that the p -value calculated is an approximate p -value as an exact p -value can not be calculated when the dataset has ties in it (i.e. two data values have the same numeric value). As was the case in Chapter 6, we should only be concerned with this particular message when the resulting p -value is near our significance level α , as in that case we are not sure whether the *actual* p -value is above or below the significance level.

7.2.6 Communicating the Results

The following is an example of the IMRaD write-up for the independent sample blood pressure example.

Introduction: Hypertension is associated with many adverse health conditions, including cardiovascular disease. Clinicians and health-service providers are interested in treatments that reduce high levels of blood pressure and improve cardiac and pulmonary health. In a randomized, placebo-controlled clinical trial, researchers aimed to determine whether a new treatment resulted in lower systolic blood pressure than a placebo.

Methods: Systolic blood pressure was measured (mm/Hg) from 15 subjects given a new treatment for reducing blood pressure, as well as 14 subjects

Program 25 Program to conduct two sample Wilcoxon Rank-Sum test on population locations for Systolic Blood Pressure Measurements.

Code:

```
# Input the data
treatment1 <- c( 121, 112, 114, 129, 121,
                117, 118, 122, 124, 118, 120,
                123, 124, 116, 122 )

placebo1 <- c ( 137, 130, 115, 128, 133,
                129, 132, 130, 137, 124, 115,
                139, 133, 123)

# Call t.test()
wilcox.test(treatment1, placebo1, mu=0, alternative="less")
```

Output:

```
Wilcoxon rank sum test with continuity correction

data:  treatment1 and placebo1
W = 36.5, p-value = 0.0008529
alternative hypothesis: true location shift is less than 0

Warning message:
In wilcox.test.default(treatment1, placebo1, mu = 0,
  alternative = "less") :
  cannot compute exact p-value with ties
```

administered a placebo. Subjects were randomly allocated into the two treatments. The systolic blood pressure measurements were summarized for each group with means, standard deviations and 95 % confidence intervals, while the observed difference between the two sample means, its standard error and 95 % confidence interval are also reported. Data were checked for normality with QQ plots, and the equality of variances between the two groups was assessed using the Brown-Forsythe test. We will test the null hypothesis that the Treatment and Placebo group means are equal (against a one-sided alternative hypothesis $H_A : \mu_T - \mu_P < 0$) with either the equal-variance or unequal-variance independent two-sample t -test. We will reject the null hypothesis in favor of the alternative if the observed p -value is less than the significance level of ($\alpha = 0.05$), otherwise we will fail to reject the null hypothesis. All calculations and analyses were conducted using the R statistical software.

Results: We assume that the data are representative and subjects are independently measured. The data are summarized in Table 7.4 below, where we see that the Treatment group mean is 8.9 mm/Hg lower than that observed for the Placebo group. Note that while the sample sizes are relatively small, the data values in both groups are reasonably normally distributed based on inspection of the QQ plots. Further, the p -value for the Brown-Forsythe test is 0.1550, so we can assume that the variances in both groups are equal. The results from the equal-variance t -test ($t_{27} = -3.94$, p -value = 0.0003) indicate that there is a significant difference between the Treatment and Placebo group means.

Table 7.4: Summary Data for Systolic Blood Pressure Measurements Based on Group Status.

Group	n	Mean	SD	95 % CI
Treatment	15	120.1	4.35	(117.7, 122.5)
Placebo	14	128.9	7.47	(124.6, 133.2)
		Mean	SE	95 %CI
Difference (T-P)	-	-8.9	2.25	(-13.5, -4.2)

Discussion: Based on our observed data, the mean SBP for the Treatment group was significantly lower than the mean SBP for the Placebo group. Provided the SBP values were comparable between the two groups at the beginning of the study, it would appear that this treatment is successful at reducing systolic blood pressure.

7.2.7 Process for Two-Sample t -Test

1. State research question in form of testable hypotheses.
2. Determine whether assumptions are met.
 - (a) Representative
 - (b) Independence
 - (c) Sample size: check distribution of data and sample size for each group.
 - (d) Determine whether variances are equal.

3. Summarize data.
 - (a) If sample size OR normality is adequate in BOTH groups: summarize groups with sample sizes, means, standard deviations and 95 % CIs, and summarize the difference with the observed difference, standard error and 95 % CI.
 - (b) If sample size AND normality are inadequate in EITHER group: summarize groups with sample sizes, medians and IQRs.
4. Perform Test.
 - (a) If variances are equal and CLT applies, then use equal-variance independent two-sample t -test.
 - (b) If variances are unequal and CLT applies, then use unequal-variance independent two-sample t -test (Welch test).
 - (c) If CLT does not apply, then use Wilcoxon rank-sum test.
5. Compare test statistic to critical value or calculate p -value.
6. Make decision (reject H_0 or fail to reject H_0).
7. Summarize with IMRaD write-up.

7.3 Paired Measurements

Now that we have covered the one-sample case (Chapter 6) and the two independent sample case (first part of this chapter), covering the two dependent-sample case will be relatively simple (and short). We basically conduct the test in this case as if it were a one-sample test, and report the results as if it were a two-sample test. Because of this, there is nothing new to report; we need only learn how to proceed.

7.3.1 Establishing Hypotheses: Independent Groups

The dependent sample case is so called because we have repeated or multiple measurements on each subject in the sample. One of the most common circumstances where this is the case is when we have a trial where subjects are measured at baseline and then again after the treatment has concluded. The paired measurements for each subject are said to be dependent (i.e. in addition to whatever effect a treatment may have, a patient's measurement at the end of the trial *depends* upon what the patient's measurement was at the beginning of the trial). The relationship between these measures may not be strong (or exist at all), but we need to proceed as if it was.

In cases like this we would be interested in (e.g.) the change in mean from baseline to end-of-trial. The simplest and most commonly accepted way of

going about this is to focus on the difference between the pre- and post-treatment values for *each subject*, rather than focusing on the overall means from both points. This difference then gives us *one value for each subject*, and allows us to use the methods outlined in Chapter 6. If we assume that the null case is where there is no change from pre- to post-treatment (or $d = 0$), then we can arrange our null and alternative hypotheses as listed in Table 7.5 below.

Table 7.5: Possible Sets of Hypotheses for a Difference in Two Dependent Population Means Based Upon Key Phrases in the Research Question.

Hypothesis	Key Phrases of μ_1 relative to μ_2			
	“less than”, “or equal to”, “at least”	“greater than”, “at least”	“greater than”, “less than”, “equal to”, “not equal to”	“equal to”, “not equal to”
Null	$H_0 : d = 0$	$H_0 : d = 0$	$H_0 : d = 0$	$H_0 : d = 0$
Alternative	$H_A : d < 0$	$H_A : d > 0$	$H_A : d > 0$	$H_A : d \neq 0$

Recalling the blood pressure example from earlier in this chapter, the researchers had also taken systolic blood pressure measurements of the subjects before the treatment was administered (the pre- and post-treatment SBP values are provided in Table 7.6). For the treatment group, we hypothesize that the active Treatment would reduce SBP, or the mean post-treatment would be lower than the mean pre-treatment. If we consider the difference as $d = \mu_{Pre} - \mu_{Post}$, we would assume that the mean change (d) from baseline to the end of the study would be positive, so our alternative hypothesis becomes $H_A : d > 0$.

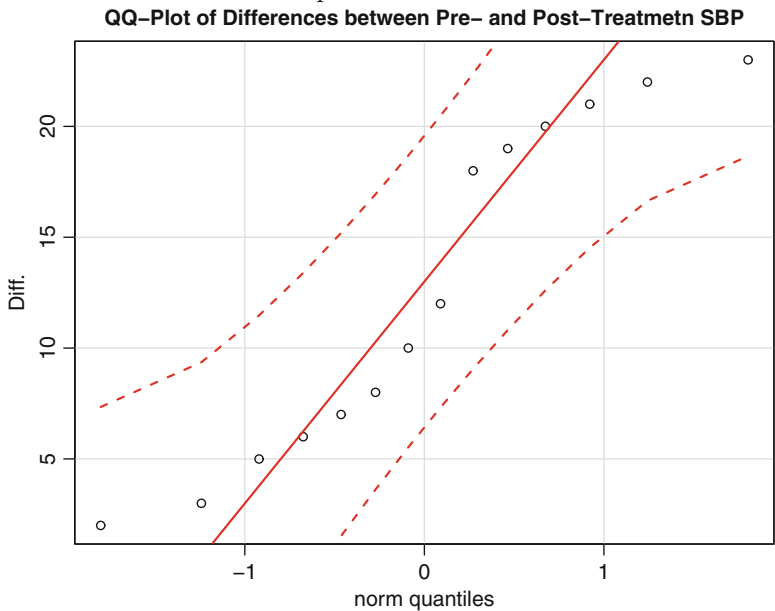
7.3.2 Assessing Assumptions (with R Code)

The statistical assumptions we require are the same as those for the one-sample case in Chapter 6, applied here to the observed differences. Importantly, we must check the observed differences for our assessment of the CLT: if there are greater than 30 observed differences or if the observed differences appear normally distributed, then we can use the parametric one-sample t -test (which in this unique instance we call a *paired t -test*); if neither of those conditions hold then we can use the non-parametric Wilcoxon signed rank test. For our blood pressure example, there are only 14 subjects who provide both the pre- and post-treatment SBP measurements (subject eight does not have a pre-treatment measurement). From those 14 subjects, we get the QQ plot shown in Figure 7.2 below, which was obtained using the `qqplot` function in R. To use this function on the differences, we created a new variable `diff=pre1-post1` for the data specified below in Program 26. Even though the 14 subjects are less than the preferred 30, the quantiles in the QQ plot indicate that the data appear nearly normally distributed. Thus, the CLT holds and we may use the parametric t -test.

Table 7.6: Pre- and Post-Treatment Systolic Blood Pressure Measurements from Subjects Enrolled in a Placebo-Controlled Clinical Trial (Measurements only provided for Treatment Group).

Patient	SBP (Pre-Trt)	SBP (Post-TRT)	Difference
1	123	121	2
2	134	129	5
3	128	118	10
4	125	118	7
5	130	124	6
6	132	112	20
7	139	121	18
8	.	122	.
9	139	120	19
10	119	116	3
11	135	114	21
12	140	117	23
13	146	124	22
14	135	123	12
15	130	122	8

Figure 7.2: QQ Plot for Difference in Pre- and Post-Treatment Systolic Blood Pressure in Clinical Trial Example.



Program 26 Program to conduct a paired two sample test on means for Systolic Blood Pressure Measurements for $H_0 : \mu_d / le0$ versus $H_a : \mu_d > 0$.

Code:

```
# Input the data
pre1 <- c( 123, 134, 128, 125, 130,
          132, 139, NA, 139, 119,
          135, 140, 146, 135, 130)

post1 <- c ( 121, 129, 118, 118, 124,
            112, 121, 122, 120, 116,
            114, 117, 124, 123, 122)

# Call t.test()
t.test(pre1, post1, mu=0, alternative="greater", paired=TRUE)
#Alternative (and Equivalent) Approach
Diff<-post1-pre1
#t.test(diff, mu=0, alternative="greater")
```

Output:

Paired t-test

```
data: pre1 and post1
t = 6.1624, df = 13, p-value = 1.709e-05
alternative hypothesis: true difference in means is greater
  than 0
95 percent confidence interval:
 8.958683      Inf
sample estimates:
mean of the differences
          12.57143
```

7.3.3 Summarizing Data (with R Code)

Since we technically have two samples, we must summarize the data as we did in the two independent samples case. If the CLT holds, the data for each repeated measure as well as the difference between those measurements must be summarized with a sample size, mean, standard deviation, and a 95% confidence interval. Note that we report the SD of the difference here (and not the SE) because we are summarizing the observed differences, not the mean difference. Of course, if the CLT does not hold, then we will summarize the two samples and the differences with sample sizes, medians and inter-quartile ranges. Please note that cases of repeated measures often involve missing

data for some subjects at one time point or another. Care must be taken in these cases, and the number of missing data must be reported, as differences cannot be calculated for subjects missing one or both measurements.

Since we've decided that the CLT holds (see Figure 7.2), we can summarize the data with means, standard deviations and confidence intervals, as shown in Table 7.7, which were obtained using the `summary` and `sd` functions for the means and standard deviations of each measure, and the `t.test` function for the confidence intervals, which as always are obtained by specifying the `alternative="two.sided"` command. Note that we have also indicated that there is a missing observation. There are a couple things worth noting. The first point is that the average difference is not the same as the difference between the two reported means (which is 12.4). This is not a problem to worry about, and is due to the missing observation in the baseline measurements. If we excluded that patient, the mean difference and the difference between the pre- and post-treatment means would match. Second, we subtracted the post-treatment measurements from the pre-treatment measures, and since the follow-up measures are smaller than at baseline, the reported mean difference is positive. You must be careful not to report this as an increase (since that's not what happened). If we had subtracted the baseline measures from the follow-up measures, then the differences and their average would be negative (though the same magnitude). Recall our research question, which states that we are expecting a decrease in SBP from baseline to follow-up. If we had taken the difference the other way (i.e. subtracted Pre-Treatment SBP from Post-Treatment SBP), then the difference would be -12.6 (and the CI would be -17.0 to -8.2 ; note the change in sign). This may be a better way of understanding the change in time, but we would have had to alter our alternative hypothesis to correspond.

Table 7.7: Summary Data for Systolic Blood Pressure Measurements Based on Group Status.

Group	n	Mean	SD	95% CI
Pre-Treatment	14	132.5	7.34	(128.3, 136.7)
Post-Treatment	15	120.1	4.35	(117.7, 122.5)
Difference (Post - Pre)	14	12.6	7.63	(8.2, 17.0)
		Pre	Post	
Missing Data	1		0	

7.3.4 Performing the Test and Decision Making (with R Code)

When the CLT holds, the one-sample t -test explained in Chapter 6 can be used. Here we use the mean difference (\bar{d}) as the basis of our test statistic, from which we subtract the hypothesized value d (which is usually zero), and

divide by the standard error of the mean difference (which is calculated from the standard deviation of the differences, s_d), and we obtain the following test statistic

$$t = \frac{\bar{d} - d}{s_d/\sqrt{n}}. \quad (7.9)$$

This test statistic will have a t -distribution with $n - 1$ degrees of freedom. From our blood pressure example, the test statistic comes out to 6.16 with 13 degrees of freedom.

We are free to choose between the critical value method and the p -value method, both of which are described in Chapter 6. However, since the t -distribution is based on the observed degrees of freedom (here based on $n - 1$, where n is the number of differences and not necessarily the sample size since there may be missing data), we have a preference to use the p -value method. Based on a test statistic of 6.16 and 13 degrees of freedom, and since our alternative hypothesis ($H_A : d > 0$) is right-tailed, our test yields a p -value < 0.0001 (see below). Since this value is less than the reported significance level of $\alpha = 0.05$, we reject the null hypothesis and conclude that the mean difference is significantly larger than zero. Note that we report the observed test statistic, degrees of freedom and p -value, and we round the test statistic to two decimal places. As always, we must translate the statistical results into statements that convey in words the meaning inherent in those results. In the dependent sample case, our analysis is centered around the observed differences, but we should interpret the results in terms of the two groups or time periods. For our example, the significant mean difference implies that the treatment reduced SBP levels from baseline to follow-up. In other words, the treatment seems to work at reducing SBP.

Program 26 provides the R code to conduct the paired t -test on the Systolic blood pressure measurements for pre and post treatments. Note that we can approach the problem in two ways: we can ask the `t.test` function to perform the paired t -test on the original variables `pre1` and `post1` by specifying the `paired=TRUE` command, or we may use the `t.test` function on the differences themselves (here using the generated variable `diff`). Both approaches lead to identical results, which are seen in Program 26. Recall that to get the 95% confidence interval on the difference, we must run the `t.test` function again, this time specifying the `alternative="two.sided"` command. Recall also that the reported difference and confidence interval depend upon the order in which we include the variables, or upon the order in which we take their difference.

7.3.5 Communicating the Results

The following is an example of the IMRaD write-up for the dependent-sample blood pressure example.

Introduction: Hypertension is associated with many adverse health conditions, including cardiovascular disease. Clinicians and health-service providers are interested in treatments that reduce high levels of blood pressure and improve cardiac and pulmonary health. In a randomized, placebo-controlled clinical trial, researchers aimed to determine whether a new treatment reduced systolic blood pressure from baseline levels.

Methods: Systolic blood pressure was measured (mm/Hg) from 15 subjects both before and after they were given a new treatment for reducing blood pressure. Subjects were randomly allocated into the active treatment. Both pre- and post-treatment systolic blood pressure measurements, as well as the observed difference from pre- to post-treatment, were summarized with means, standard deviations and 95 % confidence intervals. The observed differences were checked for normality with QQ plots, and we will test the null hypothesis that the pre- and post-treatment SBP levels are equal (against a one-sided alternative hypothesis $H_A : d > 0$) with a paired t -test. We will reject the null hypothesis in favor of the alternative if the observed p -value is less than the significance level of ($\alpha = 0.05$), otherwise we will fail to reject the null hypothesis. All calculations and analyses were conducted using the R statistical software.

Results: We assume that the data are representative and independently measured. The QQ plot showed no departure from normality for the observed differences. The data are summarized in Table 7.8 below, where we see that there was one missing pre-treatment measurement, reducing our effective sample size to 14. The mean reduction in SBP from pre- to post-treatment was 12.6, which is significantly different from zero ($t_{13} = 6.16$, p -value < 0.0001), leading us to reject the null hypothesis in favor of the alternative.

Table 7.8: Summary Data for Systolic Blood Pressure Measurements Based on Group Status.

Group	n	Mean	SD	95 % CI
Pre-Treatment	14	132.5	7.34	(128.3, 136.7)
Post-Treatment	15	120.1	4.35	(117.7, 122.5)
Difference (Post - Pre)	14	12.6	7.63	(8.2, 17.0)
		Pre	Post	
Missing Data	1	0		

Discussion: Since the mean reduction from pre- to post-treatment was significantly different from zero, we conclude that the evidence suggests that the active treatment succeeded in reducing systolic blood pressure from baseline measurements. Family practitioners interested in reducing the blood pressure of their patients may be interested in this treatment.

7.3.6 Process for Paired t -Test

1. State research question in form of testable hypotheses.
2. Determine whether assumptions are met.
 - (a) Representative
 - (b) Independence
 - (c) Sample size: check distribution of observed differences and sample size.
3. Summarize data.
 - (a) If sample size OR normality is adequate: summarize with sample sizes, means, standard deviations and 95 % CIs for both samples and difference.
 - (b) If sample size AND normality are inadequate: summarize with sample sizes, medians and IQRs.
4. Perform Test.
 - (a) If sample size is greater than 30 OR data are normally distributed, then use paired t -test on differences.
 - (b) If sample size is less than 30 AND data are not normally distributed, then use Wilcoxon signed-rank test on differences.
5. Compare test statistic to critical value or calculate p -value.
6. Make decision (reject H_0 or fail to reject H_0).
7. Summarize with IMRaD write-up.

7.3.7 Exercises

1. [Wrona \(1979\)](#) compared serum phenylalanine levels in patients with low and high exposures to that chemical. Using the data listed below, determine if there is a difference in the average phenylalanine levels between the two groups.

Group	Serum Phenylalanine (mg/dl)
Low Exposure	5.1, 9.5, 6.8, 5.5, 6.8, 9.2, 6.7, 8.9, 7.6, 4.2, 9.6, 6.2, 8.5, 4.8, 9.2, 5.7, 7.7, 9.6, 7.5, 8.9, 5.6
High Exposure	11.8, 13.5, 13.6, 11.4, 12.8, 11.5, 12.3, 13.2, 10.3, 11.2, 11.3, 15.3, 10.9, 15.3, 10.2, 13.0, 13.8, 11.0

2. [Burch et al. \(1989\)](#) performed a study on the effectiveness of shunts in helping infants with neonatal respiratory failure. Measurements of left ventricular dimension (LVD) are provided below. Determine if there was a significant increase in after treatment (extracorporeal membrane oxygenation, ECMO).

	Before	After		Before	After
Subject	ECMO	ECMO	Subject	ECMO	ECMO
1	1.6	1.6	9	1.6	1.4
2	2.0	2.0	10	1.7	1.5
3	1.2	1.2	11	1.0	1.3
4	1.6	1.6	12	1.5	1.8
5	1.6	1.5	13	1.5	1.8
6	1.7	1.6	14	1.4	1.8
7	1.6	1.5	15	1.5	2.0
8	1.6	1.7			

3. In a case-control study, [Walker et al. \(1987\)](#) compared diphtheria-tetanus-pertussis immunization rates between children who did and did not die from sudden infant death syndrome. The age (in days) at death are listed below for male and female infants. Determine if there is a difference in age at fatality between sexes.

Sex	Age at Death (days)
Females	160, 102, 117, 60, 87, 87, 56, 277, 60, 78, 134, 53
Males	167, 78, 133, 52, 80, 77, 115, 175, 84, 114, 81, 58, 59, 103, 134, 46, 175

4. Cotinine is often used as a measure of exposure to tobacco smoke. [Di Giusto and Eckhard \(1986\)](#) looked at the use of measuring saliva cotinine, as opposed to serum, as is typically done. Within the following seven subjects, compare the mean cotinine levels from 2 to 12 h, and from 2 to 24 h.

	Subject						
Time	1	2	3	4	5	6	7
2 Hours	80	83	80	87	45	37	126
12 Hours	73	58	67	93	33	18	147
24 Hours	24	27	49	59	0	11	43

5. [Winer-Muriam et al. \(2002\)](#) is interested in determining the theoretical radiation/energy absorbed by a pregnant woman during different gestational periods. Information was gathered from eight patients during the first trimester and nine patients during the second trimester.

The data collected is the theoretical dose/energy absorption to a set of points in each patient. Compare the mean dosage/energy absorptions between trimesters.

First	9.5	3.3	5.7	4.1
	13.6	4.6	20.2	20
Second	19.7	17.2	25.5	7.9
	24.7	76.7	23.5	30.4
	13.7			

6. [Austin et al. \(2006\)](#) is interested in the foraging behavior of gray seals by gender. In their study they follow gray seals and record the number of successful foraging trips per day. Over the course of a few days the average foraging success per day was recorded. Using the data below determine if there is a difference in foraging success across genders.

Males	2.9	1.6	2.8	2.7	0	2.7	1.1	0.9	1.3	2	2.2
Females	1	1.2	0.3	0.5	0	1.3	0.3	1.1			

Chapter 8

Analysis of Variance

In this chapter we cover the case where we wish to compare the means from *several groups*, where “several” is defined as more than two. The method we use in this case is called an Analysis of Variance (ANOVA), which is a well-known and widely-used statistical procedure. While this method is somewhat more complicated and involved than methods we’ve covered earlier, the process for using ANOVA modeling follows many of the same steps and rules used for the independent two-sample *t*-test covered in Chapter 7. In fact, when we have only two samples or groups, the use of ANOVA modeling produces the *same results* as the *t*-test for both the equal variance, non-equal variance and non-parametric cases. The main difference between the methods extends from the necessity to compare multiple means, meaning we will have additional steps and procedures in the multi-group case that we did not have in the two-sample case.

8.1 Establishing Hypotheses

Let us begin with a motivating example concerning student performance on an examination prior to graduation from medical school. The medical students participating in this study were randomized into one of four groups that implemented different approaches in providing supplemental material to aid students on their examination: students in the first group received supplemental material in printed form only; the second group received supplemental material through a hands-on practicum; students in the third group received neither the printed material nor participated in the practicum; the fourth group received both the printed material and participated in the practicum. For our purposes, we will ignore students in the fourth group who received both forms of supplemental learning. The point of this study was to determine what form – if any – of supplemental learning activities lead to the best

performance on the examination (note that this examination was for educational purposes only, and results were not reported to any external party).

In this case, the null hypothesis would be that the means from each group are equal. This is similar to the homogeneity of proportions hypothesis covered in Chapter 4. If the null hypothesis is that several group means are identical, then the alternative hypothesis is that *at least two* of those means differ from each other. We could get technical and think of many ways in which the null hypothesis could be contradicted, but the “at least one difference” statement is general enough to capture all other possibilities of *non-nullity*. In this case, since these hypothesis are so general, we usually state them in words rather than symbolically, and it is sometimes useful to use the context of the question for this phrasing. For our example, we could state that we wish to test the null hypothesis that the mean test scores were the same for each of the different approaches, against the alternative that at least one of the approaches lead to different test scores as compared to the other approaches.

8.2 Assessing Assumptions (with R Code)

Knowledge that subjects were selected at random into the study will help ensure that the sample is representative of the overall population and subjects are independent of one another, while knowledge that subjects were randomized into the different groups will ensure that the groups are representative and independent of one another. As for sample size, the inference we will eventually make is based around the sample means from the various groups (though ANOVA uses them differently than how they were used in the t -test), and as such, the Central Limit Theorem should apply for each sample. If the observed sample size in each group is greater than 30, then the CLT applies and we can assume that the sample size is large enough. If the sample size in some (or all) of the groups is less than 30, but the distributions of the data in those samples are more or less normal, then the CLT applies and we can again assume that the sample size is large enough. In either of these cases where the CLT applies we can use ANOVA modeling. However, if some (or all) of the groups have fewer than 30 subjects and those groups have non-normal distributions, then the CLT does not apply and we will have to use the non-parametric alternative to the ANOVA model.

As was the case with the independent two-sample t -test, we need to worry about the variances between the different groups in cases where the CLT applies. If the standard deviations are similar (based on visual inspection), or if an equality-of-variance test produces a large p -value (>0.05), then we can assume that the variances are equal and we use the equal-variance ANOVA model. If the standard deviations are visually different, or if an equality-of-variance test yields a small p -value (<0.05), then we assume that the variances are unequal and we use the unequal-variance ANOVA model. Of course, if

the CLT does not apply and we are forced to use non-parametric methods, we are not concerned about the equality of variances.

The raw test scores from the medical student examination example are found in Table 8.1 below. There are only 10 subjects in each group, so we must rely upon the distributions of the data in these samples to make our assessments. The QQ plots are provided in Figure 8.1. Here we see that two of the QQ plots (for the “none” and “practicum” groups, respectively) are more or less normal, though the QQ plot for the “paper” group shows a departure from normality. Thus, there may not be adequate sample size to conduct an ANOVA model (or at least there isn’t enough data to overcome the non-normality in the data), and we might want to consider using the non-parametric test (discussed later); however, for the sake of argument we will assume the CLT applies so we can use the ANOVA. The observed standard deviations are 3.65 for the None group, 3.67 for the Paper-only group, and 7.47 for the Practicum-only group, while the p -value from Bartlett’s test is 0.0427. It thus appears that we cannot assume that the variances are equal. Note that we used Bartlett’s test here because it is applicable for cases where we have more than two groups. We can use this option for the data set-up provided in Program 27 by using the `bartlett.test(Score1~Treat1)` command.

Table 8.1: Data for Medical Student Test Scores (Out of 100) Based on Supplemental Material Delivery Method.

No Paper or Practicum	Paper Only	Practicum Only
77	88	99
82	91	94
79	95	96
78	87	93
80	84	81
81	84	79
84	87	83
87	85	93
84	84	88
75	84	100

8.3 Summarizing Data (with R Code)

For cases in which the CLT applies, we report sample sizes, means, standard deviations and 95 % confidence intervals for each group in our study. For both cases where we have assumed equal or unequal variances between groups, we should summarize the data in the same manner, since we need to

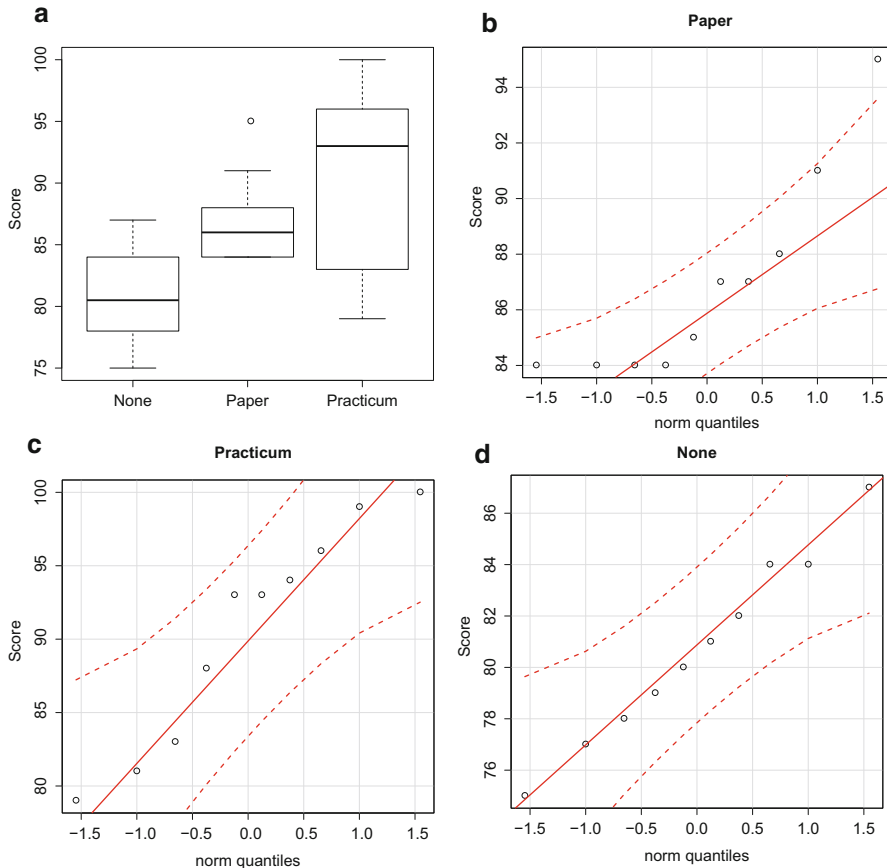


Figure 8.1: Side-by-Side Boxplot (a) and QQ Plots (b,c,d) for Medical Student Test Scores (Out of 100) Based on Supplemental Material Delivery Method.

describe the variability inherent in the data and not based on our assumptions. We should refrain from getting our summary statistics (especially the confidence intervals) from the output obtained from performing the ANOVA, which provides summaries based on our assumptions of equal- or unequal-variances. Naturally, if the CLT does not apply, then we summarize each group with its sample size, median and interquartile range.

The data from the medical school example are summarized by group according to each of the possible scenarios in Table 8.2. Recall from Chapter 7 that we can summarize these data in R using the `summary`, `sd` and `t.test` functions, where the last provides us with confidence intervals by specifying `alternative="two.sided"` and using either the `var.equal=TRUE` or `var.equal=FALSE` commands.

Program 27 Program to conduct an ANOVA test for Equal-Variance Case of Medical Student Test Scores Example.

Code:

```
# Input the data
None1 <- c( 77, 82, 79, 78, 80, 81, 84, 87, 84, 75 )
N0label1 <- rep("NoPaper", 10)

PaperOnly1 <- c ( 88, 91, 95, 87, 84, 84, 87, 85, 84, 84 )
Paper0label1 <- rep("PaperOnly", 10)

PracticumOnly1 <- c( 99, 94, 96, 93, 81, 79, 83, 93, 88, 100 )
Prac0label1 <- rep("PracticumOnly", 10)

# Organize the data appropriately.
Treat1 <- c(N0label1, Paper0label1, Prac0label1)
Score1 <- c(None1, PaperOnly1, PracticumOnly1)
data1 <- data.frame(Treatment = Treat1, Score = Score1)

# Call aov
Score.aov <- aov( Score ~ Treatment, data=data1)
summary(Score.aov)
```

Output:

```
              Df Sum Sq Mean Sq F value Pr(>F)
Treatment      2  500.5   250.23   9.088 0.00096 ***
Residuals     27  743.4    27.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 1
```

Table 8.2: Summaries for Medical Student Test Scores (Out of 100) Based on Supplemental Material Delivery Method for Three Cases: Variances Assumed Equal, Variances Assumed Unequal, Central Limit Theorem Doesn't Apply.

Group	<i>n</i>	CLT: Equal or Unequal Variances			No CLT	
		Mean	SD	95 % CI	Median	IQR
None	10	80.7	3.65	78.1, 83.3	80.5	77.8, 84.0
Paper	10	86.9	3.67	84.3, 89.5	86.0	84.0, 88.9
Practicum	10	90.6	7.47	85.3, 95.9	93.0	82.5, 96.8

8.4 Performing the Test and Decision Making (with R Code)

Recall that the test statistic in Chapter 7 for comparing two means involved the observed difference between the sample means. Indeed, an inspection of Table 8.2 will show that the Paper-group mean is 6.2 units larger than the None-group mean, while the Practicum-group mean is 9.9 units larger than the None-group mean and 3.7 units larger than the Paper-group mean. While these differences may or may not be large, we need a formal method for assessing the significance of these differences. In the multi-group case, we actually don't make the direct comparisons of those means, but instead construct a ratio of variance estimators (hence, the name *analysis of variance*). Starting with the equal-variance case, note that the variance for each group is calculated as $\Sigma_{i=1}^n (x_{ij} - \bar{x}_j)^2 / (n - 1)$, where the x_{ij} are the observed values ($i = 1, \dots, n$) and \bar{x}_j is the sample mean in the j th group (keep in mind that in this case n is the group sample size, not the total). If we momentarily ignore the degrees of freedom and add these variance terms together across the groups $\Sigma_{j=1}^a \Sigma_{i=1}^n (x_{ij} - \bar{x}_j)^2$, where a is the number of groups ($a = 3$ in our example), we get the so-called *sum-of-squares error (SSE)*. If we divide the SSE by $an - a$ (the total sample size minus the number of groups), we get the so-called *mean square error (MSE)*, which measures the variability within the groups.

We can find the variance of the group means by first calculating $\Sigma_{j=1}^a (\bar{x}_j - \bar{x})^2$, where \bar{x} is the overall mean irrespective of group membership. This measure is known as the *sum-of-squares model* or *sum-of-squares regression (SSR)*. If we divide this measure by the number of groups minus one (or $a - 1$), we get MSR, which measures the variability between groups.

Consider the null case where there are no mean differences between the groups. If the means between the groups are not different (and the variances are assumed equal), then we would expect the variability of the data within the particular groups (as measured by MSE) to be similar to the variability between the groups (as measured by MSR). In this case we would expect the ratio $F = MSR/MSE$ to be close to 1; this is statistically true because under H_0 , both MSE and MSR are unbiased estimators of the population variance σ^2 . When the null is not true and there are some mean differences, then the variability between groups is larger than the variability within groups, and the ratio $F = MSR/MSE$ will take values larger than 1. The larger the test statistic F , the more evidence is available for us to not believe the null hypothesis is true.

The test statistic F is a ratio of two variance-like measures, neither of which can take negative values. If we assume (under the null hypothesis) that (i) MSR is proportional to a chi-square distribution with $a - 1$ degrees of freedom, and (ii) MSE is proportional a chi-square distribution with $an - a$ degrees of freedom, then their ratio will have a F -distribution with $a - 1$

numerator degrees of freedom and $an - a$ denominator degrees of freedom (note that unlike every other distribution we've previously discussed, the F -distribution has two separate degrees of freedom). We will not use the critical value method for the F test statistic, and will solely rely upon p -values for inference. If the F -test yields a p -value smaller than the stated significance level, we reject the null hypothesis and claim that at least two of the group means are different; otherwise we fail to reject the null hypothesis and state that the evidence suggests the means are not different.

In the case of unequal variances, the test statistic is similar to that used for the equal variance case – in that it is a ratio of between-group variability to within-group variability – but it is complicated and we will avoid its characterization at all costs. The test statistic for the unequal variance case also follows a F -distribution with $a - 1$ numerator degrees of freedom like the equal-variance case, while the denominator degrees of freedom relies upon the degree of heterogeneity between the sample variances (and the equation is complicated).

The sum of squares for the equal-variance case of the medical student example are provided in Table 8.3 below. Note that the degrees of freedom for the group ($a - 1 = 3 - 1 = 2$) and error ($an - a = 3 \times 10 - 3 = 30 - 3 = 27$) are easily calculated. The ratio of mean squares ($MSR/MSE = 250.2/27.5$) takes the value 9.1, which is larger than what we would expect (1) if the group means were equal. The reported p -value (0.0010) is less than the nominal $\alpha = 0.05$, so we reject the null hypothesis that mean student scores are similar between the test preparation methods. We report the results – including both degrees of freedom, the value of the test statistic, and the p -value – as $F_{2,27} = 9.1$, p -value = 0.0010. If we had assumed unequal variances, we would get similar results ($F_{2,17.0} = 10.5$, p -value = 0.0011). Note that in this case the denominator degrees of freedom (17.0) is less than that for the equal-variance case (27) due to some heterogeneity in the sample variances.

Table 8.3: Sum of Squares for Equal-Variance Case of Medical Student Test Scores Example.

Source	DF	Sum of Squares	Mean Square	F Ratio	p -value
Group	2	500.4667	250.2	9.1	0.0010
Error	27	743.4000	27.5		
Total	29	1243.8667			

In order to use R to provide the ANOVA results, we first have to organize our data correctly. In Program 27 we have entered the group specific test scores into the variables `None1`, `PaperOnly1` and `PracticumOnly1`. We then create a “group” indicator, or “factor”, which we will later use to assign each test score to the appropriate group (paper only, practicum only, or none).

To do this we will use the `rep` function, which creates a vector where the same value is *repeated* a specified number of times (in this case, 10). We suggest selecting the first argument – which will become the group label – wisely, as that will be how we identify and distinguish between the particular groups. We must next organize the data properly, which is done in the `Treat1` and `Score1` commands, where we stack the group labels and test scores into single variables, which we then place into a single data set (or `data.frame`) named `data1`. Note that we have here renamed the group and outcome variables using the `Treatment=Treat1` and `Score=Score1` commands.

Once we have finished organizing the data, we can then perform the analysis of variance using the `aov` function. In Program 27 we label this function `Score.aov`, as we need to use the `summary` function to obtain the results. The `aov` function requires the first item to be a formula with the outcome or dependent variable (`Score` in our case), followed by the group or factor variable (`Treatment` in our case), while we separate the two with a tilde “~”. The next item after the comma is the name of the data frame that contains the variables in the formula (we named this dataframe is `data1`). Again, the `summary` function produces the ANOVA table we obtained earlier. The only item that differs is that R does not produce the “Total” row at the bottom of the table, which we can obtain through simple addition. The R output also includes “significance codes” which can help us determine which results are statistically significant. If we simply look at the p -value given in the `Pr(>F)` column we can determine the same information as that provided by the significance codes.

In order to obtain the unequal variance ANOVA results, we can use the `oneway.test()` function. Here we again specify the relationship between the outcome and group factor with the `Score~Treatment` option, and then specify the data set by stating `data=data1`. You may verify on your own that you obtain the same inferential material (test statistic, numerator and denominator degrees of freedom, and p -value) as was listed above.

8.4.1 Post-hoc Multiple Comparisons (with R Code)

If we have rejected either the equal-variance F -test, the unequal-variance F -test, or the non-parametric test (introduced later), then we must move on to the second part of the ANOVA model: the multiple comparisons. Though the F -test has told us that at least one difference between the various group means exists, it has not identified which groups are different (unless there are only two groups). Thus, we now need to inspect the various pairings of group means to discover – hopefully, sometimes it doesn’t happen – the significant differences. We could naively follow the methods for the independent two-sample case covered in Chapter 7, where each group mean is compared to every other group mean independently. This is logistically prohibitive, especially if the number of groups is large. The observed differences, standard errors, 95% confidence intervals and p -values from the equal-variance t -tests are provided in Table 8.4.

Table 8.4: Mean Comparisons using (Unadjusted) Independent t -Tests for Medical Student Test Scores Example.

Comparison	Difference	SE	95 %	p -value
Practicum - None	9.9	2.35	5.1, 14.7	0.0002
Paper - None	6.2	2.35	1.4, 11.0	0.0135
Practicum - Paper	3.7	2.35	-1.1, 8.5	0.1265

Program 28 shows the R code to conduct the unadjusted pairwise t -test based multiple comparison procedure for the medical student test scores. We use the data as formatted in Program 27, and now apply the `pairwise.t.test` function. The first item we include in this function is the dataset (in this case `Score1`), while the second item we include is the indicator of the group labels (in this case `Treat1`). Notice that we did not use the `data1` dataset as this function requires exactly one column of data to be the inputs. The last item we specify is the `p.adjust.method` option, which specifies what type of adjustment should be applied to the p -values. Since we do not presently want the p -values adjusted, we specify "none". Later in this Chapter we will discuss methods that will allow us to adjust the p -values.

The output of Program 28 is a table of p -values. The entries on the left-hand side of this table correspond to p -values for the comparisons between the

Program 28 Program to conduct unadjusted pairwise t -test based posthoc analysis of Equal-Variance Case of Medical Student Test Scores Example.

Code:

```
# Use data1 defined above.

pairwise.t.test(Score1, Treat1, p.adjust.method = "none")
```

Output:

Pairwise comparisons using t tests with pooled SD

```
data:  Score1 and Treat1

           None    PaperOnly
PaperOnly 0.01354 -
PracticumOnly 0.00025 0.12650
```

```
P value adjustment method: none
```

None group versus both the PaperOnly group (top value) and PracticumOnly group (bottom value). Similarly the entry in the bottom right of the table is the p -value for the difference in means between the PaperOnly group and PracticumOnly groups. Now we hope you see what it is good practice to give useful names to the groups. The p -values given in the output match those found in Table 8.4.

Based on the results in Table 8.4 from the – naive, mind you! – independent t -tests, we would – naively!! – declare that both the Practicum and Paper groups had larger means than the no-supplementary material group, and that the Practicum and Paper group means were not significantly different. However, another problem exists that we have not yet encountered. The significance level we have used in each chapter was applied to *only one test*; if there truly wasn't a difference, our error rate is fixed around the one decision we have to make. However, if we are making multiple choices or performing several tests (as we did in Table 8.4), then the probability that we make a mistake of claiming a difference when there *truly isn't a difference* is dependent upon the number of decisions or tests we have to consider. This can be seen by first noting that we would like to keep the overall error rate constant at 0.05, while at the same time acknowledging that the number of tests we are considering is also fixed due to the experimental design (i.e. number of groups). For instance, if we have three groups (A, B and C), then there are three different comparisons that can be made (see Table 8.5 below), whereas if there are four groups, then there are six non-redundant comparisons to make. Note the large number of comparisons (10) required for 5 groups (the number is 15 for 6 groups, 21 for 7 groups, and continues to increase from there). This is why most researchers try to keep the number of groups under consideration to a minimum (four or fewer is optimal).

Recall from Chapter 1 our discussion of Type I error. If the null hypothesis is true and there is no difference, we would expect to make a mistake – i.e. reject the null hypothesis – 5% of the time if our stated significance level is 0.05. If we attach a 0.05 significance level to each test or comparison we make, it remains true that the Type I error rates for each of those tests will be close to 5%. However, since we are performing multiple tests and making

Table 8.5: Number and Types of Multiple Comparisons used in ANOVA modeling based upon Number of Groups (A, B, C, D, etc.).

Number of Groups	Number of Groups	Comparisons
2	1	A–B
3	3	A–B, A–C, B–C
4	6	A–B, A–C, A–D, B–C, B–D, C–D
5	10	A–B, A–C, A–D, A–E, B–C, B–D, B–E, C–D, C–E, D–E

multiple decisions, and since some of those tests involve the same groups used for other tests, the *overall Type I error rate* of mistakenly rejecting a true null difference will be larger than 5%. We thus need to take measures to minimize our error rate.

Fortunately, there exist methods that will lower the overall Type I error rate for us. The first is the so-called *Bonferroni Correction*, where we divide the significance level (α) by the number of comparisons (k) and use the resulting value (α/k) as the significance level for each comparison. Thus, in our example, the Bonferroni-adjusted significance level is $0.05/3 = 0.0167$, meaning we reject the null hypothesis of no difference for any specific comparison (as shown in Table 8.4) if the p -value is less than 0.0167. Since the first two p -values (0.0002 and 0.0135) are less than that value, we reject those comparisons, but not the third (p -value = 0.1265). If we had four comparisons to make, the Bonferroni-adjusted significance level would be $0.05/4 = 0.0125$; if five comparisons then $0.05/5 = 0.0100$, and so on. Note that this method can be used without changing the code or output provided by R; in this case we must note that any confidence intervals on the difference we report are *unadjusted*.

While the Bonferroni method of adjustment is simple to implement, it is a conservative approach that – in making it more difficult to declare a difference significant – will increase the probability of making a Type II error (failing to reject a null hypothesis when there really is a difference). Thus, another method is preferred that more adequately manages the trade-off between Type I and Type II errors (see Chapter 9 for more details). One such method of adjustment is called the Tukey-Kramer method (often referred to as the Tukey HSD method, or simply as the Tukey test), and is already incorporated into the R software (by specifying the `TukeyHSD` function). This method is based upon the *studentized range* rather than the t -distribution, and is in general complicated to conduct hand. The Tukey-adjusted results provided in Table 8.6. Note that compared to the CIs from the unadjusted case (Table 8.4), the Tukey-adjusted CIs are wider and the p -values are larger, reflecting the decrease in significance used for each comparison. Note also that the observed differences and standard errors are the same in both cases, as only the significance level and the probabilistic measure used in the latter case changes. The Tukey-adjusted p -values are each compared to the overall significance level of 0.05, so we declare the first two differences significantly different from zero, while we fail to declare the third comparison different from zero.

Program 29 shows how to conduct the Tukey Honestly Significant Difference multiple comparison procedure in R for the medical student data as entered and formatted in Program 27. Here we use the `TukeyHSD` function, which requires the output from the `aov` function. Notice that the output from the `TukeyHSD` function is different from that of the `pairwise.t.test`, in that here the `TukeyHSD` function gives a table that lists the pairwise comparisons, the difference between group means (`diff`), the lower (`lwr`) and upper (`upr`)

Table 8.6: Mean Comparisons using Tukey-Adjusted Independent t -Tests for Medical Student Test Scores Example.

Comparison	Difference	SE	Adjusted CI	p -value
Practicum - None	9.9	2.35	4.1, 15.7	0.0007
Paper - None	6.2	2.35	0.4, 12.0	0.0350
Practicum - Paper	3.7	2.35	-2.1, 9.5	0.2727

Program 29 Program to conduct Tukey's Honestly Significant Difference (HSD) posthoc analysis of Equal-Variance Case of Medical Student Test Scores Example.

Code:

```
# Use data1 defined above.

Score.aov <- aov( Score ~ Treatment, data=data1)
TukeyHSD(Score.aov)
```

Output:

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = Score ~ Treatment, data = data1)
```

```
$Treatment
              diff          lwr          upr          p adj
PaperOnly-None  6.2  0.3817271 12.018273 0.0350147
PracticumOnly-None  9.9  4.0817271 15.718273 0.0007025
PracticumOnly-PaperOnly  3.7 -2.1182729  9.518273 0.2727302
```

bounds for the adjusted confidence intervals on the differences, as well as the adjusted p -values (`p adj`). Most importantly, note that the output matches the results presented in Table 8.6.

The previous two adjustments were made under the assumption of equal variances, but what if we were to assume unequal variances? Basically, we must ask R to perform a series of Bonferroni-adjusted, two-sample unequal-variance t -tests for each comparison that we wish to make (yet another reason to keep the number of groups small). After we have summarized the data in each group and performed our overall test of homogeneity amongst the means, we proceed with the following two-step process. First, we use the `pairwise.t.test` function with the `pool.sd=FALSE` and `p.adj="bonf"`

options to get the bonferonni-adjusted p -values without assuming equal variances. Then for each pair of groups, perform a two-sample, unequal-variance t -test to get the observed difference, SE, adjusted CI and p -value using the `t.test` function with the `alternative="two.sided"` option specified and the `conf.level` function set so that it equals *one-minus* the bonferroni-adjusted significance level. Note that here we most certainly *do not* want to specify `var.equal=TRUE`. The results for this approach are found in Table 8.7. Note that while the differences are the same as those found in Table 8.6, the SEs have changed (and are now larger for the two Practicum comparisons), and the adjusted CIs are now wider and the p -values are not larger (reflecting the greater variability due to slightly unequal variances). However, for these data the Bonferroni adjustment in the unequal-variance case does not change our interpretation of the results. We can calculate the SEs *manually* as was done in Chapter 7.

Table 8.7: Mean Comparisons using Bonferroni-Adjusted Independent, Unequal-Variance t -Tests for Medical Student Test Scores Example.

Comparison	Difference	SE	Adjusted CI	p -value
Practicum - None	9.9	2.63	2.7, 17.1	0.0023
Paper - None	6.2	1.64	1.9, 10.5	0.0013
Practicum - Paper	3.7	2.63	-3.5, 10.9	0.1830

With regards to interpreting the results from an ANOVA model, we will run into one of several cases. The first is that the F -test from the overall ANOVA model yields a large p -value, which indicates that there are no significant mean differences. In this case, we do not make any multiple comparisons, since there are no significant differences to find, and simply report the results from the F -test.

If the F -test is significant, then we use the Tukey method to make multiple comparisons (Tukey is preferred to Bonferroni for one-factor ANOVA). We report the observed differences from those comparisons (as well as the SEs, adjusted CIs, and p -values), and comment upon the direction of those differences. From our medical school example, the significant differences of both the Practicum group scores and the Paper group scores over the None group scores imply that using some form of supplementary material lead to increased test scores over those students who didn't use any supplementary material. The non-significant difference between the Practicum and Paper groups imply that the type of supplementary material made no difference, or that the difference of increased scores in the Practicum over the Paper group was not significant.

Of course, the situation may arise where there is a significant F -test but none of the multiple comparisons lead to a significant difference. This doesn't happen often, but when it does happen we are obligated to report what the tests tell us. In this case, as well as in the case of an insignificant F -test, do not force explanations that don't exist by interpreting the differences; they were not found significant, so they must be left alone.

8.5 Contingency Methods (with R Code)

If the assumptions comprising the CLT are not met, then we cannot use the parametric ANOVA model (either equal- or unequal-variance), and must instead use non-parametric methods (here we use the Kruskal-Wallis test). The output for the Medical School example yields test statistic of 11.6945 with 2 degrees of freedom and a small p -value (0.0029). We would report this information as $\chi_2^2 = 11.7$, p -value = 0.0029, and this p -value would lead us to reject the null hypothesis of equal treatment group medians. Formal comparisons of the medians is rare, and anecdotal comparisons of those medians – without attaching a level of significance or confidence to such statements – will usually suffice, though they can be done using a similar approach as reported for the unequal-variance case above. Basically, we must ask R to perform a series of Bonferroni-adjusted, two-sample Wilcoxon tests for each comparison that we wish to make (yet another reason to keep the number of groups small). After we have summarized the data in each group, we set the significance level to account for the Bonferroni adjustment perform a series of two-sample Wilcoxon tests for each pair of groups to get the corresponding p -values.

Program 30 shows the R code to create the Kruskal-Wallis test for the medical student test score example. We will again use the dataset as structured in Program 27. Here we see that the `kruskal.test` function is similar to the `aov` function, where the first item links the outcome with the group variable (`Score~Treatment`) and the second item specifies the dataset (`data1`). We do not need to define this function to a variable as the summary is directly provided. Note that the provided output matches what we reported above.

Program 30 Program to conduct an Kruskal-Wallis test for the Medical Student Test Scores Example.

Code:

```
# Call kruskal.test

kruskal.test( Score ~ Treatment, data=data1)
```

Output:

```
Kruskal-Wallis rank sum test
```

```
data:  Score by Treatment
Kruskal-Wallis chi-squared = 11.6945, df = 2, p-value = 0.002888
```

8.6 Communicating the Results

The following is an example of the IMRaD write-up for the medical school example. While we have assumed that the CLT applies, the actual evidence suggests we should not.

Introduction: Medical school administrators and educators are often in search of methods that will enhance and improve the educational experience they offer their students. One such method has been to provide supplementary material before tests, such as through providing pre-fabricated lecture notes, or offering a hands-on student practicum. Administrators at a particular medical school aimed to determine the usefulness of such methods of providing supplementary material in use for test preparation to determine whether providing supplementary material lead to increased test scores.

Methods: Volunteer students were randomized into one of three groups before taking a non-binding examination. A Practicum group consisted of students who reviewed test information through a hands-on learning experience; a Paper group consisted of students who were offered study guides in printed form; and a control group consisted of students who neither partook in the practicum nor received the printed material. Ten students were randomized into each group, for a total of 30 students. The distributions of test scores for each group were analyzed using QQ plots, and the variances from each group are compared using the Bartlett's test. The test scores for each group are summarized with sample sizes, means, standard deviations and 95% confidence intervals. A one-factor, unequal-variance analysis of variance (ANOVA) is used to test the null hypothesis that the three group means are equal (against the alternative hypothesis that at least two of the means differ) at the 0.05 significance level. If the resulting p -value from the ANOVA model is less than 0.05, then we will reject the null hypothesis, and multiple comparisons between the group-level means will be made using the Bonferroni method, and observed differences, standard errors, adjusted confidence intervals and adjusted p -values will be reported; we will otherwise fail to reject the null hypothesis. All data summaries and analyses are conducted using R statistical software.

Results: The test scores for each group are summarized in Table 8.8, and the data are assumed independent and representative. The QQ plots (not reported) showed the data were reasonably normally distributed, and Bartlett's test yielded a small p -value (0.0427), justifying the unequal variance ANOVA model. The results from the overall ANOVA model ($F_{2,17.0} = 10.5$, p -value = 0.0011) indicate that group membership has an effect on mean tests scores. Using the Bonferroni adjustment method, both the Practicum group (9.9) and Paper group (6.2) had increased mean test scores over the None group, though the Practicum and Paper group mean test scores were not significantly different.

Discussion: Providing supplementary material to medical students lead to significantly improved test scores over those students who did not receive

Table 8.8: Summaries for Medical Student Test Scores (Out of 100) Based on Supplemental Material Delivery Method.

Group	n	Mean	SD	95% CI
None	10	80.7	3.65	78.1, 83.3
Paper	10	86.9	3.67	84.3, 89.5
Practicum	10	90.6	7.47	85.3, 95.9

Comparison	Difference	SE	CI	<i>p</i> -value
Practicum - None	9.9	2.63	2.7, 17.1	0.0023
Paper - None	6.2	1.64	1.9, 10.5	0.0013
Practicum - Paper	3.7	2.63	-3.5, 10.9	0.1830

either form of supplemental material. However, the form of supplementary material (either practicum- or paper-based) did not matter much in the test score improvement. Medical school administrators may wish to provide supplementary material for students prior to testing.

8.7 Process

1. State research question in form of testable hypotheses.
2. Determine whether assumptions are met.
 - (a) Representative
 - (b) Independence
 - (c) Sample size: check distribution of data and sample size for each group.
 - (d) Equal Variances.
3. Summarize data.
 - (a) If sample size/normality is adequate for each group: summarize each group with sample size, mean, standard deviation and 95% CI.
 - (b) If sample size/normality is inadequate for AT LEAST one group: summarize each group with sample size, median and IQR.
4. Perform Test.
 - (a) If variances are equal and CLT applies for each group: use equal-variance ANOVA.
 - (b) If variances are not equal and CLT applies for each group: use unequal-variance ANOVA.

- (c) If sample size is less than 30 AND data are not normally distributed for any of the groups: use Kruskal-Wallis test.
5. Check main effect.
- (a) If main effect is significant, perform multiple comparisons.
- (b) If main effect is not significant, stop the test.
6. Summarize with IMRaD write-up.

8.8 Exercises

1. Wang et al. (2002) studied the effect of age on the mechanical integrity of the collagen network in bone tissue. In their study, the femurs from 30 human cadavers were obtained from young (19–49 years), middle aged (50–69 years) and elderly (>70 years) donors. Determine if there are differences in the force required to break the bones between the groups.

Young	Middle Age	Elderly
193.6	125.4	59.0
137.5	126.5	87.2
122.0	115.9	84.4
145.4	98.8	78.1
117.0	94.3	51.9
105.4	99.9	57.1
99.9	83.3	54.7
74.0	72.8	78.6
74.4	83.5	53.7
112.8		96.0

2. Low platelet counts (a condition known as thrombocytopenia) are often observed in infants with necrotizing enterocolitis (otherwise known as gangrene). Ragazzi et al. (2003) – as part of a study aimed to determine of neutrophil count was correlated with the extent of disease – provided log (base 10) platelet counts for infantile subjects based on their gangrenous categorization: no presence of gangrene (0); focal gangrene (1); multifocal gangrene (2); or panintestinal gangrene (3). Determine if there were any differences in log platelet counts based on gangrene categorization.

0			1		2			3	
1.97	2.33	2.23	1.38	2.18	1.87	2.16	1.36	1.77	1.75
0.85	2.6	2.51	1.86	2.53	1.9	2.17	2.48	1.68	1.86
1.79	1.88	2.38	2.26	1.98	2.43	2.12	1.4	1.46	1.26
2.3	2.33	2.31	1.99	1.93	1.32	2.27	1.75	1.53	2.36
1.71	2.48	2.08	1.32	2.42	2.06	2.37	2.67	1.36	
2.66	2.15	2.49	2.11	0.79	1.04	1.75	2.37	1.65	
2.49	1.41	2.21	2.54	1.38	1.99	2.57	1.46	2.12	
2.37	2.03	2.45	2.06		1.52	1.51	1.91	1.73	
1.81	2.59	1.96	2.41		1.99	1.08		1.91	
2.51	2.23	2.29	2.23		2.52	2.36		1.57	
2.38	1.61	2.54	2		1.93	1.58		2.27	
2.58	1.86	2.23	2.74		2.29	1.83		1	
2.58	2.33	2.78	2		1.75	2.55		1.81	
2.84	2.34	2.36	2.51		2.16	1.8		2.27	
2.55	1.38	1.89	2.08		1.81	2.44		2.43	
1.9	2.52	2.26	2.45		2.46	2.81		1.74	
2.28	2.35	1.79	2.6		1.66	2.17		1.6	
2.33	2.63	1.87	1.83		2.51	2.72		2.08	
1.77	2.03	2.51	2.47		1.76	2.44		2.34	
1.83	1.08	2.29	1.92		1.72	1.98		1.89	
1.67	2.4	2.38	2.51		2.57	1.57		1.75	
2.67	1.77		1.79		2.3	2.05		1.69	
1.8	2.48		2.17		0.7	2.3		2.49	

3. The *Health Effects Institute Research Report (Number 25)* by [Allred et al. \(1989\)](#) focused on the effect of carbon monoxide exposure in males with coronary artery disease. As a part of this study, forced expiratory volume (FEV) as measured in patients from three sites (seen below). Determine if there were any differences in FEV between the three sites.

Johns Hopkins		Rancho Los Amigos		St. Louis	
3.23	1.98	3.22	2.87	2.79	2.81
3.47	2.57	2.88	2.61	3.22	3.17
1.86	2.08	1.71	3.39	2.25	2.23
2.47	2.47	2.89	3.17	2.98	2.19
3.01	2.47	3.77		2.47	4.06
1.69	2.74	3.29		2.77	1.98
2.10	2.88	3.39		2.95	2.81
2.81	2.63	3.86		3.56	2.85
3.28	2.53	2.64		2.88	2.43
3.36		2.71		2.63	3.20
2.61		2.71		3.38	3.53
2.91		3.41		3.07	

4. Jay et al. (2009) are interested in determining how doctors of different specialties impact on their obese patients losing weight. They considered doctors from the specialties of Internal Medicine, Pediatrics and Psychiatry. They recorded the percentage of patients who lost weight by practice. Below is a sample representative of the data they collected.

Determine if differences exist across specialties and if they do exist determine which specialties differ.

IM	8.9	15.7	11.1	12.4	18.6
	12.0	11.9	8.8	18.2	11.7
	12.6	25.1	21.5	11.1	9.1
Ped	9.6	3.3	7.4	7.1	3.7
	6.7	1.7	4.0	5.1	9.7
	10.5	5.8	6.5	8.7	6.9
	9.0	7.2			
Psy	18.8	23.3	12.9	21.6	11.9
	17.4	24.8	16.3	18.9	13.0
	12.1	13.8	9.4	19.0	22.9
	17.6	11.6	17.7		

5. A researcher is interested in the effects of Benedryl on reaction time. It has been hypothesized that Benedryl slows down people's reaction time. To study this she randomly assigns people to either a control or a specified dosage. In this case "1" is the recommended dosage on the package. She gives administers the Benedryl and waits 30 min and records the Before and After reaction time for a specific reaction test. She can only administer the reaction test once before administering the drug and once after administering the drug as subjects *learn* the test. Determine if Benedryl slows reaction time and determine which dosages produce a statistically significant reduction in reaction time using the data below.

Control	Treatment				
	0.5	1	1.5	2	2.5
-0.46	-2.03	2.04	2.31	2.52	7.50
-1.09	-0.36	4.56	1.57	1.63	5.62
4.45	-1.97	0.33	5.82	1.60	3.05
1.67	-0.84	2.40	3.09	7.74	5.34
0.45	2.71	2.64	2.61	6.79	6.43
-3.41	-2.53	3.34	1.01	9.01	3.87
1.16	1.73	0.61	3.66	3.69	5.14

Chapter 9

Power

We have previously discussed the concepts of Type I and Type II errors, but we have not much considered their relevance to study design and statistical analysis. In practice, controlling either the type II error rate or the power – holding all other components fixed – will determine the sample size needed to find a significant result. Thus, the phrases “sample size determination” and “power analysis” are used interchangeably. Power determination is an active research area in the statistical sciences, mostly due to the complicated problems now at the vanguard of data analytics. Many of the more complicated problems require advanced statistical computing or simulations for sample size determination, and so these are naturally beyond the scope of this course. Fortunately, many of the commonly used statistical techniques – such as those covered in Chapters 2, 3, 6–8 – either have closed-form sample size equations, or the sample size determination is available through statistical software (such as R). We will focus mostly on the use of power in designing studies (i.e. answering the awful question “how much data do I need?” that statisticians typically dread) for both categorical and continuous outcomes in one- and two-group cases, as well as the multi-group case for continuous data. Following this we will briefly touch on the *post-hoc power analysis*, which is the statistical equivalent to conducting an autopsy (i.e. finding out why your study died). Of course, we will first need to more concretely define statistical power and its components.

9.1 Making Mistakes with Statistical Tests

As mentioned earlier, statisticians can make only two mistakes (all others must have been made by someone else): we can falsely reject a true null hypothesis (type I error), or we can falsely fail to reject the null hypothesis when the alternative hypothesis is true (type II error). Which of these mistakes we make is dependent upon the “true state of the world”, which in our parlance

means that either the null hypothesis (H_0) is true (and the alternative (H_A) false), or the alternative hypothesis is true (and the null hypothesis is false). Recall that since we create H_0 and H_A to be mutually exclusive, only one of them can be true at one time. Though we cannot know for certain which hypothesis is “true”, we do not need to “know” the truth for testing purposes. Based on our statistical evidence, we will make one of two decisions: reject H_0 in favor of the alternative, or fail to reject H_0 . Thus – given the truth – we will either make the correct decision or we will make a mistake (or an error). The consequences of our decisions are shown in Table 9.1. If the null hypothesis is true (no difference) and we reject the null and claim that there is a difference, then we have made the wrong decision (a type I error). Note that the probability of making this mistake – if the null hypothesis is true – is α , or the stated significance level. We generally aim to make this mistake no more than 5% of the time ($\alpha = 0.05$) when the null hypothesis is true. Conversely, if the alternative hypothesis is true – there is a difference – and we fail to reject the null hypothesis (we didn’t find the difference), then we have again made the wrong decision (a type II error). Note that the probability of this mistake (if the alternative hypothesis is true) is β , which we generally take to be 10% ($\beta = 0.1$) or 20% ($\beta = 0.2$). Finally, note that the probabilities of the two options in each column sum to 1, reflecting the fact that only one hypothesis can be true at one time.

Table 9.1: Types of Errors and Their Probabilities in Hypothesis Testing Based on True Hypothesis.

Decision	True State	
	H_0 is True	H_A is True
Reject H_0	Type I Error α	Correct $1 - \beta$
Fail to Reject H_0	Correct $1 - \alpha$	Type II Error β
	$\alpha + 1 - \alpha = 1$	$\beta + 1 - \beta = 1$

It is the converse of the type II error rate (β) that we refer to as power, which means that power is the probability ($1 - \beta$) that we find a difference when that difference actually exists. The higher the power, the more likely we are to reject the null hypothesis when we should, while the lower the power, the less likely we are to reject the null hypothesis when we should. It would be nice if we could simply pick the power that we want and state that we are 80 or 90% likely to find a significant difference, but unfortunately power does not work like that. Power is actually comprised of four different pieces: the stated significance level, the expected variability in the response, the desired effect size, and sample size. These four components are inter-related and each affects power in different ways.

9.2 Determinants of Sample Size

Significance level: As mentioned elsewhere, the type I error rate α is typically set at a standard value (such as 0.05), so there is not much we can do to affect power through α . However, it is important to note that power ($1 - \beta$) and significance level (α) are proportionally related, which means that the type I and type II error rates are inversely related. This means that if we want to make it easier to declare differences significant by rejecting the null hypothesis (increasing the power or decreasing our Type II error rate), we can do so by increasing the significance level – and thereby increasing our type I error rate – if we hold everything else constant. Likewise, if we want to decrease the likelihood of incorrectly rejecting the null hypothesis (decreasing the type I error rate and α), we make it more difficult to declare differences significant (and thus increase our type II error rate and decrease our power). Rather than mess with this relationship, we often fix α at 0.05 and change everything else (for that matter, we also generally fix $1 - \beta = 0.8$ or 0.9 and only change the remaining power components).

Variability: The variability in the response plays a very straightforward role in power. Recall that every test statistic we considered in Chapters 2, 3, 6 and 7 was of the same general form

$$\text{Test Statistic} = \frac{(\text{observed statistic}) - (\text{hypothesized value})}{\text{standard error of observed statistic}}. \quad (9.1)$$

Since the standard error of the observed statistic (either the sample proportion or sample mean) is in the denominator of the test statistic, it is inversely related to our ability to reject or fail to reject the null hypothesis. Large amounts of variability (and thus large standard errors) will decrease the value of the test statistic, increase the associated p -value, and thus make it more difficult to reject the null hypothesis for actual differences (i.e. power will be lower). Low variability, on the other hand, will increase the value of the test statistic, decrease the associated p -value, and thus make it easier to reject the null hypothesis for actual differences (i.e. power will increase). In more practical terms, lower variability means that the observed statistic is more accurately estimating the population parameter, and any actual difference between the observed data and the hypothesized value is more likely to be found. In summary, larger variability lowers power, while smaller variability increases power. In practice, the variability is not known (if it were, there's not much point in conducting the study), so it must be taken from previous or related studies, clinical experience, or if all else fails it must be guesstimated.

Effect size: The effect size is the smallest difference between the observed statistic and the hypothesized value that we wish to declare significant; this is effectively the numerator of the test statistic. The phrase “smallest observed difference” is often interpreted as the clinically meaningful difference, or the difference below which – if observed – a clinician or scientist would not be

interested (e.g. a physician would not prescribe a cancer treatment that was found to increase life-expectancy by 5 min, but they may prescribe a treatment that increases life-expectancy by 5 years). Effect size is proportionally related to power (large differences are easier to catch), and while this may at first appear counter-intuitive, there is a sound reason for this relationship. Large effect sizes lead to large test statistics, which lead us to reject the null hypothesis more often, and thus increase power. Small effect sizes lead to small test statistics, which prohibit us from rejecting the null hypothesis, and thus decrease power. This might actually seem like a good thing: if small effect sizes are not clinically meaningful, and they lead to lower power, then let's always have large effect sizes with the corresponding higher power (and be done with it!). However, in practice we rarely have large effect sizes, since most of the treatments we deal with don't work (if they did we'd all be famous). Rather, the effect sizes we commonly observe are small, and we are constantly in a "tug-of-war" between designing a study with a reasonable sample size that also maintains an effect size that is still clinically meaningful. The effect size is often the most difficult aspect of the power analysis, especially when there are no previous studies or clinical experience for us to draw upon.

Sample size: Sample size is also straight-forwardly related to power. If the sample size increases, the standard error of the observed statistic decreases, the test statistic increases, the p -value decreases, it becomes easier to reject the null hypothesis, and thus power increases. Likewise, smaller sample sizes lead to larger standard errors, which lead to smaller test statistics, which lead to larger p -values, which cause us to reject the null hypothesis less frequently, which reduces power. This is why statisticians usually answer the question "how much data do I need?" with "as much as possible", because more so than anything else, sample size is the most easily controllable aspect of a study. You can't change the inherent variability in the data (assuming that your methods are scientifically sound), and you can't (or shouldn't) change what is clinically meaningful, so all that is left is sample size. Note that in the context of power, sample size refers to the number of subjects needed to find a significant difference when one exists, and is not necessarily focused on whether or not our sample size assumption is met (e.g. the "rule of 5" for categorical data, or $n > 30$ for continuous data).

9.3 Categorical Outcomes

9.3.1 One-Sample Case

Naturally, when we do a power analysis, we are trying to find the sample size (n) that yields a particular level of power, given a fixed type I error rate (α), measure of variability (based on p_0) and effect size (δ). The way we do that is to set the desired power level and – if an equation exists – enter the

remaining pieces and solve for n . Recall in the one-sample proportion case (Chapter 2) that our test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (9.2)$$

was based upon the standard normal distribution. The type I error rate is fixed at $\alpha = 0.05$, and for a one-sided test the critical value we would use is $z_\alpha = 1.645$ (for a two-sided test we would use 1.96). The null value for the proportion is p_0 , which means that the expected variability in the response is $p_0(1 - p_0)$, which takes its value based on whatever null value we assign. If we want 80 % power, then $1 - \beta = 0.8$, and the 80th percentile from the standard normal distribution is $z_\beta = 0.8416$ (if we wanted 90 % power, the 90th percentile is 1.2816). Without deriving its source, other than to say it closely resembles the test statistic, the sample size formula in the one-sample proportion case is

$$n = \frac{(z_\alpha + z_\beta)^2 p_0(1 - p_0)}{\delta^2}. \quad (9.3)$$

Thus, if we assume that $p_0 = 0.2$ and we want to find a $\delta = 0.10$ -level difference, then we enter the relevant values into the equation and calculate to get

$$n = \frac{(1.645 + 0.8416)^2 (0.2)(1 - 0.2)}{(0.1)^2} = 98.92 \approx 99.$$

Note that we round these values to the next highest integer. This value ($n = 99$) is interpreted as the sample size required to detect a 0.1 difference over a hypothesized value of 0.2 with significance level 0.05 with 80 % power. Note that if we wanted 90 % power, we would need $137.02 \approx 138$ subjects, showing that, all things being equal, increased power requires additional subjects (138 vs. 99). If we wanted 80 % power for a two-sided test, then we would need $n = 125.58 \approx 126$ subjects. If we wanted 80 % power to detect a difference of 0.2, then we would need $24.73 \approx 25$ subjects, which shows how a large difference doesn't require as many subjects (25 vs. 99). Interestingly, if we change the null hypothesized value from 0.2 to 0.5 and require 80 % power to detect a 0.1 difference, we would require $154.56 \approx 155$ subjects, which is more than we needed when $p_0 = 0.2$ ($n = 99$). The implication here is that categorical responses are more variable when the hypothesized value is closer to 0.5 than they are when the hypothesized value is closer to 0 or 1 (see for yourself by calculating $p(1 - p)$ for various values of p).

9.3.2 Two-Sample Case (with R Code)

For the two-group case we do not have a closed-form sample size formula, but we can nevertheless use sophisticated algorithms to calculate the required sample size. Thus, we will rely upon R to do these calculations, with the only

added wrinkle that we now have two proportions instead of one. Recall from Chapter 3 that we are generally interested in testing the null hypothesis that the two proportions are equal, so we simply choose them so that they are desired distance δ apart. For instance, if we want to find a 0.1-level increase between proportion 1 and 2, where one of them is 0.2, we can set $p_1 = 0.3$ and $p_2 = 0.2$. Further, let's assume that this is a two-sided test (so that $\alpha = 0.025$) and that we want 80% power. To determine the required sample size for these specifications, we can use the `power.prop.test` function in R (see Program 31 below). Quite simply, we can specify the assumed proportions for each group (`p1=0.2` and `p2=0.3`), the desired power level (`power=0.8`), and the stated significance level (`sig.level=0.025`; note if we wanted a one-way test we would have used `sig.level=0.05`). With these commands, R tells us that we need $n = 355.1383 \approx 356$ subjects *in each group* to detect a difference of 0.1 with 80% power. Thus, our total required sample size is $2n = 2 \times 356 = 712$, which is a lot, and underscores the fact that categorical data often require large sample sizes to find reasonably small differences between proportions. If we could relax our assumptions and require a 0.15 increase, so that our proportions are 0.35 and 0.2, then the required sample size reduces to $n = 167.1463 \approx 168$ per group (for a total of 336), which is a substantial reduction.

Program 31 Program to determine the sample size needed for a two sample test on proportions where $p_1 = 0.2$, $p_2 = 0.3$, `power=0.8` and significance level $\alpha = 0.025$.

Code:

```
# Use power.prop.test
power.prop.test(p1=0.2, p2=0.3, power=0.8, sig.level=0.025)
```

Output:

```
Two-sample comparison of proportions power calculation

      n = 355.1383
      p1 = 0.2
      p2 = 0.3
sig.level = 0.025
  power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

9.4 Continuous Outcomes

9.4.1 One-Sample Case (with R Code)

For continuous outcomes, we again fix the type I error rate (α), assume some value for the variability (σ), and select the desired effect size (δ). For a desired level of power $1 - \beta$, we obtain the required sample size for comparing a single sample mean to a hypothesized value through the following equation

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\delta^2}. \quad (9.4)$$

Interestingly, we still use the standard normal distribution to get the two probabilistic measures, rather than the student's t -distribution that is used in the hypothesis test (there is no good reason for this; even though the percentiles from the t -distribution are dependent upon sample size, such problems haven't stopped statisticians in the past).

So assuming some σ (say 2) from a previous study, and given some clinically meaningful δ (say 1), we can calculate the required sample size for a one-sided test as

$$n = \frac{(1.645 + 0.8416)^2 (2)^2}{(1)^2} = 24.73 \approx 25.$$

Thus, given these characteristics, we would need 25 subjects to achieve 80 % power to find a difference of 1. To achieve 90 % power, we would need $n = 34.26 \approx 35$ subjects, showing that greater power requires more subjects. Of course, if the variability inherent in the population increases (to 4), we would need $n = 98.92 \approx 99$ subjects to achieve 80 % power, and if we wanted to find a smaller difference (say 0.3), then we would need $n = 274.78 \approx 275$ subjects. We can use R to calculate sample size for us as well using the `power.t.test` function. As shown in Program 32 below, we need to specify the desired effect size (`delta=1`), the assumed standard deviation (`sd=2`), the desired power and significance levels, the number of samples under consideration (here `type="one.sample"`), and the type of alternative hypothesis (`alternative="one-sided"`). For our original case, R tells us that we need $n = 26.13751 \approx 27$ subjects, which is slightly more than we calculated by hand (this is due to R using a slightly altered formula).

9.4.2 Two-Sample Case (with R Code)

Various equations exist for the two-group case of comparing two means, so we will rely solely upon R for sample size calculations. If we assume a common variance σ^2 between the two groups (as well as equal sample sizes), then in R we can again use the `power.t.test` function. For instance, say we want to detect a difference of 1 with 80 % power, assuming a 5 % significance level,

Program 32 Program to determine the sample size needed for a one sample test on a mean where $\delta = 1$, $\sigma = 2$, power= 0.8 and significance level $\alpha = 0.05$.

Code:

```
# Use power.t.test to determine sample size
power.t.test(delta=1, sd=2, type="one.sample",
             power=0.8, sig.level=0.05,
             alternative="one.sided")
```

Output:

```
One-sample t test power calculation

          n = 26.13751
    delta = 1
        sd = 2
sig.level = 0.05
   power = 0.8
alternative = one.sided
```

a (common) standard deviation of 4, and using a two-sided test. The resulting R coding is shown in Program 33, where here we are careful to specify `type="two.sample"` and `alternative="two.sided"`. With these specifications, R tells us we required $n = 252.1281 \approx 253$ subjects *per group* – for a total of 506 subjects – to detect the desired difference given our assumptions.

9.4.3 Multi-sample Case (with R Code)

If we have several group means that we wish to compare using an analysis of variance, we can again use R for the sample size analysis using the `power.anova.test` function (as shown in Program 34). This function has different specifications than the `power.t.test` function, which requires us to perform some steps manually. You may recall that the ANOVA model actually uses the ratio of two variance-like estimators: the estimated variance between the sample means (the MSR); and the estimated variance within the samples (the MSE). In order to obtain a value of MSR, we have to assume (i.e. prespecify) values of our group means that give the desired difference we want to find, and calculate their variance. For instance, we could have three groups, and we would expect that means for two of the groups are equal (with means near 5) while the mean for the third group is 2 units higher (or 7). Then the standard deviation of these means (as shown in Program 34) is calculated using the `sd` function on the vector of assumed group means (`m`). The square of our assumed (common) standard deviation for the observations will serve as our estimate of MSE, so if we again assume a standard deviation

Program 33 Program to determine the sample size needed for a two sided two sample test on a difference in means where $\delta = 1$, $\sigma = 4$, power= 0.8 and significance level $\alpha = 0.05$.

Code:

```
# Use power.t.test to determine sample size
power.t.test(delta=1, sd=4, type="two.sample",
             power=0.8, sig.level=0.05,
             alternative="two.sided")
```

Output:

```
Two-sample t test power calculation

      n = 252.1281
  delta = 1
     sd = 4
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

of 4, then we will use as MSE the variance of 16. These values are specified in the `power.anova.test` function with the options `between.var=var.m` and `within.var=16`, respectively for the MSR and MSE. If we want to detect the 2-unit difference with 80% power and a 5% significance level, then R tells us we will need $n = 58.81829 \approx 59$ subjects per group, or 177 total. Of course, this number will change based on our assumptions.

9.5 Post-hoc Power Analysis

When we perform a hypothesis test – say for comparing two group means – one of two things happens: either we reject the null hypothesis and we report a significant difference (yeah!), or we fail to reject the null hypothesis and state that there is no evidence of a difference (so sad). If we reject the null hypothesis, one thing that may not be evident – probably due to your excitement from finding a significant result – is that *by necessity* you had enough power to declare the observed difference significant. By the same logic, if you did not reject the null hypothesis then you clearly *did not have enough power* to declare the observed difference significant. Some-

Program 34 Program to determine the sample size needed in a one-factor ANOVA.

Code:

```
# Use power.anova.test
m<-c(5,5,7)
var.m<-sd(m)*sd(m)
power.anova.test(groups=3, between.var=var.m,
                 within.var=16, power=0.80, sig.level=0.05)
```

Output:

Balanced one-way analysis of variance power calculation

```
groups = 3
      n = 58.81829
between.var = 1.333333
within.var = 16
sig.level = 0.05
power = 0.8
```

NOTE: n is number in each group

times this is no worry, because the observed difference is smaller than what we would deem clinically meaningful, so our non-significant difference is truly not significant.

However, there may be other occasions where, in our power analysis, we over-estimated the difference we would expect to observe, or we underestimated the variability in the response. In these types of cases we clearly had an under-powered experiment. For example, say we had the test scores of two groups of students on a graduate program entrance examination, where students are placed into two groups based on whether or not they had taken the prerequisite course (these data are found in Table 9.2). The mean test score for students who have met the prerequisites is 90.0 (SD = 3.68), which is 4.1 units higher than the mean test score for students who have not met the prerequisites (mean = 85.9, SD = 5.93). If we assume equal variances, the resulting t -test for comparing these means gives the following results: $t_{18} = 1.86$, p -value = 0.0798, meaning that there is not enough evidence to declare the means different between the two groups.

Maybe there really is no difference in the performance on this examination based on previous course load, or maybe ten prospective students in each group is not enough to find such a difference. If the latter case is true (or if we expect it to be true), we can use R to find the sample size that would declare the observed difference significant, given that everything else stays

Table 9.2: Test Scores on a Graduate Program Entrance Examination based on Prerequisite Status.

	Prerequisite Not Met	Prerequisite Met	
80	86	91	87
88	91	94	90
83	96	98	88
81	92	91	87
85	77	87	87

the same. Thus, we could see how far off we were to having an adequately powered study. This is done in R using the `power.t.test` function. Based on the observed information significance level (α , equal to 0.050 in this case), standard deviation (σ , equal to 4.937948), effect size (δ , equal to 2.05, which is the observed difference divided by 2), and the “Total” sample size (here equal to 20), we are informed that we need 24.84753 subjects in each group to have enough power to declare the difference significant. We can calculate these results for various sample sizes, as shown in Table 9.3. Thus, it appears that our naive study was too small and thus doomed to fail.

To calculate the post-hoc power in R we can again use the `power.t.test` function. However instead of entering in the power we will enter the observed sample size; the `power.t.test` function is smart enough to know that we want power if the sample size is provided. Program 35 shows the R code to

Table 9.3: Power for Various Sample Sizes in Graduate Program Entrance Examination Example.

α	σ	δ	Number	Power
0.0500	4.937948	2.05	20	0.4199
0.0500	4.937948	2.05	22	0.4577
0.0500	4.937948	2.05	24	0.4940
0.0500	4.937948	2.05	26	0.5287
0.0500	4.937948	2.05	28	0.5617
0.0500	4.937948	2.05	30	0.5931
0.0500	4.937948	2.05	32	0.6228
0.0500	4.937948	2.05	34	0.6509
0.0500	4.937948	2.05	36	0.6773
0.0500	4.937948	2.05	38	0.7021
0.0500	4.937948	2.05	40	0.7254
0.0500	4.937948	2.05	42	0.7471
0.0500	4.937948	2.05	44	0.7674
0.0500	4.937948	2.05	46	0.7863
0.0500	4.937948	2.05	48	0.8039
0.0500	4.937948	2.05	50	0.8203

Program 35 Program to determine the post-hoc power for a two sided two sample test on a difference in means where $\delta = 4.1$, $\sigma = 4.937948$, $n = 10$ and significance level $\alpha = 0.05$.

Code:

```
# Use power.t.test to determine power
power.t.test(delta=4.1, sd=4.937948, type="two.sample",
             n=10, sig.level=0.05,
             alternative="two.sided")
```

Output:

```
Two-sample t test power calculation

      n = 10
  delta = 4.1
     sd = 4.937948
sig.level = 0.05
  power = 0.4197928
alternative = two.sided
```

NOTE: n is number in *each* group

conduct the post-hoc power calculation for this scenario based upon our observed information, including the mean difference (`delta=4.1`), the observed pooled standard deviation (`sd=4.937948`), and the sample size in each group (`n=10`). Be sure to note that when using the R function `delta` does not need to be divided by 2. We must also specify the `type="two.sample"` and `"two.sided"` commands in order to specify the correct alternative hypothesis, as well as the appropriate significance level (`sig.level=0.05`). We could then repeat this process for each subsequent sample size in order to reproduce the results found in Table 9.3.

We can also use R to find the smallest difference that would be declared significant given the present significance level, variability and sample size. This again is done using the `power.t.test` function in R, where here we start from some small effect size (say 2) and continue to some larger effect size (say 3.5), increasing each time by some fixed value (say 0.1). Proceeding as was done in Program 35, we get the powers listed in Table 9.4. From this power analysis, we can see that we would need an effect size of at least 3.3 (or a difference of $3.3 \times 2 = 6.6$) to obtain 80% power with 20 subjects. Much like was the case where we looked at sample size, we can see that this study was vastly underpowered.

Table 9.4: Power for Effect Sizes in Graduate Program Entrance Examination Example.

α	σ	δ	Number	Power
0.0500	4.937948	2	20	0.4033
0.0500	4.937948	2.1	20	0.4367
0.0500	4.937948	2.2	20	0.4705
0.0500	4.937948	2.3	20	0.5046
0.0500	4.937948	2.4	20	0.5387
0.0500	4.937948	2.5	20	0.5724
0.0500	4.937948	2.6	20	0.6057
0.0500	4.937948	2.7	20	0.6382
0.0500	4.937948	2.8	20	0.6697
0.0500	4.937948	2.9	20	0.7000
0.0500	4.937948	3	20	0.7290
0.0500	4.937948	3.1	20	0.7566
0.0500	4.937948	3.2	20	0.7825
0.0500	4.937948	3.3	20	0.8068
0.0500	4.937948	3.4	20	0.8293
0.0500	4.937948	3.5	20	0.8501

9.6 Exercises

1. A researcher is planning a study to determine if the yearly flu vaccination rates are different among insured and uninsured Americans who are between the ages of 60 to 65. An meaningful difference to find is 5 % or more. She hypothesizes that the rates will be lower for uninsured Americans. Assuming that 80 % of insured Americans get vaccinated, how large of a sample size is needed for the test to have a power of 90 % with a significance level of $\alpha = 0.03$?
2. [Hall et al. \(2012\)](#) are interested in the hospitalizations for Congestive Heart Failure (CHF) in the United States. A hospital system looks at their report and they wish to know if there is a difference in their system for hospitalization rates for CHF between males and females. The report shows that in 2000 that 42 % of CHF hospitalizations were male and 58 % were females. Based on this information determine the total sample size needed for a test with 85 % power and a significance level of $\alpha = 0.01$.
3. We again the consider example from [Green et al. \(2005\)](#) who is interested in estimating the amount of *diethylhexyl phthalate* (DEHP) that leach from IV tubing and bags into intravenous medications. Suppose they take 25 standard IV bags and standard tubing of length 1 meter and put distilled water in the bag and let it sit for 8 h and then drain

the bag through the tube into a container. From each of the containers they measure the DEHP in ng/mL and suppose they obtain the following data:

53.0, 40.4, 39.1, 39.6, 52.9,
32.8, 51.7, 42.9, 55.0, 43.8,
51.1, 44.2, 38.3, 44.3, 47.7,
43.7, 44.2, 40.0, 60.1, 42.9,
27.0, 50.8, 37.0, 47.5, 69.6

Determine the power of the test for the following hypothesis test $H_0 : \mu \leq 50$ versus $H_a \mu > 50$ based on the sample data.

Chapter 10

Association and Regression

10.1 Introduction

10.1.1 Association Between Measurements

One of the assumptions we have required for nearly every test we've conducted thus far is that of independent subject measurements. The “independence” in question implies that subjects are measured without influencing each other. In Chapter 4 we discussed the association between two *categorical measurements* in the same subject, where the values subjects took for one measurement were somehow related to the values they took on another measurement. These two cases highlight the differences between the types of dependencies that can arise: dependencies between subjects (which is generally bad), and dependencies within subjects (which is – while not bad – hard to call good). Provided that we have randomly selected subjects from the parent population (or at least randomized those conveniently available subjects we have into groups, as the case arises), we can assume that any between-subject dependence will be kept to a minimum; in other words, subjects will be assumed independent. But when we take multiple measurements on each subject, we have to at least entertain the possibility that those measurements will be related (whether we want them to be or not is beside the point). Unlike the case in Chapter 7 – where we had repeated measurements of the same thing – here we focus on cases where we take measurements of two different characteristics – like in Chapter 4, except now with continuous measurements. For instance: systolic and diastolic blood pressures (SBP and DBP) are often measured simultaneously within the same subject; lipid panels generally produce several measurements, including triglycerides (TG) and high-density lipoprotein cholesterol (HDL).

When we have such multiple measures, we generally want to see if there is a relationship between those values. For instance, we might be interested in knowing whether subjects with high SBP also have high DBP, and conversely

whether subjects with low SBP also have low DBP. Alternatively, we may wonder whether subjects with high TG have low HDL, or if subjects with low TG have high HDL (note that high TG and low HDL are considered bad). These examples both exhibit what is called association between the two measurements, where an association – in its most general form – means that there is some kind of relationship between the ways in which the two measurements take values within the same subjects. The stronger this relationship (or the more clear it is to the beholder), the more associated the two measurements are said to be. If there was no discernable relationship, then we would say that two measurements are unrelated (take, for instance, measurements of blood pressure and the color of the sphygmomanometer used to measure it).

10.1.2 Scatter Plots (with R Code)

One of the best ways of observing whether any association exists is to create a *scatter plot* of the two measurements. A scatter plot is a simple plot where the values of one measure are plotted on one axis, while the values of the other measure are plotted on the other axis. Thus, each point in the plot represents the two measurements for a given subject. Several scatter plots are shown in Figure 10.1, where the different plots show varying levels of association. The first plot shows an instance where there is no association between the two variables (creatively entitled “x” and “y”). This scatter plot represents statistical “white noise” in that the points are scattered about the plot window with no pattern, much like an old television receiving no signal. In the last (6th) of these plots, we see that the plotted points form a mostly straight line. This represents the case where the variables exhibit very strong association. In addition, this is what we call *positive association*, since large values of one measure (x) are paired with large values of the other measure (y), and likewise small values of x are paired with small values of y; this is noted by the gradual increase in y as the values for x increase. Conversely, Figure 10.2 shows us what is called *negative association*, where large values of x are paired with small values of y, and large values of y are paired with small values of x. Note that even though the association is “negative” it is still strong. In general, the “strength” of an association is indicated by how discernable the pattern is, or how fine a line is formed by the plotted measurements. Returning to Figure 10.1, we can see that the pattern in each successive plot more closely resembles a straight line, so we can tell that the associations are increasing with each successive plot. We can generate these scatter plots in R by using the `plot` function.

Program 36 shows the R code to create a scatterplot for the Blood Pressure data. We first enter the data into R using the `read.csv` function to read in the `Chp 10 BP Ex.dat` data file. The first row in this data file gives the column names and hence we need to specify `header=TRUE`. To create a scatter plot we use the `plot` function, where the first item is the column in the dataframe (BP1) that corresponds to the variable along the

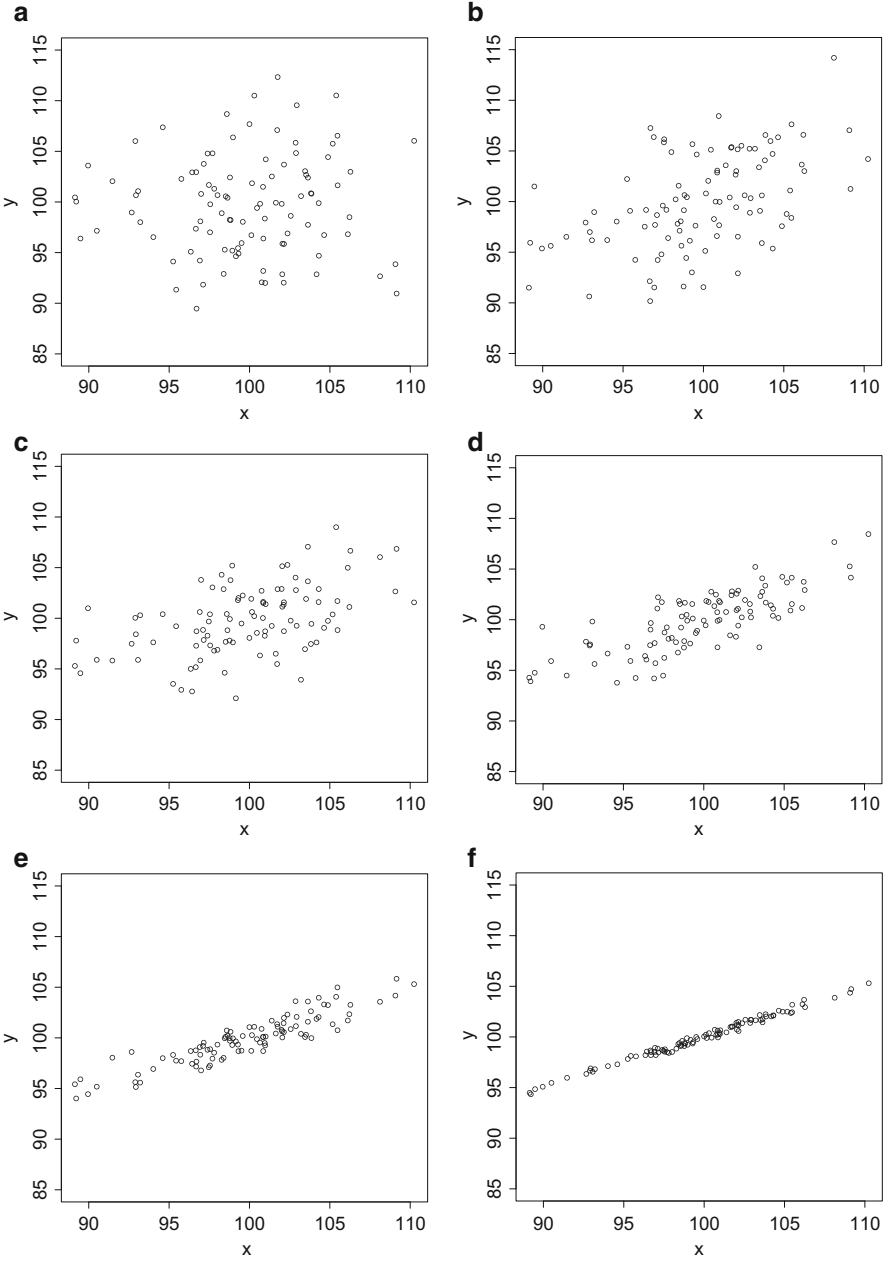


Figure 10.1: Scatter Plots Showing Increasing Association.

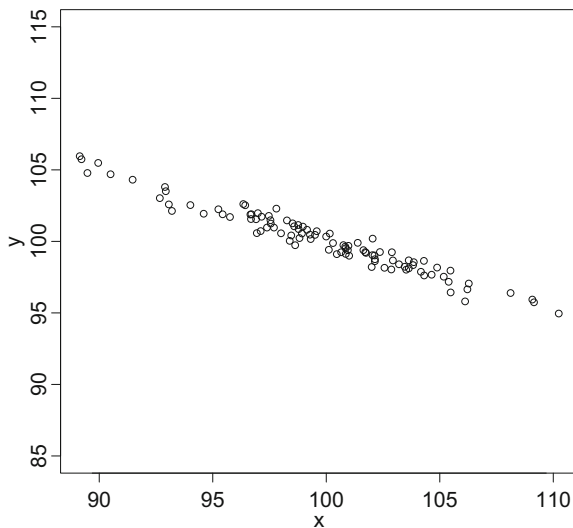


Figure 10.2: Scatter Plot Showing Negative Association.

Program 36 Program to generate a scatterplot of Systolic Blood Pressure versus Diastolic Blood Pressure.

Code:

```
# Read in the data.
BP1 <- read.csv("Chp 10 BP Ex.dat", header=TRUE)

# Create a scatter plot of SBP versus DBP
plot(BP1$DBP, BP1$SBP,
     xlab="DBP",
     ylab="SBP",
     main="Systolic versus Diastolic Blood Pressure")
```

Output:

Found in Figure 10.3.

horizontal axis (`BP1$DBP`) which is the column for Diastolic Blood Pressure, and the second item is the column corresponding to the variable on the vertical axis (`BP1$SBP`) which is the Systolic Blood Pressure. The other options in the function are similar to those in other plotting functions in R. Here `xlab='DBP'` labels the horizontal axis as DBP. Similarly, `ylab='SBP'` labels the vertical axis as SBP. And as usual the `main` statement creates a title for the plot. The resulting scatter plot can be found in Figure 10.3.

While scatter plots are an excellent tool for *seeing* whether association is present, it is a somewhat subjective tool for assessing the strength of the

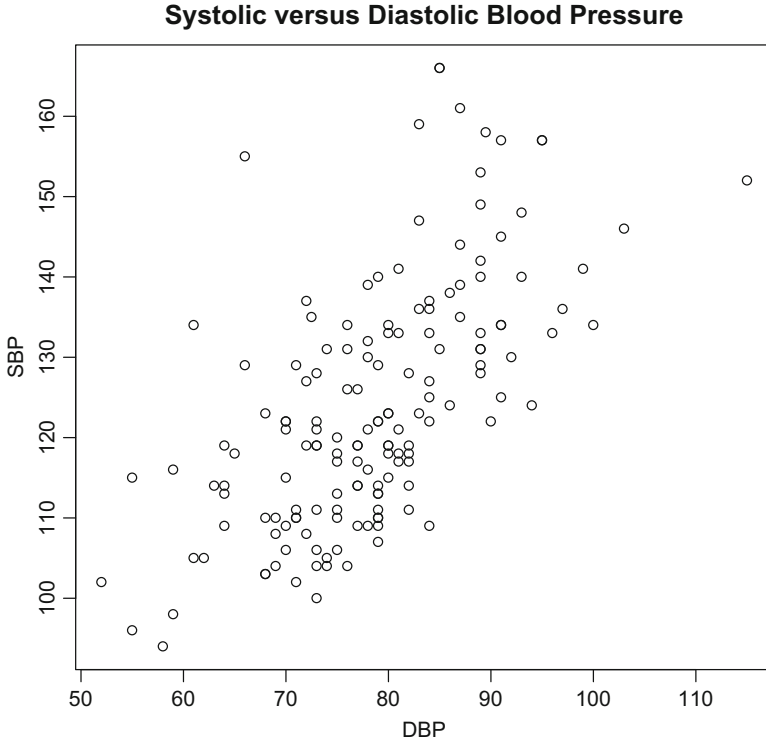


Figure 10.3: Scatter plot of Systolic Blood Pressure versus Diastolic Blood Pressure

association (though some information on strength can be ascertained). Rather than rely upon these plots, there exist several statistical methods for producing a numerical measure or estimate of the association between two measures. The simplest type of measure is the *correlation coefficient*, which produces a value between -1 and 1 that represents the strength of the association between two measurements. Correlations close to 1 represent strong and positive association, correlations close to -1 represent strong and negative association, and correlations close to 0 represent no association. Naturally, positive correlations between 0 and 1 represent varying degrees of positive association, while correlations between -1 and 0 represent varying degrees of negative association. A more sophisticated approach is to perform a *simple linear regression* between the two measurements. A regression aims to estimate the line formed from plotting two measurements in a scatter plot (like that from Figure 10.3). The estimates we obtain from a regression indicate the functional relationship (or lack thereof) between two measurements, but they do not indicate the strength, while a correlation coefficient indicates the strength of association between two measurements but not the functional

relationship. Thus, what approach we use depends upon the type of question we ask. If we are only interested in the strength of association between two measurements, then we will estimate a correlation coefficient, whereas if we want to know the relationship between two measurements, we will perform a regression analysis.

As briefly mentioned in the previous paragraph, a regression estimates the line formed by the plotting together of two measurements. This implies that regression analyses – and correlation estimates, for that matter – are only applicable when the relationship between two measurements is *linear*, or the relationship approximately forms a straight line. Thus, if the relationship between two measurements is anything other than a straight line (see Figure 10.4 for examples), then the methods outlined in this chapter do not apply.

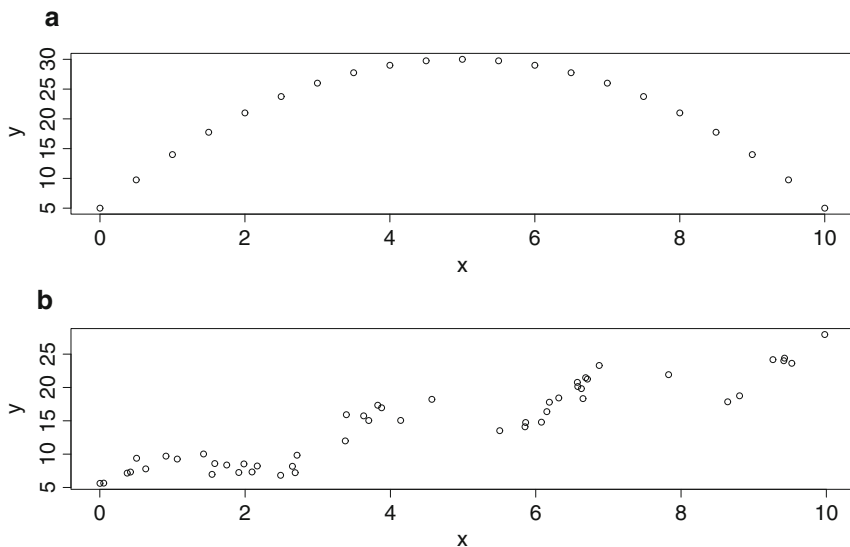


Figure 10.4: Example Scatter Plots of Non-Linear Association.

10.2 Correlation Coefficients

10.2.1 Establishing Hypotheses

When we wish to measure the strength of a linear association between two continuous measurements, we are still in effect performing a hypothesis test. The null case – by default – is that there is no association between the two measurements. This can be symbolically represented by letting ρ stand for the association, so that we get $H_0 : \rho = 0$. The alternative hypothesis must then be that the association is somehow other than zero, or $H_0 : \rho \neq 0$.

10.2.2 Assessing Assumptions (with R Code)

As always, we need our sample to be representative of the population from which it was drawn, and we need the subjects to be independently measured (though we could expect the measurements to be dependently measured *within subjects*). As stated earlier, we need the relationship between the two measures – if it exists – to be linear. Additionally, we need the variability of one measure to be more or less constant throughout the values of the other measure, and vice versa. This condition is difficult to explain in words, but it is easy to see when it occurs, as shown in Figure 10.5. This is what is known as *heteroskedasticity*, or where the variability in one measure is not constant throughout the range of the other variable, as shown by the increasing spread of the variable on the vertical axis as the variable on the horizontal axis gets larger in value. While the relationship remains linear in this case, the extra variance will not be captured by the correlation coefficient (or the regression in the latter part of this chapter). Lastly, we need sufficient sample size. Unfortunately, there is no commonly agreed upon requisite sample size needed to estimate association with correlation coefficient. A good rule of thumb is that samples consisting of fewer than 20 subjects should use non-parametric correlation coefficients, which will be shown later. Further, if either of the

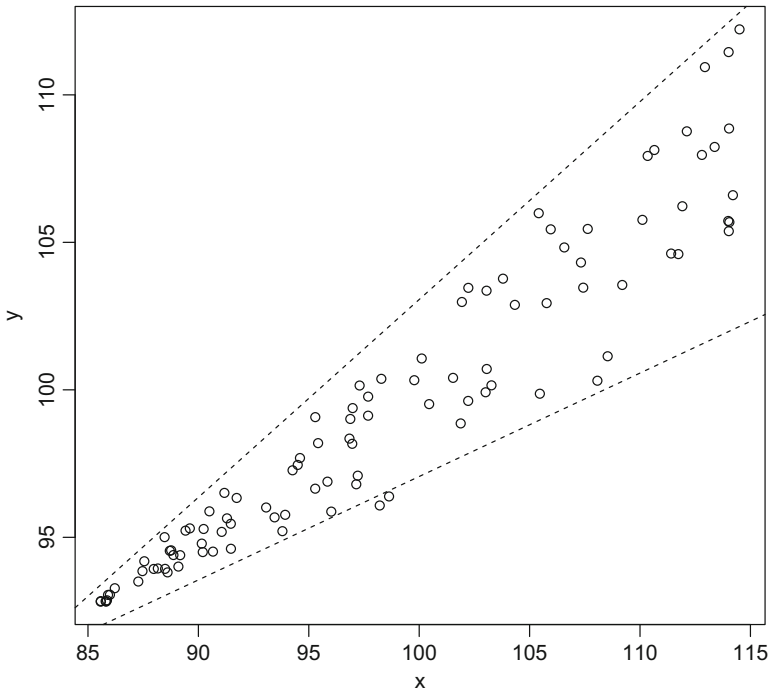


Figure 10.5: Example of Heteroskedasticity.

measurements have skewed or otherwise non-normal distributions, the non-parametric correlation coefficient should be used.

10.2.3 Summarizing Data

Quite simply, if we are estimated a correlation coefficient, then we must also summarize the two variables in question. If we have a large enough sample and our assumptions are met (linear relationship, constant variance, etc.), then we summarize the two measurements with sample sizes, means, standard deviations, and 95 % confidence intervals. If our sample size is small or if our assumptions are not met (or if the data are markedly not symmetrically distributed), then we summarize the measurements with sample sizes, medians and interquartile ranges. In either case, we note the number of missing observations for each variable.

10.2.4 Estimating Correlation, Performing the Test, and Decision Making (with R Code)

Provided we have a large enough sample (and all of our other assumptions are met), we may estimate the association between two measurements X and Y with Pearson's linear correlation coefficient, given by

$$\hat{\rho} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) s_X s_Y}. \quad (10.1)$$

For this estimator we essentially need six things, the sample means and standard deviations for each measure, the sample size, and the sum of the subject-specific products of both measures. This measure takes values between -1 and 1 with the same interpretation as given above. The distribution for $\hat{\rho}$ is complicated, and its test is more complicated still, so we will rely upon the resulting p -value for inference. Naturally, we should also provide a 95 % confidence interval along with the estimated correlation coefficient – rounded to at most two decimal places – in addition to the corresponding p -value. This information can be obtained in R using the `cor.test` function and is illustrated in Program 37 below.

As an example of the process, we will again consider data from the Fels Longitudinal Study (FLS) database, where we have measured both systolic and diastolic blood pressure (SBP and DBP) in 155 male subjects. The data summaries and histograms for both measures are provided in Figure 10.6, which shows that the two measures are reasonably normal distributed for us to continue. A scatter plot of the SBP and DBP measurements was provided earlier in Figure 10.3. We can see that the “scatter” in this plot is more or less linear and that the association seems positive (we're looking for more of a “football” shape than an actual line, though it is nice to see the latter).

The estimated coefficient provided by R in this case is 0.632937, which we report as $r = 0.63$. Note that the associated p -value is <0.0001 , which is less

Program 37 Program to conduct a hypothesis test on a correlation coefficient ρ , with $H_0 : \rho = 0$ versus $H_0 : \rho \neq 0$.

Code:

```
# Read in the BP1 dataset .
BP1 <- read.csv( "Chp 10 BP Ex.dat", header=TRUE )

# Run cor.test for DBP versus SBP
cor.test( BP1$DBP, BP1$SBP ,
          alternative="two.sided")
```

Output:

Pearson's product-moment correlation

```
data: BP1$DBP and BP1$SBP
t = 10.1123, df = 153, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5279701 0.7188565
sample estimates:
      cor
0.6329367
```

than the nominal 0.05 significance level, so we reject the null hypothesis that SBP and DBP have zero correlation. Since $r = 0.63$ is between 0 and 1, we state that SBP and DBP are significantly and positively associated. Since the value is not necessarily close to 1, it would be a stretch to claim that this is strong correlation, though we certainly wouldn't call it weak. Accept for cases of extremely large or small positive correlations (with significant p -values), we generally don't qualify the association as strong or weak, respectively.

Program 37 shows the R code to conduct the analysis above. After reading-in the dataset, we can use the `cor.test` function, which requires two columns of data: `BP1$DBP` and `BP1$SBP`, which stand for DBP and SBP, respectively. The `cor.test` functions has many options similar to those of `t.test`. We can also use the `conf.level` statement to create confidence intervals with various confidence levels. The output from this function gives the t value, the degrees of freedom (`df`), and the p -value is given in scientific notation. You will have to translate the scientific notation of the reported p -value (`p-value<2.2e-16`) to conclude that the p -value < 0.0001 . These results match what we calculated by hand. We also see the 95% CI on the correlation is 0.527 and 0.718.

10.2.5 Contingency Methods (with R Code)

If we did not have a large sample, if one or both of our measurements were not normally distributed, or if the relationship between them was not linear, then we are not allowed to estimate the strength of the linear association using Pearson's correlation coefficient. Instead, we use Spearman's rank correlation coefficient (sometimes called Spearman's ρ), which uses the same equation as Pearson's metric, except we replace the observed measures with their ranks. The ranks, rather than the measures they are ranks of, will have less skewed distributions, and this process is relatively robust to small sample sizes (still, it pays to be reasonable. If you have three subjects, you shouldn't be measuring associations).

Returning to our example, you can calculate Spearman's rank correlation coefficient in R, using the `cor.test` function with `method="spearman"`, which in this case provides an estimate of 0.6540, which we round to 0.65 (note that this is close to the Pearson value of 0.63). The small corresponding p -value (<0.0001) indicates that we again reject the null hypothesis of zero correlation, and conclude that SBP and DBP are significantly and positively associated. Program 38 shows the code to add the `method="spearman"` option in the `cor.test` function. Notice in the output that no confidence intervals are provided when Spearman's ρ is used and also notice that a **Warning message** was generated. This warning is telling us that there are "ties" in the dataset and hence the p -value is an *approximate* p -value. This doesn't pose a problem unless the generated p -value is close to our significance level α , which in this case is not a problem.

10.2.6 Communicating the Results (with IMRaD Write-Up)

The following is an example of the IMRaD write-up for the Blood Pressure example.

Introduction: It is well known that systolic and diastolic blood pressure measurements (SBP and DBP, respectively) are related in the same patients. An observational study was conducted to test for the strength of association between the two measures.

Methods: A sample of 155 paired SBP and DBP measurements was obtained from the Fels Longitudinal Study. Subjects were randomly recruited to participate in the study, and subjects are assumed to be independent. The SBP and DBP measurements are summarized with sample sizes, means, standard deviations and 95% confidence intervals, and a scatter plot is used to check for a linear relationship between SBP and DBP. The strength in association between SBP and DBP was measured using Pearson's linear correlation coefficient provided the two measurements were normally distributed, and was measured using Spearman's rank correlation coefficient otherwise. We reject the null hypothesis that SBP and DBP are not associated in fa-

Program 38 Program to conduct a hypothesis test on a correlation coefficient ρ , with $H_0 : \rho = 0$ versus $H_0 : \rho \neq 0$ using Spearman's ρ .

Code:

```
# Run cor.test for DBP versus SBP
cor.test( BP1$DBP, BP1$SBP ,
          alternative="two.sided",
          method="spearman")
```

Output:

Spearman's rank correlation rho

```
data: BP1$DBP and BP1$SBP
S = 214720, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6540233
```

Warning message:

```
In cor.test.default(BP1$DBP, BP1$SBP, alternative =
"two.sided",: Cannot compute exact p-values with ties
```

vor of the alternative that they are associated if the p -value is less than the significance level of $\alpha = 0.05$, and we will fail to reject the null hypothesis otherwise. The R statistical software was used for all data summaries and analyses.

Results: The sample is assumed representative and subjects are assumed independent. The SBP and DBP measurements are summarized in Figure 10.6, and the two measures were normally distributed. There were also no missing measurements. A scatter plot of the two measures indicates a positive linear association (Figure 10.7). The estimated correlation is $r = 0.63$, which is significantly different from zero (p -value < 0.0001).

Discussion: There is a significant positive association between DBP and SBP in the same subjects. Thus, both measures should be used as biomarkers for hypertension and its effects.

10.2.7 Process for Estimating Correlation

1. State research question in form of testable hypotheses.
2. Determine whether assumptions are met.
 - (a) Representative

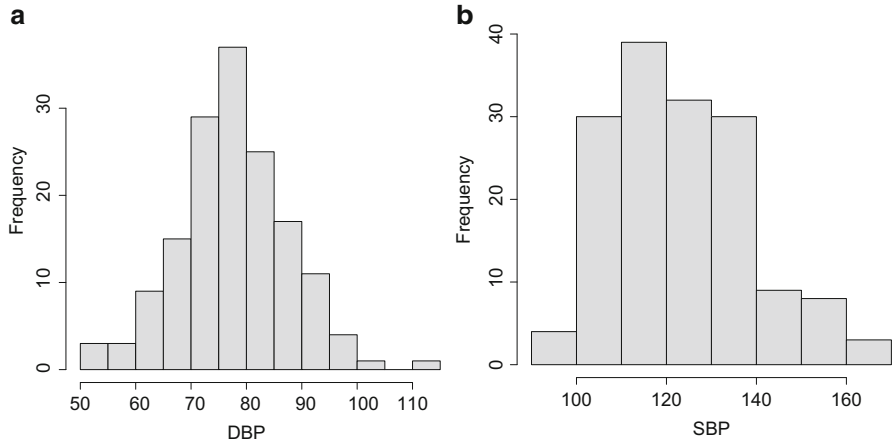


Figure 10.6: Histograms and Summaries for FLS Blood Pressure Example.

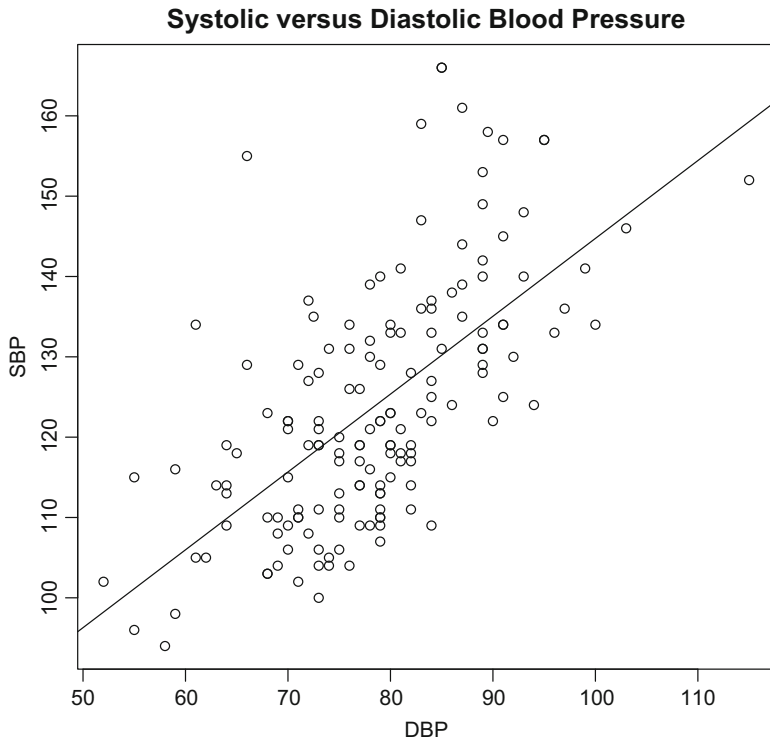


Figure 10.7: Scatter Plot of SBP and DBP for FLS Blood Pressure Example. The estimated Pearson correlation coefficient for this data is $r = 0.63$.

- (b) Independence
 - (c) Sample size: check distribution of data and sample size for each group.
 - (d) Linearity of Relationship.
 - (e) Heteroskedasticity
3. Summarize data.
 - (a) If sample size/normality is adequate for each measurement: summarize with sample size, mean, standard deviation, 95 % CI and # of missing observations.
 - (b) If sample size/normality is inadequate for each measurement: summarize with sample size, median, IQR, and # of missing observations.
 4. Estimate Correlation.
 - (a) If assumptions met: use Pearson Linear Correlation Coefficient.
 - (b) If assumptions not met: use Spearman's Rank Correlation Coefficient.
 5. Report estimated correlation coefficient, *CI* (only if using Pearson's estimator) and *p*-value.
 6. Summarize with IMRaD write-up.

10.3 Simple Linear Regression

If we would like to know more than just the strength of the association, then we can conduct a regression analysis, which aims to measure *what the association is* between two measures. A simple linear regression effectively consists of measuring the linear relationship between two measurements (two is the simplest case for which we can do this, hence the term “simple”), and does so by literally estimating a line between the paired values of the two measures. Recall – from High School, of all places – that the equation of a line is $Y = mX + b$, where m is the slope of the line (how much does Y change for a one-unit change in X), b is the y -intercept (the value of Y when X is equal to zero), and Y and X are two particular paired values on the line. The utility in such an equation comes in the sense that if we know the values of m and b , then we know the value of Y corresponding to any value of X just by placing that value of X into the equation (and we thus could replace all pairs of X and Y with our equation for the line). For instance, if $m = 10$ and $b = 100$, and we let $X = 0.74$, then we know $Y = 10 \times 0.74 + 100 = 107.4$. We could do this for every value of X , obtaining the corresponding value of

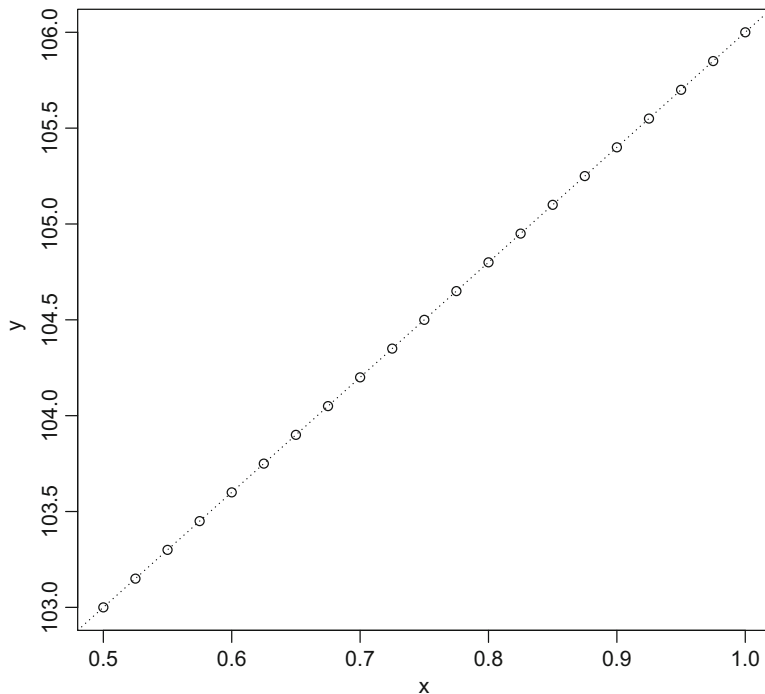


Figure 10.8: Example of a line from the Straight Line Formula found in Equation 10.2.

Y in turn, and if we were to plot these points, they would form a straight line (shown in Figure 10.8 for these choices of m and b).

Since we are limiting ourselves to cases where the relationship between the two measures is linear, a simple linear regression analysis will aim to estimate the slope and intercept of that line. Formally, we will estimate the following regression line

$$Y_i = \beta_0 + \beta_1 X_i, i = 1, \dots, n, \quad (10.2)$$

where β_0 is the y -intercept and β_1 is the slope between the X_i and the Y_i . In this set-up, the intercept β_0 is merely a place-holder, or a mathematical necessity (i.e. it is what it is because it's used to make the equation work), and as such is usually of no interest to us. Thus, all of our inference will be on the slope β_1 , which represents how the variable X and Y are associated. We interpret the slope as the change in Y for a one-unit change in X . If β_1 is positive, then Y will increase with X (they are positively associated); if β_1 is negative, then Y will decrease as X increases (they are negatively associated), and if β_1 is close to zero then there is little or no association between Y and X .

10.3.1 Establishing Hypotheses

The null hypothesis for a simple linear regression is similar to that for correlation, in that we assume there is no relationship between the two measurements. We can state that in words, or we can symbolically write it as $H_0 : \beta_1 = 0$ (though we will estimate β_0 , we don't care about it and thus won't test it). The alternative hypothesis will then be that there is a relationship between the two measures, or $H_0 : \beta_1 \neq 0$. In general, we use two-sided hypothesis because (i) it's easier to test for, and (ii) we typically care about any association in either direction (positive or negative).

10.3.2 Assessing Assumptions (with R Code)

A simple linear regression requires similar assumptions to those needed for calculating Pearson's correlation coefficient, though there are some differences. We of course need a representative sample and independently measured subjects, as well as a large-enough sample size ($n > 20$ usually suffices for a simple linear regression), and the relationship between Y and X needs to be linear. However, we do not necessarily require Y or X to be normally distributed. Instead we need the "residuals" from the regression to be normally distributed, where the estimated residuals are defined as the difference between the observed values of Y and the predicted values of Y given by the simple linear regression model (we will cover residuals and predicted values a little later). While there are non-parametric regression models, we will not cover them (as they are generally either too simple or too complicated), mostly due to the fact that you generally will be able to do the regression.

10.3.3 Summarizing Data

Data are summarized in the same manner as was done for the estimating correlations: we provide the sample size, mean, standard deviation, 95% confidence interval, and number of missing observations for each variable. Of course, the measurements should be summarized appropriately (medians and interquartile ranges) if the values are not normally distributed. In some cases – when the assumptions of the regression model are not met (e.g. linearity) – we will take transformations of the original measurements. This could consist of taking logarithms or roots of our measurements, or any other number of functions that statisticians commonly use. Though we will not get into them here, note that you should be summarizing the data on both the original and transformed scales.

10.3.4 Estimating the Regression, Performing the Test, and Decision Making (with R Code)

Provided that our assumptions are met, we will then estimate the slope and intercept of the regression line. The derivation of these equations is one of the

most fundamental aspects of statistical application, and can be accomplished through the process called ordinary least squares (OLS). While we won't show how it's done, the OLS estimators are such that they minimize the sum of the squared residuals, which is another way of saying that the OLS estimators fit the best possible straight line between Y and X . The OLS estimator for the slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x^2}. \quad (10.3)$$

Notice the similarities between this equation and that for the Pearson correlation coefficient (they are practically the same except for the denominator). Unlike $\hat{\rho}$, $\hat{\beta}_1$ does not gauge the strength of the association between X and Y , it only shows how they are related (more on this later). In fact, $\hat{\beta}_1$ can take any positive or negative value and is not restricted to take values between -1 and 1. To estimate the y -intercept, we use

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (10.4)$$

Implicit in this estimator is that the slope $\hat{\beta}_1$ must be estimated first.

In R, we can fit a regression line between the SBP and DBP measurements from the FLS database by using the `lm` function (as shown in Program 40 below). Note in Figure 10.9 that the scatter plot now has the regression line

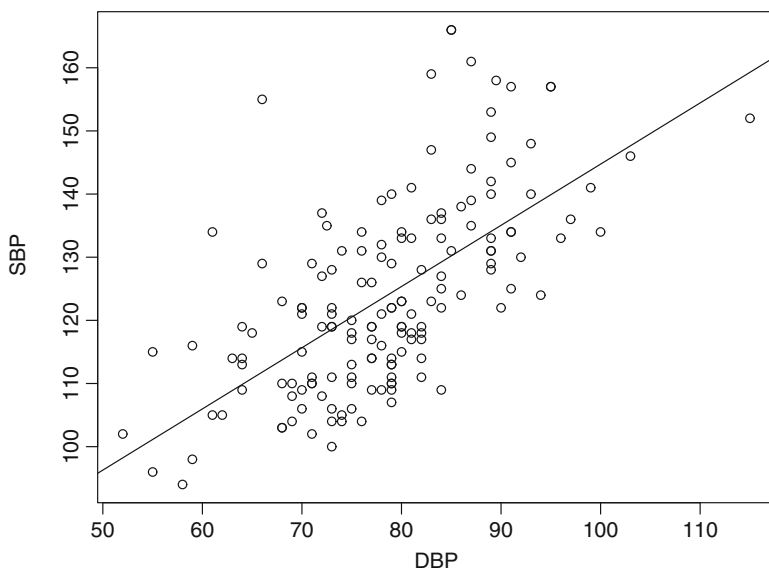


Figure 10.9: R Plot for Simple Linear Regression in Blood Pressure Example 10.9.

superimposed over the data, which increases toward the right-hand side of the plot (indicating positive association). For the components of the equation for the regression line, we see that the estimated slope is 0.9691219 (which we round to 0.97), while the estimated intercept is 47.827896 (which we round to 47.83). In Program 40 we also see the output presented by R, which provides the standard errors, test statistics (based on a t -distribution with $n - 1$ degree of freedom), and the resulting p -values. For our example, we would report that the slope between DBP and SBP was positive and significant ($\hat{\beta}_1 = 0.97$, $SE = 0.096$, $t_1 = 10.11$, p -value < 0.0001), meaning that a 1 mmHg increase in DBP leads to an expected 0.97 mmHg increase in SBP (note that the word *expected* is necessary because the regression is an estimate of the association and is not exact or true in the conventional sense). We need to report the estimates, standard errors, test statistics and p -values for both estimates somewhere (usually in a table), but we also typically report the information for the slope in text.

Program 40 shows the R code to obtain a simple linear regression for the Systolic and Diastolic Blood Pressure data. We use the `lm` function, which stands for “linear model”. The function is similar to the `aov` function in that it requires a formula as the first item, which in this case is `SBP~DBP`. The second item is the dataset we are working with which is `BP1`. We write the output of this function into the variable `BP.lm1` so that we can use the `summary` function. The `summary` function, which produces the regression estimators and hypothesis tests and organizes the output in a useful manner. The information we are typically looking for is found in the `Coefficients` section. It provides the `Estimate` (the estimated intercept and slope coefficients) the `Std. Error` of those estimates, the corresponding `t` value for the hypothesis test, and the p -value denoted with `Pr(>|t|)`. Recall that these p -values are for two sided alternative hypotheses. The output from Program 40 contains more information than we have covered thus far. The `Residuals` section gives the five-number summary of the residuals from the regression (these are described below). At the bottom of the output there are other information that may be useful, such as the `Residual standard error` the `Multiple R-squared` the `Adjusted R-squared` and the `F-Statistics` material (which will be covered later).

10.3.5 Establishing the Worth of the Regression

Once we have fit the regression, it is important for us to check to see how well fit the regression line is to the data. This also brings us back to checking to see whether our residuals are normally distributed. We can use the estimated regression line to calculate predicted values for our response, where predicted values are the values given by the regression equation if we substitute a value of X (DBP in our example) into the equation and calculate. If there was strong association between X and Y , then the straight line relationship between them would be fairly clear as well, meaning that the predicted values

of Y obtained through the equation should be close to the observed values of Y from the original data set. If we have already fitted the regression using the `lm` function then we can easily obtain the fitted values. Program 41 shows the code to obtain the fitted values for the Blood Pressure example. The output in Program 41 shows the predicted values along with the original Blood Pressure data. In this case we wish to compare the response variable SBP versus the predicted SBP or in the output SBP versus `BP.lm1$fit`. We note that while some of these values are close to what we observed, others are not that accurate. For example, consider the observation in row 4 in the output. The actual SBP value is 155 in contrast to the fitted value of 111.79. Here the fitted value and actual value differ quite a lot.

The major benefit from calculating these predicted values is that we can estimate the residuals (mentioned earlier), which are the differences of the observed values from those predicted by the regression line (or the predicted values). If the regression model was fit in R using the `lm` function then the residuals are already calculated. Program 41 shows how to obtain the residuals for the Blood Pressure example. To obtain the residuals we can simply use the `BP.lm1$residual` statement. If our regression fit perfectly (i.e. the scatter plot showed a perfect straight line relationship between X and Y), then the residuals would all be close to zero. Clearly, as seen in the Residual plot in Figure 10.10a, the residuals are not all close to zero (though some are). This is okay, as the magnitude of the residuals is of a secondary concern. We are more interested in the distribution of the residuals, which we can view by observing a histogram or QQ plot of the residuals found in Figure 10.10b–c. We look for any abnormal or clear patterns (you’ll know it when you see it), which are an indication of non-fit or heteroskedasticity. For our blood pressure example, we see in Figure 10.10 that we have more or less white noise with some outliers (recall that outliers are not bad in and of themselves). The perfect residual plot looks somewhat like a side profile of an American football: a thick spread in the center that gradually diminishes as X increases and decreases from that center. As with assessing normality of sample data, we will grant some leeway for minor departures from symmetry or slight skewness.

As an interesting exercise, watch what happens if we calculate the variances of the observed, predicted, and residual SBP values. Since we’ve had R place the original data, predicted and residual values into a new data set named `BP1.data.fit.resid`, we can calculate the means and standard deviations using the `mean` and `sd` functions respectively. These calculations are found in Table 10.1. Note that the mean of the predicted values is equal to the observed mean, while the residual mean is zero (it will always be, even for a bad model). If we square the standard deviations, we get the variances also provided in Table 10.1. However, when we add the variances of the predicted and residual SBP measurements together, the resulting sum is *exactly equal* to the variance of the observed SBP measurements. This is not a coincidence, and results from the OLS method used to estimate the regression line.

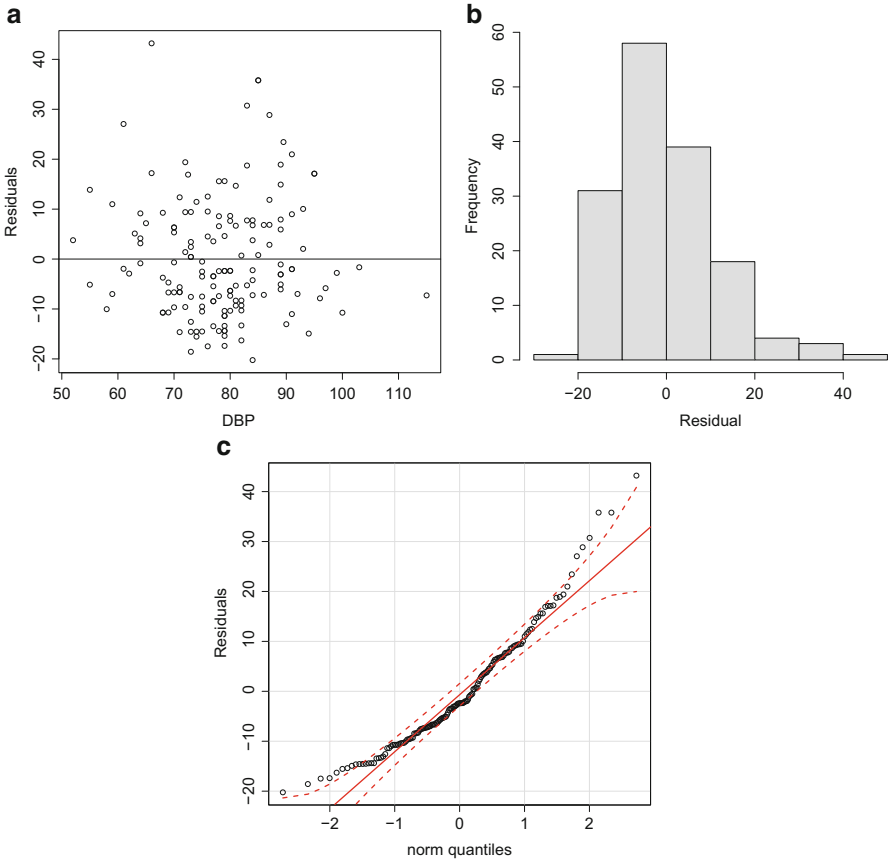


Figure 10.10: Residual plot (a) Histogram (b) and QQplot (c) for the residuals from the SBP vs DBP regression model.

These values are referred to as the “sum of squares” (in fact, they are derived from the same source as the sums of squares from the ANOVA modeling we studied in Chapter 8). The observed variance is the total sum of squares, the predicted variance is the model-based sum of squares, and the residual variance is the residual sum of squares. Naturally, we want the residual sum of squares (or residual variance) to be as small as possible *relative to the total sum of squares* (or observed variance). In this case, the ratio of residual variance is $140.20/233.90 = 0.60$, meaning 60% of the total variation in the observed SBP measurements *is not explained* by the regression equation. Conversely, the ratio of predicted variance is $93.70/233.90 = 0.40$, meaning 40% of the total variation in the observed SBP measurements is explained by the regression equation.

Table 10.1: Variability in Observed, Predicted and Residual SBP for FLS Blood Pressure Example.

	Observed	Predicted	Residual
Mean	123.83	123.83	0.00
SD	15.293725	9.6799607	11.840456
Variance	233.8980	93.7016	140.1964
		Sum = 233.8980	

Whether the observed ratio of predicted to total variance is good or bad is at best a subjective decision, and so we have a formal test that is provided by R. In practice, we call the ratio of predicted variance to total variance the *coefficient of determination*, or R^2 for short. If we look at the output in Program 40, we see the results for an F -test. The p -value listed in this part actually corresponds to a test that the R^2 value equals zero. Since the p -value is less than 0.05, we would reject the null hypothesis that the R^2 (or the predicted variance) is 0. We would report this as $F_{1,153} = 102.3$, p -value < 0.0001 (where the degrees of freedom are taken from the “Model” and “Error” lines, respectively). Note that in the simple linear regression case this test is the same as the t -test for the regression slope, but you still need to report them both somewhere in your write-up. This is because these tests will no longer be identical once we add other covariates to the regression model (the so-called multiple regression model).

Note that the square root of R^2 is $\sqrt{0.400609} = 0.6329$, and recall that the Pearson correlation coefficient between SBP and DBP was $r = 0.6329$. This is no coincidence, and is the reason why the predicted to total variance ratio is called “r-squared”. This also lends itself to describing the difference between a correlation coefficient and regression slope: the correlation measures the strength of an association between two measurements, while the regression slope measures the relationship between those measurements. The size of a regression slope does not tell us how strong the relationship is; that is given by the R^2 , which is simply the square of the correlation.

Note also (last time, we promise), that R^2 has an additional interpretation as the square of the correlation between the observed and predicted response. If we use R to create the predicted values (shown earlier), and then take the correlation between those predicted SBP values with the observed SBP values, we obtain $r = 0.6329$. The square of this value is $r^2 = (0.6329)^2 = 0.4006$, which is equivalent to the R^2 from the model. In addition, this implies that the correlation between the observed and predicted SBP values is the same as the correlation between the observed SBP and DBP values. While seemingly miraculous, this is due to the fact that the predicted SBP values have a one-to-one relationship with the observed DBP values, and is not too much of a surprise upon reflection.

In previous chapters we have provided confidence bounds for our estimates, and we can do a similar thing for our regression line as well. The

problem comes down to what we want these bounds to be for, or what we want them to do. If we are interested in the average relationship between our two measurements, then we will estimate what are called *confidence intervals*. These are obtained in R by using the `predict.lm` function, and are presented for the blood pressure example in Figure 10.11. These bounds are basically a 95% confidence interval on the predicted response that we would expect *on average* for any given DBP level. On the other hand, if we are interested in the relationship between the two measurements for a particular subject, then we will estimate what are called *prediction intervals*. These are obtained in R by using the `predict.lm` function, and are presented for the blood pressure example in Figure 10.11b. These bounds are basically a 95% confidence interval on the predicted response that we would expect for *a particular person or subject*. Note that the prediction bounds are wider than the confidence bounds, which reflects that the confidence bounds are for an average response over n subjects, while the prediction bounds are for the response of a particular subject. Our choice between these two intervals depends upon for whom we want to summarize: the average or particular response.

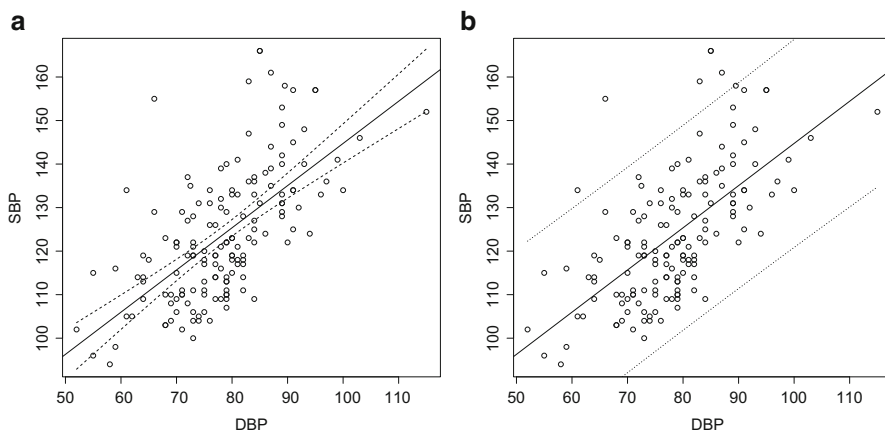


Figure 10.11: Confidence (a) and Prediction (b) Bands of the Regression Line in the Blood Pressure Example.

If we want to provide a “predicted value” of the response (Y) from a particular value of X (usually the mean of X), then the `predict.lm` function in R will give us the value in which we are interested. Returning the blood pressure example, say that we wanted to determine the predicted SBP value – according to the regression line – for the average DBP value, and also find the prediction interval for that particular predicted value. Program 39 shows how to obtain predictions from a regression model. The first step is to create a new `data.frame` that contains the new X value to be predicted. In this case we create a new `data.frame` called `new1` which contains a variable `DBP` which is assigned the value 78.4. Note the name of the variable in the new data

Program 39 Program to conduct a simple linear regression for Systolic Blood Pressure versus Diastolic Blood Pressure.

Code:

```
# For a specific value of DBP
new1 <- data.frame( DBP = 78.4 )
predict.lm( BP.lm1, new1, interval="prediction")
# Generate a new set of values to at which to evaluate the
  regression
newdata1 <- data.frame( DBP = seq(52, 115, by=0.1) )
# Generate confidence intervals at the new values
BP.CL <- predict.lm( BP.lm1, newdata1, interval="confidence" )
# Generate prediction intervals at the new values
BP.Pred <- predict.lm( BP.lm1, newdata1, interval="prediction")

# Plot the data
plot( BP1$DBP, BP1$SBP )
# Add a regression line
abline( BP.lm1 )
# Add the confidence bounds
matlines( newdata1$DBP, BP.CL[ ,c("lwr","upr")], lty=2, col=
  "black")
# Add the prediction bounds
matlines( newdata1$DBP, BP.Pred[ ,c("lwr","upr")], lty=3, col=
  "black")
```

Output:

```
      fit      lwr      upr
1 123.807 100.2633 147.3508
```

See Figure [10.11](#) for plots.

frame must be the same variable name as that in the original dataset; if these names do not match then the `predict.lm` function will not work correctly. The next step is to call the `predict.lm` function and put the regression model, `BP.lm1` as the first item, the new dataframe, `new1` and what type of `interval` we want generated. In this case we are looking for a prediction interval and hence we set `interval="prediction"`. From the output we see that for the mean DBP value of 78.4, the predicted value of SBP is 123.7 (after rounding), with a 95% prediction interval of (100.0, 147.0).

Program 39 also shows how to create scatter plots with the regression line as well as confidence bands and prediction bands. Whenever we use the `predict.lm` function we will need to have a `data.frame` of values that contains the values of X that we wish to predict Y. Here we create `newdata1`

Program 40 Program to conduct a simple linear regression for Systolic Blood Pressure versus Diastolic Blood Pressure.

Code:

```
# Run lm for SBP versus DBP
BP.lm1 <- lm( SBP ~ DBP, data=BP1)
summary( BP.lm1 )

# Code for Creating Scatter Plot with Regression Line
plot(SBP~DBP)
abline(lm(SBP~DBP))
```

Output:

Call:

```
lm(formula = SBP ~ DBP, data = BP1)
```

Residuals:

Min	1Q	Median	3Q	Max
-20.234	-8.389	-2.389	7.019	43.210

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	47.82790	7.57570	6.313	2.81e-09 ***
DBP	0.96912	0.09584	10.112	< 2e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 11.88 on 153 degrees of freedom

Multiple R-squared: 0.4006, Adjusted R-squared: 0.3967

F-statistic: 102.3 on 1 and 153 DF, p-value: < 2.2e-16

as a `data.frame` that contains the variable `DBP` as a sequence of values from 52 to 115 at increments of 0.1, which is coded in R as `seq(52, 115, by=0.1)`. In order to create a scatter plot with confidence or prediction bands we will need to create one dataset that contains the confidence band values and one dataset that contains the prediction band values. In this case we use the `predict.lm` function to create the `BP.CL` dataset which contains the confidence band values, since the `interval="confidence"` option was used. The `BP.Pred` dataset is created similarly, except the `interval="prediction"` option is used. To plot these we first start with a base scatterplot given by the `plot` function. We add the regression line to the plot using the `abline` function, which can accept the fitted regression object, `BP.lm1`, and `overlays` the

Program 41 Program obtain diagnostic quantities for the Systolic Blood Pressure versus Diastolic Blood Pressure regression model. This requires that Program 40 be run first.

Code:

```
# Obtain useful quantities for model diagnostics
BP1.fit <- BP.lm1$fitted           #Fitted values
BP1.resid <- BP.lm1$residuals     #Residuals

#Combine it all together
BP1.data.fit.resid <- cbind(DBP=BP1$DBP,
                             SBP=BP1$SBP,
                             BP1.fit,
                             BP1.resid)

BP1.data.fit.resid                #Print out the information
```

A portion of the output:

	DBP	SBP	BP1.fit	BP1.resid
1	89.0	128	134.07974	-6.0797409
2	72.5	135	118.08923	16.9107697
3	78.0	139	123.41940	15.5805995
4	66.0	155	111.78994	43.2100617
5	71.0	129	116.63555	12.3644524
6	77.0	126	122.45028	3.5497213
7	78.0	132	123.41940	8.5805995
8	73.0	121	118.57379	2.4262087
9	87.0	135	132.14150	2.8585028
10	78.0	121	123.41940	-2.4194005

regression line on to the existing plot. To add the confidence and prediction bands we will use the `matlines` function, which allow us to *overlay* many different lines simultaneously on an existing plot. To add the confidence bands we first need to give the values at which the bands are generated from, which in this case this is `newdata1\DBP`. The second item is the dataset with the confidence or prediction bands in it (either `BP.CL` or `BP.Pred`, respectively). Since we want all the rows and only the columns "lwr" and "upr" we will subset the dataset using the bracket notation. Hence `BP.CL[c("lwr", "upr")]` will subset the dataset this way. Next we need to specify the *line type*, `lty` (of which there are many). In this case we will use `lty=2` and `lty=3` for confidence and prediction bands, respectively. Finally we specify the color of the lines using the `col="black"` statement. The output of these can be found in Figure 10.11.

10.3.6 Communicating the Results (with IMRaD Write-Up)

The following is an example of the IMRaD write-up for the Blood Pressure example.

Introduction: It is well known that systolic and diastolic blood pressure measurements (SBP and DBP, respectively) are related in the same patients. An observational study was conducted to determine the association between the two measures.

Methods: A sample of 155 paired SBP and DBP measurements was obtained from the Fels Longitudinal Study. Subjects were randomly recruited to participate in the study, and subjects are assumed to be independent. The SBP and DBP measurements are summarized with sample sizes, means, standard deviations and 95% confidence intervals, and a scatter plot is used to check for a linear relationship between SBP and DBP. The association between SBP and DBP was estimated using a simple linear regression line. The residuals were checked for normality using a QQ-plot. The predicted SBP at the mean level of DBP and corresponding 95% prediction interval are presented. The null hypothesis that the slope between SBP and DBP is zero is rejected in favor of the alternative that the slope is non-zero if the resulting p -value was less than the significance level of $\alpha = 0.05$, otherwise we fail to reject the null hypothesis. The R statistical software was used for all data summaries and analyses.

Results: The sample is assumed representative and subjects are assumed independent. The SBP and DBP measurements are summarized in Table 10.2. The estimated intercept and slope of the regression line are reported in Table 10.3 and shown graphically in Figure 10.11b (with 95% prediction intervals), with the displayed scatter plot showing a linear relationship. The regression fit reasonably well with $R^2 = 40.0\%$ of the total variability in SBP explained ($F_{1,153} = 102.3$, p -value < 0.0001), and the estimated residuals were normally distributed based on the QQ-plot. The slope (0.97) is positive and significantly different from zero (p -value < 0.0001), meaning that a 1-unit increase in DBP leads to an expected 0.97 mmHg increase in SBP. For the mean DBP value of 78.4, the predicted value of SBP is 123.7 (95%PI : 100.0, 147.0) (Figure 10.11(b)).

Table 10.2: Data Summary for SBP and DBP for FLS Blood Pressure Example.

	n	Mean	SD	95% CI
SBP	155	123.8	15.29	121.4, 126.3
DBP	155	78.4	9.99	76.8, 80.0

Discussion: There is a significant positive association between DBP and SBP in the same subjects. Thus, both measures should be used as biomarkers for hypertension and its effects.

Table 10.3: Regression Results between SBP and DBP for FLS Blood Pressure Example.

	Estimate	SE	Test Statistics	<i>p</i> -value
Intercept	47.8	7.58	6.31	<0.0001
Slope	0.97	0.10	10.11	< 0.0001

10.3.7 Process for Simple Linear Regression

1. State research question in form of testable hypotheses.
2. Determine whether assumptions are met.
 - (a) Representative
 - (b) Independence
 - (c) Sample size: check distribution of data and sample size for each group.
 - (d) Linearity of Relationship.
 - (e) Heteroskedasticity
3. Summarize data.
 - (a) If sample size/normality is adequate for each measurement: summarize with sample size, mean, standard deviation, 95 % CI and # of missing observations.
 - (b) If sample size/normality is inadequate for each measurement: summarize with sample size, median, IQR, and # of missing observations.
4. Perform Regression.
5. Check Regression Diagnostics.
 - (a) Check for Normality of Residuals: if not reasonably normal, disregard the output.
 - (b) Test R^2 using F -test: if not significant, then do not proceed.
 - (c) If appropriate, test slope using t -test.
6. Report Test Results.
 - (a) Report F -test results and R^2 .
 - (b) Report estimated regression coefficients, standard errors, CI s and p -values.
7. Summarize with IMRaD write-up.

10.4 Exercises

1. [Jestoi et al. \(2009\)](#) are interested in how much Furan (a potential carcinogen) is in different commercially available baby food. The researcher is interested if there is a relationship between the amount of protein found in the food as a predictor of the level of Furan found. Below is the data from their experiment:

Protein	Furan	Protein	Furan	Protein	Furan
0.7	29.9	1.1	4.7	2.2	54.6
0.4	14.1	0.5	22.6	3.0	30.3
0.5	14.1	1.0	39.4	3.0	12.8
0.2	8.1	0.5	22.9	4.0	38.6
0.6	8.6	0.8	33.5	3.2	90.3
2.5	74.8	0.7	5.5	1.1	37.0
3.9	45.6	1.0	9.4	0.6	73.4

2. In the study by [Jacobus et al. \(1992\)](#) mentioned in the exercises of Chapter 6, the authors collected measurements of both calcium and albumen (among others) on eight patients (data listed below). Estimate the correlation between these measurements, and determine whether the association is significant.

Measure	Patient							
	1	2	3	4	5	6	7	8
Calcium	2.92	3.84	2.37	2.99	2.67	3.17	3.74	3.44
Albumen	43	42	42	40	42	38	34	42

3. A study by [Weeks and Fox \(1983\)](#) examined fatality rates from mining accidents before and after the passage of the Federal Coal Mine Health and Safety Act of 1969. The total number of fatalities per year from 1959 to 1982 are listed in the table below. Determine whether the number of fatalities has decreased over time.

Year	Fatalities	Year	Fatalities
1959	227	1971	148
1960	267	1972	127
1961	260	1973	104
1962	245	1974	95
1963	230	1975	111
1964	202	1976	109
1965	232	1977	91
1966	197	1978	76
1967	178	1979	114
1968	275	1980	99
1969	155	1981	121
1970	219	1982	81

4. [Bache et al. \(1972\)](#) are interested in the amount of PCB in lake trout for fish of different ages from Cayuga Lake, NY. The PCB measurements are in parts per million (ppm) and age is years. Determine whether or not the amount of PCBs increase with the age of the fish. Be sure to comment on the appropriateness of a line and any deficiencies in assumptions.

Age	PCB	Age	PCB
1	0.6	6	3.4
1	1.6	6	9.7
1	0.5	6	8.6
1	1.2	7	4.0
2	2.0	7	5.5
2	1.3	7	10.5
2	2.5	8	17.5
3	2.2	8	13.4
3	2.4	8	4.5
3	1.2	9	30.4
4	3.5	11	12.4
4	4.1	12	13.4
4	5.1	12	26.2
5	5.7	12	7.4

5. A forestry researcher is interested in the height of Loblolly pine trees. Measure the height of these trees can be a perilous endeavor. Hence, he would like to be able to measure the diameter of the tree at breast height (DBH) in inches and generate an estimate of the height in feet. Use the data below to fit a regression model and evaluate the appropriateness of the model.

DBH	Height	DBH	Height
6.4	15.0	4.5	3.4
10.3	19.7	7.3	9.7
6.4	15.6	9.4	8.6
6.6	16.0	6.9	4.0
9.2	18.0	7.1	5.5
6.0	15.1	6.6	10.5
7.3	16.1	5.7	17.5
6.7	15.9	13.3	13.4
8.2	16.5	8.2	4.5
6.9	15.8	9.0	30.4
5.3	15.0	6.5	12.4
5.8	16.5	7.4	13.4
5.5	15.1	5.2	26.2
6.1	15.8	7.1	7.4
7.4	16.2	7.8	16.8
7.1	16.3	10.5	19.6

6. [Cloquet et al. \(2005\)](#) is interested in the dispersion of lead (Pb) in the air near an industrial French city. Lichens' have the ability to absorb atmospheric pollutants which allows them to be used as biological monitors of air pollution. The authors record the distance to the center of the industrial center of the city in (km) where they take sample of lichens. They then analyze the lichens for the concentration of Pb in ($\mu\text{g/g}$). Using their data determine if Pb concentration decreases as you move away from the industrial center of the city.

Pb	km	Pb	km	Pb	km	Pb	km
6	15.5	5	12	3	11.75	6	10.25
5	7.25	5	4.25	7	3.5	3	15
8	12	6	9	5	9	7	8
11	6	7	4.75	34	3.3	19	3.3
26	3.3	22	2.75	7	2.5	16	1.5
36	1.5	26	1.5	14	1.7	11	1.7
13	1.7	18	11	9	5.5	72	1.5
49	1.5	49	1.5	61	1.5	8	2.6
25	4.25	12	4.25	19	4.25	11	6.25
6	12.75	31	10.75	29	10.75	10	5

7. [Choi et al. \(2009\)](#) is interested in genotyping the Human Papillomavirus. Specifically they are interested in the melting temperature ($^{\circ}\text{C}$) of each genetic sequence and the percent Genetic Content (GC%) of the resulting sample. Using the data below determine if higher genetic content leads to a higher melting temperature.

GC%	Temp	GC%	Temp	GC%	Temp	GC%	Temp
38	68	38	69	41	67	51	71
52	74	44	78	38	73	37	71
46	72	47	72	41	68	53	67
33	71	29	72	36	69	41	71
71	74	53	69	27	67	57	71
46	65	37	71	35	74	31	69
53	78	31	71	33	68	38	70
37	69	40	63	41	70	40	68

Bibliography

- Allred EN, Bleecker ER, Chaitman BR, Dahms TE, Bottlieb SO, Hackney JD, Hayes D, Pagano M, Selvester RH, Walden SM, Warren J (1989). Acute effects of carbon monoxide exposure on individuals with coronary artery disease. *Health Effects Institute Research Report Number 25*, November.
- Austin D, Bowen WD, McMillian JI and Iversion SJ (2006). Linking movement, diving and habitat to foraging success in large marine predators. *Journal of Ecology* 87: 3095–3108.
- Bache CA, Serum JW, Youngs WD, and Lisk DJ (1972). Polychlorinated biphenyl residues: Accumulation in Cayuga Lake trout with age. *Science* 117: 1192–1193.
- Barrison AF, Jarboe LA, Weinberg BM, Nimmagadda K, Sullivan LM, Wolfe MM (2001). Patterns of proton pump inhibitor use in clinical practice. *The American Journal of Medicine* 111(6): 469–473.
- Boles SM, Johnson PB (2001). Gender, weight concerns, and adolescent smoking. *Journal of Addictive Diseases*. 20(2): 5–14.
- Burch KD, Covitz W, Lovett EJ, Howell C, Kanto WP (1989). The significance of ductal shunting during extracorporeal membrane oxygenation. *Journal of Pediatric Surgery*. 24(9): 855–859.
- Choi J, Kim C and Park H (2009). Peptide Nucleic Acid-Based Array for Detecting and Genotyping Human Papillomaviruses. *Journal of Clinical Microbiology* 47(6): 1785–1790.
- Cloquet C, Cariganan J and Libourel G (2005). Atmospheric pollutant dispersion around an urban area using trace metal concentrations and Pb isotopic compositions in epiphytic lichens. *Atmospheric Environment* 40:574–587.
- Di Giusto E, Eckhard I (1986). Some properties of saliva cotinine measurements in indicating exposure to tobacco smoking. *American Journal of Public Health* 76(10): 1245–1246.
- Ekstrom CT (2012). *The R Primer*. CRC Press, New York.

- Engs RC, Hanson DJ (1988). University students' drinking patterns and problems: examining the effects of raising the purchase age. *Public Health Reports*. 103: 667–673.
- Flynn MB and Allen DA (2004). The operative note as billing documentation: A preliminary report. *American Surgeon* 70: 570–575.
- Green R, Hauser R, Calafat AM, Weuve J, Schettler T, Ringer S, Huttner K and Hu H (2005). Use of Di(2-ethylhexyl) Phthalate-Containing Medical Products and Urinary Levels of Mono(2-ethylhexyl) Phthalate in Neonatal Intensive Care Unit Infants. *Environmental Health Perspectives* 113(9): 1222–1225.
- Hall MJ, Levant S, DeFrances CJ (2012). Hospitalization for congestive heart failure: United States, 2000–2010. NCHS data brief, no 108. Hyattsville, MD: National Center for Health Statistics.
- Jay M, Kalet A, Ark T, McMacken M, Messito MJ, Richter R, Schlair S, Sherman S, Zabar S and Gillespie C (2009). Physicians' attitudes about obesity and their associations with competency and specialty: A cross-sectional study *BMC Health Serv Res*. 9: 106.
- Jestoi M, Järvinen T, Järvenpää E, Tapnainen H, Virtanen S and Peltonen K (2009). Furan in the baby-food samples purchased from the Finnish markets – Determination with SPME–GC–MS. *Food Chemistry* 117: 522–528.
- Jacobus CH, Holick MF, Shao Q, Chen TC, Holm IA, Kolodny JM, Fuleihan GE-H, Seely EW (1992). Hypervitaminosis D associated with drinking milk. *The New England Journal of Medicine* 326(18): 1173–1177.
- Justesen US, Levring AM, Thomsen A, Lindberg JA, Pedersen C, Tauris P (2003). Low-dose indinavir in combination with low-dose ritonavir: steady-state pharmacokinetics and long-term clinical outcome follow-up. *HIV Medicine*. 4: 250–254.
- Jade Keightley, Anna Chur-Hansen, Rita Princi, Gary A. Wittert (2011). Perceptions of obesity in self and others. *Obesity Research & Clinical Practice* 5(4):341–349.
- Kreutzer, J., Stejskal, T., Ketchum, J., Marwitz, J., Taylor, L., & Menzel, J. (2009). A preliminary investigation of the Brain Injury Family Intervention: Impact on family members. *Brain Injury*, 23(6), 535–547.
- Lansford JE, Yu T, Erath SA, Pettit GS, Bates JE and Dodge KA (2010). Developmental precursors of number of sexual partners from ages 16 to 22. *Journal of Research on Adolescence*. 20(3): 651–677.

- Norman GJ, Carlson JA, O'Mara S, Sallis JF, Patrick K, Frank LD, Godbole SV (2013). Neighborhood preference, walkability and walking in overweight/obese men. *American Journal of Health Behavior* 37(2): 277–282.
- Ragazzi S, Pierro A, Peters M, Fasoli L, Eaton S (2003). Early full blood count and severity of disease in neonates with necrotizing enterocolitis. *Pediatric Surgery International* 19: 376–379.
- Roche A (1992). Growth, Maturation and Body Composition: the Fels Longitudinal Study 1929–1991. Cambridge University Press, Cambridge.
- Rossi MD, Eberie T, Roche M, Waggoner M, Blake R, Burwell B and Baxter A (2009). Delaying knee replacement and implications on early postoperative outcomes: a pilot study. *Orthopedics* 32(12): 885.
- Salerno J, Darling-Fisher C, Hawkins NM, and Fraker E (2013). Identifying relationships between high-risk sexual behaviors and screening positive for Chlamydia and Gonorrhea in school-wide screening events. *Journal of School Health* 83(2): 99–104.
- Saudek CD, Selam J-L, Pitt HA, Waxman K, Rubio M, Jeandidier N, Turner D, Fischell R, Charles MA (1989). A preliminary trial of the programmable implantable medication system for insulin delivery. *New England Journal of Medicine* 321(9): 574–579.
- Schottenfeld D, Eaton M, Sommers SC, Alonso DR, Wilkinson C (1982). The autopsy as a measure of accuracy of the death certificate. *Bulletin of the New York Academy of Medicine*. 58: 778–794.
- Walker AM, Jick H, Perera DR, Thompson RS, Knauss TA (1987). Diphtheria-tetanus-pertussis immunization and sudden infant death syndrome. *American Journal of Public Health* 77(8): 945–951.
- Wang X, Shen X, Li X, Agrawal CM (2002). Age-related changes in the collagen network and toughness of bone. *Bone* 31(1): 1–7.
- Weeks JL, Fox M (1983). Fatality rates and regulatory policies in bituminous coal mining, United States, 1959–1981. *American Journal of Public Health* 73(11): 1278–1280.
- Winer-Muriam HT, Boone JM, Borwn HL, Jennings SG, Mabie WC and Lombardo GT (2002). Pulmonary embolism in pregnant patients: fetal radiation dose with helical CT. *Radiology* 224:487–492.
- Wrona RM (1979). A clinical epidemiologic study of hyperphenylalaninemia. *The American Journal of Public Health* 69(7): 673–679.

- Yoshinaga S, Mabuchi K, Sigurdson AJ, Doody MM, Ron E (2004). Cancer Risks among Radiologists and Radiologic Technologists: Review of Epidemiologic Studies. *Radiology* 233: 313–321.
- Zuliani, U., Mandras, A., Beltrami, G. F., Bonetti, A., Montani, G., and Novarini, A. (1983). Metabolic modifications caused by sport activity: effect in leisure-time cross-country skiers. *Journal of Sports Medicine and Physical Fitness*, 23, 385–392.