

Introduction to Statistics for Psychology

Introduction to Statistics for Psychology

ALISA BEYER



Introduction to Statistics for Psychology by Alisa Beyer is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/), except where otherwise noted.

The contents of this work have been adapted from the following Open Access

Resources:

Poldrack, Russell A. (2021). Statistical Thinking for the 21st Century, Available at: <https://open.umn.edu/opentextbooks/textbooks/statistical-thinking-for-the-21st-century>

Foster, Garrett C.; Lane, David; Scott, David; Hebl, Mikki; Guerra, Rudy; Osherson, Dan; and Zimmer, Heidi, "An Introduction to Psychological Statistics" (2018). *Open Educational Resources Collection*. 4. Available at: <https://irl.umsl.edu/oer/4>

Online Statistics Education: A Multimedia Course of Study (<http://onlinestatbook.com/>). Project Leader: David M. Lane, Rice University.

Some of the text in the chapter also came from Lisa Degiorgio Worthy.

Changes to the previous works to tailor the text to fit the needs of the introductory statistics course for Maricopa County Community College students. Materials from the original sources have been combined, reorganized, and added to by the current author, and any conceptual, mathematical, or typographical errors are the responsibility of the current author.

Contents

| | |
|---|-----|
| Introduction | 1 |
| Part I. <u>Main Body</u> | |
| 1. Chapter 1: Introduction to Statistics | 5 |
| 2. Chapter 2 Types of Data, How to Collect Them & More Terminology | 22 |
| 3. Chapter 3: Describing Data using Distributions and Graphs | 82 |
| 4. Chapter 4: Measures of Central Tendency | 134 |
| 5. Chapter 5: Measures of Dispersion | 156 |
| 6. Chapter 6: z-scores and the Standard Normal Distribution | 179 |
| 7. Chapter 7: Probability | 200 |
| 8. Chapter 8: Sampling Distributions | 224 |
| 9. Chapter 9 Hypothesis testing | 252 |
| 10. Chapter 10: Hypothesis Testing with Z | 279 |
| 11. Chapter 11: Introduction to t-tests | 307 |
| 12. Chapter 12: Repeated Measures t-test | 332 |
| 13. Chapter 13: Independent Samples | 364 |
| 14. Chapter 14: Analysis of Variance | 405 |
| 15. Chapter 15: 2 - Factor ANOVA | 444 |
| 16. Chapter 16: Correlations | 457 |
| 17. Chapter 17: Linear Regression | 508 |
| 18. Chapter 18. Chi-square | 536 |

| | |
|---|-----|
| 19. Chapter 19. Doing reproducible research | 562 |
| Appendix | 575 |
| References | 576 |

Introduction

What will you learn?

- What is statistical thinking?
- How data is organized, described, and how inferences are made about data
- What are the different statistical tests out there? When can they be used? How do you interpret them?
- How to critically evaluate statistics
- How to be a wise consumer of psychological information, by understanding statistics, to make better decisions for your health and well-being!

Many careers use statistics! Although you are likely taking this course as your degree path requires it, you will find this class is important for you to be a good consumer of statistics and data literacy. This book takes on a more traditional approach to teaching statistics laying the foundation with computational formulas for descriptive and inferential statistics.

Quick overview of the 3 units in this textbook

Unit 1: The first unit in this course will introduce you to the principles of statistics and why we study and use them in the behavioral sciences. It covers the basic terminology and notation used for statistics, as well as how behavioral sciences think about, use, interpret, and communicate information and data. The unit concludes with a brief introduction to concepts in probability and sampling that underlie how scientists perform data analysis. The material in this unit serves as the building blocks for the logic and application of hypothesis

testing, which is introduced in unit 2 and comprises the rest of the material in the course. Unit 1 reviews material in 8 chapters.

Unit 2: In unit 1, we learned the basics of statistics – what they are, how they work, and the mathematical and conceptual principles that guide them. In this unit, we will learn to use everything from the previous unit to test hypotheses, formal statements of research questions that form the backbone of statistical inference and scientific progress. This unit focuses on hypothesis tests about means, and unit 3 will continue to use hypothesis testing for other types of data, statistics, and relations. Unit 2 covers chapters 9 to 15.

Unit 3: The last unit in this course introduces you to analyzing data beyond having the predictor (independent) variable as categorical (nominal) with a continuous (interval/ratio) dependent variable. In this final unit we continue to use the same hypothesis testing logic and procedures on new types of data. We start with hypotheses using only continuous data and then look at a different kind of test statistic: a non-parametric statistic for only categorical data. With the basics of statistics covered in the textbook, the last chapter focuses on considerations for doing reproducible research.

Attribution:

The contents of this work have been adapted from the following Open Access Resources:

Poldrack, Russell A. (2021). Statistical Thinking for the 21st Century, Available at: <https://open.umn.edu/opentextbooks/textbooks/statistical-thinking-for-the-21st-century>

Foster, Garrett C.; Lane, David; Scott, David; Hebl, Mikki; Guerra, Rudy; Osherson, Dan; and Zimmer, Heidi, "An Introduction to Psychological Statistics" (2018). *Open Educational Resources Collection*. 4. Available at: <https://irl.umsi.edu/oer/4>

Online Statistics Education: A Multimedia Course of Study

(<http://onlinestatbook.com/>). Project Leader: David M. Lane, Rice University.

Some of the text in the chapter also came from readings written by Lisa Degiorgio Worthy, Glendale Community College (AZ).

Changes to the previous works to tailor the text to fit the needs of the introductory statistics course for Maricopa County Community College students. Materials from the original sources have been combined, reorganized, and added to by the current author, and any conceptual, mathematical, or typographical errors are the responsibility of the current author.

1. Chapter 1: Introduction to Statistics

What are statistics?

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.
- One in every 8 South Africans is HIV positive.
- By the year 2020, there will be 15 people aged 65 and over for every new baby born.

The study of statistics involves math and relies upon calculations of numbers. *But it also relies heavily on how the numbers are chosen and how the statistics are interpreted.*

Statistical reasoning involves how numbers are chosen and how statistics are interpreted. Consider the following three scenarios and the interpretations based upon the presented statistics. You

*will find that the numbers may be right,
but the interpretation may be wrong.*

*Try to identify a major flaw with each
interpretation before we describe it.*

1. A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.
2. The more churches in a city, the more crime there is. Thus, churches lead to crime.
3. 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages.

What did you come up with?

1. A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective. A major flaw is that ice cream consumption generally increases in the months of June, July, and August regardless of advertisements. This effect is called a history effect and leads people to interpret outcomes as the result of one variable when another variable (in this case, one having to do with the passage of time) is actually responsible.
2. The more churches in a city, the more crime

there is. Thus, churches lead to crime. A major flaw is that both increased churches and increased crime rates can be explained by larger populations. In bigger cities, there are both more churches and more crime. This problem, which we will discuss in more detail in unit 2, refers to the third-variable problem. Namely, a third variable can cause both situations; however, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.

3. 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages. A major flaw is that we don't have the information that we need. What is the rate at which marriages are occurring? Suppose only 1% of marriages 25 years ago were interracial and so now 1.75% of marriages are interracial (1.75 is 75% higher than 1). But this latter number is hardly evidence suggesting the acceptability of interracial marriages. In addition, the statistic provided does not rule out the possibility that the number of interracial marriages has seen dramatic fluctuations over the years and this year is not the highest. Again, there is simply not enough information to understand fully the impact of the statistics.

As a whole, the three examples above show that

statistics are *not only facts and figures*; they are something more than that—they are numbers measured for some purpose. In the broadest sense, “**statistics**” refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.

Statistics is the language of science and data.

The ability to understand and communicate using statistics enables researchers from different labs, different languages, and different fields to articulate to one another exactly what they have found in their work. *It is an objective, precise, and powerful tool in science and in everyday life.*

What statistics are not.

Many psychology, social science, and nursing students dread the idea of taking a statistics course, and more than a few have changed majors upon learning that it is a requirement. That is because many students view statistics as a math class, which is actually not true. While many of you will not believe this or agree with it, statistics isn't math. Although math is a central component of it, *statistics is a broader way of organizing, interpreting, and communicating information in an objective manner.* Indeed, great care has been taken to eliminate as much math from this course as possible. Statistics is a way of viewing reality as it exists around us in a way that we otherwise could not.

Why do we study statistics?

Virtually every student of the behavioral sciences takes some form of statistics class. This is because statistics is how we communicate in science. It serves as the link between a research idea and usable conclusions. Without statistics, we would be unable to interpret the massive amounts of information contained in data. Even small datasets contain hundreds – if not thousands – of numbers, each representing a specific observation we made. Without a way to organize these numbers into a more interpretable form, we would be lost, having wasted the time and money of our participants, ourselves, and the communities we serve.

Beyond its use in science, however, there is a more personal reason to study statistics. Like most people, you probably feel that it is important to “take control of your life.” But what does this mean? Partly, it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study.

To be more specific, here are some claims that we have heard on several occasions. (We are not saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentine.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.

- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are. They come from psychology, health, law, sports, business, etc. Indeed, data and data interpretation show up in discourse from virtually every facet of contemporary life.

Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. (It is not, of course, the only step needed for this purpose.) The purpose of this course is to help you learn statistical essentials and help prepare you for a career in psychology, nursing, counseling, physical therapy, occupational therapy, or other fields that use evidence-based decision making. Most importantly,

taking this course will make you into an intelligent consumer of statistical claims.

You can take the first step right away. **To be an intelligent consumer of statistics, your first reflex must be to question the statistics that you encounter.** The British Prime Minister Benjamin Disraeli is quoted by Mark Twain as having said, “There are three kinds of lies — lies, damned lies, and statistics.” This quote reminds us why it is so important to understand statistics. So let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

The above section puts an emphasis on defending ourselves against fraudulent claims wrapped up as statistics, but let us look at a more positive note. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases. In doing so, statistics might be the course you use most in your day to day life, even if you do not ever run a formal analysis again. You will use statistical thinking!

What is statistical thinking?

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.” – H.G. Wells

Statistical thinking is a way of understanding a complex world by describing it in relatively simple terms that nonetheless capture essential aspects of its structure or function, and that

also provide us some idea of how uncertain we are about that knowledge. The foundations of statistical thinking come primarily from mathematics and statistics, but also from computer science, psychology, and other fields of study.

We can distinguish statistical thinking from other forms of thinking that are less likely to describe the world accurately. In particular, human intuition often tries to answer the same questions that we can answer using statistical thinking, but often gets the answer wrong. For example, in recent years most Americans have reported that they think that violent crime was worse compared to the previous year ([Pew Research Center](#)). However, a statistical analysis of the actual crime data shows that in fact violent crime has steadily *decreased* since the 1990's. Intuition fails us because we rely upon best guesses (which psychologists refer to as *heuristics*) that can often get it wrong. For example, humans often judge the prevalence of some event (like violent crime) using an *availability heuristic* – that is, how easily can we think of an example of violent crime. For this reason, our judgments of increasing crime rates may be more reflective of increasing news coverage, in spite of an actual decrease in the rate of crime. Statistical thinking provides us with the tools to more accurately understand the world and overcome the biases of human judgment

Dealing with statistics anxiety

Many people come to their first statistics class with a lot of trepidation and anxiety. Learning statistics, like learning in general, takes knowledgeable teachers, willing students, and, most importantly, a great deal of time and practice. Learning statistics is like learning a language. The symbols and notation make up the rules of grammar and the terminology is the vocabulary. Doing the homework is like practicing the

conversation of statistics. Becoming fluent (and staying fluent) in statistics requires practice and continuous use.

Questionnaires can be used to survey students prior to the first class in order to measure their attitude towards statistics, asking them to rate a number of statements on a scale of 1 (strongly disagree) to 7 (strongly agree). One of the items on the statistical attitudes survey is “The thought of being enrolled in a statistics course makes me nervous”. In a recent class, almost two-thirds of the class responded with a five or higher, and about one-fourth of the students said that they strongly agreed with the statement. So if you feel nervous about starting to learn statistics, you are not alone.

Anxiety feels uncomfortable, but psychology tells us that some emotional arousal can actually help us perform *better* on some tasks, by focusing our attention. So if you start to feel anxious about the material in this book, remind yourself that many other students are feeling similarly, and that this emotional arousal could actually help you learn the material better (even if it doesn’t seem like it!).

Tips for Statistics Anxiety

1. Learn stress management and relaxation techniques.

Techniques such as deep breathing and meditation that help you to relax in any stressful situation can also be helpful when dealing with the nervousness and tension that affect students with math anxiety.

2. Combat negative thinking.

Lack of confidence can be a major impediment for students with math anxiety. Replace those negative thoughts (“I can’t do this”, “I’ve never been good at

math", "I won't finish in time") with confidence-building affirmations ("I know this", "I'm prepared", "I can do this").

3. Visualize yourself succeeding.

Athletes use the technique of "visualization" to prepare for major competitions. Imagine yourself being relaxed doing math and during a test and confidently solving the problems.

4. Do "easiest" problems first.

Build up your confidence by first doing those problems in an assignment or on a test that you "know" best. It'll help you relax when you tackle the "harder" stuff.

5. Channel your stress into something else.

Free up your mind by relieving some of your physical responses to stress. Get up and run around the hall for a minute before the test or squeeze a stress ball like crazy during the test.

6. Start preparing early.

If you try to "cram" the material quickly, you are likely to forget it quickly too. If you practice the material over a period of time, you will have a better understanding of it and are less likely to forget it when under stress.

7. Take care of yourself.

Although it's not easy when you're in school, eating and sleeping well helps your body and mind function to their fullest potential.

8. Try to understand the "why" of statistical concepts rather than memorizing them.

The first thing to go when you are under stress is your short-term memory. This is one reason it is so

important to understand that math is not just a set of rules that you have to memorize but that each concept builds on what came before. If you understand the reason behind the rules, you will remember the concepts better and be able to apply them in many different types of problems (not just ones you've seen before).

9. Reward yourself for hard work.

After completing a difficult assignment or an exam, it's time to give yourself a break.

What can statistics do for us?

There are three major things that we can do with statistics:

1. *Describe*: The world is complex and we often need to describe it in a simplified way that we can understand.
2. *Decide*: We often need to make decisions based on data, usually in the face of uncertainty.
3. *Predict*: We often wish to make predictions about new situations based on our knowledge of previous situations.

Let's look at an example of these in action, centered on a question that many of us are interested in: How do we decide what's healthy to eat? There are many different sources of guidance; government dietary guidelines, diet books, and bloggers, just to name a few. Let's focus in on a specific question: Is saturated fat in our diet a bad thing?

One way that we might answer this question is *common sense*. If we eat fat, then it's going to turn straight into fat in our bodies, right? And we have all seen photos of arteries clogged with fat, so eating fat is going to clog our arteries, right?

Another way that we might answer this question is by listening to *authority figures*. The Dietary Guidelines from the US Food and Drug Administration have as one of their Key Recommendations that “A healthy eating pattern limits saturated fats”. You might hope that these guidelines would be based on good science, and in some cases they are, but as Nina Teicholz outlined in her book “Big Fat Surprise”(Teicholz [2014](#)), this particular recommendation seems to be based more on the longstanding dogma of nutrition researchers than on actual evidence.

Finally, we might look at actual *scientific research*. Let’s start by looking at a large study called the PURE (Prospective Urban Rural Epidemiology) study, which has examined diets and health outcomes (including death) in more than 135,000 people from 18 different countries. In one of the analyses of this dataset (published in *The Lancet* in 2017; Dehghan et al. ([2017](#))), the PURE investigators reported an analysis of how intake of various classes of macronutrients (including saturated fats and carbohydrates) was related to the likelihood of dying during the time that people were followed. People were followed for a *median* of 7.4 years, meaning that half of the people in the study were followed for less and half were followed for more than 7.4 years. Figure 1 plots some of the data from the study (extracted from the paper), showing the relationship between the intake of both saturated fats and carbohydrates and the risk of dying from any cause.

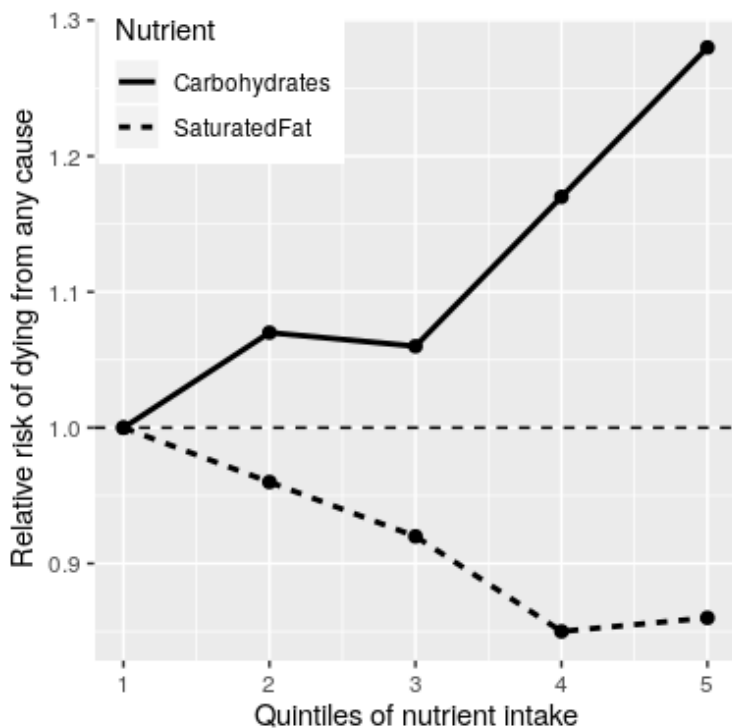


Figure 1: A plot of data from the PURE study, showing the relationship between death from any cause and the relative intake of saturated fats and carbohydrates.

This plot is based on ten numbers. To obtain these numbers, the researchers split the group of 135,335 study participants (which we call the “sample”) into 5 groups (“quintiles”) after ordering them in terms of their intake of either of the nutrients; the first quintile contains the 20% of people with the lowest intake, and the 5th quintile contains the 20% with the highest intake.

The researchers then computed how often people in each of those groups died during the time they were being followed. The figure expresses this in terms of the *relative risk* of dying in comparison to the lowest quintile: If this number is greater

than one, it means that people in the group are *more* likely to die than are people in the lowest quintile, whereas if it's less than one, it means that people in the group are *less* likely to die. Figure 1.1 is pretty clear: People who ate more saturated fat were *less* likely to die during the study, with the lowest death rate seen for people who were in the fourth quintile (that is, who ate more fat than the lowest 60% but less than the top 20%). The opposite is seen for carbohydrates; the more carbs a person ate, the more likely they were to die during the study. *This example shows how we can use statistics to describe a complex dataset in terms of a much simpler set of numbers; if we had to look at the data from each of the study participants at the same time, we would be overloaded with data and it would be hard to see the pattern that emerges when they are described more simply.*

The numbers in Figure 1 seem to show that deaths decrease with saturated fat and increase with carbohydrate intake. This large-scale study also had some [methodological challenges](#) controlling for socioeconomic factors and measurement of dietary intake data. We also know that there is a lot of uncertainty in the data; there are some people who died early even though they ate a low-carb diet, and, similarly, some people who ate a ton of carbs but lived to a ripe old age. Given this variability, we want to *decide* whether the relationships that we see in the data are large enough that we wouldn't expect them to occur randomly if there was not truly a relationship between diet and longevity. Statistics provide us with the tools to make these kinds of decisions. But as we will see throughout the book, this need for black-and-white decisions based on fuzzy evidence can lead researchers astray.

Based on the data we would also like to make predictions about future outcomes. For example, a life insurance company might want to use data about a particular person's intake of fat and carbohydrate to predict how long they are likely to live. *An important aspect of prediction is that it requires us*

to generalize from the data we already have to some other situation, often in the future; if our conclusions were limited to the specific people in the study at a particular time, then the study would not be very useful. In general, researchers must assume that their particular sample is representative of a larger *population*, which requires that they obtain the sample in a way that provides an unbiased picture of the population. For example, if the PURE study had recruited all of its participants from religious sects that practice vegetarianism, then we probably wouldn't want to generalize the results to people who follow different dietary standards.

The big ideas of statistics

One way to think of statistics is as a set of tools that enable us to learn from data.

There are two main branches of statistical analysis, descriptive statistics and inferential statistics.

- Statistics that are used to *organize and summarize the information* so that the researcher can see what happened during the research study and can also communicate the results to others are called **descriptive statistics**. The first unit of the book is focused on descriptive statistics.
- Statistics that help the researcher to *answer the general research question* by determining exactly what conclusions are justified based on the results that were obtained are referred to as **inferential statistics**.

Statistics provides us with the tools to characterize uncertainty, to make decisions under uncertainty, and to make predictions whose uncertainty we can quantify. For example, we now know that cigarette smoking causes lung

cancer, but this causation is probabilistic: A 68-year-old man who smoked two packs a day for the past 50 years and continues to smoke has a 15% (1 out of 7) risk of getting lung cancer, which is much higher than the chance of lung cancer in a nonsmoker. However, it also means that there will be many people who smoke their entire lives and never get lung cancer.

One often sees journalists write that scientific researchers have “proven” some hypothesis. But statistical analysis can never “prove” a hypothesis, in the sense of demonstrating that it must be true (as one would in a logical or mathematical proof). **Statistics can provide us with evidence, but it's always tentative and subject to the uncertainty that is always present in the real world.**

The concept of aggregation implies that we can make useful insights by collapsing across data – but how much data do we need? **The idea of *sampling* says that we can summarize an entire population based on just a small number of samples from the population, as long as those samples are obtained in the right way.** As we already discussed above, the way that the study sample is obtained is critical, as it determines how broadly we can generalize the results. Another fundamental insight about sampling is that while larger samples are always better (in terms of their ability to accurately represent the entire population), there are diminishing returns as the sample gets larger. In fact, the rate at which the benefit of larger samples decreases follows a simple mathematical rule, growing as the square root of the sample size, such that in order to double the precision of our estimate we need to quadruple the size of our sample.

Study design is also important part of statistical thinking — remember correlation and causation. Any introduction to psychology course and introductory statistics will often teach that “correlation does not imply causation”, though the

renowned data visualization expert Edward Tufte has added, “but it sure is a hint.”

We will examine more about study design and types of data in our next chapter!

Learning Objectives

1. Define statistical thinking and why we use statistics.
2. Practice ways to reduce statistical anxiety.
3. Identify how statistical techniques fit into the general process of science.

Exercises – Chapter 1

1. Reflect on a statistics that you have encountered in daily life. How can you apply statistical thinking?
2. What are two reasons that you identified for why taking a course in statistics is important?
3. How would you define statistics to a friend, neighbor, family member? Define statistics from what you have learned so far.
4. Review the tips for statistical anxiety and reflect on how you can implement at least one tip to help you succeed in the course.

2. Chapter 2 Types of Data, How to Collect Them & More Terminology

Types of Data and How to Collect Them

In order to use statistics, we need data to analyze. Data come in an amazingly diverse range of formats, and each type gives us a unique type of information. In virtually any form, data represent the measured value of variables. A **variable** is simply a characteristic or feature of the thing we are interested in understanding. Let's imagine we want to conduct a study to measure the stress level of students who are taking PSY 230. We will administer the survey during the first week of the course. One question we will ask is, "How stressed have you been in the last 2 weeks, on a scale of 0 to 10, with 0 being not at all stressed and 10 being as stressed as possible?"

- **Variable** is a condition or characteristic that can take on different values. In our example, the variable was stress, which can take on any value between 0 and 10. Height is a variable. Social class is a variable. One's score on a creativity test is a variable. The number of people absent from work on a given day is a variable. In psychology, we are interested in people, so we might get a group of people together and measure their levels of anxiety (a variable) or their physical health (another variable). You get the point. Pretty much anything we can count or measure

can be a variable.

- Once we have data on different variables, we can use statistics to understand if and how they are related.
- A **value** is just a number, such as 4, – 81, or 367.12. A value can also be a category (word), such as male or female, or a psychological diagnosis (major depressive disorder, post-traumatic stress disorder, schizophrenia).
 - We will learn more about values and types of data a little later in this chapter.
- Each person studied has a particular **score** that is his or her value on the variable. As we've said, your score on the stress variable might have a value of 6. Another student's score might have a value of 8.

We also need to understand the nature of our data: what they represent and where they came from. We will briefly review some keys to understanding statistical studies.

Tips to understanding statistical studies

Here are a few key considerations for evaluating studies using statistics.

1. Know the basic components of a statistical investigation.
2. Know the sample. Identify if using a representative sample.
3. Identify the sample size. Evaluate if using a large enough sample.
4. Understand and evaluate the study design.
5. Identify type of data working with.
6. Understand the statistics used.
7. Evaluate conclusions made from statistical findings.

The basic components to a statistical investigation

- **Planning the study:** Start by asking a testable research question and deciding how to collect data. For example, how long was the study period of the study? How many people were recruited for the study, how were they recruited, and from where? How old were they? What other variables were recorded about the individuals, such as smoking habits, on the comprehensive lifestyle questionnaires?
- **Examining the data:** What are appropriate ways to examine the data? What graphs are relevant, and what do they reveal? What descriptive statistics can be calculated to summarize relevant aspects of the data, and what do they reveal? What patterns do you see in the data? Are there any individual observations that deviate from the overall pattern, and what do they reveal?
- **Inferring from the data:** What are valid statistical methods for drawing inferences “beyond” the data you collected? Is a 10%–15% reduction in risk of death something that can happen just by chance?
- **Drawing conclusions:** Based on what you learned from your data, what conclusions can you draw? Who do you think these conclusions apply to? Can you draw a cause-and-effect conclusion about your treatment? (note: we are about to learn more about the study design needed for this)

Notice that the *numerical analysis* (“crunching numbers” on the computer) comprises only a small part of the overall statistical investigation. In this module, you will see how we can answer some of these questions and what questions you

should be asking about any statistical investigation you read about. In the end, statistics provides us a way to give a very objective “yes” or “no” answer to the question, “is this treatment or intervention effective and, if so, how effective is it?” Nearly all statistical techniques boil down to answering these questions. Statistics is all about helping make correct and reliable decisions in our chosen field of study. But even if you never plan on conducting research or pursuing a career where you have to use statistics, the material in this course will help you in your daily life. In today’s world of instant gratification, information overload, and the 24-hour news cycle, statistics are thrown at us nonstop. Soon, you will be able to determine if the person or group providing these statistics is being honest or manipulating the data to suit their ideas.

Let’s learn a little bit more about what is needed to know to better understand statistics.

Who are your participants? Who is your population?

Research in psychology typically begins with a general question about a specific group (or groups) of individuals or animals. For example, a researcher might want to know how many homeless people live on the streets of Phoenix. Or a researcher might want to know how often married people have sex, as reported by partners separately. In the first example, the researcher is interested in the group of homeless people. In the second example, the researcher may study heterosexual couples and compare the group of men with the group of women. In statistics, we call the entire group that a researcher wishes to study **a population**. As you can well imagine, a population can be quite large; for example, any student enrolled in college. A researcher might be more specific,

limiting the population for a study to college students who have successfully completed a statistics course and who live in the United States.

Populations can obviously vary in size from extremely large to very small, depending on how the researcher defines the population. The population being studied should always be identified by the researcher. In addition, the population can include more than people and animals. A population could be corporations, parts produced in a factory, or anything else a researcher wants to study. Because populations tend to be very large it usually is impossible for a researcher to examine every individual in the population of interest. It is typically not feasible to collect data from an entire population. Therefore, researchers typically select a smaller, more manageable group from the population and limit their studies to the individuals in the selected group. A smaller more manageable group, known as a **sample**, is used to measure populations.

The participants in the research are the **sample**, and the larger group the sample represents is the **population**. In statistical terms, a set of individuals selected from a population is called a **sample**. A sample is intended to be representative of its population, and a sample should always be identified in terms of the population from which it was selected. As with populations, samples can vary in size. For example, one study might examine a sample of only 10 autistic children, and another study might use a sample of more than 10,000 people who take specific cholesterol medication. The sample is intended to represent the population in a research study.

When describing data it is necessary to distinguish whether the data come from a population or a sample.

- If data describe a **sample** it is called a **statistic**.
- If data describe a **population** it is called a **parameter**.

If I had given a statistical attitudes survey to the class, the class

would be my sample. I might be interested in all students taking a statistics class for the first time, generalizing my findings to all statistics students would be applying information from my sample to a population. While it might be convenient for me to ask my class, does my class best represent all students taking statistics? I would need to carefully consider selecting the best sample for a population or critically think about the limits for generalizing my findings to a population. While our results would be most accurate if we could study the entire population rather than a sample from it, in most research situations this is not practical. Moreover, research usually is to be able to make generalizations or predictions about events beyond your reach. Additionally, sampling is an important concept to consider with the big picture of understanding statistics.

Imagine that we wanted to see if statistics anxiety was related to procrastination. We could measure everyone's levels of statistics anxiety and procrastination and observe how strongly they were related to each other. This would, however, be prohibitively expensive. A more convenient way is to select a number of individuals randomly from the population and find the relationship between their statistics anxiety and procrastination levels. We could then generalize the findings from this sample to the population. We use statistics, more specifically inferential statistics, to help us generalize from a particular sample to the whole population. Understanding the relationship between populations and their samples is the first vital concept to grasp in this course. Remember that the research started with a general question about the population but to answer the question, a researcher studies a sample and then generalizes the results from the sample to the population.

As we move into further concepts in statistics, we will see that how you get your participants (sampling) and sample size are important. The general rule is to get a large enough

sample size and have the sample be a good representation of your population.

Representative Sample

Because we are using samples to generalize to the larger population it is important, vital, that the samples look like the population they came from. When the sample closely matches the population from which it was selected we call this a **representative** sample. An unrepresentative (**biased**) sample is a subset of the population that does not have the characteristics typical of the target population.

Random Sampling

Usually, the ideal method of picking out a sample to study is called random selection or sampling. The researcher starts with a **complete list of the population** and randomly selects some of them to study. . Random sampling is considered a fair way of selecting a sample from a given population since every member is given equal opportunities of being selected. This process also helps to ensure that the sample selected is more likely to be representative of the larger population. Theoretically, the only thing that can compromise its representativeness is luck. If the sample is not representative of the population, the random variation is called **sampling error**.

Example #1: You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Who will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The

Americans actually queried constitute our sample of the larger population of all Americans.

A **sample** is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferences from statistics are based on the assumption that sampling is representative of the population. If the sample is not representative, then the possibility of sampling bias occurs. Sampling bias means that our conclusions apply only to our sample and are not generalizable to the full population.

Example #2: We are interested in examining how many math classes have been taken on average by current graduating seniors at American colleges and universities during their four years in school.

Whereas our population in the last example included all US citizens, now it involves just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls 48,000 students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make inferences to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our

sample were 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors.

To solidify your understanding of sampling bias, consider the following example. Try to identify the population and the sample, and then reflect on whether the sample is likely to yield the information desired.

Example #3: A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In Example #3, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example #4: A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

In Example #4, the population is the class of all freshmen at

the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who can't do cartwheels probably did not volunteer! In the example, we are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed).

Simple Random Sampling

Researchers adopt a variety of sampling strategies. The most straightforward is simple random sampling. Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. To check your understanding of simple random sampling, consider the following example. What is the population? What is the sample? Was the sample picked by simple random sampling? Is it biased?

Example #5: A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the National Twin Registry, and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, however, our researcher decides to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of

twins raised apart versus together.

In Example #5, the population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the Z's may bias the sample.

Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the B's is that the "every-other-one" procedure disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

Sample size matters

Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the sampling procedure rather than the results of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small, are not necessarily representative of the entire population. For example, if a random sample of 20 subjects were taken from a population with an equal number

of males and females, there would be a nontrivial probability (0.06) that 70% or more of the sample would be female. Such a sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population. For this reason, inferential statistics take into account the sample size when generalizing results from samples to populations. In later chapters, you'll see what kinds of mathematical techniques ensure this sensitivity to sample size.

More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competed to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For example, how will you contact those individuals who don't vote and don't have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional 4,212 people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

Stratified Sampling

Since simple random sampling often does not ensure a

representative sample, a sampling method called **stratified random sampling** is sometimes used to make the sample more representative of the population. This method can be used if the population has a number of distinct “strata” or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Let’s take an example: Suppose you were interested in views of capital punishment at an urban university. You have the time and resources to interview 200 students. The student body is diverse with respect to age; many older people work during the day and enroll in night courses (average age is 39), while younger students generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. If 70% of the students were day students, it makes sense to ensure that 70% of the sample consisted of day students. Thus, your sample of 200 students would consist of 140 day students and 60 night students. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would therefore be more secure.

Convenience Sampling

Unfortunately, it is often impractical or impossible to study a truly random sample. Much of the time, in fact, studies are conducted with whoever is willing or available to be a research participant – this is commonly referred to as **convenience sampling**. At best, as noted, a researcher tries to study a sample that is not systematically unrepresentative of the population in any known way. For example, suppose a study is about a process that is likely to differ for people of different age groups.

In this situation, the researcher may attempt to include people of all age groups in the study. Alternatively, the researcher would be careful to draw conclusions only about the age group studied. Remember that one of the goals of research is to make conclusions about the population from the sample results. An unbiased random sample and a representative sample are important when drawing conclusions from the results of a study.

“WEIRD” Culture Samples

Psychologists have been guilty of largely recruiting samples of convenience from the thin slice of humanity—students—found at universities and colleges (Sears, 1986). This presents a problem when trying to assess the social mechanics of the public at large. Aside from being an overrepresentation of young, middle-class Caucasians, college students may also be more compliant and more susceptible to attitude change, have less stable personality traits and interpersonal relationships, and possess stronger cognitive skills than samples reflecting a wide range of age and experience (Peterson & Merunka, 2014; Visser, Krosnick, & Lavrakas, 2000). Put simply, these traditional samples (college students) may not be sufficiently representative of the broader population. Furthermore, considering that 96% of participants in psychology studies come from western, educated, industrialized, rich, and democratic countries (so-called **WEIRD cultures**; Henrich, Heine, & Norenzayan, 2010), and that the majority of these *are also psychology students*, the question of non-representativeness becomes even more serious. Of course, when studying a basic cognitive process (like working memory capacity) or an aspect of social behavior that appears to be fairly universal (e.g., even cockroaches exhibit social facilitation), a non-representative sample may not be a big deal. Over time research has repeatedly demonstrated the important role that

individual differences (e.g., personality traits, cognitive abilities, etc.) and culture (e.g., individualism versus collectivism) play in shaping social behavior. For instance, even if we only consider a tiny sample of research on aggression, we know that narcissists are more likely to respond to criticism with aggression (Bushman & Baumeister, 1998); conservatives, who have a low tolerance for uncertainty, are more likely to prefer aggressive actions against those considered to be 'outsiders' (de Zavala et al., 2010); countries, where men hold the bulk of the power in society, have higher rates of physical aggression directed against female partners (Archer, 2006); and males from the southern part of the United States are more likely to react with aggression following an insult (Cohen et al., 1996).

Why does random sampling work?

Below is an example showing how many credit hours students are currently enrolled in at a community college. This data represents the entire population of interest, all students currently enrolled in classes at Chandler-Gilbert Community College. Let's say we randomly selected one student out of the population and asked them how many credit hours they are currently taking. How likely would it be for this one student to represent the entire population? This is the first line showing 1 student reported taking 12 hours while the average credit hours for a CGCC student was 8 (population average).

| <u>Sample size (n)</u> | <u>Sample average</u> | <u>Population average</u> | <u>Difference between Sample & Population</u> |
|------------------------|-----------------------|---------------------------|---|
| 1 | 12 | 8 | 4 |
| 2 | 15 | 8 | 7 |
| 5 | 9.8 | 8 | 1.8 |
| 25 | 9.5 | 8 | 1.5 |
| 250 | 8.3 | 8 | 0.3 |
| 2500 | 7.9 | 8 | 0.1 |

The larger the sample size, the more closely it represents the population.

As we can see from this activity, the larger our sample is, the more accurately it will represent the population from which it was drawn. This brings up a very important rule in research design. The larger the sample size is, the more accurately the sample will represent the population from which it was drawn. Also, if you are comparing groups, consider that the more diverse, or variable, individuals in each group are, the larger the sample needs to be to detect real differences between groups. We will further dive into the importance of sample sizes with inferential statistics, but for now, consider that the larger the sample, the more likely the researcher will represent the population.

Type of Research Designs

Research studies come in many forms, and, just like with the

different types of data we have, different types of studies tell us different things. The choice of research design is determined by the research question and the logistics involved. Though a complete understanding of different research designs is the subject for at least one full class, if not more, a basic understanding of the principles is useful here. There are three types of research designs we will discuss: non-experimental, quasi-experimental, and random experimental.

Non-Experimental Designs

Non-experimental research (sometimes called correlational research) involves observing things as they occur naturally and recording our observations as data. In **observational studies**, information is gathered from observing. This could include self-report as well as interviews.

Consider this example: A data scientist wants to know if there is a relation between how conscientious a person is and whether that person is a good employee. She hopes to use this information to predict the job performance of future employees by measuring their personality when they are still job applicants. She randomly samples volunteer employees from several different companies, measuring their conscientiousness and having their bosses rate their performance on the job. She analyzes this data to find a relation. Conscientiousness is a person-based variable that researcher must gather data from employees as they are in order to find a relation between her variables.

This type of research design cannot establish causality, it can still be quite useful. If the relation between conscientiousness and job performance is consistent, then it doesn't necessarily

matter is conscientiousness causes good performance or if they are both caused by something else – she can still measure conscientiousness to predict future performance. Additionally, these studies have the benefit of reflecting reality as it actually exists since we as researchers do not change anything.

Experimental Designs

If we want to know if a change in one variable causes a change in another variable, we must use a true experiment. *A true experiment is an experimental design with random assignment.* In an **experimental design** a researcher assigns or manipulates, the group's participants will be in. Further, each participant is **randomly assigned** to a group. If there is no random assignment, the experiment can not have cause-effect conclusions.

Types of Variables in an Experiment

When conducting research, experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is “type of antidepressant.” When a variable is manipulated by an experimenter, it is called an **independent variable**. The experiment seeks to determine the effect of the independent variable on relief from depression. In this example, relief from depression is called a **dependent variable**. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

To understand what this means, let's look at an example: A clinical researcher wants to know if a newly developed drug is effective in treating the flu. Working with collaborators at several local hospitals, she

randomly samples 40 flu patients and randomly assigns each one to one of two conditions: Group A receives the new drug and Group B received a placebo. She measures the symptoms of all participants after 1 week to see if there is a difference in symptoms between the groups.

In the example, the *independent variable* is the drug treatment; we manipulate it into 2 levels: new drug or placebo. Without the researcher administering the drug (i.e. manipulating the independent variable), there would be no difference between the groups. Each person, after being randomly sampled to be in the research, was then randomly assigned to one of the 2 groups. That is, random sampling and random assignment are *not* the same thing and cannot be used interchangeably. *For research to be a true experiment, random assignment must be used.* For research to be representative of the population, random sampling must be used. The use of both techniques helps ensure that there are no systematic differences between the groups, thus eliminating the potential for sampling bias. The *dependent variable* in the example is flu symptoms. Barring any other intervention, we would assume that people in both groups, on average, get better at roughly the same rate. Because there are no systematic differences between the 2 groups, if the researcher does find a difference in symptoms, she can confidently attribute it to the effectiveness of the new drug.

Can you identify the independent and dependent variables?

Example #1: Can blueberries slow down

aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month- old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder (randomly assigned). After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

- What is the independent variable? (dietary supplement: none, blueberry, strawberry, and spinach)
- What are the dependent variables? (memory test and motor skills test)

Example #2: Does beta-carotene protect against cancer? Beta-carotene supplements have been thought to protect against cancer. However, a study published in the Journal of the National Cancer Institute suggests this is false. The study was conducted with 39,000 women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a placebo, and their health was studied over their lifetime. Cancer rates for women taking the beta-carotene supplement did not differ

systematically from the cancer rates of those women taking the placebo.

- What is the independent variable? (supplements: beta-carotene or placebo)
- What is the dependent variable? (occurrence of cancer)

Example #3: How bright is right? An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of the following car to realize that the car in front is stopping and to hit the brakes.

- What is the independent variable? (brightness of brake lights)
- What is the dependent variable? (time to hit brakes)

Levels of an Independent Variable

In order to establish that one variable must cause a change in another variable and so a researcher will likely use two groups or levels in order to observe the changes and make comparisons.

- **Experimental (treatment) group** is the group who are exposed to the independent variable (or the manipulation) by the researcher; the experimental group represents the treatment group.
- **Control group** is the group who are not exposed to the treatment variable; the control group serves as the comparison group.

If an experiment compares an experimental treatment group with a control group, then the independent variable (type of treatment) has two levels: experimental and control. Further, if an experiment were comparing five types of diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions. *Another term for levels for the independent variable is groups, treatments, or conditions.*

Scores from the experimental group are compared to scores in the control group and if there is a systematic difference between groups then there is evidence of a relationship between variables. Let's use our earlier example of stress as a way to illustrate the experimental method. Let's assume that a researcher examining stress wants to test the impact of a stress reduction program on the stress levels of students and recruits 100 students to participate. Students are randomly assigned to either the experimental group or the control group. The experimental group participates in the stress reduction program but the control group does not. The stress-reduction program is the independent variable and stress level is the dependent variable. At the end of the training program each group, the experimental group, and the control group complete a stress test, and the scores are compared. If the stress reduction program worked, then the stress levels for the experimental group should be lower than the stress levels for the control group.

Quasi-Experimental Designs

Quasi-experimental research involves getting as close as possible to the conditions of a true experiment when we cannot meet all requirements. Specifically, a **quasi-experiment** involves manipulating the independent variable but *not* randomly assigning people to groups. There are several

reasons this might be used. First, it may be unethical to deny potential treatment to someone if there is good reason to believe it will be effective and that the person would unduly suffer if they did not receive it. Alternatively, it may be impossible to randomly assign people to groups.

Consider the following example: A professor wants to test out a new teaching method to see if it improves student learning. Because he is teaching two sections of the same course, he decides to teach one section the traditional way and the other section using the new method. At the end of the semester, he compares the grades on the final for each class to see if there is a difference.

In this example, the professor has manipulated his teaching method, which is the independent variable, hoping to find a difference in student performance, the dependent variable. *However, because students enroll in courses, he cannot randomly assign the students to a particular group, thus precluding using a true experiment to answer his research question.* Because of this, we cannot know for sure that there are no systematic differences between the classes other than teaching style and therefore cannot determine causality.

Extraneous and Confounding Variables

Sometimes in a research study things happen that make it difficult for a researcher to determine whether the independent variable caused the change in the dependent variable. These have special names.

- An **extraneous variable** is something that occurs in the environment or happens to the participants that unintentionally (accidentally) influences the outcome of the study. An extraneous variable affects everyone in a study. In an experiment on the effect of expressive writing on health, for example, extraneous variables would include participant variables (individual differences) such as their

writing ability, their diet, and their shoe size. They would also include situation or task variables such as the time of day when participants write, whether they write by hand or on a computer, and the weather. Extraneous variables pose a problem because many of them are likely to have some effect on the dependent variable. For example, participants' health will be affected by many things other than whether or not they engage in expressive writing. This can make it difficult to separate the effect of the independent variable from the effects of the extraneous variables, which is why it is important to control extraneous variables by holding them constant.

- A **confounding variable** is a type of extraneous variable that changes at the same time as the independent variable, making it difficult to discern which one is causing changes in the dependent variable.

Working with data

What are data?

The first important point about data is that data *are* – meaning that the word “data” is plural (though some people disagree with me on this). You might also wonder how to pronounce “data” – I say “day-tah”, but I know many people who say “dah-tah”, and I have been able to remain friends with them in spite of this. Now, if I heard them say “the data is” then that would be a bigger issue...

Operationalizing Variables

We need to have specifically defined how we are measuring our construct or our variable. The act of defining how to measure your data is to **operationalize**. Some variables are easier to define, like height or weight. I can measure height in inches tall or weight in pounds. Some other variables can be more open to measurement, like happiness or love. How would I measure happiness? Would I simply ask are you happy (yes or no)? Would I use a questionnaire for a self-report measure? Would I rate individuals from observing them for happiness? Would I ask their partner, teacher, parent, best friend about the person's happiness? Researchers' decisions on how to measure data is an important factor and helps to determine what kind of data is being used.



How would you measure happiness in a research study? [Image Source](#)

Qualitative and Quantitative Variables

Data are composed of *variables*, where a variable reflects a unique measurement or quantity. An important distinction between variables is between qualitative variables and quantitative variables. **Qualitative variables** are those that

express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. Qualitative means that they describe a quality rather than a numeric quantity. *Qualitative variables are sometimes referred to as categorical variables.* For qualitative variables, response options are usually limited or fixed to a set of possible values. Assigning a person, animal or event to a category is done on the basis of some qualitative property. For example, in my stats course, I generally give an introductory survey, both to obtain data to use in class and to learn more about the students. One of the questions that I ask is “What is your favorite food?”, to which some of the answers have been: blueberries, chocolate, tamales, pasta, pizza, and mango. Those data are not intrinsically numerical; we could assign numbers to each one (1=blueberries, 2=chocolate, etc), but we would just be using the numbers as labels *rather than as real numbers*.

Personality type, gender, and shirt sizes are all categorical, or qualitative, variables. The values of a qualitative variable do not necessarily imply order and do not produce numerical responses or use real numbers. For example, there is an order to shirt size but shirt size is categorical and not number based. Another example is postal Zip Code data. Those numbers are represented as integers, but they don’t actually refer to a numeric scale; each zip code basically serves as a label or category representing a different region. Because this data is not using real numbers, what we do with those numbers is constrained; for example, it wouldn’t make sense to compute the average of those numbers.

More commonly in statistics we will work with *quantitative* data, meaning data that are numerical. For example, here Table 1 shows the results from another question that I ask in my introductory class, which is “Why are you taking this class?”

Table 1: Counts of the prevalence of different responses to the question

Why are you taking this class?

It fulfills a degree plan requirement

It fulfills a General Education Breadth Requirement

It is not required but I am interested in the topic

Other

Note that the students' answers were qualitative, but we generated a quantitative summary of them by counting how many students gave each response. **Quantitative variables** are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

Experimental studies can involve qualitative and quantitative data. In the study on the effect of diet discussed previously, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

Discrete and Continuous Variables

Variables such as number of children in a household are called **discrete variables** since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as "time to respond to a question" are **continuous variables** since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds. Of course, the practicalities

of measurement preclude most measured variables from being truly continuous.

Levels of Measurement

Numbers mean different things in different situations. Consider three answers below that appear to be the same, but they really are not. All three questions pertain to a running race that you just finished. The three 5s all look the same. However, the three variables (identification number, finish place, and time) are quite different. Because of these different variables, the way we interpret 5 is unique for each variable.

- What number were you wearing in the race? 5
- What place did you finish in? 5
- How many minutes did it take you to finish the race?

To illustrate the difference, consider your friend who also ran the race. Their answers to the same three questions were 10, 10, and 10. If we take the first question by itself and know that you had a score of 5, and your friend had a score of 10, what can we conclude? We can conclude that your race identification number is different from your friend's number. That is all we can conclude. On the second question, with scores of 5 and 10, what can we conclude regarding the place you and your friend finished in the race? We can state that you were faster than your friend in the race and, of course, that your finishing places are different. Comparing the 5 and 10 on the third question, what can we conclude? We could state that you ran the race twice as fast as your friend, you ran the race faster than your friend and that your time was different than your friend's time. The point of this discussion is to demonstrate the ***relationship between the questions we ask, and what the answers to those questions can tell us*** . Chances are, much of your past

experience with numbers has been with pure numbers or with measurements such as time, length, and amount. “Four is twice as much as two” is true for the pure numbers themselves and for time, length, and amount –but this statement would not be true for finish places in a race. Fourth place is not twice anything in relation to 2nd place. Fourth place is not twice as slow or twice as far behind the 2nd place runner. The types of descriptive and inferential statistics we can use depend on the type of variable measured. Remember, a variable is defined as a characteristic we can measure that can assume more than one value.

For statistical analysis, exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stopwatch. Stopwatches are of no use, of course, when it comes to measuring someone’s attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like “very favorable,” “somewhat favorable,” etc.). For a dependent variable such as “favorite color,” you can simply note the color-word (like “red”) that the subject offers.

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. The psychologist S. S. Stevens suggested that scores can be assigned to individuals so that they communicate more or less quantitative information about the variable of interest (Stevens, 1946). Stevens actually suggested four different levels of measurement(which he called “scales of measurement”) that correspond to four different levels of quantitative information that can be communicated by a set of scores. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called “scale types,” or just “scales,” and are described in this section.

Nominal scales

When measuring using a nominal scale, one simply names or categorizes responses. Gender, handedness, favorite color, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favorite color, there is no sense in which green is placed “ahead of” blue. Responses are merely categorized. Nominal scales embody the lowest level of measurement.

Ordinal scales

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either “very dissatisfied,” “somewhat dissatisfied,” “somewhat satisfied,” or “very satisfied.” The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the dependent variable. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens. Such an assertion reflects the first person's use of a verbal label that comes later in the list than the label chosen by the second person.

On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale,

for example, the difference between the responses “very dissatisfied” and “somewhat dissatisfied” is probably not equivalent to the difference between “somewhat dissatisfied” and “somewhat satisfied.” Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction.

Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

What if the researcher had measured satisfaction by asking consumers to indicate their level of satisfaction by choosing a number from one to four? Would the difference between the responses of one and two necessarily reflect the same difference in satisfaction as the difference between the responses two and three? The answer is No. Changing the response format to numbers does not change the meaning of the scale. We still are in no position to assert that the mental step from 1 to 2 (for example) is the same as the mental step from 3 to 4.

(Equal) Interval scales

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and

90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name “zero.” The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label “zero” is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the “zero” label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is “twice as hot” as 40 degrees. Such a claim would depend on an arbitrary decision about where to “start” the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

Ratio scales (Absolute zero)

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates

the absence of the quantity being measured. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

The Fahrenheit scale for temperature has an arbitrary zero point and is therefore not a ratio scale. However, zero on the Kelvin scale is absolute zero. This makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents (or that Bill Gates has a million times more money than you do).

**Digging deeper: What about the number value?
It is important to know what number values
mean. Is the number meaningful or it is a**

category? This section briefly reviews how numbers can be categorized according to meaning.

Binary numbers. The simplest are binary numbers – that is, zero or one. We will often use binary numbers to represent whether something is true or false, or present or absent. For example, I might ask 10 people if they have ever experienced a migraine headache, recording their answers as “Yes” or “No”. It’s often useful to instead use *logical* values, which take the value of either `TRUE` or `FALSE`. This can be especially useful for programming languages to analyze data, since these languages already understand the concepts of `TRUE` and `FALSE`. In fact, most programming languages treat truth values and binary numbers equivalently. The number 1 is equal to the logical value `TRUE`, and the number zero is equal to the logical value `FALSE`.

Integers. Integers are whole numbers with no fractional or decimal part. We most commonly encounter integers when we count things, but they also often occur in psychological measurement. For example, in my introductory survey I administer a set of questions about attitudes towards statistics (such as “Statistics seems very mysterious to me.”), on which the students respond with a number between 1 (“Disagree strongly”) and 7 (“Agree strongly”). Integers are discontinuous.

Real numbers. Most commonly in statistics we work with real numbers, which have a fractional/

decimal part. For example, we might measure someone's weight, which can be measured to an arbitrary level of precision, from kilograms down to micrograms. Real numbers can be discontinuous or continuous.

What level of measurement is used for psychological variables?

Rating scales are used frequently in psychological research. For example, experimental subjects may be asked to rate their level of pain, how much they like a consumer product, their attitudes about capital punishment, their confidence in an answer to a test question. Typically these ratings are made on a 5-point or a 7-point scale. These scales are ordinal scales since there is no assurance that a given difference represents the same thing across the range of the scale. For example, there is no way to be sure that a treatment that reduces pain from a rated pain level of 3 to a rated pain level of 2 represents the same level of relief as a treatment that reduces pain from a rated pain level of 7 to a rated pain level of 6.

In memory experiments, the dependent variable is often the number of items correctly recalled. What scale of measurement is this? You could reasonably argue that it is a ratio scale. First, there is a true zero point; some subjects may get no items correct at all. Moreover, a difference of one represents a difference of one item recalled across the entire scale. It is certainly valid to say that someone who recalled 12 items recalled twice as many items as someone who recalled only 6 items.

But number-of-items recalled is a more complicated case

than it appears at first. Consider the following example in which subjects are asked to remember as many items as possible from a list of 10. Assume that (a) there are 5 easy items and 5 difficult items, (b) half of the subjects are able to recall all the easy items and different numbers of difficult items, while (c) the other half of the subjects are unable to recall any of the difficult items but they do remember different numbers of easy items. Some sample data are shown below.

| Subject | Easy Items | | | | | Difficult Items | | | | | Score |
|---------|------------|---|---|---|---|-----------------|---|---|---|---|-------|
| A | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| B | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| C | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 7 |
| D | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 8 |

Let’s compare (i) the difference between Subject A’s score of 2 and Subject B’s score of 3 and (ii) the difference between Subject C’s score of 7 and Subject D’s score of 8. The former difference is a difference of one easy item; the latter difference is a difference of one difficult item. Do these two differences necessarily signify the same difference in memory? We are inclined to respond “No” to this question since only a little more memory may be needed to retain the additional easy item whereas a lot more memory may be needed to retain the additional hard item. *The general point is that it is often inappropriate to consider psychological measurement scales as either interval or ratio. You will often see in statistical software that the distinction is between nominal, ordinal, and interval/ratio.*

Consequences of level of measurement

Why are we so interested in the type of scale that measures a dependent variable? The crux of the matter is the relationship between the variable’s level of measurement and the statistics that can be meaningfully computed with that variable. For example, consider a hypothetical study in which 5 children are asked to choose their favorite color from blue, red, yellow, green, and purple. The researcher codes the results as follows:

| Color | Code |
|--------|------|
| Blue | 1 |
| Red | 2 |
| Yellow | 3 |
| Green | 4 |
| Purple | 5 |

This means that if a child said her favorite color was “Red,” then the choice was coded as “2,” if the child said her favorite color was “Purple,” then the response was coded as 5, and so forth. Consider the following hypothetical data:

| Subject | Color | Code |
|---------|--------|------|
| 1 | Blue | 1 |
| 2 | Blue | 1 |
| 3 | Green | 4 |
| 4 | Green | 4 |
| 5 | Purple | 5 |

Each code is a number, so nothing prevents us from computing the average code assigned to the children. The average happens to be 3, but you can see that it would be senseless to conclude that the average favorite color is yellow (the color with a code of 3). Such nonsense arises because favorite color is a nominal scale, and taking the average of its numerical labels is like counting the number of letters in the name of a snake to see how long the beast is.

Does it make sense to compute the mean of numbers measured on an ordinal scale? This is a difficult question, one that statisticians have debated for decades. The prevailing (but by no means unanimous) opinion of statisticians is that for almost all practical situations, the mean of an ordinal-measured variable is a meaningful statistic. However, there are extreme situations in which computing the mean of an ordinal-measured variable can be very misleading.

What makes a good measurement?

In many fields such as psychology, the thing that we are measuring is not a physical feature, but instead is an unobservable theoretical concept, which we usually refer to as a *construct*. For example, let's say that I want to test how well you understand the distinction between the different types of numbers described above. I could give you a pop quiz that would ask you several questions about these concepts and count how many you got right. This test might or might not be a good measurement of the construct of your actual knowledge – for example, if I were to write the test in a confusing way or use language that you don't understand, then the test might suggest you don't understand the concepts when really you do. On the other hand, if I give a multiple-choice test with very obvious wrong answers, then

you might be able to perform well on the test even if you don't actually understand the material.

It is usually impossible to measure a construct without some amount of error. In the example above, you might know the answer, but you might misread the question and get it wrong. In other cases, there is error intrinsic to the thing being measured, such as when we measure how long it takes a person to respond on a simple reaction time test, which will vary from trial to trial for many reasons. We generally want our measurement error to be as low as possible, which we can achieve either by improving the quality of the measurement (for example, using a better time to measure reaction time), or by averaging over a larger number of individual measurements.

Sometimes there is a standard against which other measurements can be tested, which we might refer to as a “gold standard” – for example, measurement of sleep can be done using many different devices (such as devices that measure movement in bed), but they are generally considered inferior to the gold standard of polysomnography (which uses measurement of brain waves to quantify the amount of time a person spends in each stage of sleep). Often the gold standard is more difficult or expensive to perform, and the cheaper method is used even though it might have greater error.

When we think about what makes a good measurement, we usually distinguish two different aspects of a good measurement: it should be *reliable*, and it should be *valid*.

Reliability

Reliability refers to the consistency of our measurements. One common form of reliability, known as “test-retest reliability”, measures how well the measurements agree if the same measurement is performed twice. For example, I might give

you a questionnaire about your attitude towards statistics today, repeat this same questionnaire tomorrow, and compare your answers on the two days; we would hope that they would be very similar to one another, unless something happened in between the two tests that should have changed your view of statistics (like reading this book!).

Another way to assess reliability comes in cases where the data include subjective judgments. For example, let's say that a researcher wants to determine whether a treatment changes how well an autistic child interacts with other children, which is measured by having experts watch the child and rate their interactions with the other children. In this case we would like to make sure that the answers don't depend on the individual rater — that is, we would like for there to be high *inter-rater reliability*. This can be assessed by having more than one rater perform the rating, and then comparing their ratings to make sure that they agree well with one another.

Reliability is important if we want to compare one measurement to another, because the relationship between two different variables can't be any stronger than the relationship between either of the variables and itself (i.e., its reliability). This means that an unreliable measure can never have a strong statistical relationship with any other measure. For this reason, researchers developing a new measurement (such as a new survey) will often go to great lengths to establish and improve its reliability.

A: Reliable and valid



B: Unreliable but valid



C: Reliable but invalid



D: Unreliable and invalid



Figure 1: A figure demonstrating the distinction between reliability and validity, using shots at a bullseye. Reliability refers to the consistency of location of shots, and validity refers to the accuracy of the shots with respect to the center of the bullseye.

Validity

Reliability is important, but on its own it's not enough: After all, I could create a perfectly reliable measurement on a personality test by re-coding every answer using the same number, regardless of how the person actually answers. We want our measurements to also be *valid* — that is, we want to make sure that we are actually measuring the construct that we think

we are measuring (Figure 1). There are many different types of validity that are commonly discussed; we will focus on three of them.

Face validity. Does the measurement make sense on its face? If I were to tell you that I was going to measure a person's blood pressure by looking at the color of their tongue, you would probably think that this was not a valid measure on its face. On the other hand, using a blood pressure cuff would have face validity. This is usually a first reality check before we dive into more complicated aspects of validity.

Construct validity. Is the measurement related to other measurements in an appropriate way? This is often subdivided into two aspects. *Convergent validity* means that the measurement should be closely related to other measures that are thought to reflect the same construct. Let's say that I am interested in measuring how extroverted a person is using either a questionnaire or an interview. Convergent validity would be demonstrated if both of these different measurements are closely related to one another. On the other hand, measurements thought to reflect different constructs should be unrelated, known as *divergent validity*. If my theory of personality says that extraversion and conscientiousness are two distinct constructs, then I should also see that my measurements of extraversion are *unrelated* to measurements of conscientiousness.

Predictive validity. If our measurements are truly valid, then they should also be predictive of other outcomes. For example, let's say that we think that the psychological trait of sensation seeking (the desire for new experiences) is related to risk taking in the real world. To test for predictive validity of a measurement of sensation seeking, we would test how well scores on the test predict scores on a different survey that measures real-world risk taking.

Critical Evaluation of Statistical Results

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include:

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions.
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of a participant to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.

- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Types of Statistical Analyses

Now that we understand the nature of our data, let's turn to the types of statistics we can use to interpret them. As mentioned at the end of chapter 1, there are 2 types of statistics: descriptive and inferential.

Descriptive Statistics

Descriptive statistics are numbers that are used to summarize and describe data. The word "data" refers to the information that has been collected from an experiment, a survey, an historical record, etc. (By the way, "data" is plural. One piece of information is called a "datum.") If we are analyzing birth certificates, for example, a descriptive statistic might be the percentage of certificates issued in New York State, or the average age of the mother. Any other number we choose to compute also counts as a descriptive statistic for the data from

which the statistic is computed. Several descriptive statistics are often used at one time to give a full picture of the data.

Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand. Generalizing from our data to another set of cases is the business of inferential statistics, which you'll be studying in another section. Here we focus on (mere) descriptive statistics.

Some descriptive statistics are shown in Table 2. The table shows the average salaries for various occupations in the United States in 1999.

| Salary 1999 | Salary 2019 | Occupation |
|-------------|-------------|--|
| \$112,760 | \$175,310 | pediatricians |
| \$106,130 | \$155,600 | dentists |
| \$100,090 | \$126,240 | podiatrists |
| \$76,140 | \$97,152 | physicists |
| \$53,410 | \$80,750 | architects |
| \$49,720 | \$78,200 | school, clinical, and counseling psychologists |
| \$47,910 | \$56,640 | flight attendants |
| \$39,560 | \$59,670 | elementary school teachers |
| \$38,710 | \$65,170 | police officers |
| \$18,980 | \$28, 040 | floral designers |

Table 2. Average salaries for various occupations in 1999 and 2019 (median salaries reported by Bureau of Labor Statistics).

Descriptive statistics like these offer insight into American society. It is interesting to note, for example, that we pay the people who educate our children and who protect our citizens a great deal less than we pay people who take care of our feet or our teeth.

For more descriptive statistics, consider Table 3. It shows the number of employed single young men to single young women for large metro areas in the US (reported in 2014). From this table we see that men outnumber women most in the San Jose, CA area, and women outnumber men most in the Memphis, TN area. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner between the ages of 25-34 years old! (These data come from Pew Research)

| Highest Ratios of Employed Single Men to Single Women (25-34 y/o) | Men per 100 Women | Lowest Ratios of Employed Single Men to Single Women (25-34 y/o) |
|---|-------------------|--|
| 1. San-Jose-Sunnyvale-Santa Clara, CA | 114 | 1. Memphis, TN-MS |
| 2. Denver-Aurora-Lakewood, CO | 101 | 2. Jacksonville, FL |
| 3. San Diego-Carlsbad, CA | 99 | 3. Detroit-Warren-Livonia, MI |
| 4. Minneapolis-St. Paul-Bloomington, MN-WI | 98 | 4. Charlotte-Concord-Gastonia, NC |
| 5. Seattle-Tacoma-Bellevue, WA | 96 | 5. Philadelphia-Camden-Wilmington, PA-NJ-DE-MD |
| 6. San Francisco-Oakland-Hayward, CA | 93 | 6. Kansas City, MO |
| 7. Washington-Arlington-Alexandria, DC-VA-MD-WV | 92 | 7. Nashville-Davidson--DeS Moines, TN |
| 8. Los Angeles-Long Beach-Anaheim, CA | 91 | 8. Miami-Fort Lauderdale-Boca Raton, FL |
| 9. Pittsburgh, PA | 90 | 9. New Orleans-Metairie, LA |
| 10. Orlando-Kissimmee-Sanford, FL | 90 | 10. Cincinnati, OH |

Table 3. Number of employed, 25-34 year old ratio of men to women in large metro areas of the U.S. (Pew Research, 2014)

These descriptive statistics may make us ponder why there are ratio differences in these large metropolitan areas. You

probably know that descriptive statistics are central to the world of sports. Every sporting event produces numerous statistics such as the shooting percentage of players on a basketball team. For the Olympic marathon (a foot race of 26.2 miles), we possess data that cover more than a century of competition. (The first modern Olympics took place in 1896.) The following table shows the winning times for both men and women (the latter have only been allowed to compete since 1984).

| Women | | | |
|-------|--------------------------|---------|---------|
| Year | Winner | Country | Time |
| 1984 | Joan Benoit | USA | 2:24:52 |
| 1988 | Rosa Mota | POR | 2:25:40 |
| 1992 | Valentina Yegorova | UT | 2:32:41 |
| 1996 | Fatuma Roba | ETH | 2:26:05 |
| 2000 | Naoko Takahashi | JPN | 2:23:14 |
| 2004 | Mizuki Noguchi | JPN | 2:26:20 |
| 2008 | Constantina Dita-Tomescu | Romania | 2:26:44 |
| 2012 | Tiki Gelana | ETH | 2:23:07 |
| 2016 | Jemima Sumgong | Kenya | 2:24:04 |
| 2020 | Peres Jepchirchir | Kenya | 2:27:20 |
| Men | | | |
| Year | Winner | Country | Time |
| 1896 | Spiridon Louis | GRE | 2:58:50 |
| 1900 | Michel Theato | FRA | 2:59:45 |
| 1904 | Thomas Hicks | USA | 3:28:53 |
| 1906 | Billy Sherring | CAN | 2:51:23 |
| 1908 | Johnny Hayes | USA | 2:55:18 |
| 1912 | Kenneth McArthur | S. Afr. | 2:36:54 |

| | | | |
|------|--------------------|-----|---------|
| 1920 | Hannes Kolehmainen | FIN | 2:32:35 |
| 1924 | Albin Stenroos | FIN | 2:41:22 |

| | | | |
|------|---------------------|---------|---------|
| 1928 | Boughra El Ouafi | FRA | 2:32:57 |
| 1932 | Juan Carlos Zabala | ARG | 2:31:36 |
| 1936 | Sohn Kee-Chung | JPN | 2:29:19 |
| 1948 | Delfo Cabrera | ARG | 2:34:51 |
| 1952 | Emil Ztopek | CZE | 2:23:03 |
| 1956 | Alain Mimoun | FRA | 2:25:00 |
| 1960 | Abebe Bikila | ETH | 2:15:16 |
| 1964 | Abebe Bikila | ETH | 2:12:11 |
| 1968 | Mamo Wolde | ETH | 2:20:26 |
| 1972 | Frank Shorter | USA | 2:12:19 |
| 1976 | Waldemar Cierpinski | E.Ger | 2:09:55 |
| 1980 | Waldemar Cierpinski | E.Ger | 2:11:03 |
| 1984 | Carlos Lopes | POR | 2:09:21 |
| 1988 | Gelindo Bordin | ITA | 2:10:32 |
| 1992 | Hwang Young-Cho | S. Kor | 2:13:23 |
| 1996 | Josia Thugwane | S. Afr. | 2:12:36 |
| 2000 | Gezahenge Abera | ETH | 2:10:10 |
| 2004 | Stefano Baldini | ITA | 2:10:55 |
| 2008 | Samuel Wanjiru | Kenya | 2:06:32 |
| 2012 | Stephen Kiprotich | Uganda | 2:08:01 |

| | | | |
|------|----------------|-------|---------|
| 2016 | Eliud Kipchoge | Kenya | 2:08:44 |
| 2020 | Eliud Kipchoge | Kenya | 2:08:38 |

Table 4. Winning Olympic marathon times.

There are many descriptive statistics that we can compute from the data in the table. To gain insight into the improvement in speed over the years, let us divide the men’s times into two pieces, namely, the first 13 races (up to 1952) and the second 13 (starting from 1956). The mean winning time for the first 13 races is 2 hours, 44 minutes, and 22 seconds (written 2:44:22). The mean winning time for the second 13 races is 2:13:18. This is quite a difference (over half an hour). Does this prove that the fastest men are running faster? Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year? We can’t answer this question with descriptive statistics alone. All we can affirm is that the two means are “suggestive.”

Examining Table 4 leads to many other questions. We note that Takahashi (the lead female runner in 2000) would have beaten the male runner in 1956 and all male runners in the first 12 marathons. This fact leads us to ask whether the gender gap will close or remain constant. When we look at the times within each gender, we also wonder how far they will decrease (if at all) in the next century of the Olympics. Might we one day witness a sub-2 hour marathon? The study of statistics can help you make reasonable guesses about the answers to these questions.

It is also important to differentiate what we use to describe populations vs what we use to describe samples. A population is described by a parameter; the parameter is the true value of the descriptive in the population, but one that we can never know for sure. For example, the Bureau of Labor Statistics reports that the average hourly wage of chefs or head cooks

is \$25.66¹. However, even if this number was computed using information from every single chef in the United States (making it a parameter), it would quickly become slightly off as one chef retires and a new chef enters the job market. Additionally, as noted above, there is virtually no way to collect data from every single person in a population. In order to understand a variable, we estimate the population parameter using a sample statistic. Here, the term “statistic” refers to the specific number we compute from the data (e.g. the average), not the field of statistics. A sample statistic is an estimate of the true population parameter, and if our sample is representative of the population, then the statistic is considered to be a good estimator of the parameter.

Even the best sample will be somewhat off from the full population, earlier referred to as sampling bias, and as a result, there will always be a tiny discrepancy between the parameter and the statistic we use to estimate it. This difference is known as sampling error, and, as we will see throughout the course, understanding sampling error is the key to understanding statistics. Every observation we make about a variable, be it a full research study or observing an individual's behavior, is incapable of being completely representative of all possibilities for that variable.

Knowing where to draw the line between an unusual observation and a true difference is what statistics is all about.

Inferential Statistics

Descriptive statistics are wonderful at telling us what our data look like. However, what we often want to understand is how our data behave. What variables are related to other variables?

1. BLS, 5/2020, median value reported--to be explained in chapter 4

Under what conditions will the value of a variable change? Are two groups different from each other, and if so, are people within each group different or similar? These are the questions answered by inferential statistics, and inferential statistics are how we generalize from our sample back up to our population. Units 2 and 3 are all about inferential statistics, the formal analyses and tests we run to make conclusions about our data.

For example, we will learn how to use a t statistic to determine whether people change over time when enrolled in an intervention. We will also use an F statistic to determine if we can predict future values on a variable based on current known values of a variable. There are many types of inferential statistics, each allowing us insight into a different behavior of the data we collect. This course will only touch on a small subset (or a *sample*) of them, but the principles we learn along the way will make it easier to learn new tests, as most inferential statistics follow the same structure and format.

Mathematical Notation

As noted above, statistics is not math. It does, however, use math as a tool. Many statistical formulas involve summing numbers. Fortunately there is a convenient notation for expressing summation. This section covers the basics of this summation notation.

Let's say we have a variable X that represents the weights (in grams) of 4 grapes:

| Grape | X |
|-------|-----|
| 1 | 4.6 |
| 2 | 5.1 |
| 3 | 4.9 |
| 4 | 4.4 |

$$\sum_{i=1}^4 X_i$$

We label Grape 1's weight X_1 , Grape 2's weight X_2 , etc. The following formula means to sum up the weights of the four grapes:

The Greek letter Σ indicates summation. The "i = 1" at the bottom indicates that the summation is to start with X_1 and the 4 at the top indicates that the summation will end with X_4 . The " X_i " indicates that X is the variable to be summed as i goes from 1 to 4. Therefore,

$$\sum_{i=1}^4 X_i = X_1 + X_2 + X_3 + X_4 = 4.6 + 5.1 + 4.9 + 4.4 = 19$$

The symbol

$$\sum_{i=1}^3 X_i$$

indicates that only the first 3 scores are to be summed. The index variable i goes from 1 to 3.

When all the scores of a variable (such as X) are to be summed, it is often convenient to use the following abbreviated notation:

$$\sum X$$

Thus, when no values of i are shown, it means to sum all the values of X.

Many formulas involve squaring numbers before they are summed. This is indicated as

$$\sum X^2 = 4.6^2 + 5.1^2 + 4.9^2 + 4.4^2$$
$$= 21.16 + 26.01 + 24.01 + 19.36 = 90.54$$

Notice that:

$$(\sum X)^2 \neq \sum X^2$$

because the expression on the left means to sum up all the values of X and then square the sum ($19^2 = 361$), whereas the expression on the right means to square the numbers and then sum the squares (90.54, as shown).

Some formulas involve the sum of cross products. Below are the data for variables X and Y. The cross products (XY) are shown in the third column. The sum of the cross products is $3 + 4 + 21 = 28$.

| | X | Y | XY |
|---|---|----|----|
| 1 | 3 | 3 | |
| 2 | 2 | 4 | |
| 3 | 7 | 21 | |

In summation notation, this is written as:

$$\sum XY = 28$$

Three key concepts for statistical formulas:

1. Perform summation in the correct order following the order of operations (PEMDAS).

2. Typically we will use a set of scores for the mathematical operations/formulas used in statistics.
3. Each operation, except for summation, creates a new column of numbers (we will see this in action in chapter 4). Summation adds up the sum for the column and is typically seen as the last row.

Learning Objectives

Having read this chapter, you should be able to:

- Familiarize with terminology and special notations of statistics.
- Differentiate and identify different types of research design.
- Differentiate and identify different types of sampling.
- Distinguish between different types of variables and given examples of each of these kinds of variables.
- Distinguish between the concepts of reliability and validity and apply each concept to a particular dataset.
- Understand and list the three key concepts for using summation.

Exercises – Ch. 2

1. In your own words, describe why we study statistics.

2. For each of the following, determine if the variable is continuous or discrete:
 1. Time taken to read a book chapter
 2. Favorite food
 3. Cognitive ability
 4. Temperature
 5. Letter grade received in a class
3. For each of the following, determine the level of measurement:
 1. T-shirt size
 2. Time taken to run 100 meter race
 3. First, second, and third place in 100 meter race
 4. Birthplace
 5. Temperature in Celsius
4. What is the difference between a population and a sample? Which is described by a parameter and which is described by a statistic?
5. What is sampling bias? What is sampling error?
6. What is the difference between a simple random sample and a stratified random sample?
7. What are the two key characteristics of a true experimental design?
8. When would we use a quasi-experimental design?
9. Use the following dataset for the computations below:

| X | Y |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 5 | 5 |
| 7 | 1 |

Computations to use for the above data set:

1. ΣX

2. ΣY^2
 3. ΣXY
 4. $(\Sigma Y)^2$
10. What are the most common measures of central tendency and spread?

Answers to Odd-Numbered Exercises – Ch. 2

1. Your answer could take many forms but should include information about objectively interpreting information and/or communicating results and research conclusions

3. For each of the following, determine the level of measurement:

1. Ordinal
2. Ratio
3. Ordinal
4. Nominal
5. Interval

5. Sampling bias is the difference in demographic characteristics between a sample and the population it should represent. Sampling error is the difference between a population parameter and sample statistic that is caused by random chance due to sampling bias.

7. Random assignment to treatment conditions and manipulation of the independent variable

9. Use the following dataset for the computations below:

1. 15
2. 36
3. 39
4. 100

3. Chapter 3:

Describing Data using Distributions and Graphs

Statistics that are used to organize and summarize the information so that the researcher can see what happened during the research study and can also communicate the results to others are called *descriptive statistics*. Let us assume that the data are quantitative and consist of scores on one or more variables for each of several study participants. Although in most cases the primary research question will be about one or more statistical relationships between variables, it is also important to describe each variable individually. We will look at some of the most common techniques for describing single variables including:

- Frequency distributions
- Measures of Central Tendency
- Measures of Dispersion

The first step in understanding data is using tables, charts, graphs, plots, and other visual tools to see what our data look like. This is known as data visualization.

We will begin with frequency distributions which are visual representations and include tables and graphs. We will conclude with some tips for making graphs some principles for good data visualization!

Data Visualization

On January 28, 1986, the Space Shuttle Challenger exploded 73 seconds after takeoff, killing all 7 of the astronauts on board. As when any such disaster occurs, there was an official investigation into the cause of the accident, which found that an O-ring connecting two sections of the solid rocket booster leaked, resulting in failure of the joint and explosion of the large liquid fuel tank (see figure 1).¹

The investigation found that many aspects of the NASA decision-making process were flawed, and focused in particular on a meeting between NASA staff and engineers from Morton Thiokol, a contractor who built the solid rocket boosters. These engineers were particularly concerned because the temperatures were forecast to be very cold



on the morning of the launch, and they had data from previous launches showing that performance of the O-rings was compromised at lower temperatures. In a meeting on the evening before the launch, the engineers presented their data to the NASA managers, but were unable to convince them to

1. Figure 1: An image of the solid rocket booster leaking fuel, seconds before the explosion. The small flame visible on the side of the rocket is the site of the O-ring failure. By NASA (Great Images in NASA Description) [Public domain], via Wikimedia Commons

postpone the launch. Their evidence was a set of hand-written slides showing numbers from various past launches.

The visualization expert Edward Tufte has argued that with a proper presentation of all of the data, the engineers could have been much more persuasive. In particular, they could have shown a figure like the one in Figure 2, which highlights two important facts. First, it shows that the amount of O-ring damage (defined by the amount of erosion and soot found outside the rings after the solid rocket boosters were retrieved from the ocean in previous flights) was closely related to the temperature at takeoff. Second, it shows that the range of forecasted temperatures for the morning of January 28 (shown in the shaded area) was well outside of the range of all previous launches. While we can't know for sure, it seems at least plausible that this could have been more persuasive.

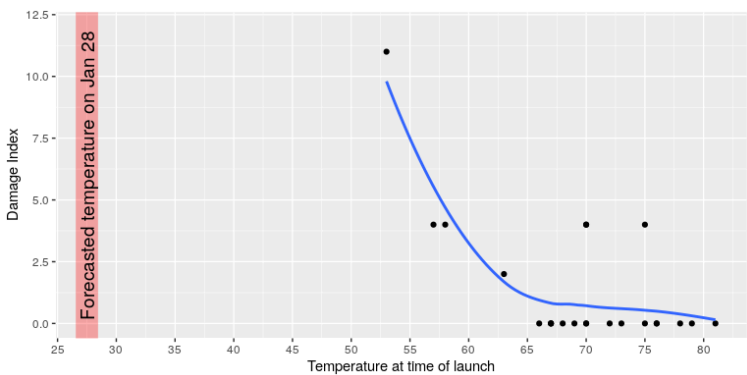


Figure 2: A replotting of Tufte's damage index data. The line shows the trend in the data, and the shaded patch shows the projected temperatures for the morning of the launch.

Graphing Qualitative & Quantitative Variables

We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. Qualitative variables can be summarized by frequency (how often) and researchers can then use frequency tables and bar charts to show frequencies for categorized responses, but we are limited in graphing them due to the data not be numerically based. *The key point about the qualitative data is they do not come with a pre-established ordering (the way numbers are ordered).*

We are focused on *quantitative variables*. Quantitative data, such as a person's weight, are naturally ordered with respect to people of different weights. Often we wish to know if there are any scores that might look a bit out of place. A **frequency distribution** is a way to take a disorganized set of scores and places them in order from highest to lowest and at the same time grouping everyone with the same score. Frequency distributions can help researchers identify outliers. An **outlier** is an observation of data that does not fit the rest of the data. An *outlier* is sometimes called an *extreme value*. When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening.

Frequency Tables

All of the graphical methods shown in this section are derived from **frequency tables**. Table 1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for “none” of 0.17 = 85/500.

| Previous Ownership | Frequency | Relative Frequency |
|--------------------|-----------|--------------------|
| None | 85 | 0.17 |
| Windows | 60 | 0.12 |
| Macintosh | 355 | 0.71 |
| Total | 500 | 1 |

Table 1. Frequency Table for the iMac Data.

Below is a table (Table 2) showing a hypothetical distribution of scores on the Rosenberg Self-Esteem Scale for a sample of 40 college students. The Rosenberg Self-Esteem Scale is one way to operationalize (define) self-esteem in a quantitative way. Participants rate each of the 10-items from strongly disagree to strongly agree. All items are then scored yielding an overall self-esteem score that would be a numerical value to represent one’s self-esteem.

- Column one lists the values of the variable – the possible scores on the Rosenberg scale
- Column two lists the frequency of each score

| <u>Self-Esteem Scores</u> | <u>Frequency</u> |
|----------------------------------|-------------------------|
| 24 | 3 |
| 23 | 5 |
| 22 | 10 |
| 21 | 8 |
| 20 | 5 |
| 19 | 3 |
| 18 | 3 |
| 17 | 0 |
| 16 | 2 |
| 15 | 1 |

Table 2. Frequency Table for Rosenberg Self-Esteem Scale Scores.

Table 2 shows that there were three students who had self-esteem scores of 24, five who had self-esteem scores of 23, and so on. From a frequency table like this, one can quickly see several important aspects of a distribution, including the range of scores (from 15 to 24), the most and least common scores (22 and 17, respectively), and any extreme scores that stand out from the rest.

Considerations

There are a few other points worth noting about frequency tables. First, the levels listed in the first column usually go from the highest at the top to the lowest at the bottom, and they usually do not extend beyond the highest and lowest scores in the data. For example, although scores on the Rosenberg scale can vary from a high of 30 to a low of 0 only includes levels from 24 to 15 because that range includes all the scores in this particular data set. All scores within the data set must

be presented. For example, no one received a score of 17 on the Rosenberg Self-esteem scale; it is still represented in the table.

Additionally, when there are many different scores across a wide range of values, it is often better to create a grouped frequency table, in which the first column lists ranges of values and the second column lists the frequency of scores in each range. In a **grouped frequency table**, the ranges must all be of equal width, and there are usually between five and 15 of them. Finally, frequency tables can also be used for categorical variables, in which case the levels are category labels. The order of the category labels is somewhat arbitrary, but they are often listed from the most frequent at the top to the least frequent at the bottom. Table 3 shows an example for majors where majors is a categorical (nominal) variable.

| Majors | Frequency |
|----------------------|-----------|
| Business | 30 |
| Psychology | 50 |
| Nursing | 102 |
| Nutritional Sciences | 10 |
| Communications | 5 |
| English | 3 |
| Computer Science | 13 |

Table 3. Frequency Table for Majors

Graphs

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are

only a few data values. Statisticians often graph data first to get a picture of the data; then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize quantitative data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this lesson, we will briefly look at bar graphs, histograms, and frequency polygons.

Bar charts

Bar charts can also be used to represent frequencies of different categories. Bar charts may be appropriate for qualitative data (categorical variables) that use a nominal or ordinal scale of measurement. A bar chart of the iMac purchases is shown in Figure 2. Frequencies are shown on the Y- axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category (rather than the percentage of observations in each category as is typical in pie charts).

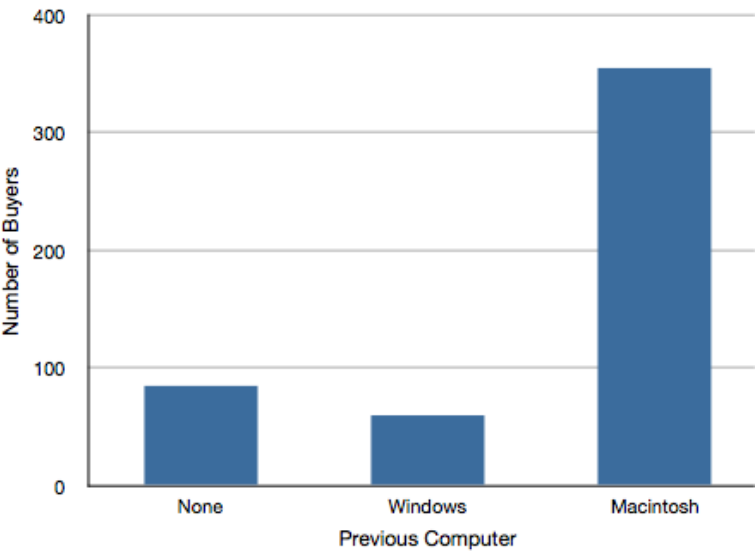


Figure 2. Bar chart of iMac purchases as a function of previous computer ownership.

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the “distributions” of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 3 shows the number of people playing card games at the Yahoo website on a Sunday and on a Wednesday in the spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

Comparing Distributions

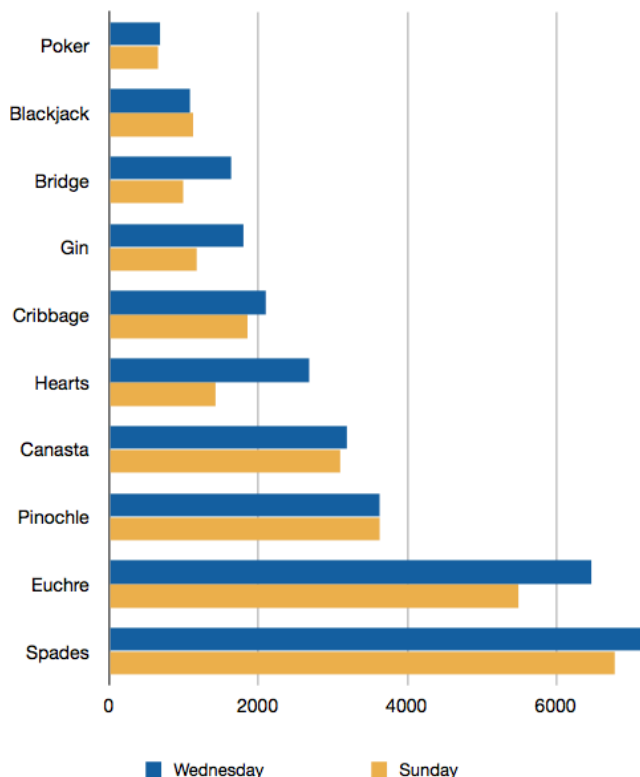
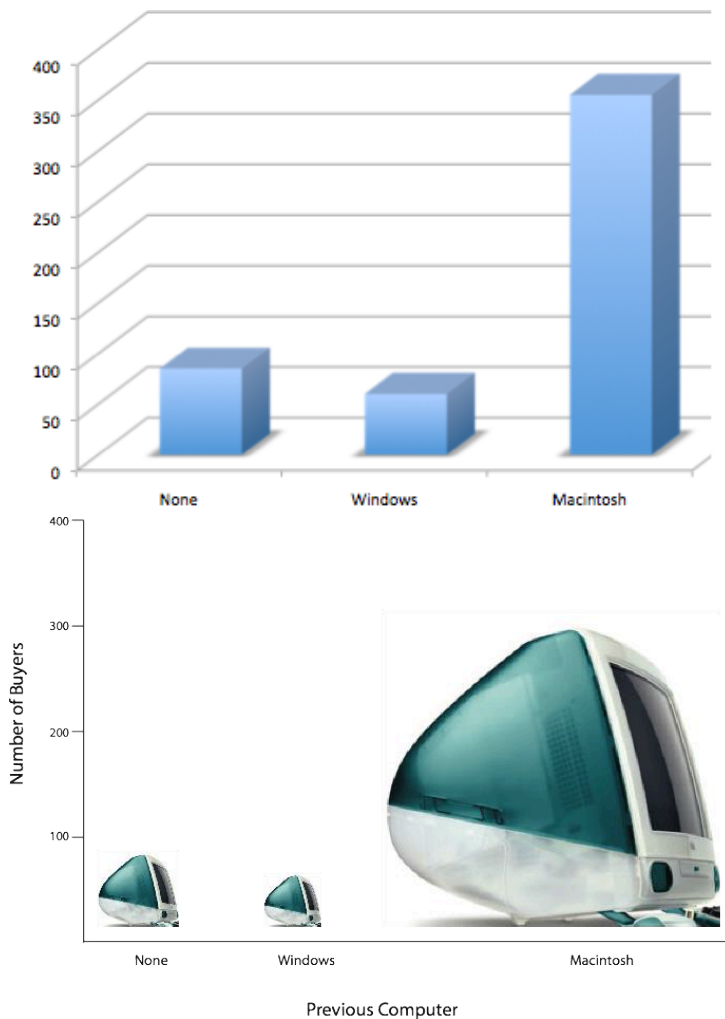


Figure 3. A bar chart of the number of people playing different card games on Sunday and Wednesday.

The bars in Figure 3 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in this chapter.

Some graphical mistakes to avoid with bar charts

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. See the examples below as things not to do! Three-dimensional figures are less clear than 2-d. Further, don't get creative as show below! Use plain bars, as tempting as it is to substitute meaningful images. The MacIntosh is out of proportion to the None and Windows categories. Edward Tufte coined the term "lie factor" to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. If a graphic has a lie factor near 1, then it is appropriately representing the data, whereas lie factors far from one reflect a distortion of the underlying data. The computer monitor bar figure has a lie factor of about 8! He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion-so just keep it simple with plain bars!



Figures 4 & 5. A three-dimensional version of Figure 2 and a redrawing of Figure 2 with disproportionate bars.

Here is another example, Figure 3.6 (created using Microsoft Excel) plots the relative popularity of different religions in the

United States. There are at least three things wrong with this figure -can you identify them?

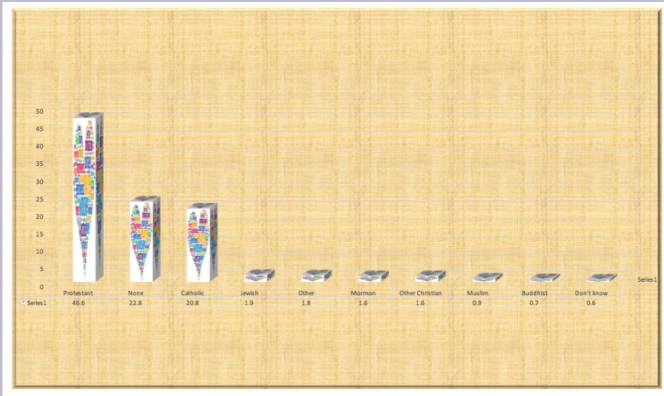


Figure 6. A bad chart graph

Did you figure it what is wrong?

- it has graphics overlaid on each of the bars that have nothing to do with the actual data
- it has a distracting background texture
- it uses three-dimensional bars, which distort the data

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure 7 shows the iMac data with a baseline of 50. Once again, the differences in areas suggests

a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.

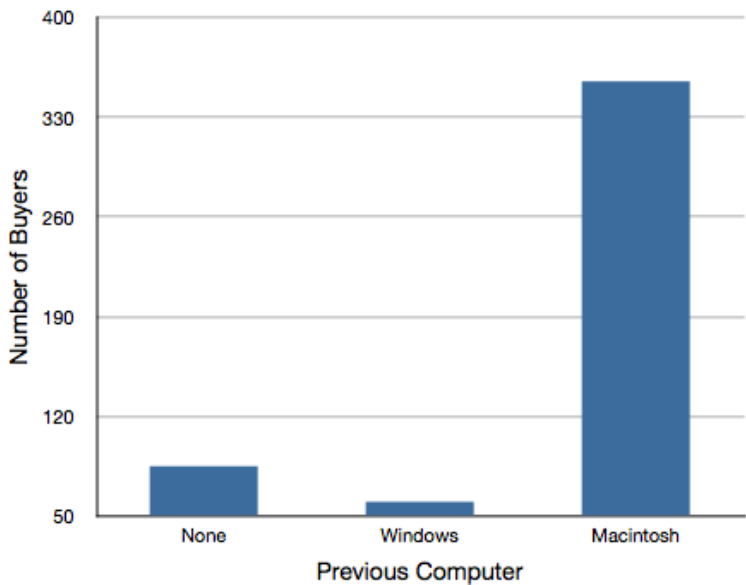


Figure 7. A redrawing of Figure 2 with a baseline of 50.

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative (or categorical) variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 8 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 8 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

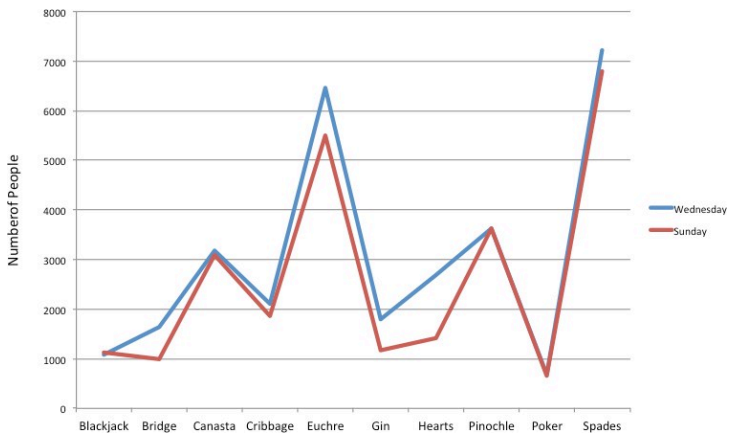


Figure 8. A line graph used *inappropriately* to depict the number of people playing different card games on Sunday and Wednesday.

Recap

Bar charts can be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

Graphing Quantitative Variables

As discussed in the section on variables in Chapter 1, *quantitative variables* are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, favorite sport in which there is no ordering or measuring involved.

There are many types of graphs that can be used to portray distributions of quantitative variables. We already reviewed bar charts. The upcoming sections cover the following types of graphs: (1) histograms, (2) frequency polygons, (3) stem and leaf displays, (4) box plots, (5) more bar charts, (6) line graphs, and (7) scatter plots (discussed in a different chapter). Some graph types such as stem and leaf displays are best suited for small to moderate amounts of data, whereas others such as histograms are best- suited for large amounts of data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

Histograms

A **histogram** is a graphic version of a frequency distribution. It helps to display the shape of a distribution. The graph consists of bars of equal width drawn adjacent to each other and has both a horizontal axis and a vertical axis. The horizontal axis (x-axis) is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The histogram shows the distribution of the values including the highest, middle, and lowest values.

Sometimes we need to group scores if the data has a large distribution. For example, if I wanted to create a frequency distribution of 642 students' scores on a psychology test, that would be a big frequency table. For reference, the test consists of 197 items each graded as "correct" or "incorrect." The students' scores ranged from 46 to 167. A simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 4. A basic rule for grouping data is to make sure each group (or class) has the same grouping amount (in this example it is grouped in

10s), and to make sure you have the lowest category including your lowest value to make sure all scores are included.

| Interval's Lower Limit | Interval's Upper Limit | Class Frequency |
|------------------------|------------------------|-----------------|
| 39.5 | 49.5 | 3 |
| 49.5 | 59.5 | 10 |
| 59.5 | 69.5 | 53 |
| 69.5 | 79.5 | 107 |
| 79.5 | 89.5 | 147 |
| 89.5 | 99.5 | 130 |
| 99.5 | 109.5 | 78 |
| 109.5 | 119.5 | 59 |
| 119.5 | 129.5 | 36 |
| 129.5 | 139.5 | 11 |
| 139.5 | 149.5 | 6 |
| 149.5 | 159.5 | 1 |
| 159.5 | 169.5 | 1 |

Table 4. Grouped Frequency Distribution of Psychology Test Scores

To create this table, the range of scores was broken into intervals, called **class intervals**. The first interval is from 39.5 to

49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc. For this data set, class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too “choppy.” More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals. If you add up all the scores for the class frequencies you will get to the total number of scores (in this example 642).

In a *histogram*, the class intervals are represented by bars. The height of each bar corresponds to its *class frequency*. A histogram of these data is shown in Figure 9.

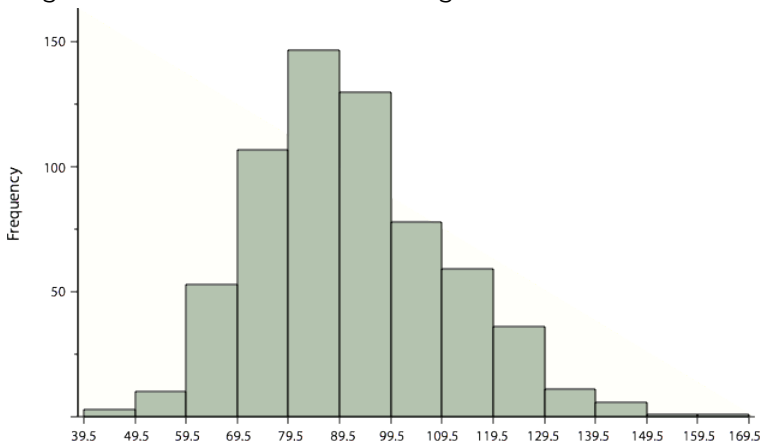


Figure 9. Histogram of scores on a psychology test.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left. The distribution is therefore said to be skewed. (We’ll have more to say about shapes of distributions a little later in the chapter). In our example, the observations are whole numbers.

Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms.

Histograms can be based on *relative frequencies* instead of actual frequencies. Histograms based on **relative frequencies** show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

There is more to be said about the widths of the class intervals, sometimes called *bin widths*. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

Frequency Polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

A frequency polygon for 642 psychology test scores shown in Figure 12 was constructed from the frequency table shown in Table 5.

| Lower Limit | Upper Limit | Count | Cumulative |
|-------------|-------------|-------|------------|
| 29.5 | 39.5 | 0 | 0 |
| 39.5 | 49.5 | 3 | 3 |
| 49.5 | 59.5 | 10 | 13 |
| 59.5 | 69.5 | 53 | 66 |
| 69.5 | 79.5 | 107 | 173 |
| 79.5 | 89.5 | 147 | 320 |
| 89.5 | 99.5 | 130 | 450 |
| 99.5 | 109.5 | 78 | 528 |
| 109.5 | 119.5 | 59 | 587 |
| 119.5 | 129.5 | 36 | 623 |
| 129.5 | 139.5 | 11 | 634 |
| 139.5 | 149.5 | 6 | 640 |
| 149.5 | 159.5 | 1 | 641 |
| 159.5 | 169.5 | 1 | 642 |
| 169.5 | 170.5 | 0 | 642 |

Table 5. Frequency Distribution of Psychology Test Scores
The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45

represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 147 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure 10. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). We call this skew and we will study shapes of distributions more systematically later in this chapter.

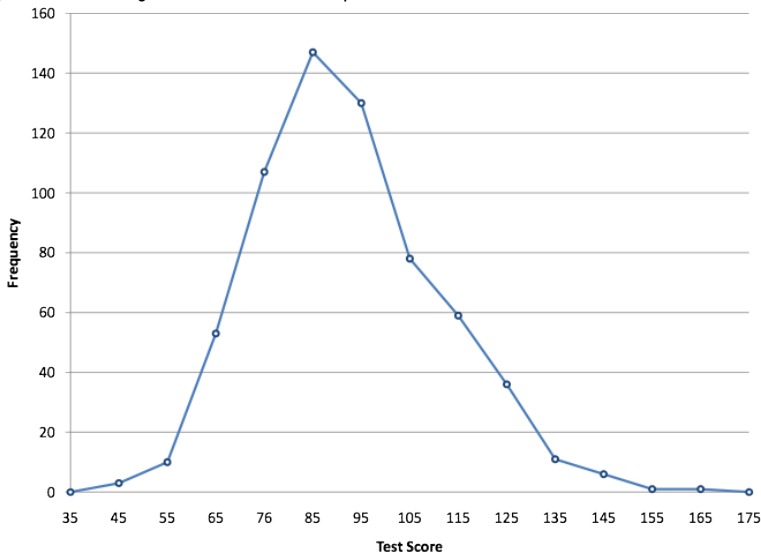


Figure 10. Frequency polygon for the psychology test scores.

A cumulative frequency polygon for the same test scores is shown in Figure 11. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled “35,” three in the interval “45,” and 10 in the interval “55.” Therefore, the Y value corresponding to “55” is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

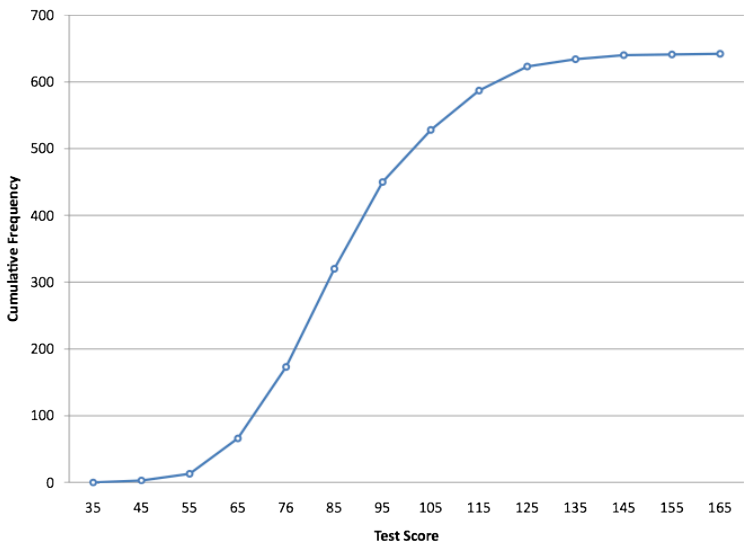


Figure 11. Cumulative frequency polygon for the psychology test scores.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 12 provides an example. The data come from a task in which the goal is to move a computer cursor to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in Figure 15. The figure shows that, although there is some overlap in times, it generally took longer to move the cursor to the small target than to the large one.

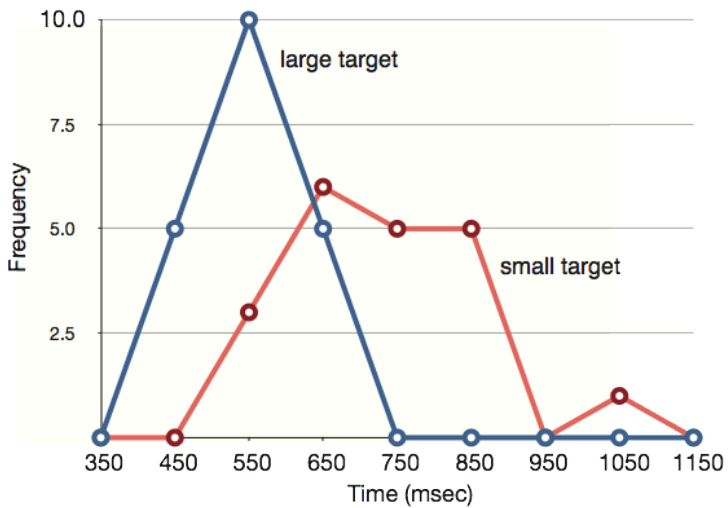


Figure 12. Overlaid frequency polygons.

It is also possible to plot two *cumulative frequency distributions* in the same graph. This is illustrated in Figure 13 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

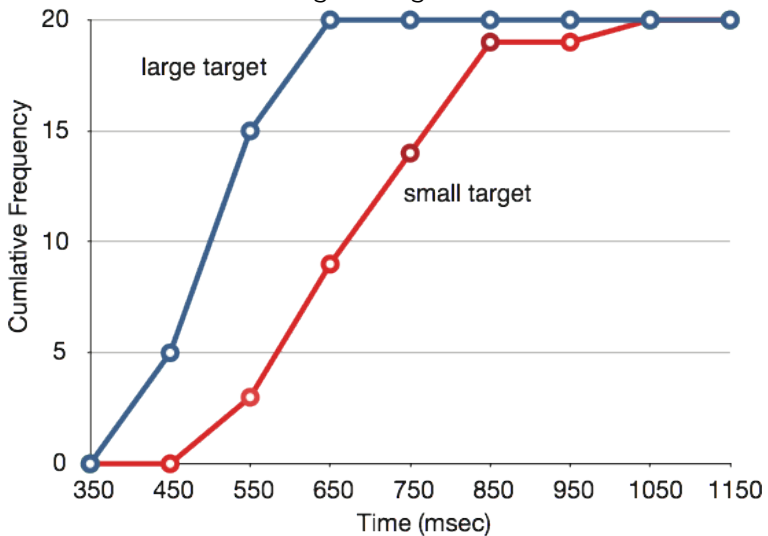


Figure 13. Overlaid cumulative frequency polygons.

Stem and Leaf

The **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a *final significant digit*. For example, 23 has stem two and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Assume the data on the left represents scores from a statistics exam last spring. It is random and unorganized. On the right, you can see we have separated the scores into the stems and leaves. The stemplot shows that most scores were in the 70s. The lowest score was 32 and the highest score was 97.

| Data | | | Stem and Leaf | |
|------|----|----|---------------|---------|
| 83 | 82 | 63 | 3 | 23 |
| 62 | 93 | 78 | 4 | 26 |
| 71 | 68 | 33 | 5 | 6279 |
| 76 | 52 | 97 | 6 | 283 |
| 85 | 42 | 46 | 7 | 1643846 |
| 32 | 57 | 59 | 8 | 3521 |
| 56 | 73 | 74 | 9 | 37 |
| 74 | 81 | 76 | | |

Figure 14. Stem and Leaf Plot

Box Plots

We have already discussed techniques for visually representing data (see histograms and frequency polygons). In this section, we present another important graph, called a box plot. **Box**

plots are useful for identifying outliers (extreme scores) and for comparing distributions. We will explain box plots with the help of data from an in-class experiment. Students in Introductory Statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve parallel box plots.

There are several steps in constructing a box plot. The first relies on the 25th, 50th, and 75th percentiles in the distribution of scores. Figure 15 shows how these three statistics are used. For each gender we draw a box extending from the 25th percentile to the 75th percentile. The 50th percentile is drawn inside the box.

Therefore, the bottom of each box is the 25th percentile, the top is the 75th percentile, and the line in the middle is the 50th percentile. The data for the women in our sample are shown in Table 6.

| |
|--|
| 14, 15, 16, 16, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 19, 19, 19 |
| 20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 24, 29 |

Table 6. Women's times.

For these data, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20. For the men (whose data are not shown), the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5.

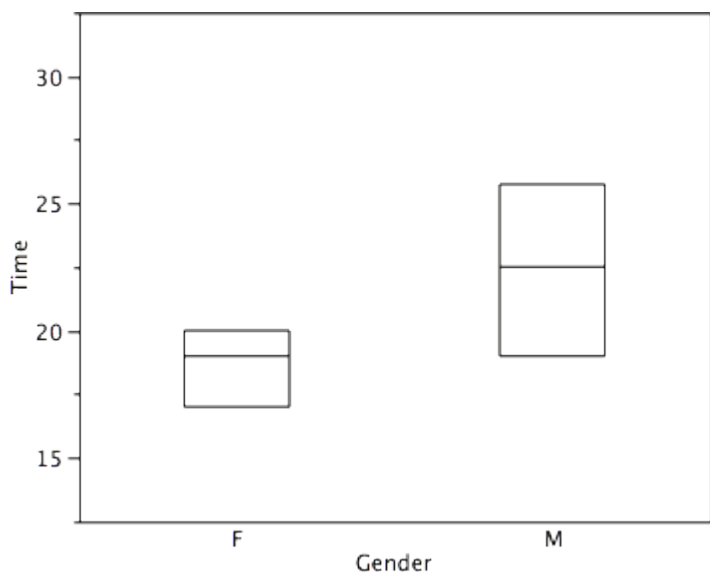


Figure 15. The first step in creating box plots is to identify appropriate quartiles.

Before proceeding, the terminology in Table 7 is helpful.

| Name | Formula | Value from example |
|---------------------------|---|--------------------|
| Upper Hinge | 75th Percentile | 20 |
| Lower Hinge | 25th Percentile | 17 |
| H-Spread | Upper Hinge – Lower Hinge | 3 |
| Step | 1.5 x H-Spread | 4.5 |
| Upper Adjacent | Largest value below Upper Hinge + 1 Step | 24 |
| Lower Adjacent | Smallest value above Lower Hinge + 1 Step | 14 |
| Outside value/ Outlier | Value beyond “whiskers” | 29 |

Table 7. Box plot terms and values for women’s times.

Continuing with the box plots, we put “whiskers” above and below each box to give additional information about the spread of data. *Whiskers* are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women’s data), as shown in Figure 16.

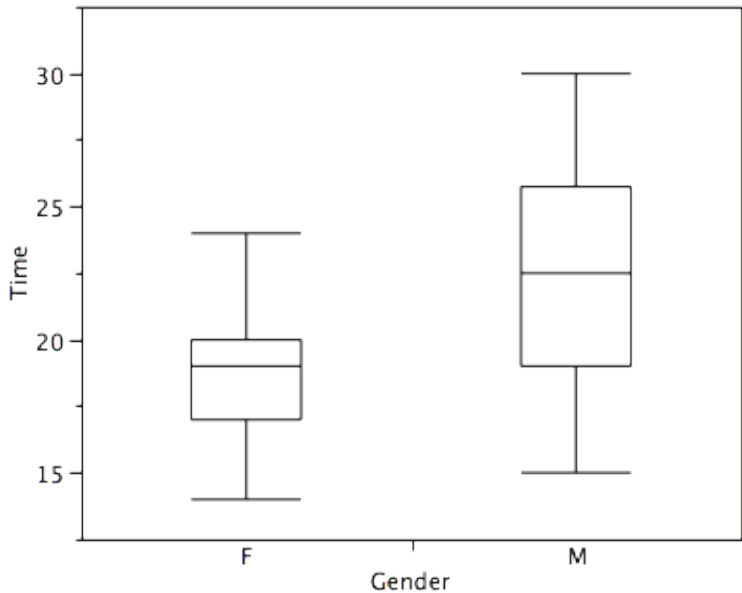


Figure 16. The box plots with the whiskers drawn.

Although whiskers may not cover all data points, we still wish to represent data outside whiskers in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small “o’s” and outlier values are indicated by asterisks (*). In our data, there are no far-out values and just one outside value. This outside value of 29 is for the women and is shown in Figure 17.

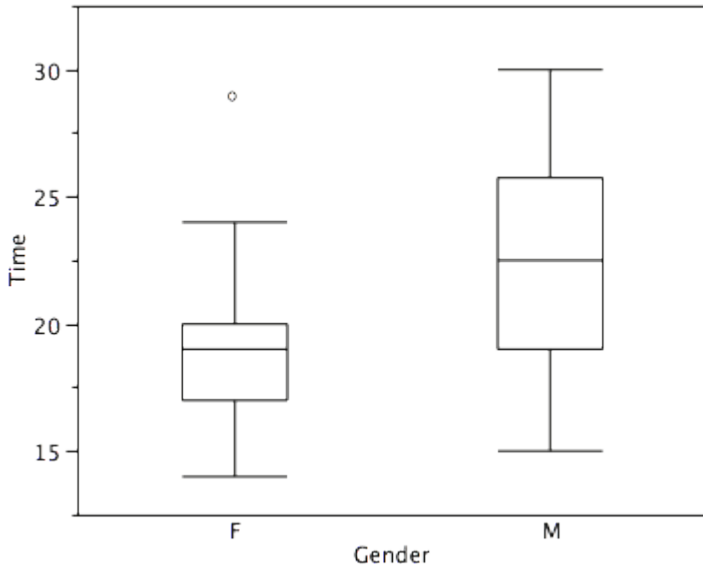


Figure 17. The box plots with the outside value shown.

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. A mean is one type of average we will learn about calculating in the next chapter. Figure 18 shows the result of adding means to our box plots.

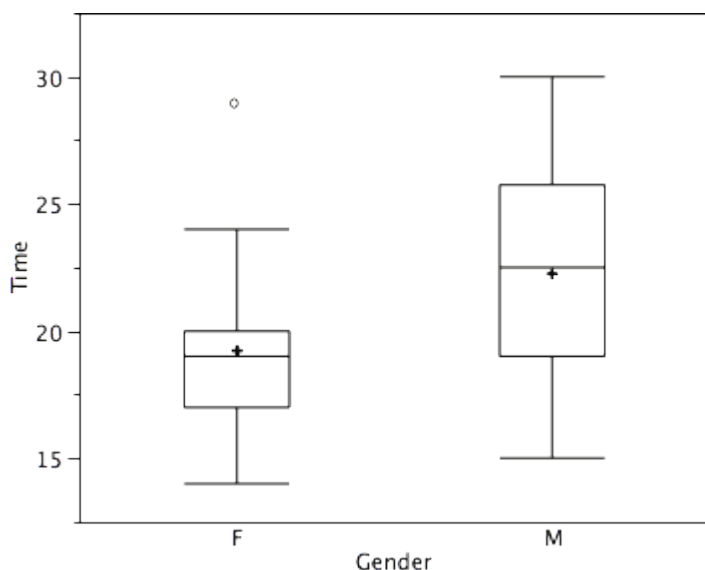


Figure 18. The completed box plots.

Figure 18 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25th and 75th percentiles), we see that half the women's times are between 17 and 20 seconds whereas half the men's times are between 19 and 25.5 seconds. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men.

The Shape of Distribution

Finally, it is useful to present discussion on how we describe the shapes of distributions, which we will revisit in the next chapter to learn how different shapes affect our numerical descriptors of data and distributions.

The primary characteristic we are concerned about when assessing the shape of a distribution is whether the distribution

is symmetrical or skewed. A symmetrical distribution, as the name suggests, can be cut down the center to form 2 mirror images. Although in practice we will never get a perfectly symmetrical distribution, we would like our data to be as close to symmetrical as possible for reasons we delve into in Chapter 3. Many types of distributions are symmetrical, but by far the most common and pertinent distribution at this point is the normal distribution, shown in Figure 19. Notice that although the symmetry is not perfect (for instance, the bar just to the right of the center is taller than the one just to the left), the two sides are roughly the same shape. The normal distribution has a single peak, known as the center, and two tails that extend out equally, forming what is known as a bell shape or bell curve.

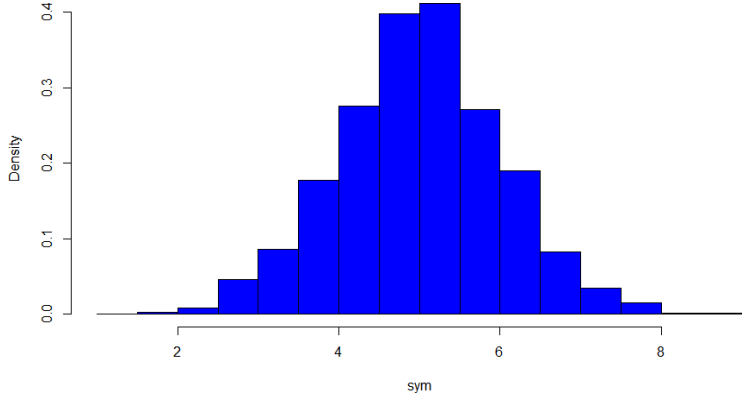


Figure 19. A symmetrical distribution

Symmetrical distributions can also have multiple peaks. Figure 20 shows a bimodal distribution, named for the two peaks that lie roughly symmetrically on either side of the center point. As we will see in the next chapter, this is not a particularly desirable characteristic of our data, and, worse, this is a relatively difficult characteristic to detect numerically. Thus, it is important to visualize your data before moving ahead with any formal analyses.

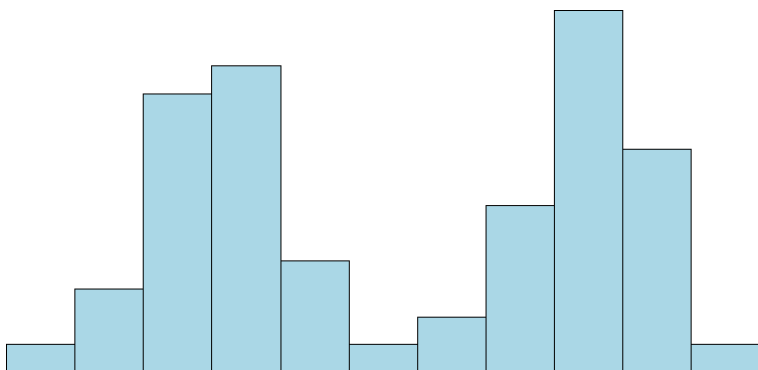


Figure 20. A bimodal distribution

Distributions that are not symmetrical also come in many forms, more than can be described here. The most common asymmetry to be encountered is referred to as skew, in which one of the two tails of the distribution is disproportionately longer than the other. This property can affect the value of the averages we use in our analyses and make them an inaccurate representation of our data, which causes many problems.

Skew can either be positive or negative (also known as right or left, respectively), based on which tail is longer. It is very easy to get the two confused at first; many students want to describe the skew by where the bulk of the data (larger portion of the histogram, known as the body) is placed, but the correct determination is based on which tail is longer. You can think of the tail as an arrow: whichever direction the arrow is pointing is the direction of the skew. Figures 21 and 22 show positive (right) and negative (left) skew, respectively.

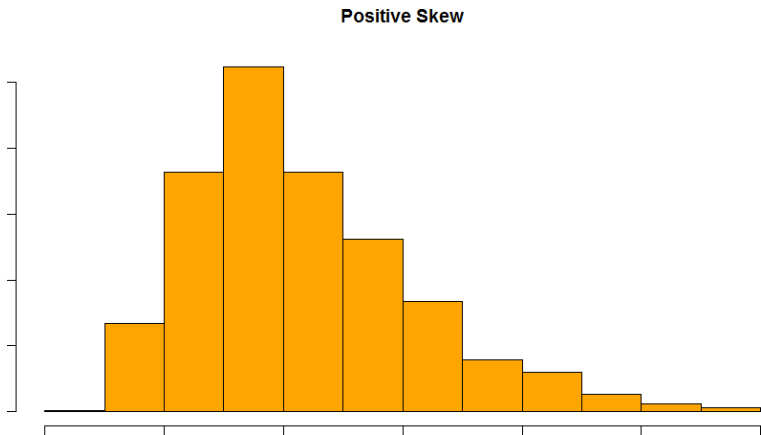


Figure 21. A positively skewed distribution

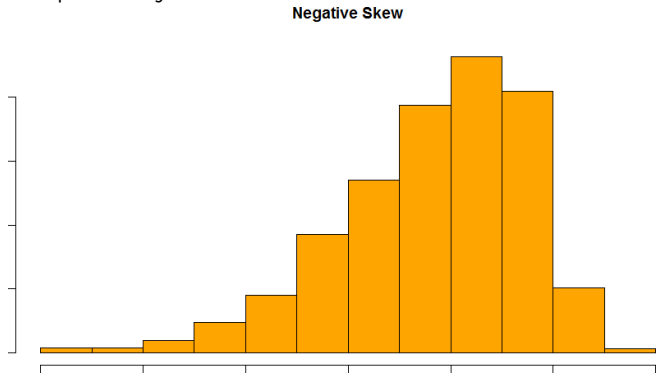


Figure 22. A negatively skewed distribution

A tip to remember skewness

Tip: Take a look down at your feet!



The left foot shows a negative skew (tail is pinky). The right foot is a positive skew.

Recap

Whether you are using a **table** or a **graph** the same two elements of frequency distribution must be present:

1. the entire set of categories that make-up the original distribution must be included
2. a record of the frequency, or number of individuals in each category within the distribution must be included

Examining our data graphically is useful and there are different choices in graphing depending on what is needed and the type of data you have. The scale of measurement determines the most appropriate graph to use. Bar charts are used to display qualitative data along a nominal or ordinal scale of measurement. Histograms, frequency polygons, stem and leaf

plots, and box plots are most appropriate when using interval or ratio scales of measurement.

Box plots provide basic information about the distribution, examining data according to quartiles. By examining a box plot you are able to identify more about the distribution (see Figure X). For example, a distribution with a positive skew would have a longer box and whisker above the 50th percentile (median) in the positive direction than in the negative direction (middle boxplot in Figure 23). Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot and to examine these details one should use create a histogram and/or a stem and leaf plot.

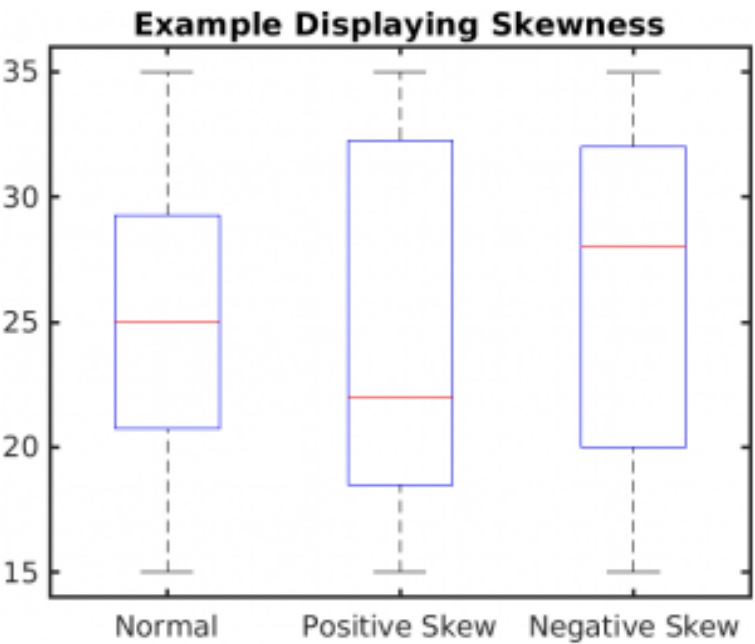


Figure 23. Examples of distributions in Box plots.

Graphing Beyond Frequency

In this section, we will briefly review some graphing techniques that extend beyond reporting frequencies.

Bar charts beyond frequency

In this section we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure 24 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24th 2000 to May 24th 2001. Notice that both the S & P and the Nasdaq had “negative increases” which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity *percentage increase*.

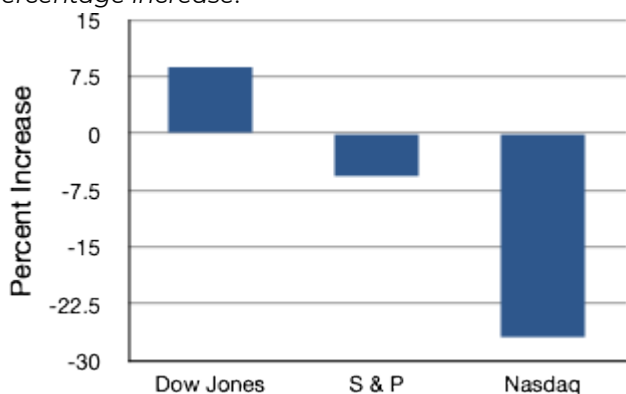


Figure 24. Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

Bar charts are particularly effective for showing change over time. Figure 25, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

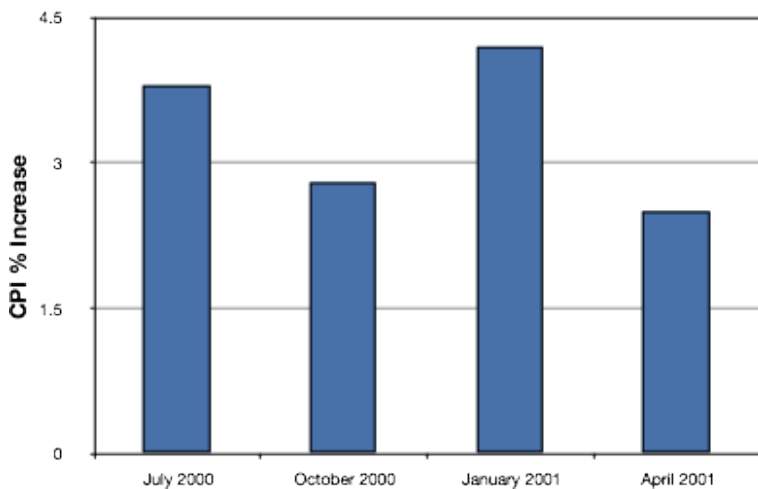


Figure 25. Percent change in the CPI over time. Each bar represents a percent increase for the three months ending at the date indicated.

Bar charts are often used to compare the means of different experimental conditions. Figure 26 shows the mean time it took one of us (DL) to move the cursor to either a small target or a large target. On average, more time was required for small targets than for large ones.

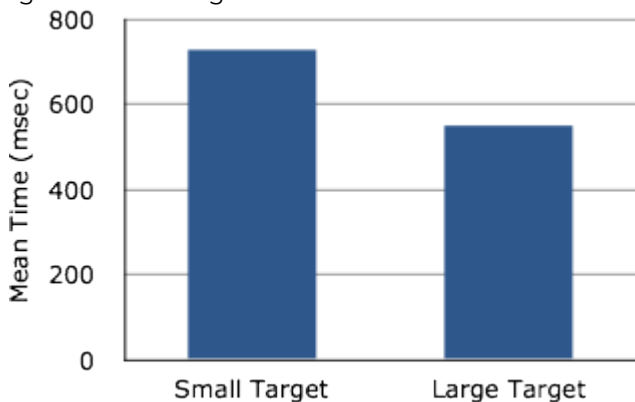


Figure 26. Bar chart showing the means for the two conditions.

Although bar charts can display means, we do not

recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the cursor-movement data is shown in Figure 27. You can see that Figure 27 reveals more about the distribution of movement times than does Figure 26.



Figure 27. Box plots of times to move the cursor to the small and large targets.

Line Graphs Beyond Frequency

A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). For example, Figure 28 was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

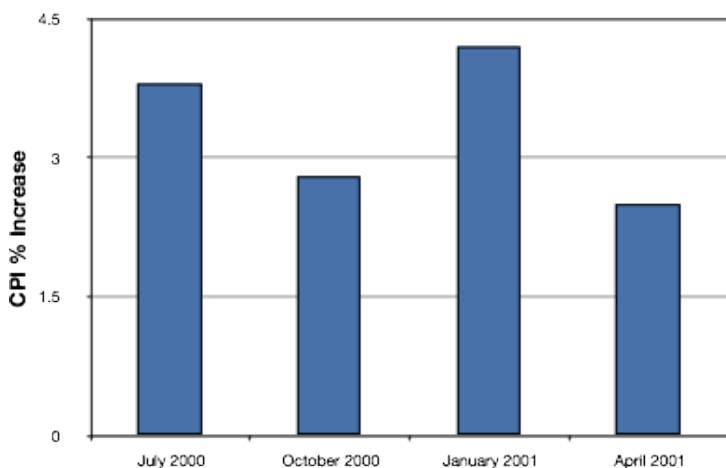


Figure 28. A bar chart of the percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

A line graph of these same data is shown in Figure 29. Although the figures are similar, the line graph emphasizes the change from period to period.

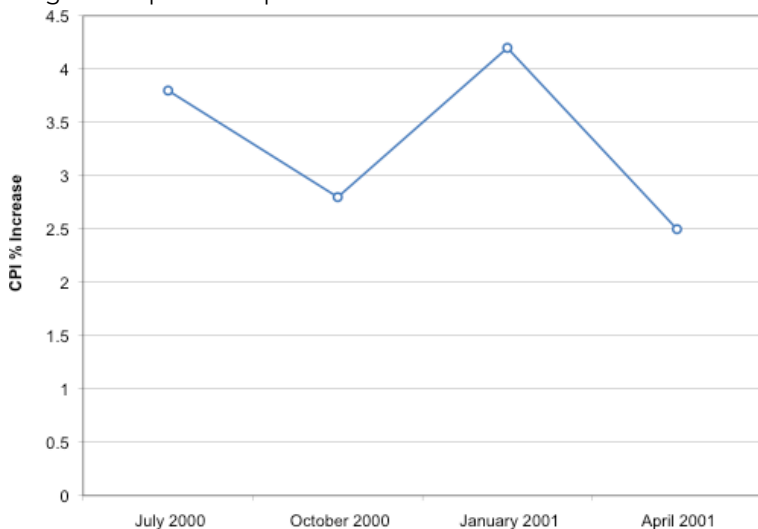


Figure 29. A line graph of the percent change in the CPI

over time. Each point represents percent increase for the three months ending at the date indicated.

Line graphs are appropriate only when both the X- and Y-axes display ordered (rather than qualitative) variables. Although bar charts can also be used in this situation, line graphs are generally better at comparing changes over time. Figure 30, for example, shows percent increases and decreases in five components of the CPI. The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its interpretation would not be as easy. Again, let us stress that it is misleading to use a line graph when the X-axis contains merely categorical variables.

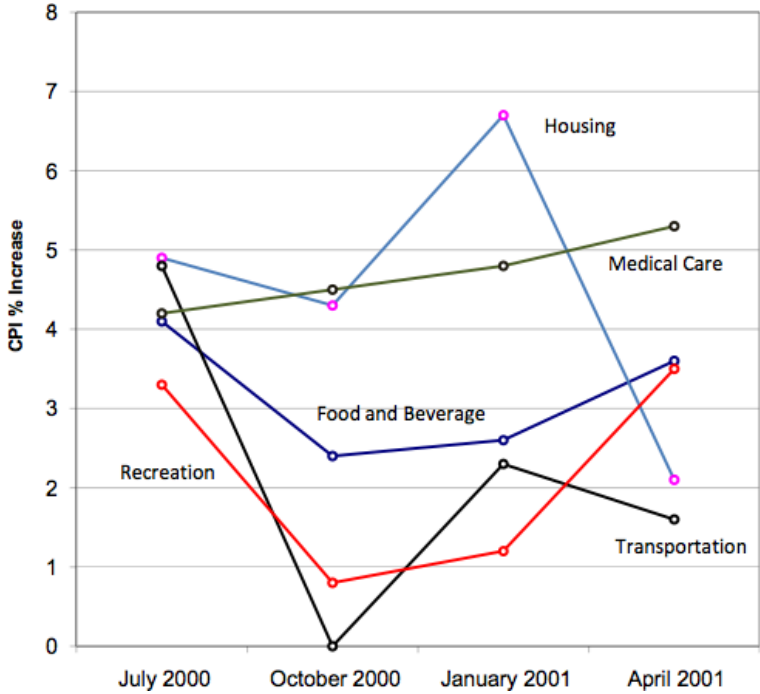


Figure 30. A line graph of the percent change in five components of the CPI over time

Beyond Frequencies: Which graph to use?

There are many different types of plots that we can use, which have different advantages and disadvantages. Let's say that we are interested in characterizing the difference in height between men and women in the NHANES dataset. Figure [31](#) shows four different ways to plot these data.

1. The bar graph in panel A shows the difference in means (a type of average), but doesn't show us how much spread there is in the data around these means – and as we will see later, knowing this is essential to determine whether we think the difference between the groups is large enough to be important.
2. The second plot shows the bars with all of the data points overlaid – this makes it a bit clearer that the distributions of height for men and women are overlapping, but it's still hard to see due to the large number of data points.

In general we prefer using a plotting technique that provides a clearer view of the distribution of the data points.

3. In panel C, we see one example of a *violin plot*, which plots the distribution of data in each condition (after smoothing it out a bit).
4. Another option is the *box plot* shown in panel D, which shows the median (another type of average, central line), a measure of variability (the width of the box, which is based on a measure called the interquartile range), and any outliers (noted by the points at the ends of the lines). These are both effective ways to show data that provide a good feel for the distribution of the data.

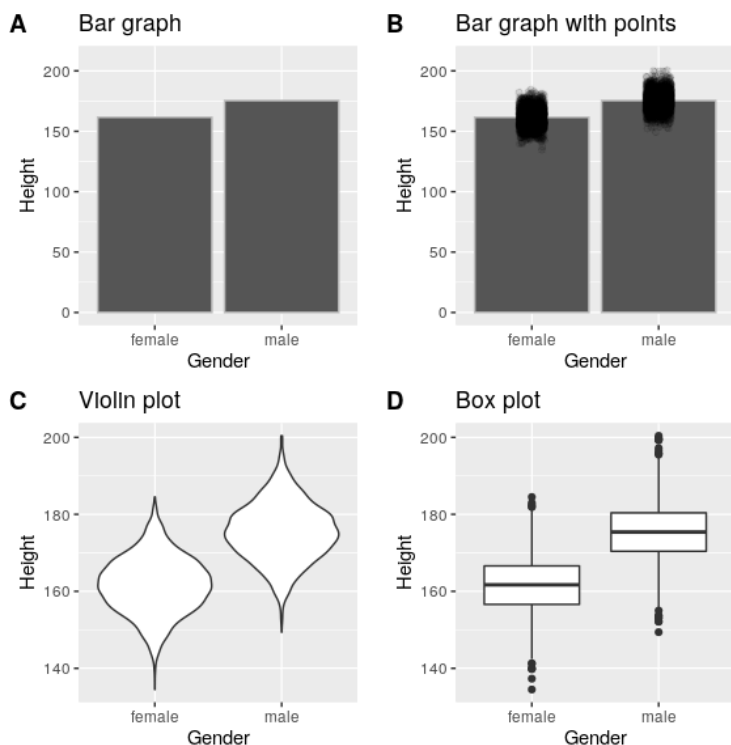


Figure 34: Four different ways of plotting the difference in height between men and women in the NHANES dataset. Panel A plots the means of the two groups, which gives no way to assess the relative overlap of the two distributions. Panel B shows the same bars, but also overlays the data points, jittering them so that we can see their overall distribution. Panel C shows a violin plot, which shows the distribution of the datasets for each group. Panel D shows a box plot, which highlights the spread of the distribution along with any outliers (which are shown as individual points).

Avoid distorting the data

It's often possible to use visualization to distort the message of a dataset. A very common one is use of different axis scaling to

either exaggerate or hide a pattern of data. For example, let's say that we are interested in seeing whether rates of violent crime have changed in the US. In Figure 35, we can see these data plotted in ways that either make it look like crime has remained constant, or that it has plummeted. The same data can tell two very different stories!

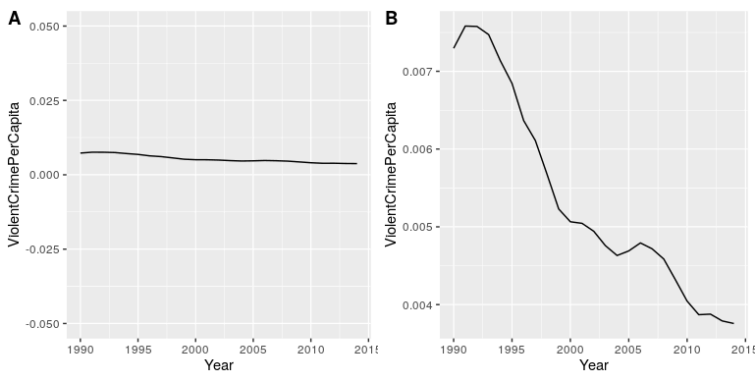


Figure 35: Crime data from 1990 to 2014 plotted over time. Panels A and B show the same data, but with different ranges of values along the Y axis. Data obtained from <https://www.ucrdatatool.gov/Search/Crime/State/RunCrimeStatebyState.cfm>

Choose the Y-axis wisely

We mentioned this tip when we went over bar charts, but it is worth reviewing again. One of the major controversies in statistical data visualization is how to choose the Y-axis, and in particular whether it should always include zero. In his famous book “How to lie with statistics”, Darrell Huff argued strongly that one should always include the zero point in the Y axis. On the other hand, Edward Tufte has argued against this:

“In general, in a time-series, use a baseline that shows

the data not the zero point; don't spend a lot of empty vertical space trying to reach down to the zero point at the cost of hiding what is going on in the data line itself.”
(from <https://qz.com/418083/its-ok-not-to-start-your-y-axis-at-zero/>)

There are certainly cases where using the zero point makes no sense at all. Let's say that we are interested in plotting body temperature for an individual over time. In Figure 36 we plot the same (simulated) data with or without zero in the Y-axis. It should be obvious that by plotting these data with zero in the Y-axis (Panel A) we are wasting a lot of space in the figure, given that body temperature of a living person could never go to zero! By including zero, we are also making the apparent jump in temperature during days 21-30 much less evident. In general, my inclination for line plots and scatterplots is to use all of the space in the graph, unless the zero point is truly important to highlight.

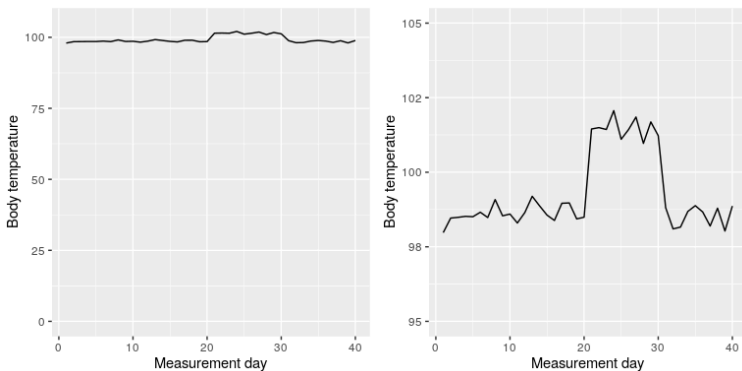


Figure 36: Body temperature over time, plotted with or without the zero point in the Y axis.

Avoid pie charts

This is one reason why statisticians never use pie charts: It can

be very difficult for humans to accurately perceive differences in the volume of shapes. *Pie charts are not recommended when you have a large number of categories.* Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted “The only worse design than a pie chart is several of them.” The pie chart in Figure 37 (presenting the same data on religious affiliation that we showed above) shows how tricky this can be. Can you spot the issues in reading this graph?

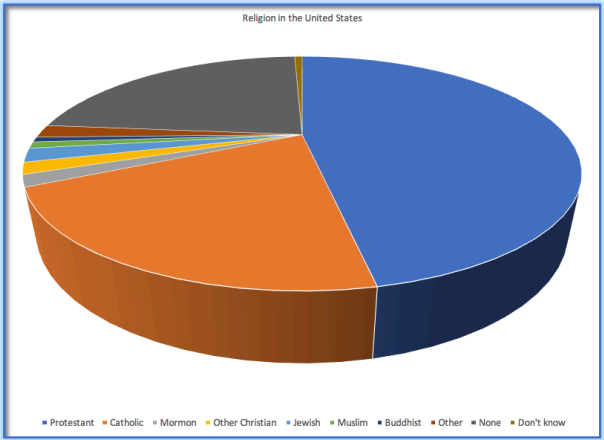


Figure 37: An example of a pie chart, highlighting the difficulty in apprehending the relative volume of the different pie slices.

This plot is terrible for several reasons. First, it requires distinguishing a large number of colors from very small patches at the bottom of the figure. Second, the visual perspective distorts the relative numbers, such that the pie wedge for Catholic appears much larger than the pie wedge for None, when in fact the number for None is slightly larger (22.8 vs 20.8 percent), as was evident in Figure 37. Third, by separating the legend from the graphic, it requires the viewer

to hold information in their working memory in order to map between the graphic and legend and to conduct many “table look-ups” in order to continuously match the legend labels to the visualization. And finally, it uses text that is far too small, making it impossible to read without zooming in.

Plotting the data using a more reasonable approach (Figure 38), we can see the pattern much more clearly. This plot may not look as flashy as the pie chart generated using Excel, but it’s a much more effective and accurate representation of the data.

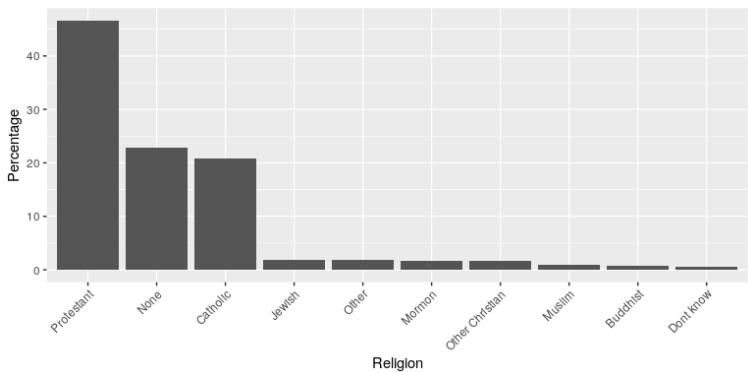


Figure 38: A clearer presentation of the religious affiliation data (obtained from <http://www.pewforum.org/religious-landscape-study/>).

This plot allows the viewer to make comparisons based on the length of the bars along a common scale (the y-axis). Humans tend to be more accurate when decoding differences based on these perceptual elements than based on area or color.

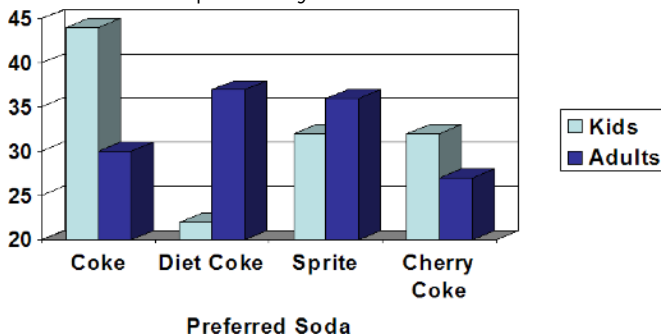
Learning objectives

Having read this chapter, you should be able to:

- Identify different types of graphs and when we would use them based on the type of data
- Differentiate between different types of frequency graphs
- Identify the shape of a distribution in a frequency graph.
- Identify good versus bad graphs using some basic tips and principles
- Promise to never create a pie chart.

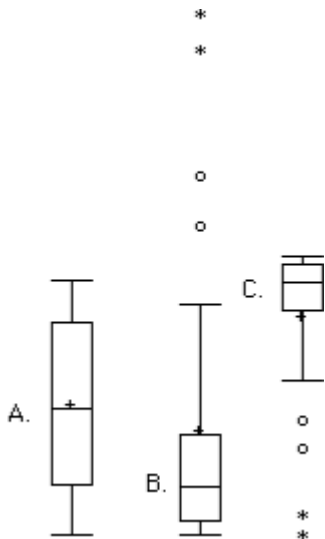
Exercises – Ch. 3

1. Name some ways to graph quantitative variables and some ways to graph qualitative variables.
2. Given the following data, construct a pie chart and a bar chart. Which do you think is the more appropriate or useful way to display the data?
3. Pretend you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired.
 1. What is on the Y-axis? Explain.
 2. What is on the X-axis? Explain.
 3. What would be the probable shape of the salary distribution? Explain why.



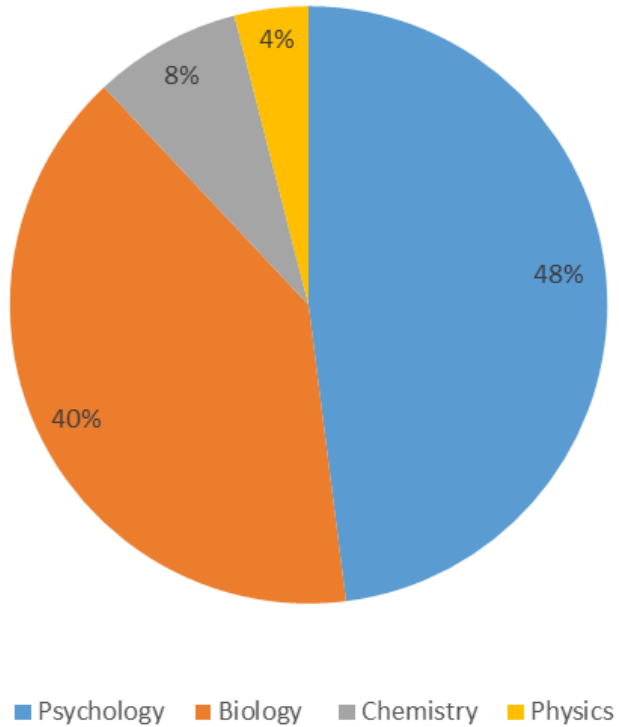
4. A graph appears below showing the number of adults and children who prefer each type of soda. There were 130

adults and kids surveyed. Discuss some ways in which the graph below could be improved.



5. Which of the box plots on the graph has a large positive skew? Which has a large negative skew?
6. Create a histogram of the following data representing how many shows children said they watch each day.
7. Explain the differences between bar charts and histograms. When would each be used
8. Draw a histogram of a distribution that is
 - Negatively skewed
 - Symmetrical
 - Positively skewed
9. Based on the pie chart below, which was made from a sample of 300 students, construct a frequency table of college majors.

College Majors



10. Create a histogram of the following data. Label the tails and body and determine if it is skewed (and direction, if so) or symmetrical.

| Hours worked per week | Proportion |
|-----------------------|------------|
| 0-10 | 4 |
| 10-20 | 8 |
| 20-30 | 11 |
| 30-40 | 51 |
| 40-50 | 12 |
| 50-60 | 9 |
| 60+ | 5 |

Answers to Odd-Numbered Exercises – Ch. 3

1. Qualitative variables are displayed using pie charts and bar charts. Quantitative variables are displayed as box plots, histograms, etc.

3. [You do not need to draw the histogram, only describe it below]

- The Y-axis would have the frequency or proportion because this is always the case in histograms
- The X-axis has income, because this is out quantitative variable of interest
- Because most income data are positively skewed, this histogram would likely be skewed positively too

5. Chart b has the positive skew because the outliers (dots and asterisks) are on the upper (higher) end; chart c has the negative skew because the outliers are on the lower end.

7. In bar charts, the bars do not touch; in histograms, the bars do touch. Bar charts are appropriate for qualitative variables, whereas histograms are better for quantitative variables.

9. Use the following dataset for the computations below:

| Major | Freq |
|------------|------|
| Psychology | 144 |
| Biology | 120 |
| Chemistry | 24 |
| Physics | 12 |

4. Chapter 4: Measures of Central Tendency

Now that we have visualized our data to understand its shape, we can begin with numerical analyses! The descriptive statistics presented in this chapter serve to start to describe the distribution of our data objectively and mathematically – our first step into statistical analysis! *The topics here will serve as the basis for everything we do in the rest of the course.*

Review: There are four different scales of measurement that go along with these different ways that values of a variable can differ.

Nominal scale. A nominal variable satisfies the criterion of identity, such that each value of the variable represents something different, but the numbers simply serve as qualitative labels as discussed above. For example, we might ask people for their political party affiliation, and then code those as numbers: 1 = “Republican”, 2 = “Democrat”, 3 = “Libertarian”, and so on. However, the different numbers do not have any ordered relationship with one another.

Ordinal scale. An ordinal variable satisfies the criteria of identity and magnitude, such that the values can be ordered in terms of their magnitude. For example, we

might ask a person with chronic pain to complete a form every day assessing how bad their pain is, using a 1-7 numeric scale. Note that while the person is presumably feeling more pain on a day when they report a 6 versus a day when they report a 3, it wouldn't make sense to say that their pain is twice as bad on the former versus the latter day; the ordering gives us information about relative magnitude, but the differences between values are not necessarily equal in magnitude.

Interval scale. An (equal) interval scale has all of the features of an ordinal scale, but in addition, the intervals between units on the measurement scale can be treated as equal. A standard example is physical temperature measured in Celsius or Fahrenheit; the physical difference between 10 and 20 degrees is the same as the physical difference between 90 and 100 degrees, but each scale can also take on negative values.

Ratio scale. A ratio scale variable has all four of the features outlined above: identity, magnitude, equal intervals, and absolute zero. The difference between a ratio scale variable and an interval scale variable is that the ratio scale variable has a true zero point. Examples of ratio scale variables include physical height and weight, along with temperature measured in Kelvin.

There are two important reasons that we must pay attention to the scale of measurement of a variable. First, the scale determines what kind of mathematical operations we can apply to the data (see Table 1). A nominal variable can only be compared for equality; that is, do two observations on that

variable have the same numeric value? It would not make sense to apply other mathematical operations to a nominal variable, since they don't really function as numbers in a nominal variable, but rather as labels. With ordinal variables, we can also test whether one value is greater or lesser than another, but we can't do any arithmetic. Interval and ratio variables allow us to perform arithmetic; with interval variables we can only add or subtract values, whereas with ratio variables we can also multiply and divide values.

| Table 1: Different scales of measurement admit different types of numeric operations | | | | |
|--|-----------------|-----|-----|-----------------|
| | Equal/not equal | >/< | +/- | Multiply/divide |
| Nominal | OK | | | |
| Ordinal | OK | OK | | |
| Interval | OK | OK | OK | |
| Ratio | OK | OK | OK | OK |

These constraints also imply that there are certain kinds of statistics that we can compute on each type of variable. Statistics that simply involve counting different values (such as the most common value, known as the *mode*), can be calculated on any of the variable types. Other statistics are based on ordering or ranking of values (such as the *median*, which is the middle value when all of the values are ordered by their magnitude), and these require that the value at least be on an ordinal scale. Finally, statistics that involve adding up values (such as the average, or *mean*), require that the variables be at least on an interval scale. Having said that, we should note that it's quite common for researchers to compute the mean of variables that are only ordinal (such as responses on personality tests), but this can sometimes be problematic.

What is Central Tendency?

Therefore, a measure of central tendency is a way to summarize a large set of numbers using one single score. We can use measures of central tendency to describe a single distribution or compare multiple sets of scores but we have to figure out which measure of central tendency best represents a given distribution.

You might be thinking this is simple. After all, finding the “center” of a distribution involves just looking at it but let’s look at the 3 frequency distributions below and decide subjectively what the most typical or representative “center” score would be.

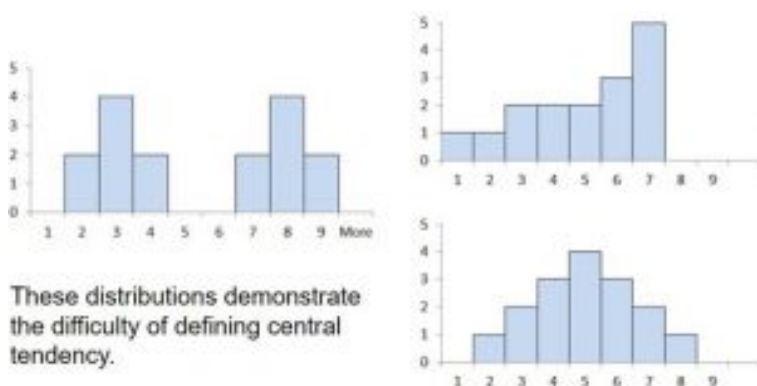


Figure 1. Three different distributions

These distributions demonstrate that finding the center of a distribution may be more challenging than first thought.

Let’s consider another example. Imagine this situation: You are in a class with just four other students, and the five of you took a 5-point pop quiz. Today your instructor is walking around the room, handing back the quizzes. She stops at your desk and hands you your paper.

Written in bold black ink on the front is “**3/5.**” How do you react? Are you happy with your score of 3 or disappointed?

How do you decide? You might calculate your percentage correct, realize it is 60%, and be appalled. But it is more likely that when deciding how to react to your performance, you will want additional information. What additional information would you like?

If you are like most students, you will immediately ask your neighbors, “Whad’ja get?” and then ask the instructor, “How did the class do?” In other words, the additional information you want is how your quiz score compares to other students’ scores. You therefore understand the importance of comparing your score to the class distribution of scores. Should your score of 3 turn out to be among the higher scores, then you’ll be pleased after all. On the other hand, if 3 is among the lower scores in the class, you won’t be quite so happy.

This idea of comparing individual scores to a distribution of scores is fundamental to statistics. So let’s explore it further, using the same example (the pop quiz you took with your four classmates). Three possible outcomes are shown in Table 2. They are labeled “Dataset A,” “Dataset B,” and “Dataset C.” *Which of the three datasets would make you happiest? In other words, in comparing your score with your fellow students’ scores, in which dataset would your score of 3 be the most impressive?*

In Dataset A, everyone’s score is 3. This puts your score at the exact center of the distribution. You can draw satisfaction from the fact that you did as well as everyone else. But of course, it cuts both ways: everyone else did just as well as you.

| Student | Dataset A | Dataset B | Dataset C |
|-------------|-----------|-----------|-----------|
| You | 3 | 3 | 3 |
| John's | 3 | 4 | 2 |
| Maria's | 3 | 4 | 2 |
| Shareecia's | 3 | 4 | 2 |
| Luther's | 3 | 5 | 1 |

Table 2. Three possible datasets for the 5-point make-up quiz. Now consider the possibility that the scores are described as in Dataset B. This is a depressing outcome even though your score is no different than the one in Dataset A. The problem is that the other four students had higher grades, putting yours below the center of the distribution. Finally, let's look at Dataset C. This is more like it! All of your classmates score lower than you so your score is above the center of the distribution.

Now let's change the example in order to develop more insight into the center of a distribution. For this example, there is a quasi-experiment with 2 groups (levels of the IV), tournament players and novices (people who don't play chess). Subjects were shown a chess position and then asked to reconstruct it on an empty chessboard. The number of pieces correctly placed was recorded for three chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions (the DV). The maximum possible score was 89. Figure 2 shows the results of an experiment on memory for chess positions. This is a type of stem and leaf plot called a *back-to-back stemplot*. There are two groups being compared. On the left are people who don't play chess (novice).

On the right are people who play a great deal (tournament players). It is clear that the location of the center of the distribution for the non-players is much lower than the center of the distribution for the tournament players.

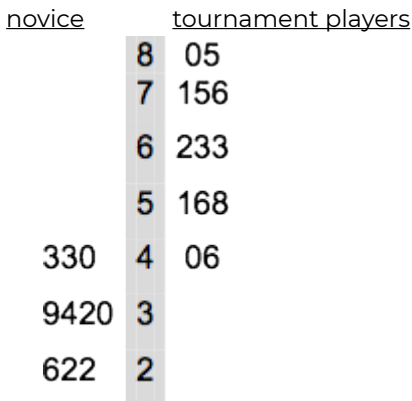


Figure 2. Back-to-back stem and leaf display. The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.

We're sure you get the idea now about the center of a distribution. It is time to move beyond intuition. We need a formal definition of the center of a distribution. In fact, we'll offer you three definitions! This is not just generosity on our part.

There turn out to be (at least) three different ways of thinking about the center of a distribution, all of them useful in various contexts. In the remainder of this section, we will give statistical measures for these concepts of central tendency. These are the three **measures of central tendency**:

- Mean
- Median
- Mode

Mean

One definition of central tendency is the point at which the distribution is in balance. Figure 3 shows the distribution of the five numbers 2, 3, 4, 9, 16 placed upon a balance scale. If each number weighs one pound, and is placed at its position along the number line, then it would be possible to balance them by placing a fulcrum at 6.8. The fulcrum or balancing point is calculated as the **arithmetic mean or mean**.

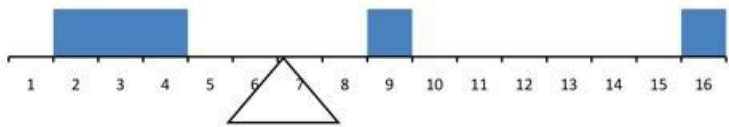


Figure 3. A balance scale demonstrating the mean as the fulcrum.

The arithmetic *mean* is the most common measure of central tendency. The mean is essentially the balancing point of a distribution of scores. This means the distance to all scores below the mean equals the distance to all scores above the mean. The mathematical definition of the mean is the point in a distribution at which the total distance to all the scores above that point equals the total distance to all scores below that point. It is simply the sum of the numbers divided by the number of numbers. The symbol “ μ ” (pronounced “mew”) is used for the mean of a *population*. The symbol “ \bar{X} ” (pronounced “X-bar”) or M is used for the mean of a *sample*.

Mean

The formula for μ (population) and \bar{x} or M (sample):

$$\mu = \frac{\sum x}{N} \qquad \bar{x} = \frac{\sum x}{n}$$

For the μ formula, $\sum x$ is the sum of all the numbers in the population and N is the number of numbers in the population. The formula for \bar{x} or M is essentially identical where $\sum x$ is the sum of all the numbers in the sample and n is the number of numbers in the sample.

The only distinction between these two equations is whether we are referring to the population (in which case we use the parameter μ) or a sample of that population (in which case we use the statistic \bar{x}).

Example: The mean of the numbers 2,3,4,9,16 = $34/5 = 6.8$ (regardless if sample or population)

Example: The mean for 1, 2, 3, 6, 8 is $20/5 = 4$

Table 3 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20,
19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6

Table 3. Number of touchdown passes.

The mean number of touchdown passes thrown is 20.45 as shown below. First, all X values were added up, then divided by the total number of teams.

$$\bar{x} = \sum x / n = 634 / 31 = 20.45$$

By the way, although the arithmetic mean is not the only “mean” (there is also a geometric mean, a harmonic mean, and many others that are all beyond the scope of this course), it is by far the most commonly used. Therefore, if the term “mean” is used without specifying whether it is the arithmetic mean, it is assumed to refer to the arithmetic mean.

Median

The median is also a frequently used measure of central tendency. The **median** is the midpoint of a distribution: the same number of scores is above the median as below it. Think of how a median is in the middle of the road (figure 4). You can also consider the median as the 50th percentile.



Figure 4. [Road median of German Road](#)

The midpoint is the middle score ranging from lowest to highest values. In figure 5, the median is in the geometric middle as there is a similar distribution of higher and lower scores. In this case, the mean value and the median, middle point, value are the same.

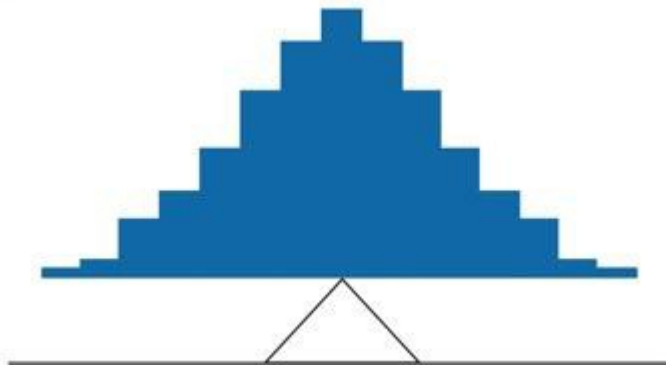


Figure 5. A distribution balanced on the tip of a triangle where the middle point, the median, is also the mean, the point of balance.

More on the Mean and Median

The mean is the point on which a distribution would balance, the median is the value that minimizes the sum of absolute deviations, and the mean is the value that minimizes the sum of the squared deviations. Figure 6 shows the numbers 2, 3, 4, 9, and 16. We calculated the mean as 6.8. The median would be the middle-value number. From the 5 scores, the median is 4.

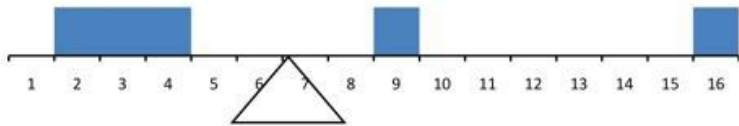


Figure 6. The distribution balances at the mean of 6.8 and the median of 4.0.

Median

In order to calculate median:

1. Arrange the numbers in the set from smallest to largest.
2. Determine N or n (number of scores)
3. If N or n is odd then the median is the middle number.
4. If N or n is even then the median is the average of the middle two numbers

For the data in Table 3 (an example earlier in the chapter with football scores), there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. Again, the median can also be thought of as the 50th percentile.

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 (3 scores for N or n) is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is: $\frac{4 + 7}{2} = 11/2 = 5.5$

When there are numbers with the same values, each appearance of that value gets counted. For example, in the set of numbers 1, 3, 4, 4, 5, 8, and 9, the median is 4 because there are three numbers (1, 3, and 4) below it and three numbers (5, 8, and 9) above it. If we only counted 4 once, the median would incorrectly be calculated at 4.5 (4+5 divided by 2). When in doubt, writing out all of the numbers in order and marking them off one at a time from the top and bottom will always lead you to the correct answer.

Mode

The **mode** is the most frequently occurring value in the dataset. If there are multiple values “tied” for most frequently occurring, the data set can have more than one mode. If all the values occur at the same rate, then there is no mode.

Mode

In order to find the mode, create a frequency table. Identify the score with the highest frequency. It is the score and not the frequency value that is the mode.

Example: 2,3,4,9,16

There is no mode as each score only has a frequency of 1.

Example: 11, 12, 12, 13, 14

| score | f |
|-------|---|
| 11 | 1 |
| 12 | 2 |
| 13 | 1 |
| 14 | 1 |

The mode is 12.

For the data in Table 3, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data, such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2 shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650). Though the mode is not frequently used for continuous data, it is nevertheless an important measure of central tendency as it is the only measure we can use on qualitative or categorical data.

| Range | Frequency |
|-----------|-----------|
| 500-600 | 3 |
| 600-700 | 6 |
| 700-800 | 5 |
| 800-900 | 5 |
| 900-1000 | 0 |
| 1000-1100 | 1 |

Table 5. Grouped frequency distribution

Recap

All measures of central tendency reflect something about the middle of a distribution; but each of the three most common measures of central tendency represents a different concept:

Mean: average, where μ is for the population and \bar{x} or M is for the sample (both same equation).

$$\mu = \frac{\sum X}{N}$$

$$M = \frac{\sum X}{n}$$

Median: middle or 50th percentile. If N or n is odd then the

median is the middle number. If N or n is even then the median is the average of the middle two numbers

Mode: most common, or most frequent value, where there can be a tie or there can be no mode.

Comparing Measures of Central Tendency

A distribution is a graph that shows how scores are distributed along a measurement scale. The mean is the point on the x-axis that falls directly at the “balancing point” for the distribution. The median is the point on the x-axis at which half the area under the distribution curve lies below the median and half lies above the median. The mode is the point on the x-axis that falls directly below the tallest point on the distribution.

In a perfectly symmetrical (normal) distribution, all three measures of central tendency are located at the same value. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. In a perfectly symmetrical distribution, the mean and the median are the same. This example has one mode (unimodal), and the mode is the same as the mean and median. How do the various measures of central tendency compare with each other? In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

A skewed distribution has one side that is long and spread out, somewhat like a tail. The side with the fewer scores (the side that looks more like a tail) is considered the direction of the skew. A distribution that is skewed to the right is called a *positive skewed*. A distribution skewed to the left is called a *negative skew*.

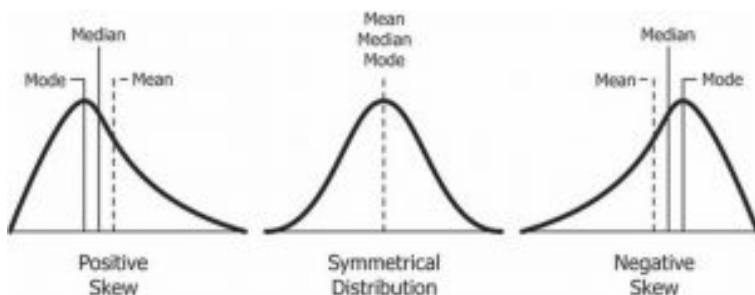


Figure 7. Distributions with mean, median and mode

Differences among the measures occur with skewed distributions. Figure 8 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

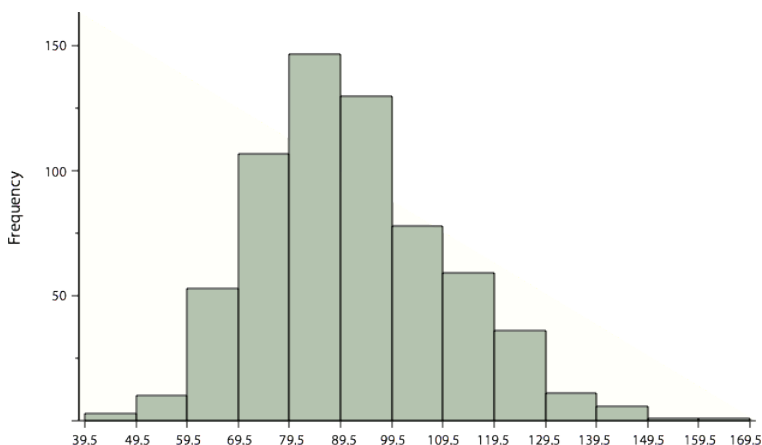


Figure 8. A distribution with a positive skew.

Measures of central tendency are shown in Table 6. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions. For these data, the mean of 91.58 is higher than the median of 90. This pattern holds true for any skew:

the mode will remain at the highest point in the distribution, the median will be pulled slightly out into the skewed tail (the longer end of the distribution), and the mean will be pulled the farthest out. Thus, the mean is more sensitive to skew than the median or mode, and in cases of extreme skew, the mean may no longer be appropriate to use.

| Measure | Value |
|---------|-------|
| Mode | 84 |
| Median | 90 |
| Mean | 91.58 |

Table 6. Measures of central tendency for the test scores.

The distribution of baseball salaries (in 1994) shown in Figure 9 has a much more pronounced skew than the distribution in Figure 8.

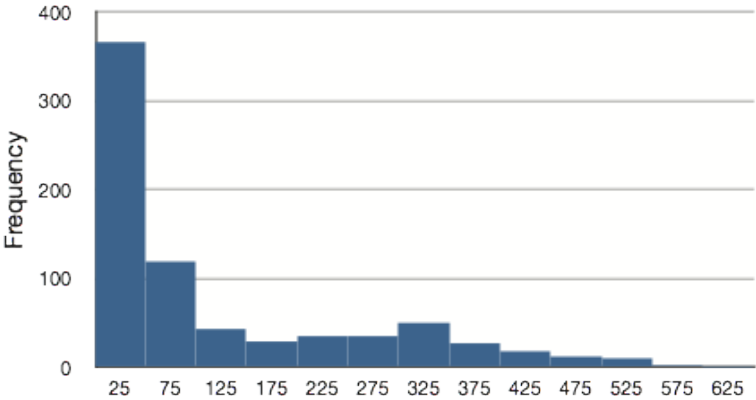


Figure 9. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Table 7 shows the measures of central tendency for these data.

The large skew results in very different values for these measures. No single measure of central tendency is sufficient for data such as these. If you were asked the very general question: “So, what do baseball players make?” and answered with the mean of \$1,183,000, you would not have told the whole story since only about one-third of baseball players make that much. If you answered with the mode of \$250,000 or the median of \$500,000, you would not be giving any indication that some players make many millions of dollars. Fortunately, there is no need to summarize a distribution with a single number. When the various measures differ, our opinion is that you should report the mean and median. Sometimes it is worth reporting the mode as well. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

| Measure | Value (in thousands) |
|---------|----------------------|
| Mode | 250 |
| Median | 500 |
| Mean | 1,183 |

Table 7. Central tendency measures for baseball salary data.

Summary

Remember that measures of central tendency summarize and organize large sets of data that allow researchers to communicate information with just a few numbers. There are three main considerations when determining which measure of central tendency to use:

| Type | Appropriate | Not Appropriate |
|--------|--|---|
| Mean | Interval/Ratio | Extreme scores Skewed distribution Ordinal Nominal |
| Median | Extreme scores Skewed distribution Ordinal | Nominal |
| Mode | Nominal Discrete Describe shape – bimodal | Interval/Ratio |

Before deciding to report a mean, median or mode ask yourself what the data are trying to convey, what is the shape of the distribution (e.g., normal or skewed) and the level of measurement for the data.

The level of measurement of a particular variable will determine which measure(s) of central tendency can be used. For example:

- Mean is preferred when using ratio level data unless distribution includes outliers
- Median is the preferred when using ordinal data
- Median is preferred when data include outliers
- Mode is preferred when using nominal data

The goal of descriptive statistics is to summarize and organize large amounts of data and measures of central tendency tell us about the middle of a distribution but we need to select the measure that is most representative of the distribution.

Generally, if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean. The mean will inaccurately describe a skewed (non-symmetrical) distribution. You have seen this happen if you've ever received one very low grade in a class after receiving many high grades; your average drops like a rock. The one low grade

produces a negatively skewed distribution, and the mean gets pulled away from where most of your grades are, toward that low grade. What hurts is then telling someone your average because it's misleading. It gives the impression that all of your grades are relatively low, even though you have only that one F.

Learning Objectives

Having read this chapter, you should be able to:

- explain the purpose of measuring central tendency
- define and compute the three measures of central tendency (mean, median, mode)
- list the circumstances where each of the three measures of central tendency are appropriate
- explain how the three measures of central tendency are related to distribution (positive skew, negative skew, normal)

Exercises – Ch. 4

1. If the mean time to respond to a stimulus is much *higher* than the median time to respond, what can you say about the shape of the distribution of response times?
2. Compare the mean, median, and mode in terms of their sensitivity to extreme scores.
3. Your younger brother comes home one day after taking a science test. He says that some- one at school told him that “60% of the students in the class scored above the median test grade.” What is wrong with this statement? What if he had said “60% of the students scored above the mean?”
4. Make up three data sets with 5 numbers each that have:
 1. the same mean but different standard deviations.
 2. the same mean but different medians.
 3. the same median but different means.

4. Compute the population mean for the following scores: 5, 7, 8, 3, 4, 4, 2, 7, 1, 6
5. Compute the sample mean for the following scores: -8, -4, -7, -6, -8, -5, -7, -9, -2, 0
6. For the following problem, use the following scores: 5, 8, 8, 8, 7, 8, 9, 12, 8, 9, 8, 10, 7, 9, 7, 6, 9, 10, 11, 8
 1. Create a histogram of these data. What is the shape of this histogram?
 2. How do you think the three measures of central tendency will compare to each other in this dataset?
 3. Compute the sample mean, the median, and the mode
 4. Draw and label lines on your histogram for each of the above values. Do your results match your predictions?

Answers to Odd-Numbered Exercises – Ch. 4

1. If the mean is higher, that means it is farther out into the right-hand tail of the distribution. Therefore, we know this distribution is positively skewed. Review Figure 7.

3. The median is defined as the value with 50% of scores above it and 50% of scores below it; therefore, 60% of score cannot fall above the median. If 60% of scores fall above the mean, that would indicate that the mean has been pulled down below the value of the median, which means that the distribution is negatively skewed

5. $\mu = 4.80$

7. $M = -6.6$

5. Chapter 5: Measures of Dispersion

Measure of central tendency (a value around which other scores in the set cluster) and a measure of variability (an indicator of how spread out about the mean scores are in a data set) are used together to give a description of the data.

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this chapter we will discuss measures of the variability of a distribution.

Measures of dispersion describe the spread of scores in a distribution. The more spread out the scores are, the higher the dispersion or spread. In Figure 1, the y-axis is frequency and the x-axis represents values for a variable. There are two distributions, labeled as small and large. You can see both are normally distributed (unimodal, symmetrical), and the mean, median, and mode for both fall on the same point. What is different between the two is the spread or dispersion of the scores. The taller-looking distribution shows a smaller dispersion while the wider distribution shows a larger dispersion. For the “small” distribution in Figure 1, the data values are concentrated closely near the mean; in the “large” distribution, the data values are more widely spread out from the mean.

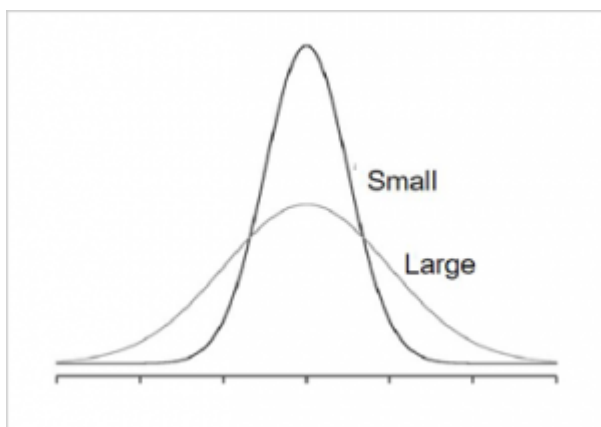


Figure 1. Examples of 2 normal (symmetrical, unimodal) distributions.

In this chapter, we will look at three measures of variability: range, variance, and standard deviation. An important characteristic of any set of data is the variation in the data. Imagine that students in two different sections of statistics take Exam 1 and the mean score in both classrooms is a 75. If that is the only descriptive statistic I report you might assume that both classes are identical – but that is not necessarily true. Let's examine the scores for each section.

Section A

Scores = 70, 70, 70, 70, 85, 85

Mean = 75

Section B

Scores = 70, 72, 73, 75, 75, 85

Mean = 75

Table 1. Exam scores for 2 sections of a class.

Comparing both sections you can see that the scores for Section A very few scores are represented (e.g., 70 and 85) and they are very far from the mean, while in Section B more scores are represented and clustered close to the mean. We would say that the spread of scores for Section A is greater than Section B.

Range

The range is the simplest measure of variability and is really easy to calculate.

Range: We calculate it by subtracting the smallest score from the largest score in the data set.

You can see in our statistics course example (Table 1) that Section A scores have a range of 15 and Section B scores have a range of 15. That means all the other scores are not included and may not give an unbiased description of the data.

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The simplicity of calculating range is appealing but it can be a very unreliable measure of variability. We noticed earlier that the spread of score for each section was very different for each section.

Let's take a few examples. What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so $10 - 2 = 8$. The range is 8. Let's take another example. Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so $99 - 23$ equals 76; the range is 76. Again, the problem with using range is that it is extremely sensitive to outliers, and one number far away from the rest of the data will greatly alter the value of the range. For example, in the set of numbers 1, 3, 4, 4, 5, 8, and 9, the range is 8 ($9 - 1$). However, if we add a single person whose score is nowhere close to the rest of the scores, say, 20, the range more than doubles from 8 to 19.

Interquartile Range

A special take on range, is to identify values in terms of quartiles of the distribution (remember chapter 2 with boxplots). The **interquartile range** (IQR) is the range of the middle 50% of the scores in a distribution and is sometimes used to communicate where the bulk of the data in the distribution are located. It is computed as follows: $IQR = 75\text{th percentile} - 25\text{th percentile}$. Recall that in the discussion of box plots in chapter 2, the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Using this terminology, the interquartile range is referred to as the H-spread.

The Mean Needed to Further Examine Dispersion

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the center of the distribution, we can see how far, on average, each data point is from the center. Remember that the mean is the point on which a distribution would balance. We can examine spread by identifying how far each value is from the mean. This is known as the **deviation from the mean** (or differences from the mean). The sum of deviations is the smallest for the mean value. Interestingly, the sum of deviation from the mean is zero because the mean is the fulcrum or balance point.

Let's revisit an example from chapter 4 (Figure 2). We had a set of 5 numbers: 2, 3, 4, 9, and 16 with a mean of 6.8.

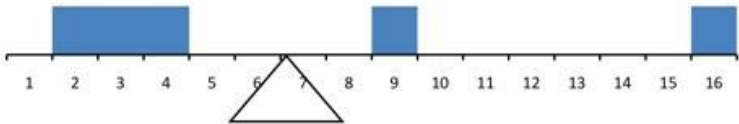


Figure 2. The distribution balances at the mean of 6.8.

In Table 2, the value is represented as \bar{X} , the column

“deviation from the mean or $\diamond - \text{mean}$ ” contains deviations (how far each score deviates from the mean), here calculated as the score minus 6.8.

| Value (X) | Deviation from the Mean (X – mean) |
|--------------------|------------------------------------|
| 2 | -4.8 |
| 3 | -3.8 |
| 4 | -2.8 |
| 9 | 2.2 |
| 16 | 9.2 |
| Total (Σ) | 0 |

Table 2 (from chapter 4) shows the deviations of the numbers 2, 3, 4, 9, and 16 from their mean of 6.8.

Moving toward variance: sum of squares deviations

For us to get to a value that would represent the dispersion, we will add another step from calculating the deviation of the mean scores. You can see in Table 3 there is now a third column, the squared deviations column. The column “ $(\diamond - \text{mean})^2$ ” has the “Squared Deviations” and is simply the previous column squared.

| Value (X) | Deviation from the Mean (X – mean) | Squared Dev |
|------------|------------------------------------|--------------------|
| 2 | -4.8 | 23.04 |
| 3 | -3.8 | 14.44 |
| 4 | -2.8 | 7.84 |
| 9 | 2.2 | 4.84 |
| 16 | 9.2 | 84.64 |
| Total (Σ) | 0 | 134.8 (←Sum of Squ |

Table 3. Adding on a squared deviations column to create “sum of squares” or SS

There are a few things to note about how Table 4 is formatted, as this is the format you will use to calculate variance (and, soon, standard deviation). The raw data scores (X) are always placed in the left-most column. This column is then summed at the bottom to facilitate calculating the mean (simply divided this number by the number of scores in the table). Remember that mean can be μ (population) or if a sample, M or \bar{X} . Once you have the mean, you can easily work your way down the middle column calculating the deviation scores. This column is also summed and has a very important property: it will always sum to 0 (or close to zero if you have rounding error due to many decimal places). *This step is used as a check on your math to make sure you haven’t made a mistake. If this column sums to 0, you can move on to filling in the third column of squared deviations.* This column is summed as well and has its own name: the **Sum of Squares** (abbreviated as SS and given the formula $\sum(X-\text{mean})^2$). As we will see, the Sum of

Squares appears again and again in different formulas – it is a very important value, and this table makes it simple to calculate without error.

Here is another example of calculating SS with 20 data points where the mean = 7:

| X | X - mean | (X - mean) ² |
|---|----------|-------------------------|
| 9 | 2 | 4 |
| 9 | 2 | 4 |
| 9 | 2 | 4 |
| 8 | 1 | 1 |
| 8 | 1 | 1 |
| 8 | 1 | 1 |
| 8 | 1 | 1 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 7 | 0 | 0 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |
| 6 | -1 | 1 |

| | | |
|-----------------------|--------------|---------------|
| 5 | -1 | 4 |
| 5 | -1 | 4 |
| total (Σ) | $\Sigma = 0$ | $\Sigma = 30$ |

Table 4. Calculations for Sum of Squares.

Variance

Now that we have the Sum of Squares calculated, we can use it to compute our formal measure of average distance from the mean, the variance. Informally, it measures how far a set of (random) numbers are spread out from their average value. The **variance** is defined as the average squared difference of the scores from the mean. The mathematical definition of the variance is the sum of the squared deviations (distances) of each score from the mean divided by the number of scores in the data set. Remember that we square the deviation scores because, as we saw in the Sum of Squares table, the sum of raw deviations is always 0, and there's nothing we can do mathematically without changing that.

Variance

The population parameter for variance is σ^2 ("sigma-squared") and is calculated as:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Example from Table 3: If we assume that the values in Table 3 represent the full population, then we can take our value of Sum of Squares and divide it by N to get our population variance:

$$\sigma^2 = 134.8/5 = 26.96$$

Example from Table 4: If we assume that the values in Table 4 represent the full population, then we can take our value of Sum of Squares and divide it by N to get our population variance:

$$\sigma^2 = 30/20 = 1.5$$

Notice that the numerator that formula is identical to the formula for Sum of Squares presented in Tables 3 and 4 with mean replaced by μ . Thus, we can use the Sum of Squares table to easily calculate the numerator then simply divide that value by N to get variance. Remember variance for a population is noted as σ^2 (sigma-squared). So, on average, scores in this population from our quiz example in Table 4 are 1.5 squared units away from the mean. Variance as a measure of spread is much more robust (a term used by statisticians to mean resilient or resistant to outliers) than the range, so it is a much more useful value to compute. Additionally, as we will see in

future chapters, variance plays a central role in inferential statistics.

The variance formula for a sample is very similar to the formula for the population variance with a few changes as shown below. First, for mean we use M or \bar{X} , and most importantly, we now divide by $n - 1$ instead of N . The value $n - 1$ has a special name: the **degrees of freedom** (abbreviated as *df*). You don't need to understand in depth what degrees of freedom are (essentially they account for the fact that we have to use a sample statistic to estimate the mean (\bar{X} or M) before we estimate the variance) in order to calculate variance, but knowing that the denominator is called *df* provides a nice shorthand for the variance formula: SS/df . Variance is also known as S^2 for a sample.

Variance

The sample statistic used to estimate the variance is s^2 ("s-squared"):

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$

$$s^2 = \frac{SS}{df} = \frac{SS}{n - 1}$$

Note: the sum of square deviations is abbreviated as *SS* and the degrees of freedom abbreviated as *df*. The shorthand for sample variance is *SS/df*.

Example from Table 3: If we assume that the values in Table 4 represent a sample, then we can take our value of Sum of Squares and divide it by *n*-1 to get our sample variance:

$$s^2 = 134.8/(5-1) = 33.7$$

Example from Table 4 (treating those scores as a sample): we can estimate the sample variance as:

$$s^2 = 30/(20 - 1) = 1.58$$

Notice that the sample variance values are slightly larger than the one we calculated when we assumed these scores were the full population. This is because our value in the denominator is slightly smaller, making the final value larger.

In general, as your sample size n gets bigger, the effect of subtracting 1 becomes less and less.

Comparing a sample size of 10 to a sample size of 1000; $10 - 1 = 9$, or 90% of the original value, whereas $1000 - 1 = 999$, or 99.9% of the original value. Thus, larger sample sizes will bring the estimate of the sample variance closer to that of the population variance. This is a key idea and principle in statistics that we will see over and over again: *larger sample sizes better reflect the population*. The variance is “the sum of the squared distances of each score from the mean divided by total scores,” according to the definitional formula. This means the final answer is always in the original units of measurement, squared. This means that if we had been measuring reaction time, the units would have been seconds squared. If we had been measuring height, the units would have been inches squared. These units are not very useful because we do not talk about inches squared in day to day language.

Standard Deviation

The **standard deviation** is simply the square root of the variance. This is a useful and interpretable statistic because taking the square root of the variance (recalling that variance is the average squared difference) puts the standard deviation back into the original units of the measure we used. Thus, when reporting descriptive statistics in a study, scientists virtually always report mean and standard deviation. Standard deviation is therefore the most commonly used measure of spread for our purposes.

The population parameter for standard deviation is σ (“sigma”), which, intuitively, is the square root of the variance parameter σ^2 (on occasion, the symbols work out nicely that way). The formula is simply the formula for variance under a square root sign.

Standard Deviation

population standard deviation is given as σ :

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Example from Table 3 (data represent the full population):

$$\sigma = \sqrt{(134.8/5)} = \sqrt{26.96} = 5.19$$

Example from table 4 (data represent the full population):

$$\sigma = \sqrt{(30/20)} = \sqrt{1.5} = 1.22$$

The sample statistic follows the same conventions and is given as s . It is the square root of the sample variance.

Standard Deviation

sample standard deviation is given as s :

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

It can be noted shorthand as $s = SS/df$

Example from Table 3 (treat as sample):

$$S^2 = 134.8/(5-1) = 33.7$$

Example from Table 4 (treating those scores as a sample):

$$S^2 = 30/(20 - 1) = 1.58$$

The Normal Distribution and Standard Deviation

The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal because the proportion of the distribution within a given number of standard deviations from the mean can be calculated.

For a normal distribution,

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.

- More than 99% of the data is within three standard deviations of the mean.

This is known as the **Empirical Rule or the 68-95-99 Rule**, as shown in Figure 3.

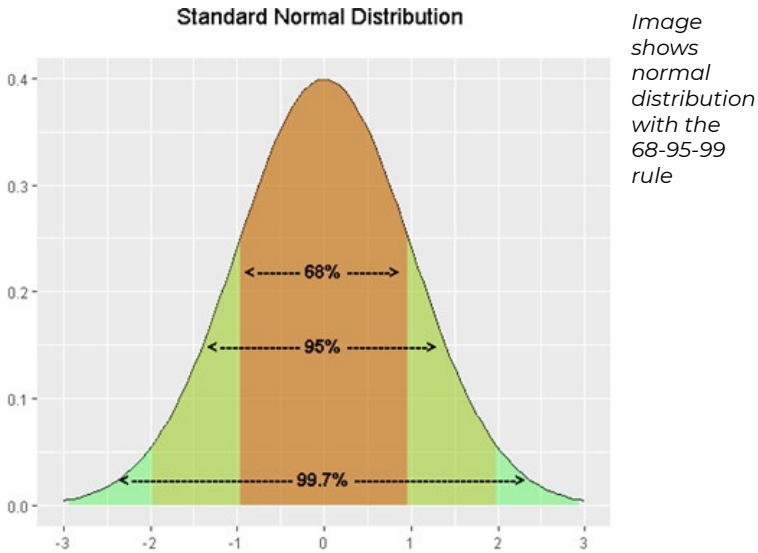


Figure 3: Percentages of the normal distribution showing the 68-95-99 rule.

For example, if you had a normal distribution with a mean of 50 and a standard deviation of 10, then 68% of the distribution would be between $50 - 10 = 40$ and $50 + 10 = 60$. Similarly, about 95% of the distribution would be between $50 - 2 \times 10 = 30$ and $50 + 2 \times 10 = 70$.

Figure 4 shows two normal distributions. The red distribution has a mean of 40 and a standard deviation of 5; the blue distribution has a mean of 60 and a standard deviation of 10. For the red distribution, 68% of the distribution is between 45 and 55; for the blue distribution, 68% is between 50 and 70. Notice that as the standard deviation gets smaller, the

distribution becomes much narrower, regardless of where the center of the distribution (mean) is. Figure 5 presents several more examples of this effect.

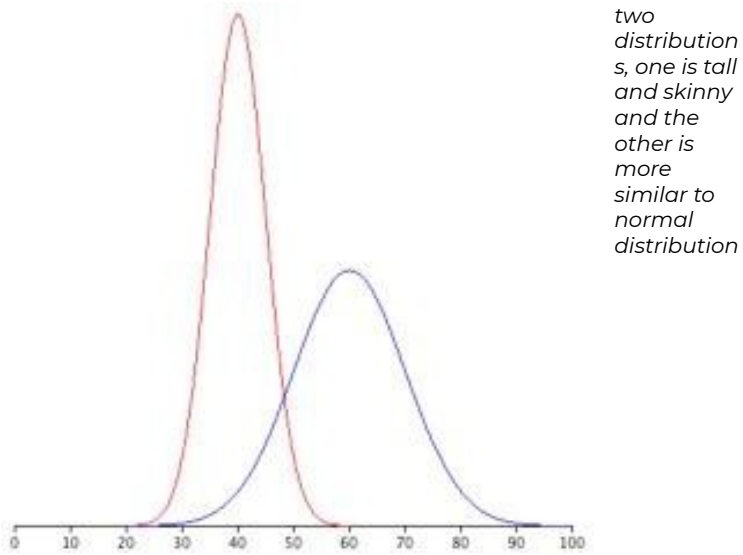


Figure 4. Normal distributions with standard deviations of 5 and 10.

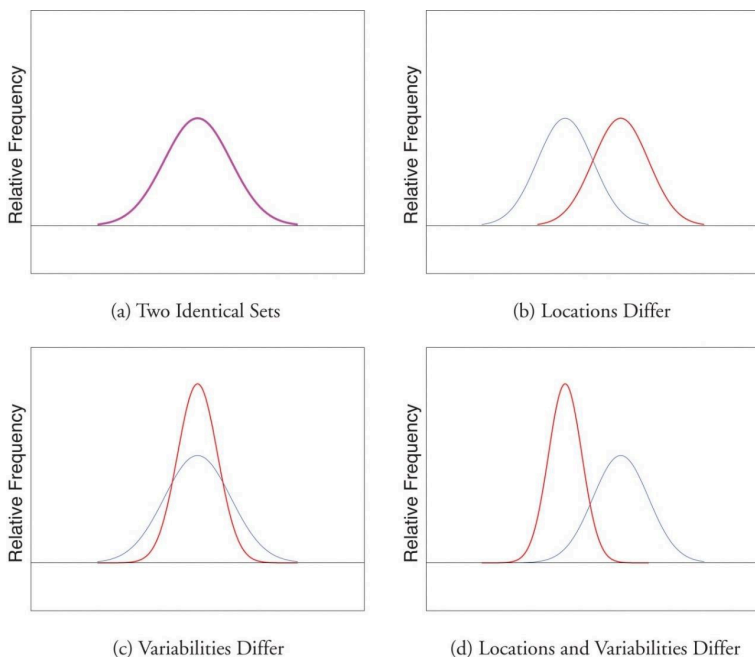


Figure 5. Differences between two datasets.

The image below represents IQ scores as measured by the Wechsler Intelligence test and has a $\mu = 100$ and $\sigma = 15$. This means that about 70% of the scores are between 85 and 115 and that 95% of the scores are between 70 and 130.

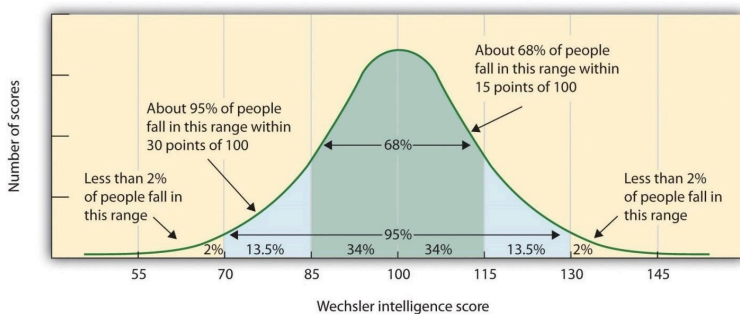


Figure 6. Weshcler IQ Score distribution. [Photo credit](#)

A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average, and can define outside two

standard deviations as an **outlier** (extreme score). Considering data to be far from the mean if it is more of an approximate “rule of thumb” than a rigid rule. Some researchers may define an outlier as greater than 3 standard deviations from the mean. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations.

Example

Let's use an example to help us understand how we can use standard deviation. Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B.

The average wait time at both supermarkets is five minutes. At supermarket A, the standard deviation for the wait time is two minutes. At supermarket B the standard deviation for the wait time is four minutes

Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the average; wait times at supermarket A are more concentrated near the average.

Suppose that Rosa and Binh both shop at supermarket A. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. Rosa waits for seven minutes – Seven is two minutes longer than the average of five; two. Binh waits for one minute

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.



A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be an outlier. Again, we would also want to do more about the distributions for this data.

Recap

| Formula for Sum of Squares for population (using μ) | Symbol | Meaning |
|--|---------------------|---|
| $\Sigma(X - \mu)^2$ | X | Raw score |
| | $X - \mu$ | Deviation score that is calculated by subtracting raw scores from population mean |
| | $(X - \mu)^2$ | The deviation scores are squared |
| | $\Sigma(X - \mu)^2$ | The squared deviation scores are added up to calculate sum of squares |

The formula for variance expresses the mathematical definition in symbols. Recall that the symbol for the variance

(like any statistic) changes depending on whether the statistic is referring to a sample or a population. Although the variance is itself a measure of variability, it generally plays a larger role in inferential statistics than in descriptive statistics.

| Formulas | |
|------------------------------------|------------------------------------|
| Sample variance (s^2) | Population variance (σ^2) |
| SS/df Note: sample $df = n-1$ | SS/N |

The standard deviation is the most commonly used measure of variability because it includes all the scores of the data set in the calculation, and it is reported in the original units of measurement. It tells us the average (or standard) distance of each score from the mean of the distribution.

| Sample Standard Deviation (s) | Population Standard Deviation (σ) |
|-----------------------------------|--|
| $\sqrt{SS/df}$ | $\sqrt{SS/N}$ |

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

It is important to note that the Empirical Rule or the 68-95-99

rule only applies when the shape of the distribution of the data is bell-shaped and symmetric.

Factors Affecting Variability

Before we close out the chapter, we wanted to make you aware that there are several things that can impact the spread of scores.

Extreme Scores. Range is affected most by extreme scores or outliers but standard deviation and variance are also affected by extremes because they are based on squared deviations. One extreme score can have a disproportionate effect on the overall statistic or parameter.

Sample size. Increased sample size is associated with an increase in range because of the potential to increase or decrease values in a set of data.

Stability under sampling. If you take multiple samples from the same population you expect similar values because the samples come from the same source. This accounts for their stability and we would expect samples to have the same variability as the population from which it was selected.

Learning Objectives

Having read this chapter, you should be able to:

- Explain the purpose of measuring variability and differences between scores with high versus low variability
- Define and calculate measures of spread/dispersion: range, variance, standard deviation for sample and for population
- Define and calculate sum of squared deviations (SS)

Exercises – Ch. 5

1. Compute the population standard deviation for the following scores (remember to use the Sum of Squares table and this is the same data from chapter 4):

5, 7, 8, 3, 4, 4, 2, 7, 1, 6

2. For the following problem, use the following scores: 5, 8, 8, 8, 7, 8, 9, 12, 8, 9, 8, 10, 7, 9, 7, 6, 9, 10, 11, 8

3. Compute the range, sample variance, and sample standard deviation for the following scores: 25, 36, 41, 28, 29, 32, 39, 37, 34, 34, 37, 35, 30, 36, 31, 31 (same data from chapter 4)

4. Using the same values from problem 3, calculate the range, sample variance, and sample standard deviation, but this time include 65 in the list of values. How did each of the three values change?

5. Two normal distributions have exactly the same mean, but one has a standard deviation of 20 and the other has a standard deviation of 10. How would the shapes of the two distributions compare?

6. Compute the sample standard deviation for the following scores: -8, -4, -7, -6, -8, -5, -7, -9, -2, 0 (same data from chapter 4)

Answers to Odd-Numbered Exercises – Ch. 5

1. ($\mu = 4.80$) $\sigma^2 = 2.36$

3. range = 16, $s^2 = 18.40$, $s = 4.29$

5. If both distributions are normal, then they are both symmetrical, and having the same mean causes them to overlap with one another. The distribution with the standard deviation of 10 will be narrower than the other distribution

6. Chapter 6: z-scores and the Standard Normal Distribution

We now understand how to describe and present our data visually and numerically. These simple tools, and the principles behind them, will help you interpret information presented to you and understand the basics of a variable. Moving forward, we now turn our attention to how scores within a distribution are related to one another, how to precisely describe a score's location within the distribution, and how to compare scores from different distributions.

Revisiting Percentiles

In many situations, it is useful to have a way to describe the location of an individual score within its distribution. One approach is the percentile rank. The percentile rank of a score is the percentage of scores in the distribution that are lower than that score. Percentiles are useful for comparing values.

Consider, for example, the distribution of Rosenberg Self-esteem scores we used in chapter 2. For any score in the distribution, we can find its *percentile rank* by counting the number of scores in the distribution that are lower than that score and converting that number to a percentage of the total number of scores.

| Self-Esteem Scores | Frequency | Cumulative Frequency | Cumulative Percentage |
|--------------------|-----------|----------------------|-----------------------|
| 24 | 3 | 40 | 100 |
| 23 | 5 | 37 | 92.5 |
| 22 | 10 | 32 | 80 |
| 21 | 8 | 22 | 55 |
| 20 | 5 | 14 | 35 |
| 19 | 3 | 9 | 22.5 |
| 18 | 3 | 6 | 15 |
| 17 | 0 | 3 | 7.5 |
| 16 | 2 | 3 | 7.5 |
| 15 | 1 | 1 | 2.5 |

Table 1. Frequency table for Rosenberg self-esteem scores

Notice, for example, that five of the students represented by the data in the table had self-esteem scores of 23. In this distribution, 32 of the 40 scores (80%) are lower than 23 (note that you can see for score 22 showing cumulative frequency as 32). Thus, for students with a score of 23, they have a percentile rank of 80 percent. It can also be said that they scored at the 80th percentile. Remember that *percentile rank* by counting the number of scores in the distribution that are *lower* than that score and converting that number to a percentage of the total number of scores ($32/40 = 80\%$). Percentile ranks are often used to report the results of standardized tests of ability or achievement. If your percentile rank on a test of verbal ability were 40, for example, this would mean that you scored higher than 40% of the people who took the test.

Normal Distributions

The normal distribution is the most important and most widely

used distribution in statistics. It is sometimes called the “bell curve,” although the tonal qualities of such a bell would be less than pleasing. It is also called the “Gaussian curve” of Gaussian distribution after the mathematician Carl Friedrich Gauss. Let’s review a little bit about the normal distribution. The normal distribution is described in terms of two parameters: the mean (which you can think of as the location of the peak), and the standard distribution (which specifies the width of the distribution). The bell-like shape of the distribution never changes, only its location and width. The normal distribution is commonly observed in data collected in the real world, as we have already seen in Chapter 3 — and in chapter 7 we will learn more about why this occurs.



Photo of Gauss Monument dedicated to the mathematician, geodesist and astronomer [Carl Friedrich Gauß](#). It is placed in his place of birth Brunswick. [Photo credit](#)

Strictly speaking, it is not correct to talk about “the normal distribution” since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. Figure 1 shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5 , the distribution in red (the middle distribution)

has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails. What is consistent about all normal distribution is the shape and the proportion of scores within a given distance along the x-axis. We will focus on the **Standard Normal Distribution** (also known as the *Unit Normal Distribution*), which has a mean of 0 and a standard deviation of 1 (i.e. the red distribution in Figure 1).

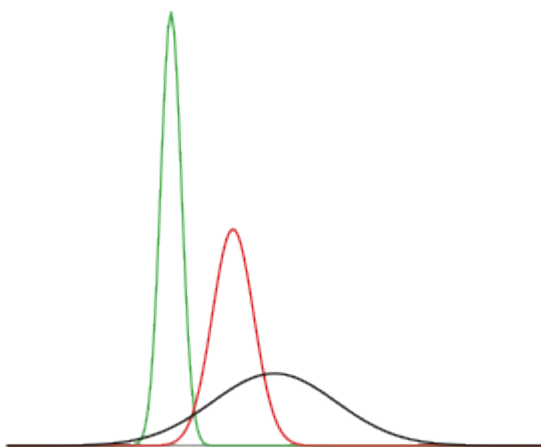


Figure 1. Normal distributions differing in mean and standard deviation.

Seven features of normal distributions are listed below.

- Normal distributions are symmetric around their mean.
- The mean, median, and mode of a normal distribution are equal.
- The area under the normal curve is equal to 1.0.
- Normal distributions are denser in the center and less dense in the tails.
- Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).

- 68% of the area of a normal distribution is within one standard deviation of the mean.
- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

These properties enable us to use the normal distribution to understand how scores relate to one another within and across a distribution. But first, we need to learn how to calculate the standardized score than make up a standard normal distribution.

Z-scores

As we learned in earlier lessons, population mean (μ) and population standard deviation (σ) are methods for describing an entire distribution of scores using individual scores. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading. A **z-score** is a standardized version of a raw score (x) that gives information about the relative location of that score within its distribution. Z-scores are standardized scores that identify and describe the *exact location of every score* within a distribution. By transforming our values (raw score) we can compare z-scores across different samples or groups and make meaningful comparisons. Each value in the distribution has a z-score that can be calculated to standardize for comparison.

Let's say that you received a score of 76 on your Chemistry exam and your friend receives a score of 76 on her Physics exam. Who is doing better in class? It is hard to say because we do not have enough information. This is an example of how z-scores can facilitate meaningful comparisons. The z score for a particular individual is the difference between that individual's

score and the mean of the distribution, divided by the standard deviation of the distribution.

Formulas to calculate Z scores

Population

$$z = \frac{x - \mu}{\sigma}$$

Sample

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

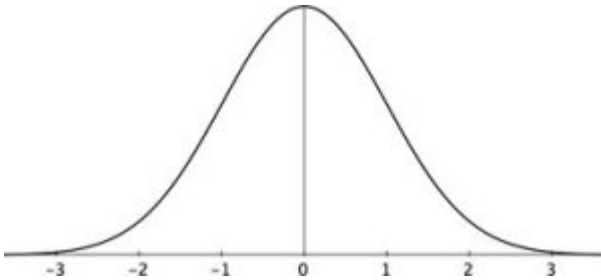
Note that it is essentially the same formula where the appropriate symbols for mean and standard deviation have been used depending on if working with population or sample data.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

As you can see, z-scores combine information about where the distribution is located (the mean/center) with how wide the distribution is (the standard deviation/spread) to interpret a raw score (x). Specifically, z-scores will tell us how far the score is away from the mean in units of standard deviations and in what direction. Z-scores transforms raw scores into units of standard deviation above or below the mean. This transformation provides a reference using the standard normal distribution. If we are given a Z score we know where, relative to the mean, the Z score and raw score lies.

The value of a z-score has two parts: the sign (positive or negative) and the magnitude (the actual number). The sign of the z-score tells you in which half of the distribution the z-score falls: a positive sign (or no sign) indicates that the score is above the mean and on the right hand-side or upper end of the distribution, and a negative sign tells you the score is below the mean and on the left-hand side or lower end of the distribution. The magnitude of the number tells you, in units of standard deviations, how far away the score is from the center or mean. The magnitude can take on any value between negative and positive infinity, but for reasons we will see soon, they generally fall between -3 and 3.

Z-scores & the standard normal distribution



Z-score distribution

- If the Z score is negative, then the score falls **below** the mean
- If the Z score is 0, then the score falls **at** the mean
- If the Z score is positive, then the score falls **above** the mean

Let's look at some examples. A z-score value of -1.0 tells us that this z-score is 1 standard deviation (because of the magnitude 1.0) below (because of the negative sign) the mean.

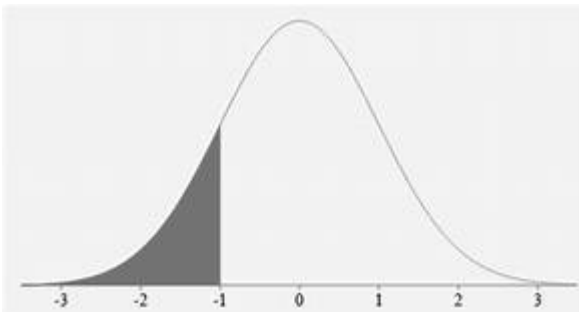


Figure 2. z-score of -1

Similarly, a z-score value of 1.0 tells us that this z-score is 1

standard deviation above the mean. Thus, these two scores are the same distance away from the mean but in opposite directions. A z-score of -2.5 is two-and-a-half standard deviations below the mean and is therefore farther from the center than both of the previous scores, and a z-score of 0.25 is closer than all of the ones before.

We can convert raw scores into z-scores to get a better idea of where in the distribution those scores fall. Let's say we get a score of 68 on an exam ($X=68$). We may be disappointed to have scored so low, but perhaps it was just a very hard exam. Having information about the distribution of all scores in the class would be helpful to put some perspective on ours. We find out that the class got an average score (M) of 54 with a standard deviation (s) of 8. To find out our relative location within this distribution, we simply convert our test score into a z-score.

$$z = (X - M)/s = (68 - 54)/8 = 1.75$$

We find that we are 1.75 standard deviations above the average, above our rough cut off for close and far. Suddenly our 68 is looking pretty good!

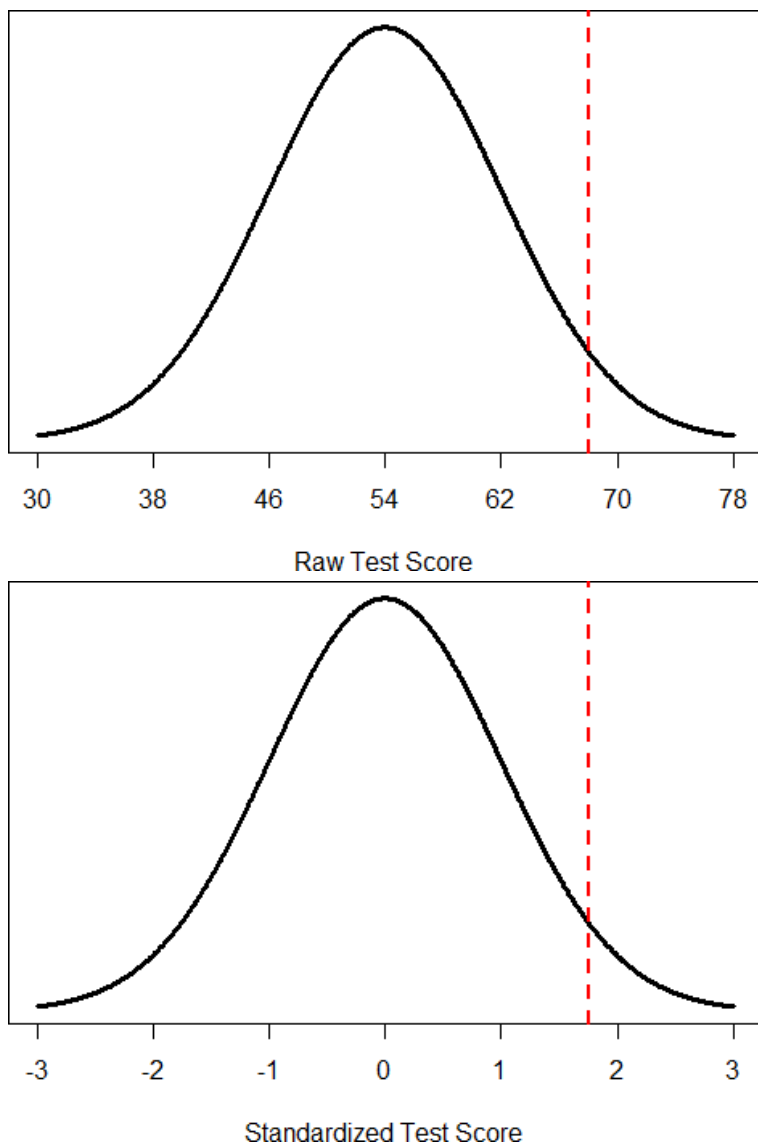


Figure 3. Raw and standardized versions of a single score

Figure 3 shows both the raw score and the z-score on their respective distributions. Notice that the red line indicating where each score lies is in the same relative spot for both. This is because transforming a raw score into a z-score does

not change its relative location, it only makes it easier to know precisely where it is.

Example comparing raw scores

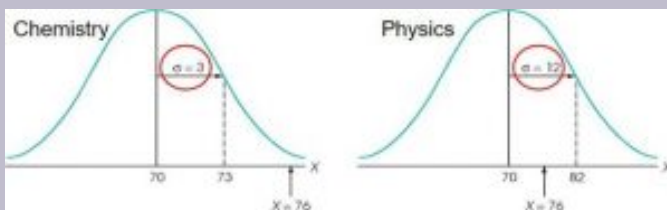
Let's go back to our Chemistry and Physics exam score comparisons. Each student received a score of $x = 76$ on the exam. Assume that:

- The average score for *both* exams is $M = 70$.
- The standard deviation for Chemistry is $s = 3$
- The standard deviation for Physics is $s = 12$.

A z score indicates how far above or below the mean a raw score is, but it expresses this in terms of the standard deviation. The z-scores for our example are above the mean.

- Chemistry z-score is $z = (76-70)/3 = +2.00$
- Physics z-score is $z = (76-70)/12 = +0.50$

When we compare the position of the test score $X = 76$ it is clear that these two distributions are very different and that the Chemistry score has a higher position in the distribution.



As mentioned earlier, z-scores are also useful for comparing scores from different distributions. Let’s say we take the SAT and score 501 on both the math and critical reading sections. Does that mean we did equally well on both? Scores on the math portion are distributed normally with a mean of 511 and standard deviation of 120, so our z- score on the math section is $z_{\text{math}} = 501 - 511/120 = -0.08$ which is just slightly below average (note that use of “math” as a subscript; subscripts are used when presenting multiple versions of the same statistic in order to know which one is which and have no bearing on the actual calculation). The critical reading section has a mean of 495 and standard deviation of 116, so $z_{\text{reading}} = 501 - 495/116 = 0.05$. So even though we were almost exactly average on both tests, we did a little bit better on the critical reading portion relative to other people.

Finally, z-scores are incredibly useful if we need to combine information from different measures that are on different scales. Let’s say we give a set of employees a series of tests on things like job knowledge, personality, and leadership. We may want to combine these into a single score we can use to rate employees for development or promotion, but look what happens when we take the average of raw scores from different scales, as shown in Table 2:

| Raw Scores | Job Knowledge (0 – 100) | Personality (1 – 5) | Leadership (1 – 5) | Average |
|------------|-------------------------|---------------------|--------------------|---------|
| Employee 1 | 98 | 4.2 | 1.1 | 34.43 |
| Employee 2 | 96 | 3.1 | 4.5 | 34.53 |
| Employee 3 | 97 | 2.9 | 3.6 | 34.50 |

Table 2. Raw test scores on different scales (ranges in parentheses).

Because the job knowledge scores were so big and the scores were so similar, they overpowered the other scores and removed almost all variability in the average. However, if we standardize these scores into z-scores, our averages retain more variability and it is easier to assess differences between employees, as shown in Table 3.

| z-scores | Job Knowledge (0 – 100) | Personality (1 – 5) | Leadership (1 – 5) | Average |
|------------|----------------------------|------------------------|-----------------------|---------|
| Employee 1 | 1.00 | 1.14 | -1.12 | 0.34 |
| Employee 2 | -1.00 | -0.43 | 0.81 | -0.20 |
| Employee 3 | 0.00 | -0.71 | 0.30 | -0.14 |

Table 3. Standardized scores

Setting the scale of a distribution

Another convenient characteristic of z-scores is that they can be converted into any “scale” that we would like. Here, the term scale means how far apart the scores are (their spread) and where they are located (their central tendency). In other words, we can convert that value into its original raw score (X) if the mean and standard deviation are known. We can still do this using the Z score formula we have been using so far this lesson. We just need to rearrange the variables so we are solving for X instead of Z.

The formulas for transforming z to x are:

Z SCORE TO RAW SCORE - SAMPLE

$$X = (Z)(SD) + M$$

Z SCORE TO RAW SCORE - POPULATION

$$X = (Z)(\sigma) + \mu$$

Note: these are just simple rearrangements of the original formulas for calculating z from raw scores.

A problem is that these new z-scores aren't exactly intuitive for many people. We can give people information about their relative location in the distribution (for instance, the first person scored well above average). Another route we can do is to take the z-scores and transform them to a known distribution, like the traditional IQ distribution.

Let's say we have z-scores of 1.71, .43, and -.80 after converting their raw intelligence test score to a z-score. We can translate these z-scores into the more familiar metric of IQ scores, which have a mean of 100 and standard deviation of 16. We can use the transforming formula from above: $X = z * SD + M$

$$X = 1.71 * 16 + 100 = 127.36, \text{ so}$$

IQ score of 127

$$X = 0.43 * 16 + 100 = 106.88, \text{ so}$$

IQ score of 107

$$X = -0.80 * 16 + 100 = 87.20, \text{ so}$$

IQ score of 100

We rounded the values to 127, 107, and 87, respectively, for convenience.

Z-scores and the Area under the Curve

Even though we can use a z-score as a measure of relative standing for any shape of frequency distribution, we commonly use z-scores in this class when discussing normal distributions. They provide a way of describing where an individual's score is located within a distribution and are sometimes used to report the results of standardized tests.

Z-scores and the standard normal distribution go hand-in-hand. A z-score will tell you exactly where in the standard normal distribution a value is located, and any normal distribution can be converted into a standard normal distribution by converting all of the scores in the distribution into z-scores, a process known as standardization. We will also see that one can identify the percentile for each z-score in a normal distribution.

Since a z-score tells us how far above or below the mean a particular raw score lies (in standard deviation units), we can use z-scores in conjunction with the empirical rule. We can use z-scores to simplify the earlier statements we made regarding the *Empirical Rule (68-95-99 rule)*:

- 68% of all scores will fall between a Z score of -1.00 and +1.00
- 95% of all scores will fall between a Z score of -2.00 and +2.00
- 99.7% of all scores will fall between a Z score of -3.00 and +3.00
- 50% of all scores lie above/below a Z score of 0.00

Take a minute to examine Figure 4 to identify these areas. For

example, you can see adding up the 2 areas between $z = -1$ to $z = 1$, you get 68.2%. Because z-scores are in units of standard deviations, this means that 68% of scores fall between $z = -1.0$ and $z = 1.0$ and so on. We call this 68% (or any percentage we have based on our z-scores) the proportion of the area under the curve. Remember, *these percentages remain true only if our sample or population is normally distributed!*

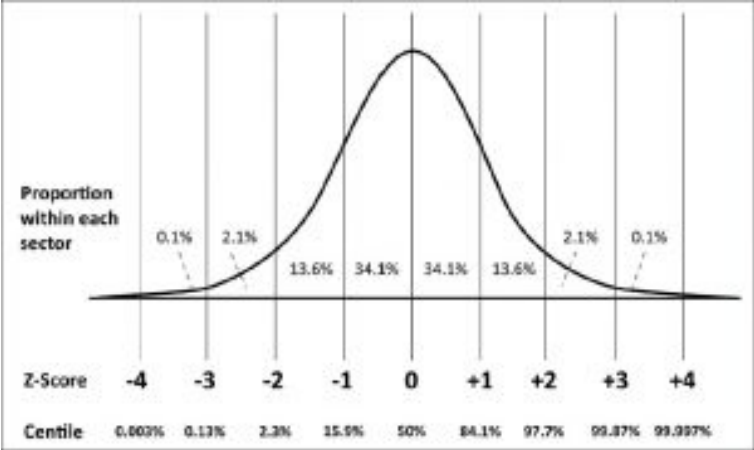


Figure 4. Z-score indicating percentiles in a standardized normal distribution.

Any area under the curve is bounded by (defined by, delineated by, etc.) by a single z-score or pair of z-scores. An important property to point out here is that, by virtue of the fact that the total area under the curve of a distribution is always equal to 1.0 (see section on Normal Distributions at the beginning of this chapter), these areas under the curve can be added together or subtracted from 1 to find the proportion in other areas. For example, we know that the area between $z = -1.0$ and $z = 1.0$ (i.e. within one standard deviation of the mean) contains 68% of the area under the curve, which can be represented in decimal form at 0.6800 (to change a percentage to a decimal, simply move the decimal point 2 places to the left). Because the total area under the curve is equal to 1.0, that means that the proportion of the area outside

$z = -1.0$ and $z = 1.0$ is equal to $1.0 - 0.6800 = 0.3200$ or 32% (see Figure 5 below). This area is called the area in the tails of the distribution. Because this area is split between two tails and because the normal distribution is symmetrical, each tail has exactly one-half, or 16%, of the area under the curve.

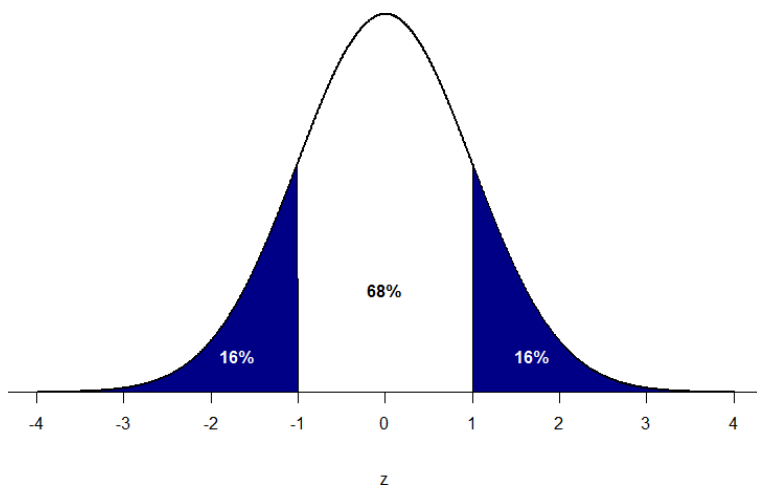


Figure 5. Shaded areas represent the area under the curve in the tails

Additionally, z-scores provide one way of defining *outliers*. For example, outliers are sometimes defined as scores that have z scores less than -3.00 or greater than $+3.00$. In other words, they are defined as scores that are more than three standard deviations from the mean. Some researchers will define outliers as greater than 2 standard deviations from the mean.

We will have much more to say about percentiles in a distribution in the coming chapters. As it turns out, this is a quite powerful idea that enables us to make statements about how likely an outcome is and what that means for research questions we would like to answer and hypotheses we would like to test. But first, we need to make a brief foray into some ideas about probability.

Learning Objectives

Having read this chapter, you should be able to:

- Identify uses for z-score
- Compute and transform z-scores and x-values
- Describe the effects of standardizing a distribution
- Identify the z-score location on a normal distribution

Exercises – Chapter 6

1. What are the two pieces of information contained in a z-score?
2. A z-score takes a raw score and standardizes it into units of.
3. Assume the following 5 scores represent a sample: 2, 3, 5, 5, 6. Transform these scores into z-scores.
4. True or false:
 1. All normal distributions are symmetrical
 2. All normal distributions have a mean of 1.0
 3. All normal distributions have a standard deviation of 1.0
 4. The total area under the curve of all normal distributions is equal to 1
5. Interpret the location, direction, and distance (near or far) of the following z- scores:
 1. -2.00
 2. 1.25
 3. 3.50
 4. -0.34
6. Transform the following z-scores into a distribution with a mean of 10 and standard deviation of 2: -1.75, 2.20, 1.65, -0.95

7. Calculate z-scores for the following raw scores taken from a population with a mean of 100 and standard deviation of 16: 112, 109, 56, 88, 135, 99
8. What does a z-score of 0.00 represent?
9. For a distribution with a standard deviation of 20, find z-scores that correspond to:
 1. One-half of a standard deviation below the mean
 2. 5 points above the mean
 3. Three standard deviations above the mean
 4. 22 points below the mean
10. Calculate the raw score for the following z-scores from a distribution with a mean of 15 and standard deviation of 3:
 1. 4.0
 2. 2.2
 3. -1.3
 4. 0.46
11. Let's say we create a new measure of intelligence, and initial calibration finds that our scores have a mean of 40 and a standard deviation of 7. Three people who have scores of 52, 43, and 34 want to know how well they did on the measure. Convert their raw scores into z-scores.

Answers to Odd-Numbered Exercises – Ch. 6

1. The location above or below the mean (from the sign of the number) and the distance in standard deviations away from the mean (from the magnitude of the number).
3. $X = 4.2$, $s = 1.64$; $z = -1.34, -0.73, 0.49, 0.49, 1.10$
 5. 1.2 standard deviations below the mean, far
 2. 1.25 standard deviations above the mean, near
 3. 3.5 standard deviations above the mean, far

4. 0.34 standard deviations below the mean, near

7. $z = 0.75, 0.56, -2.75, -0.75, 2.19, -0.06$

9. 1. -0.50, 2. 0.25, 3. 3.00, 4. 1.10

11. $Z = (52 - 40)/7 = 1.71$

$Z = (43 - 40)/7 = 0.43$

$Z = (34 - 40)/7 = -0.80$

7. Chapter 7:

Probability

In this lesson, we start to move away from descriptive statistics and begin our transition into inferential statistics. Recall that the goal of inferential statistics is to draw conclusions or make predictions about large populations by using data from smaller samples that represent that population. *Probability is the underlying concept of inferential statistics and forms a direct link between samples and the population that they come from.* In this chapter we will focus only on the principles and ideas necessary to lay the groundwork for future inferential statistics. We accomplish this by quickly tying the concepts of probability to what we already know about normal distributions and z-scores.

What is probability?

Informally, we usually think of **probability** as a number that describes the likelihood of some event occurring, which ranges from zero (impossibility) to one (certainty). Probability can be discussed more vaguely. Chances are low it will rain today. Given there are no clouds, chances are low it will rain today. “Given” is the word we use to state what the conditions are. As the conditions change, so does the probability. Thus, if it were cloudy and windy outside, we might say, “given the current weather conditions, there is a high probability that it is going to rain.” It should also be noted that the terms “low” and “high” are relative and vague, and they will likely be interpreted differently by different people (in other words: given how vague the terminology was, the probability of different interpretations is

high). In statistics, most of the time we try to use more precise language or, even better, numbers to represent the probability of our event. Regardless, the basic structure and logic of our statements are consistent with how we speak about probability using numbers and formulas.

Sometimes probabilities will instead be expressed in percentages, which range from zero to one hundred, as when the weather forecast predicts a twenty percent chance of rain today. In each case, these numbers are expressing how likely that particular event is, ranging from absolutely impossible (0%) to absolutely certain (100%). When we speak of the probability of something happening, we are talking how likely it is that “thing” will happen based on the conditions present. To formalize probability theory, we first need to define a few terms:

Probability theory is the branch of mathematics that deals with chance and uncertainty. It forms an important part of the foundation for statistics, because it provides us with the mathematical tools to describe uncertain events. The study of probability arose in part due to interest in understanding games of chance, like cards or dice. These games provide useful examples of many statistical concepts, because when we repeat these games the likelihood of different outcomes remains (mostly) the same.

To calculate probability, we need an activity that produces or observes an outcome is needed. This is typically an experiment or any situation or activity in which the result is not known in advance. Examples are the outcome for: the weather outside, flipping a coin, rolling a 6-sided die, or trying a new route to work to see if it's faster than the old route.

We also need to know the known outcomes of the activity/event/experiment. The **sample space** is the set of possible outcomes for an activity. We represent these by listing them within a set of squiggly brackets.

- For rain is {rain, not rain}
- For a coin flip, the sample space is {heads, tails}.
- For a six-sided die, the sample space is each of the possible numbers that can appear: {1,2,3,4,5,6}.
- For the amount of time it takes to get to work, the sample space is all possible real numbers greater than zero (since it can't take a negative amount of time to get somewhere, at least not yet). We won't bother trying to write out all of those numbers within the brackets.

An **outcome or event** is a subset of the sample space to examine specific probability. In principle, it could be one or more of possible outcomes in the sample space, but here we will focus primarily on *elementary events* which consist of exactly one possible outcome. An event is a catch-all term to talk about any specific thing happening.

- For example, this could be it rains, obtaining heads in a single coin flip, rolling a 4 on a throw of the die, or taking 21 minutes to get home by the new route.

In statistics, we usually define probability as the **expected relative frequency of a particular outcome**. The **relative frequency** is the number of times an event takes place relative to the number of times it could have taken place.

Let's look at a slightly deeper example before we learn a basic probability formula. Say we have a regular, six-sided die (note that "die" is singular and "dice" is plural, a distinction that can be hard to get correct on the first try) and want to know how likely it is that we will roll a 1. That is, what is the probability of rolling a 1, given that the die is not weighted (which would introduce what we call a bias, though that is beyond the scope of this chapter). We could roll the die and see if it is a 1 or not, but that won't tell us about the probability, it will only tell us a single result. We could also roll the die hundreds or thousands

of times, recording each outcome and seeing what the final list looks like, but this is time-consuming, and rolling a die that many times may lead down a dark path to gambling or, worse, playing Dungeons & Dragons. What we need is a simple equation that represents what we are looking for and what is possible.

To calculate the probability of an event, which here is defined as rolling a 1 on an unbiased die, we need to know two things: how many outcomes satisfy the criteria of our event (stated different, how many outcomes would count as what we are looking for) and the total number of outcomes possible. In our example, only a single outcome, rolling a 1, will satisfy our criteria, and there are a total of six possible outcomes (rolling a 1, rolling a 2, rolling a 3, rolling a 4, rolling a 5, and rolling a 6). Thus, the probability of rolling a 1 on an unbiased die is 1 in 6 or 1/6.

Put into an equation using generic terms, we get:

Probability

$$\text{Probability} = \frac{\text{Number of favorable (desired) outcomes}}{\text{Total number of possible outcomes}}$$

We can also use $P()$ as shorthand for probability and we can use A as shorthand for an event:

$$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of outcomes}}$$

Probability is usually symbolized by the letter p . The actual probability number is usually written as a decimal, though sometimes fractions or percentages are used. An event with a 50-50 chance is usually written as $p = .5$ but it could be written as $p = 1/2$ or $p = 50\%$. It is also common to see probability written as being less than some value, using the less than ($<$) sign. For example, $p < .05$ means the probability of the event taking place is less than .05 or less than 5%.

Using the above equation, let's calculate the probability of rolling an even number on this die:

$P(\text{even number}) = 2, 4, \text{ or } 6 / 1, 2, 3, 4, 5, \text{ or } 6 = 3/6 = .5$. So we have a 50% chance of rolling an even number of this die.

Let's look at another example, let's say that we are interested in knowing the probability of rain in Phoenix. We first have to define the activity — let's say that we will look at the National Weather Service data for each day in 2020 and determine whether there was any rain at the downtown Phoenix weather station. According to these data, in 2020 there were 15 rainy days. To compute the probability of rain in Phoenix, we simply divide the number of rainy days by the number of days counted (365), giving $p(\text{rain in PHX in 2020}) = 0.04$.

Now that we have a probability formula, we can outline the

formal features of probability (first defined by the Russian mathematician Andrei Kolmogorov). These are the features that a value *has* to have if it is going to be a probability.

- Probability cannot be negative.
- The total probability of all outcomes in the sample space is 1; that is, if we take the probability of each event and add them up, they must sum to 1. This is interpreted as saying “Take all of the possible events and add up their probabilities. These must sum to one.”
- The probability of any individual event cannot be greater than one. This is implied by the previous point; since they must sum to one, and they can’t be negative, then any particular probability cannot exceed one.

To summarize, the probability that an event happens is the number of outcomes that qualify as that event (i.e. the number of ways the event could happen) compared to the total number of outcomes (i.e. how many things are possible). The principles laid out here operate under a certain set of conditions and can be elaborated into ideas that are complex yet powerful and elegant. However, such extensions are not necessary for a basic understanding of statistics, so we will end our discussion on the math of probability here. We will now return to a more familiar topic. This idea then brings us back around to our normal distribution, which can also be broken up into regions or areas, each of which are bounded by one or two z-scores and correspond to all z-scores in that region. The probability of randomly getting one of those z-scores in the specified region can then be found on the **Standard Normal Distribution Table**. Thus, the larger the region, the more likely an event is, and vice versa. Because the tails of the distribution are, by definition, smaller and we go farther out into the tail, the likelihood or probability of finding a result out in the extremes becomes small.

Probability & Frequency Distributions

For our purposes, we will see shortly that the normal distribution is the key to how probability works. If you toss a fair coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $1/2$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions (law of large numbers). The pattern tends to resemble a symmetrical normal distribution.

To help us think about probability, population, and inferential statistics we are going to use a frequency distribution because it can be seen as representing an entire population. This can be seen as a parallel concept because if all scores are represented in a frequency distribution it can function as a normal distribution. Using the empirical rule we know that different portions of the histogram can represent different proportions of the population and the terms proportions and probabilities mean the same thing. This means that a proportion of the histogram can correspond to the probability of a population.

Probability in Normal Distributions

If the language at the end of the last section sounded familiar, that's because it's exactly the language used in the last chapter to describe the normal distribution. Recall that the normal distribution has an area under its curve that is equal to 1 and that it can be split into sections by drawing a line through it that corresponds to a given z-score. Because of this, we can interpret areas under the normal curve as probabilities that correspond to z-scores. In this section, we are going to link

together the concepts of population, probability, and z-scores. We learned earlier that a frequency distribution can represent an entire population of scores. The shape of a frequency distribution for an entire population forms a symmetrical normal curve and certain proportions can be assigned to specific parts of the distribution.

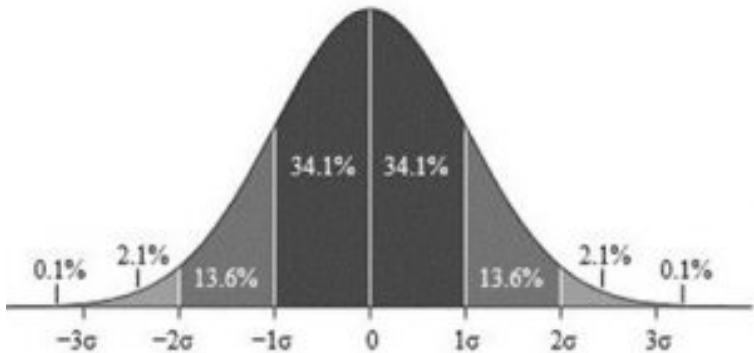


Figure 1. Z-Score Distribution. Photo Credit: M. W. Toews, via Wikimedia Commons

The graph above (Figure 1) only shows us some of the proportions associated with specific z-score values but the Unit Normal Table lists all possible values for a normal distribution. This means that we can use z-scores to help us find the specific probability for a specific outcome or event.

First, let's look back at the area between $z = -1.00$ and $z = 1.00$ presented in Figure 2.

- We were told earlier that this region contains 68% of the area under the curve. Thus, if we randomly chose a z-score from all possible z-scores, there is a 68% chance that it will be between $z = -1.00$ and $z = 1.00$ because those are the z-scores that satisfy our criteria.

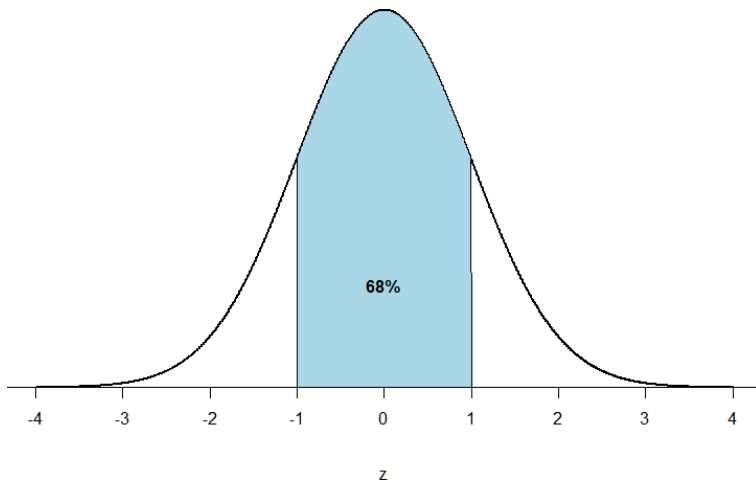


Figure 2: There is a 68% chance of selection a z-score from the blue-shaded region

Just like a pie chart is broken up into slices by drawing lines through it, we can also draw a line through the normal distribution to split it into sections. Take a look at the normal distribution in Figure 3 which has a line drawn through it as $z = 1.25$. This line creates two sections of the distribution: the smaller section called the tail and the larger section called the body. Differentiating between the body and the tail does not depend on which side of the distribution the line is drawn. All that matters is the relative size of the pieces: bigger is always body.

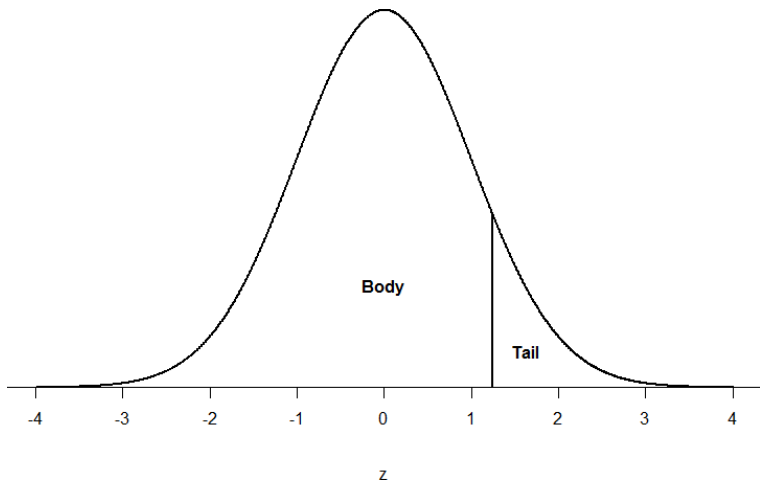


Figure 3. Body and tail of the normal distribution

As you can see, we can break up the normal distribution into 3 pieces (lower tail, body, and upper tail) as in Figure 2 or into 2 pieces (body and tail) as in Figure 3. We can then find the proportion of the area in the body and tail based on where the line was drawn (i.e. at what z-score). Mathematically this is done using calculus.

Fortunately, the exact values are given to you in the *Standard Normal Distribution Table*, also known as the **z-table**. Using the values in this table, we can find the area under the normal curve in any body, tail, or combination of tails no matter which z-scores are used to define them.

The z-table presents the values for the area under the curve to the left of the positive z-scores from 0.00-3.00 (technically 3.09), as indicated by the shaded region of the distribution at the top of the table. To find the appropriate value, we first find the row corresponding to our z-score then follow it over until we get to the column that corresponds to the number in the hundredths place of our z-score. For example, suppose we want to find the area in the body for a z-score of 1.62.

We would first find the row for 1.60 then follow it across to the column labeled

- ($1.60 + 0.02 = 1.62$) and find 0.9474 (see Figure 4). Thus, the odds of randomly selecting someone with a z-score less than (to the left of) $z = 1.62$ is 94.74% because that is the proportion of the area taken up by values that satisfy our criteria.

Standard Normal Distribution Table



| Area in the Body to the Left of Z | | | | | | | | | | |
|-----------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
| 0.00 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.10 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.20 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.30 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.40 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.50 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.60 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.70 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.80 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.90 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.00 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.10 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.20 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.30 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.40 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.50 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.60 | 0.9452 | 0.9464 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.70 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.80 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.90 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |

Figure 4. Using the z-table to find the area in the body to the left of $z = 1.62$

The z-table only presents the area in the body for positive z-scores because the normal distribution is symmetrical. Thus, the area in the body of $z = 1.62$ is equal to the area in the body for $z = -1.62$, though now the body will be the shaded area to the right of z (because the body is always larger). When in doubt, drawing out your distribution and shading the area you

need to find will always help. The table also only presents the area in the body because the total area under the normal curve is always equal to 1.00, so if we need to find the area in the tail for $z = 1.62$, we simply find the area in the body and subtract it from 1.00 ($1.00 - 0.9474 = 0.0526$).

Let's look at another example. This time, let's find the area corresponding to z -scores more extreme than $z = -1.96$ and $z = 1.96$. That is, let's find the area in the tails of the distribution for values less than $z = -1.96$ (farther negative and therefore more extreme) and greater than $z = 1.96$ (farther positive and therefore more extreme). This region is illustrated in Figure 5.

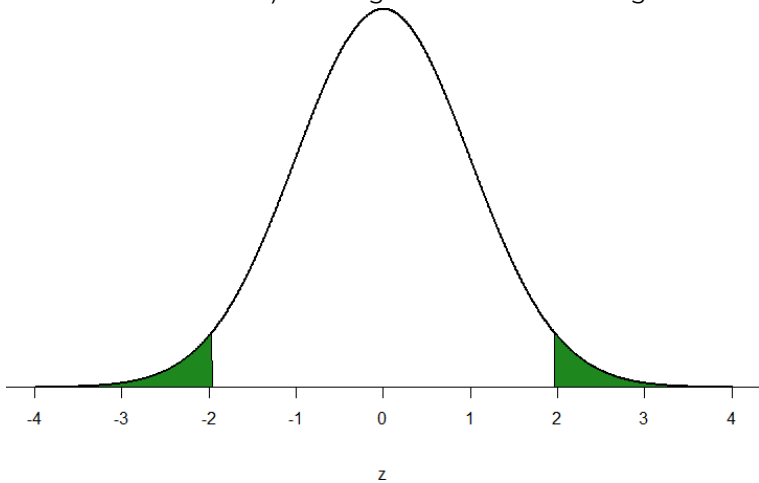


Figure 5. Area in the tails beyond $z = -1.96$ and $z = 1.96$

Let's start with the tail for $z = 1.96$. If we go to the z -table we will find that the body to the left of $z = 1.96$ is equal to 0.9750. To find the area in the tail, we subtract that from 1.00 to get 0.0250. Because the normal distribution is symmetrical, the area in the tail for $z = -1.96$ is the exact same value, 0.0250.

Finally, to get the total area in the shaded region, we simply add the areas together to get 0.0500. Thus, there is a 5% chance of randomly getting a value more extreme than $z = -1.96$ or $z = 1.96$ (this

particular value and region will become incredibly important in Unit 2).

Finally, we can find the area between two z-scores by shading and subtracting. Figure 6 shows the area between $z = 0.50$ and $z = 1.50$. Because this is a subsection of a body (rather than just a body or a tail), we must first find the larger of the two bodies, in this case the body for $z = 1.50$, and subtract the smaller of the two bodies, or the body for $z = 0.50$. Aligning the distributions vertically, as in Figure 6, makes this clearer. From the z-table, the area in the body for $z = 1.50$ is 0.9332 and the area in the body for $z = 0.50$ is 0.6915. Subtracting these gives us $0.9332 - 0.6915 = 0.2417$.

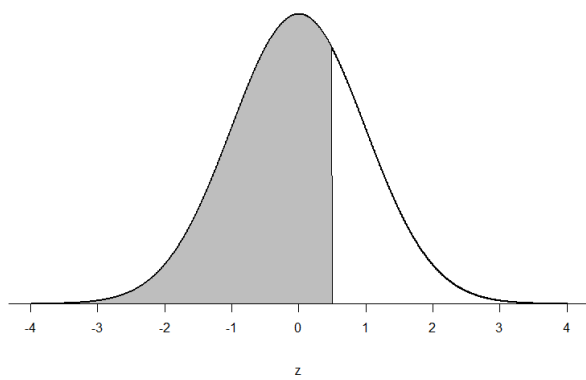
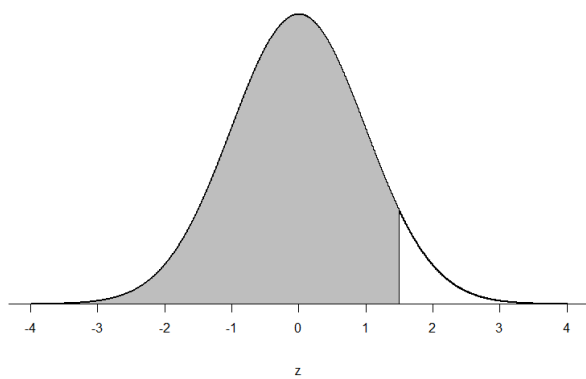
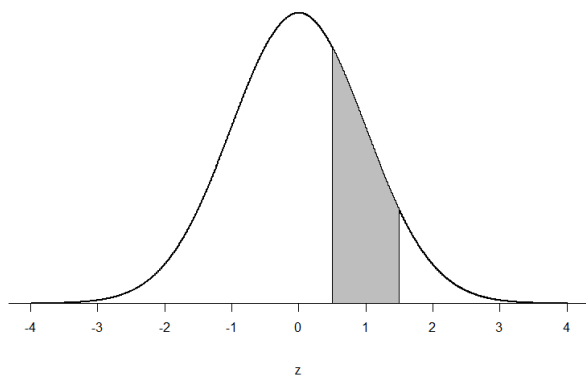


Figure 6. Area between $z = 0.50$ and 1.50 , along with the corresponding areas in the body

Considerations in understanding probability and inferential statistics: sampling

Recall that the goal of inferential statistics is to draw conclusions or make predictions about large populations by using data from smaller samples that represent that population. Probability is the underlying concept of inferential statistics and forms a direct link between samples and the population that they come from.

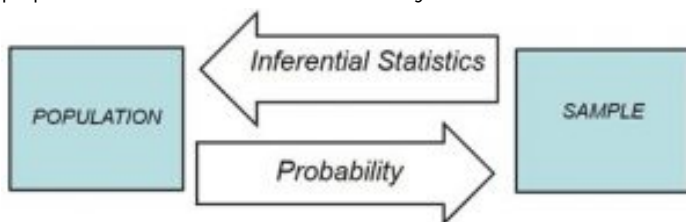


Figure 7. The relationship between inferential statistics and probability

As we learned earlier, gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. A sample should have the same characteristics as the population it is representing. **Random sampling** is one method that may ensure representativeness of a sample. For our definition of probability to be consistent and accurate we must ensure two elements:

1. Every person in the population has an equal chance of being selected
2. Sampling occurs with replacement

For **random sampling** a researcher starts with a **complete list of the population** (sometimes referred to as a sampling frame) and randomly selects some of them to an experiment. In this way every member of the population has an equal chance of being selected to participate. In order for the total number of outcomes to remain constant we need sampling with replacement – this means as one person is selected, another person must be added to keep the total number of possible outcomes the same. This is the full definition of random sampling.

For example, if the population is 25 people, the sample is ten, and you are sampling with replacement for any particular sample, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person). If you sample without replacement, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions to four decimal places:

- $9/25 = 0.3600$
- $9/24 = 0.3750$

It is clear that these numbers are not equivalent. Since we are using small samples as a stand-in for large populations in a research experiment, there will always be a certain level of uncertainty in our conclusions. How do we know that calculating statistical probability gives us the right number? The answer to this question comes from the **law of large numbers**, which shows that the empirical probability will approach the true probability as the sample size increases. We can see this by simulating a large number of coin flips, and looking at our estimate of the probability of heads after each flip.

The left panel of Figure [8](#) shows that as the number of samples

(i.e., coin flip trials) increases, the estimated probability of heads converges onto the true value of 0.5. It's unlikely that any of us has ever flipped a coin tens of thousands of times, but we are nonetheless willing to believe that the probability of flipping heads is 0.5. *However, note that the estimates can be very far off from the true value when the sample sizes are small.* A real-world example of this was seen in the 2017 special election for the US Senate in Alabama, which pitted the Republican Roy Moore against Democrat Doug Jones. The right panel of Figure 8 shows the relative amount of the vote reported for each of the candidates over the course of the evening, as an increasing number of ballots were counted. Early in the evening the vote counts were especially volatile, swinging from a large initial lead for Jones to a long period where Moore had the lead, until finally Jones took the lead to win the race.

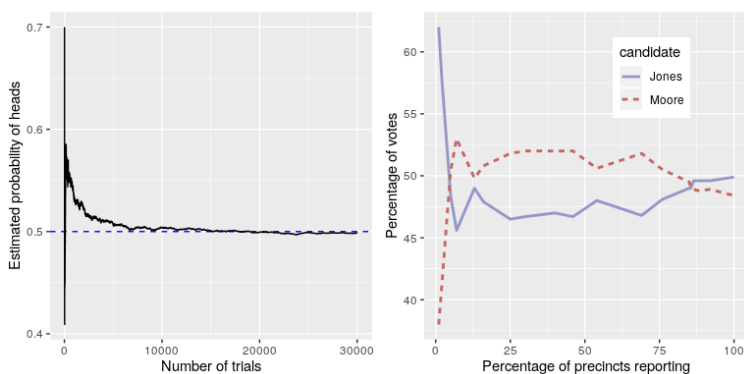


Figure 8: Left: A demonstration of the law of large numbers. A coin was flipped 30,000 times, and after each flip the probability of heads was computed based on the number of heads and tail collected up to that point. It takes about 15,000 flips for the probability to settle at the true probability of 0.5. Right: Relative proportion of the vote in the Dec 12, 2017 special election for the US Senate seat in Alabama, as a function of the percentage of precincts reporting. These data were transcribed from <https://www.ajc.com/news/national/alabama-senate->

[race-live-updates-roy-moore-doug-jones/
KPRfkdaewoiXICW3FHjXql/](https://www.kprfkdaewoiXICW3FHjXql/)

These two examples show that while large samples will ultimately converge on the true probability, the results with small samples can be far off. Unfortunately, many people forget this and overinterpret results from small samples. This was referred to as the *law of small numbers* by the psychologists Danny Kahneman and Amos Tversky, who showed that people (even trained researchers) often behave as if the law of large numbers applies even to small samples, giving too much credence to results based on small datasets. *We will see examples throughout the course of just how unstable statistical results can be when they are generated on the basis of small samples. Furthermore, a very important point: even if we are able to make an accurate prediction, we can never prove with 100% certainty that this prediction will hold true in all possible situations.* This is because samples never have exactly the same characteristics of the population from which they come from. Therefore, we will always have some level of uncertainty about whether the results found in our sample are also found in the population. The best we can do is infer what is most likely to be found. Our level of confidence in an inferential statistic or the outcome of an experiment is represented through *probability theory*.

More Probability Terms

A **probability distribution** describes the probability of all of the possible outcomes in an activity. For example, on Jan 20 2018, the basketball player Steph Curry hit only 2 out of 4 free throws in a game against the Houston Rockets. We know that Curry's overall probability of hitting free throws across the entire season was 0.91, so it seems pretty unlikely that he would hit

only 50% of his free throws in a game, but exactly how unlikely is it? We can determine this using a theoretical probability distribution; throughout this book we will encounter a number of these probability distributions, each of which is appropriate to describe different types of data. In this case, we use the *binomial* distribution, which provides a way to compute the probability of some number of successes out of a number of trials on which there is either success or failure and nothing in between (only 2 outcomes), given some known probability of success on each trial. In this example, if we calculated the probability of Curry making 2 of 4 free throws when his season average was .91, it would come out to a 4% chance¹. This shows that given Curry's overall free throw percentage, it is very unlikely that he would hit only 2 out of 4 free throws. *This just goes to show that unlikely things do actually happen in the real world.*

Often though, we might want to know how likely it is to find a value that is as extreme or more than a particular value versus a specific value; this will become very important when we discuss hypothesis testing in Chapter 8. To answer this question, we can use a **cumulative probability distribution**; whereas a standard probability distribution tells us the probability of some specific value, the cumulative distribution tells us the probability of a value as large or larger (or as small or smaller) than some specific value. For the Curry free-throw example, this would be .043². In many cases the number of possible outcomes would be too large for us to compute the

1. The fine print calculating Steph Curry's free throws probability:
 $P(2;4,0.91) = \binom{4}{2} 0.91^2 (1-0.91)^{4-2} = 0.040$
2. Fine print: To determine this, we could use the binomial probability equation and plug in all of the possible values of outcomes and add them together: $P(k \leq 2) = P(k=2) + P(k=1) + P(k=0) = 6e^{-5} + .002 + .040 = .043$
 $P(k \leq 2) = P(k=2) + P(k=1) + P(k=0) = 6e^{-5} + .002 + .040 = .043$

cumulative probability by enumerating all possible values; fortunately, it can be computed directly for any theoretical probability distribution.

Although we have information, sometimes we can be misled. Let's use the example of a coin toss. Each toss is independent of one another with only 2 outcomes {heads, tails}. What if I want to calculate the probability of getting heads for a coin toss? I might base that on my previous 12 coin flips. A coin might come up heads 8 times out of 12 flips, for a *relative frequency* of $8/12$ or $2/3$. Again, the probability is usually calculated as the proportion of successful outcomes divided by the number of all possible outcomes. *What happens if I toss the coin again?* What is the probability that it will be heads again? Do I use information from my previous 12 flips? Well, the answer is $\frac{1}{2}$ or 50%. That is because in our scenario, each toss is independent and each outcome is independent. This means that the outcome of the ninth toss is not related to the previous toss. To assume otherwise is committing what we call **gambler's fallacy**.

The term "independent" has a very specific meaning in statistics, which is somewhat different from the common usage of the term. **Statistical independence** between two variables means that knowing the value of one variable doesn't tell us anything about the value of the other. This can be expressed as: $P(A|B)=P(A)$. *That is, the probability of A given some value of B is just the same as the overall probability of A (because they are independent).* Again, while independence in common language often refers to sets that are exclusive, statistical independence refers to the case where one cannot predict anything about one variable from the value of another variable. For example, knowing a person's hair color is unlikely to tell you whether they prefer chocolate or strawberry ice cream. *Later in the book we will discuss statistical tools that will let us directly test whether two variables are independent.*

Sometimes we want to quantify the relation between probabilities more directly, which we can do by converting them into *odds* which express the relative likelihood of something happening or not. We can use odds to compare different probabilities, by computing what is called an **odds ratio** – which is exactly what it sounds like. For example, let's say that we want to know how much the positive test increases the individual's odds of having cancer. We can first compute the *prior odds* – that is, the odds before we knew that the person had tested positively. *An odds ratio is an example of what we will later call an effect size, which is a way of quantifying how relatively large any particular statistical effect is.*

A final point relates to how probabilities have been interpreted. Historically, there have been two different ways that probabilities have been interpreted. The first (known as the *frequentist* interpretation) interprets probabilities in terms of long-run frequencies. For example, in the case of a coin flip, it would reflect the relative frequencies of heads in the long run after a large number of flips. While this interpretation might make sense for events that can be repeated many times like a coin flip, it makes less sense for events that will only happen once, like an individual person's life or a particular presidential election; and as the economist John Maynard Keynes famously said, "In the long run, we are all dead." The other interpretation of probabilities (known as the *Bayesian* interpretation) is as a degree of belief in a particular proposition. If I were to ask you "How likely is it that the US will return to the moon by 2040", you can provide an answer to this question based on your knowledge and beliefs, even though there are no relevant frequencies to compute a frequentist probability. One way that we often frame subjective probabilities is in terms of one's willingness to accept a particular gamble. For example, if you think that the probability of the US landing on the moon by 2040 is 0.1 (i.e. odds of 9 to 1), then that means that you should

be willing to accept a gamble that would pay off with anything more than 9 to 1 odds if the event occurs. As we will see, these two different definitions of probability are very relevant to the two different ways that statisticians think about testing statistical hypotheses, which we will encounter in later chapters.

Recap

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Learning objectives

Having read this chapter, you should be able to:

- Describe the sample space for a selected random experiment.
- Describe the law of large numbers.
- Describe the difference between a probability and a conditional probability

- Describe the relationship between z-scores and the standard unit normal table (z-table)
- Probability is a tough topic for everyone, but the tools it gives us are incredibly powerful and enable us to do amazing things with data analysis. They are the heart of how inferential statistics work.

Exercises – Ch. 7

1. In your own words, what is probability?
2. There is a bag with 5 red blocks, 2 yellow blocks, and 4 blue blocks. If you reach in and grab one block without looking, what is the probability it is red?
3. Under a normal distribution, which of the following is more likely? (Note: this question can be answered without any calculations if you draw out the distributions and shade properly)
 1. Getting a z-score greater than $z = 2.75$
 2. Getting a z-score less than $z = -1.50$
4. The heights of women in the United States are normally distributed with a mean of 63.7 inches and a standard deviation of 2.7 inches. If you randomly select a woman in the United States, what is the probability that she will be between 65 and 67 inches tall?
5. The heights of men in the United States are normally distributed with a mean of 69.1 inches and a standard deviation of 2.9 inches. What proportion of men are taller than 6 feet (72 inches)?
6. You know you need to score at least 82 points on the final exam to pass your class. After the final, you find out that the average score on the exam was 78 with a standard deviation of 7. How likely is it that you pass the class?
7. What proportion of the area under the normal curve is

greater than $z = 1.65$?

8. Find the z -score that bounds 25% of the lower tail of the distribution.
9. Find the z -score that bounds the top 9% of the distribution.
10. In a distribution with a mean of 70 and standard deviation of 12, what proportion of scores are lower than 55?

Answers to Odd-Numbered Exercises – Ch. 7

1. Your answer should include information about an event happening under certain conditions given certain criteria. You could also discuss the relation between probability and the area under the curve or the proportion of the area in a chart.

3. Getting a z -score less than $z = -1.50$ is more likely. $z = 2.75$ is farther out into the right tail than $z = -1.50$ is into the left tail, therefore there are fewer more extreme scores beyond 2.75 than -1.50, regardless of the direction

5. 15.87% or 0.1587

7. 4.95% or 0.0495

9. $z = 1.34$ (the top 9% means 9% of the area is in the upper tail and 91% is in the body to the left; finding the value in the normal table closest to .9100 is .9099, which corresponds to $z = 1.34$)

8. Chapter 8: Sampling Distributions

People, Samples, and Populations

Most of what we have dealt with so far has concerned individual scores grouped into samples, with those samples being drawn from and, hopefully, representative of a population. We saw how we can understand the location of individual scores within a sample's distribution via z-scores, and how we can extend that to understand how likely it is to observe scores higher or lower than an individual score via probability.

Inherent in this work is the notion that an individual score will differ from the mean, which we quantify as a z-score. All of the individual scores will differ from the mean in different amounts and different directions, which is natural and expected. We quantify these differences as variance and standard deviation.

Measures of spread and the idea of variability in observations is a key principle in inferential statistics. We know that any observation, whether it is a single score, a set of scores, or a particular descriptive statistic will differ from the center of whatever distribution it belongs in.

This is equally true of things outside of statistics and format data collection and analysis. Some days you hear your alarm and wake up easily, other days you need to hit snooze a few [dozen] times. Some days traffic is light, other days it is very heavy. Some classes you are able to focus, pay attention, and take good notes, but other days you find yourself zoning out the entire time. Each individual observation is an insight but is

not, by itself, the entire story, and it takes an extreme deviation from what we expect for us to think that something strange is going on. Being a little sleepy is normal, but being completely unable to get out of bed might indicate that we are sick. Light traffic is a good thing, but almost no cars on the road might make us think we forgot it is Saturday. Zoning out occasionally is fine, but if we cannot focus at all, we might be in a stats class rather than a fun one.

All of these principles carry forward from scores within samples to samples within populations. Just like an individual score will differ from its mean, an individual sample mean will differ from the true population mean. We encountered this principle in earlier chapters: sampling error. As mentioned way back in chapter 1, sampling error is an incredibly important principle. We know ahead of time that if we collect data and compute a sample, the observed value of that sample will be at least slightly off from what we expect it to be based on our supposed population mean; this is natural and expected. However, if our sample mean is extremely different from what we expect based on the population mean, there may be something going on.

Sampling

One of the foundational ideas in statistics is that we can make inferences about an entire population based on a relatively small sample of individuals from that population. In this chapter, we will introduce the concept of statistical sampling and discuss why it works.

Anyone living in the United States will be familiar with the concept of sampling from the political polls that have become a central part of our electoral process. In some cases, these

polls can be incredibly accurate at predicting the outcomes of elections. The best known example comes from the 2008 and 2012 US Presidential elections when the pollster Nate Silver correctly predicted electoral outcomes for 49/50 states in 2008 and for all 50 states in 2012. Silver did this by combining data from 21 different polls, which vary in the degree to which they tend to lean towards either the Republican or Democratic side. Each of these polls included data from about 1000 likely voters – meaning that Silver was able to almost perfectly predict the pattern of votes of more than 125 million voters using data from only about 21,000 people, along with other knowledge (such as how those states have voted in the past).

How do we sample?

Our goal in sampling is to determine the value of a statistic for an entire population of interest, using just a small subset of the population. We do this primarily to save time and effort – why go to the trouble of measuring every individual in the population when just a small sample is sufficient to accurately estimate the statistic of interest?

In the election example, the population is all registered voters in the region being polled, and the sample is the set of 1000 individuals selected by the polling organization. The way in which we select the sample is critical to ensuring that the sample is *representative* of the entire population, which is a main goal of statistical sampling. It's easy to imagine a non-representative sample; if a pollster only called individuals whose names they had received from the local Democratic party, then it would be unlikely that the results of the poll would be representative of the population as a whole. In general, we would define a representative poll as being one in which every member of the population has an equal chance of being selected. When this fails, then we have to worry about

whether the statistic that we compute on the sample is *biased* – that is, whether its value is systematically different from the population value (which we refer to as a *parameter*). Keep in mind that we generally don't know this population parameter, because if we did then we wouldn't need to sample! But we will use examples where we have access to the entire population, in order to explain some of the key ideas.

It's important to also distinguish between two different ways of sampling: with replacement versus without replacement. In sampling *with replacement*, after a member of the population has been sampled, they are put back into the pool so that they can potentially be sampled again. In *sampling without replacement*, once a member has been sampled they are not eligible to be sampled again. It's most common to use sampling without replacement.

Sampling error

Regardless of how representative our sample is, it's likely that the statistic that we compute from the sample is going to differ at least slightly from the population parameter. We refer to this as *sampling error*. If we take multiple samples, the value of our statistical estimate will also vary from sample to sample; we refer to this distribution of our statistic across samples as the *sampling distribution*.

Sampling error is directly related to the quality of our measurement of the population. Clearly, we want the estimates obtained from our sample to be as close as possible to the true value of the population parameter. However, even if our statistic is unbiased (that is, we expect it to have the same value as the population parameter), the value for any particular estimate will differ from the population value, and those differences will be greater when the sampling error is greater.

Thus, reducing sampling error is an important step towards better measurement.

Until now we used z-scores and probability we were only looking for the probability of finding one score ($n = 1$) but most research involves looking at larger samples. Using samples allows us to make generalizations to the larger population but there are some limitations. We know that the samples will look different even when they come from the same population and the difference between the sample and population is known as **sampling error**.

Suppose you randomly sampled 10 people from the population of women in Houston, Texas, between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

It is possible to get thousands of samples from one population. These samples will each look different but the sample means, when placed in a frequency distribution form a simple, predictable pattern. The pattern makes it possible to predict sample characteristics with some degree of accuracy. These predictions are based on the **distribution of sample means**, which is a collection of *all possible random samples* of a particular size that can be obtained from a population.

The concept of a sampling distribution is perhaps the most basic concept in inferential statistics but it is also a difficult concept because a sampling distribution is a *theoretical distribution* rather than an empirical distribution. The distribution is based on sample statistics (sample means) not on individual scores. The distribution of sample means is

formed by statistics obtained by selecting all possible samples of a specific size from a population.

This can feel very abstract and confusing so let's use an example to illustrate what we mean. Let's look at a rather unique population of scores. This population is very small and consists of only four scores: one 2, one 4, one 6, and one 8. Next, we are going to take a bunch of samples from this population. Each of our samples will consist of two scores. That is, the sample size is 2 ($n = 2$). Because this population is so small we can take every sample possible from the population. Below is a table showing all 16 possible samples of $n = 2$.

| <u>Sample #</u> | <u>First Score</u> | <u>Second Score</u> | <u>Sample Mean (M)</u> |
|------------------------|---------------------------|----------------------------|-------------------------------|
| 1 | 2 | 2 | 2 |
| 2 | 2 | 4 | 3 |
| 3 | 2 | 6 | 4 |
| 4 | 2 | 8 | 5 |
| 5 | 4 | 2 | 3 |
| 6 | 4 | 4 | 4 |
| 7 | 4 | 6 | 5 |
| 8 | 4 | 8 | 6 |
| 9 | 6 | 2 | 4 |
| 10 | 6 | 4 | 5 |
| 11 | 6 | 6 | 6 |
| 12 | 6 | 8 | 7 |
| 13 | 8 | 2 | 5 |
| 14 | 8 | 4 | 6 |
| 15 | 8 | 6 | 7 |
| 16 | 8 | 8 | 8 |

Table 1. Possible sampling with $n=2$ with 4 possible scores
The far right column shows the mean of each sample. These

16 sample means have been used to create their own frequency distribution. Therefore, we can call this frequency distribution a **distribution of sample means** (see Figure 1).

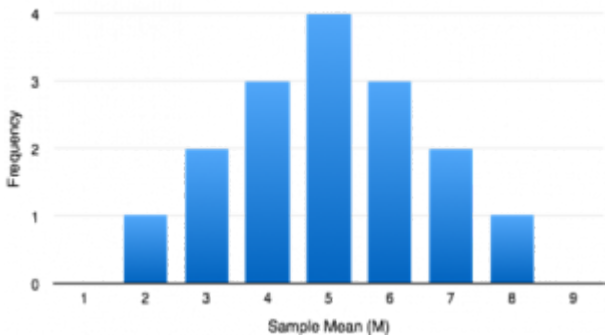


Figure 1. Distribution of sample means for $n=2$ from Table 1.

In our example, a population was specified ($N = 4$) and the sampling distribution was determined. In practice, the process actually moves the other way: you collect sample data and from these data you estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful. For example, knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. This information is directly available from a sampling distribution.

We will use the NHANES dataset (National Health and Nutrition Examination Study mentioned in chapter 1) as another example; we are going to assume that the NHANES dataset is the entire population of interest, and then we will draw random samples from this population. We will have more to say in the next chapter about exactly how the generation of “random” samples works in a computer.

In this example, we know the adult population mean (168.35) and standard deviation (10.16) for height because we are assuming that the NHANES dataset *is* the population. Table

[2](#) shows the statistics computed from a few samples of 50 individuals from the NHANES population.

Table 2: Example means and standard deviations for several samples of Height variable from NARPS

| <u>sampleMean</u> | <u>sampleSD</u> |
|-------------------|-----------------|
| 167 | 9.1 |
| 171 | 8.3 |
| 170 | 10.6 |
| 166 | 9.5 |
| 168 | 9.5 |

The sample mean and standard deviation are similar but not exactly equal to the population values. Now let’s take a large number of samples of 50 individuals, compute the mean for each sample, and look at the resulting sampling distribution of means. We have to decide how many samples to take in order to do a good job of estimating the sampling distribution – in this case, we will take 5000 samples so that we are very confident in the answer. Note that simulations like this one can sometimes take a few minutes to run, and might make your computer huff and puff. The histogram in [Figure 2](#) shows that the means estimated for each of the samples of 50 individuals vary somewhat, but that overall they are centered around the population mean. The average of the 5000 sample means (168.3463) is very close to the true population mean (168.3497).

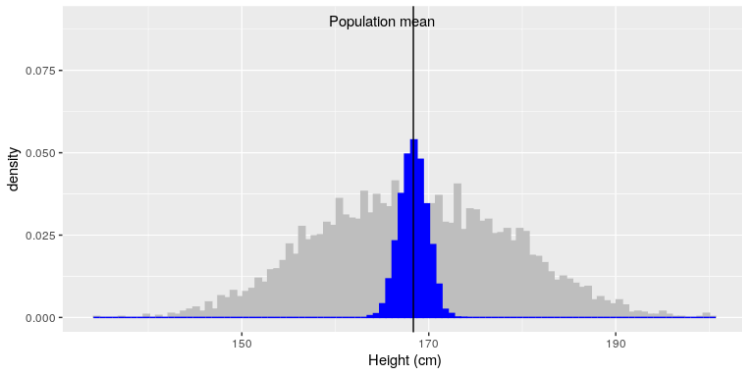


Figure 2: The blue histogram shows the sampling distribution of the mean over 5000 random samples from the NHANES dataset. The histogram for the full dataset is shown in gray for reference.

The Sampling Distribution of Sample Means

To see how we use sampling error, we will learn about a new, theoretical distribution known as the sampling distribution. In the same way that we can gather a lot of individual scores and put them together to form a distribution with a center and spread, if we were to take many samples, all of the same size, and calculate the mean of each of those, we could put those means together to form a distribution. This new distribution is, intuitively, known as the distribution of sample means. It is one example of what we call a sampling distribution, we can be formed from a set of any statistic, such as a mean, a test statistic, or a correlation coefficient (more on the latter two in Units 2 and 3). For our purposes, understanding the distribution of sample means will be enough to see how all other sampling distributions work to enable and inform our inferential analyses, so these two terms will be used

interchangeably from here on out. Let's take a deeper look at some of its characteristics.

The sampling distribution of sample means can be described by its shape, center, and spread, just like any of the other distributions we have worked with. The shape of our sampling distribution is normal: a bell-shaped curve with a single peak and two tails extending symmetrically in either direction, just like what we saw in previous chapters.

Guidelines of a Distribution of Sample Means (DSM)

In order for a distribution of sampling means (DSM) to be accurate we must draw every possible sample out of the population and plot its mean. In the first example, with our population of just four scores, it wasn't very difficult to do this. Most populations are much larger than our simple example. In real life, to make a complete, accurate DSM we should have collected over 12 trillion samples in order to plot every sample mean possible. This is not efficient and fortunately we do not have to do it because we can estimate parameters.

If we know the mean (μ) and standard deviation (σ) of the distribution of individual scores, then we can estimate the mean and standard deviation of the distribution of sample means without having to make thousands, millions and trillions of calculations. We can follow three simple guidelines:

1. Shape

The shape of the distribution of sample means (DSM) will be normal if either one of the following two conditions are met:

- Population from which the sample is selected is normal or

- The size of the sample is relatively large ($30 >$)

We will talk more about this later, but note that with samples larger than 30, the sampling distribution of the mean is normal *even if the data within each sample are not normally distributed*. This is an important concept because if we want to apply the proportions and probabilities of a normal distribution then the shape of the distribution of sample means must approximate the shape. Also, not all parameters of a population follow a normal distribution and not all research questions are interested in these parameters – but with a large enough sample, even skewed population parameters can approximate the normal distribution.

2. Mean

The mean of the distribution of sampling means is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ , then the mean of the sampling distribution of the mean is also μ . The symbol μ_M is used to refer to the mean of the sampling distribution of the mean.

The formula for the mean of the sampling distribution of the mean can be written as: $\mu_M = \mu$

This can also be written as $\mu_{\bar{x}}$ to denote it as the mean of the sample means. As you can see the mean for the distribution of the sample means is exactly the same as the population mean. We would expect these values to be the same. The center of the sampling distribution of sample means – which is, itself, the mean or average of the means – is the true population mean, μ .

3. Standard Deviation

The most common measure of how much sample means differ

from each other is the standard deviation of the distribution of sampling means. This standard deviation is called the **standard error of the mean** and it measures the expected difference between the sample mean (M) and the population mean (μ). Standard error is a valuable tool because it is a measure of how closely the sample matches the population. This is important because our overall goal in inferential statistics is to use samples to talk about the larger population. We estimate the standard error by taking the standard deviation of the original sample of population and divide it by the square root of the sample size.

Standard Error of the Mean (noted as $\sigma_{\bar{x}}$ or σ_M).

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

Notice that the sample size is in the standard error equation. As stated above, the sampling distribution refers to samples of a specific size. That is, all sample means must be calculated from samples of the same size n , such as $n = 10$, $n = 30$, or $n = 100$. This sample size refers to how many people or observations are in each individual sample, *not* how many samples are used to form the sampling distribution. This is because the sampling distribution is a theoretical distribution, not one we will ever actually calculate or observe. Note that we have to be careful about computing standard error using the estimated standard deviation if our sample is small (less than about 30).

The formula for the standard error of the mean implies that the quality of our measurement involves two quantities: the population variability, and the size of our sample. Because the

sample size is the denominator in the formula for standard error a larger sample size will yield a smaller standard error when holding the population variability constant. We have no control over the population variability, but we *do* have control over the sample size. Thus, if we wish to improve our sample statistics (by reducing their sampling variability) then we should use larger samples. The expected difference between sample means and population mean is closely related to the elements formula – sample size and standard deviation.

- A large sample size will result in small error because we expect that a large sample to be more representative of the population.
- Small standard deviation will result in small error because the variability of scores within the population are clustered more closely around the mean.

The formula also tells us something very fundamental about statistical sampling – namely, that the utility of larger samples diminishes with the square root of the sample size. In chapter 11, we will discuss statistical power, which is intimately tied to this idea.

Figure 3 displays three guidelines for the distribution of sample means in graphical form.

Sampling Distribution of the Sample Mean

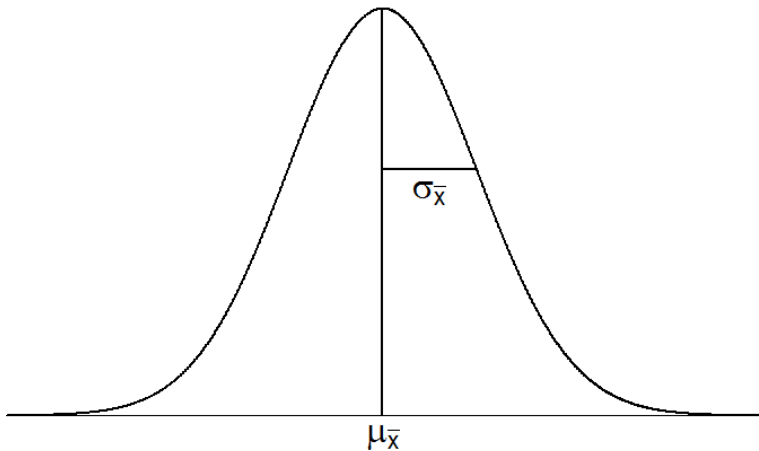


Figure 3. The sampling distribution of sample means

We can also compare this sampling distribution of the means to that of a population. Because the standard error takes the standard deviation and divides it by a number greater than 1, the value of the standard error will always be smaller than the value of the standard deviation. Also, by taking the mean of all sample means, we are automatically reducing the variability between scores. Each sample mean is a summary statistic that represents the center of that sample. A sample mean “washes out” individual high and low scores in that sample by creating the average. Further, because each sample mean (M) is an approximation of the population mean (μ), the sample means cluster more tightly around μ than individual scores. The amount of variability among sample means (σ_M) depends on the amount of variability among the individual scores in the population (σ) and on the size of samples (n) used to create the DSM. This is shown in Figure 4 and connects back to our example in Figure 2. In other words, variability among sample means in a distribution of sample means will be reduced as sample size is increased, and/or as population variability is

reduced. One way to remember this is to think about the formula we use to calculate the standard error. We calculate the standard error (σ_M) by dividing the standard deviation (σ) by the square root of the sample size (n).

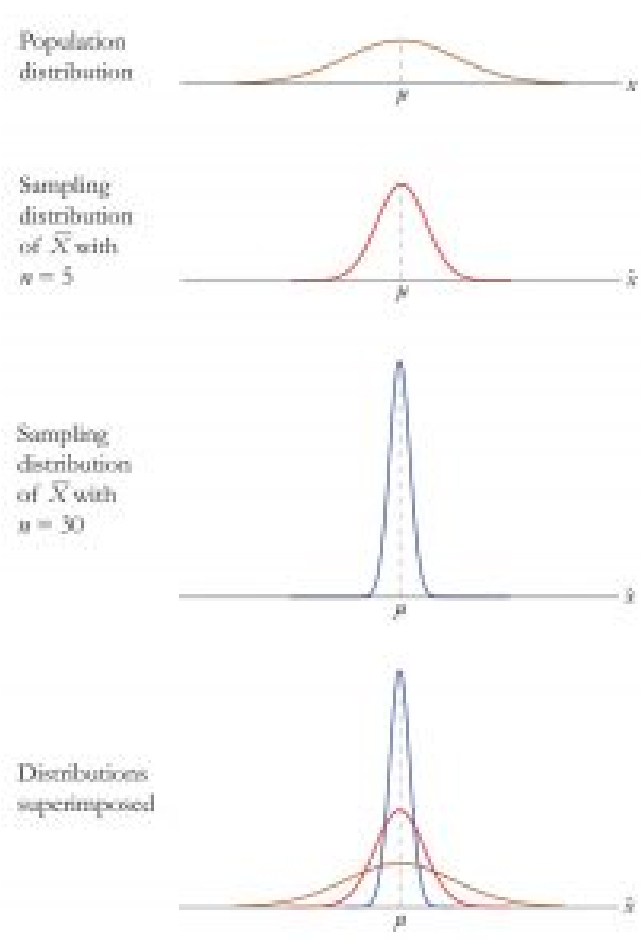


Figure 4. Distributions based on population, sample size of 5 and sample size of 30. [Image credit](#)

Two Important Axioms

We just learned that the sampling distribution that are connected to two very important mathematical facts: the central limit theorem and the law of large numbers. We will not go into the math behind how these statements were derived, but knowing what they are and what they mean is important to understanding why inferential statistics work and how we can draw conclusions about a population based on information gained from a single sample.

Central Limit Theorem

The **Central Limit Theorem**:

For samples of a single size n , drawn from a population with a given mean μ and variance σ^2 , the sampling distribution of sample means will have a mean $\mu_{\bar{x}} = \mu$ and variance $\sigma^2 = \sigma^2/n$. This distribution will approach normality as n increases tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed. The last sentence of the central limit theorem states that the sampling distribution will be normal as the sample size of the samples used to create it increases. What this means is that bigger samples will create a more normal distribution, so we are better able to use the techniques we developed for normal distributions and probabilities. So how large is large enough? In general, a sampling distribution will be normal if either of two characteristics is true: 1) the population from which the samples are drawn is normally distributed or 2) the sample size is equal to or greater than 30. This second criteria is very important because it enables us to use methods developed for normal distributions *even if the true population distribution is skewed*.

To see the central limit theorem in action, let's work with the

variable AlcoholYear from the NHANES dataset, which is highly skewed, as shown in the left panel of Figure 5. This distribution is, for lack of a better word, funky – and definitely not normally distributed. Now let’s look at the sampling distribution of the mean for this variable. Figure 5 shows the sampling distribution for this variable, which is obtained by repeatedly drawing samples of size 50 from the NHANES dataset and taking the mean. Despite the clear non-normality of the original data, the sampling distribution is remarkably close to the normal.

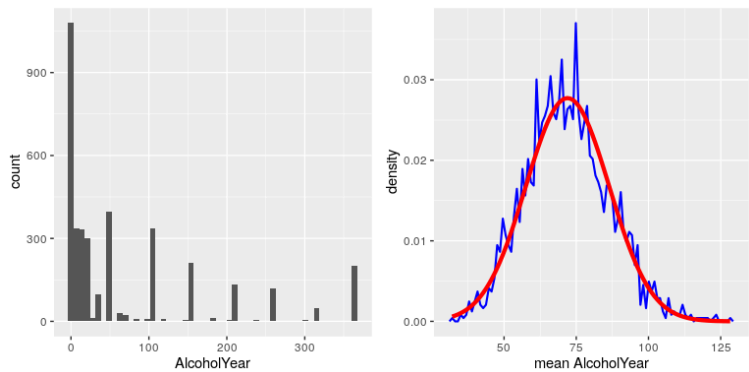


Figure 5: Left: Distribution of the variable AlcoholYear in the NHANES dataset, which reflects the number of days that the individual drank in a year. Right: The sampling distribution of the mean for AlcoholYear in the NHANES dataset, obtained by drawing repeated samples of size 50, in blue. The normal distribution with the same mean and standard deviation is shown in red.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section. It’s also important because it tells us why normal distributions are

so common in the real world; any time we combine many different factors into a single number, the result is likely to be a normal distribution. For example, the height of any adult depends on a complex mixture of their genetics and experience; even if those individual contributions may not be normally distributed, when we combine them the result is a normal distribution.

Law of Large Numbers

The **law of large numbers** simply states that as our sample size increases, the probability that our sample mean is an accurate representation of the true population mean also increases. It is the formal mathematical way to state that larger samples are more accurate. The law of large numbers is related to the central limit theorem, specifically the formulas for variance and standard error. Notice that the sample size appears in the denominators of those formulas. A larger denominator in any fraction means that the overall value of the fraction gets smaller (i.e. $1/2 = 0.50$, $1/3 = 0.33$, $1/4 = 0.25$, and so on). Thus, larger sample sizes will create smaller standard errors. We already know that standard error is the spread of the sampling distribution and that a smaller spread creates a narrower distribution. Therefore, larger sample sizes create narrower sampling distributions, which increases the probability that a sample mean will be close to the center and decreases the probability that it will be in the tails. This is illustrated in Figures 6 and 7.

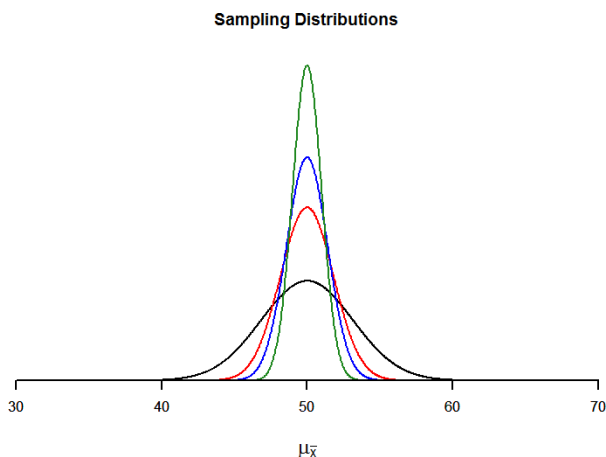


Figure 6. Sampling distributions from the same population with $\mu = 50$ and $\sigma = 10$ but different sample sizes ($n = 10$, $n = 30$, $n = 50$, $n = 100$)

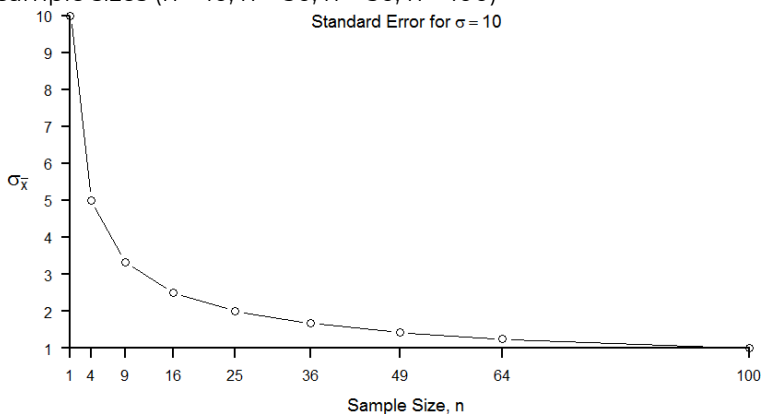


Figure 7. Relation between sample size and standard error for a constant $\sigma = 10$

Using Standard Error for Probability

We saw in chapter 7 that we can use z-scores to split up a normal distribution and calculate the proportion of the area under the curve in one of the new regions, giving us the probability of randomly selecting a z-score in that range. We

can follow the exact sample process for sample means, converting them into z-scores and calculating probabilities. The only difference is that instead of dividing a raw score by the standard deviation, we divide the sample mean by the standard error.

$$Z = \frac{M - \mu}{\sigma_M}$$

Remember that

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

Let's say we are drawing samples from a population with a mean of 50 and standard deviation of 10 (the same values used in Figure 6). What is the probability that we get a random sample of size 10 with a mean greater than or equal to 55? That is, for $n = 10$, what is the probability that $\bar{X} \geq 55$? First, we need to convert this sample mean score into a z-score:

$$Z = (55 - 50) / 10 / \sqrt{10} = 5 / 3.16 = 1.58$$

Now we need to shade the area under the normal curve corresponding to scores greater than $z = 1.58$ as in Figure 8:

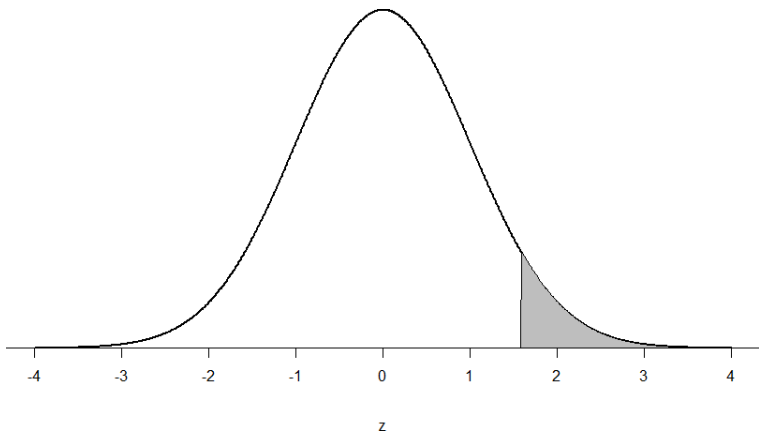


Figure 8: Area under the curve greater than $z = 1.58$

Now we go to our z-table and find that the area to the left of $z = 1.58$ is 0.9429. Finally, because we need the area to the right (per our shaded diagram), we simply subtract this from 1 to get $1.00 - 0.9429 = 0.0571$. So, the probability of randomly drawing a sample of 10 people from a population with a mean of 50 and standard deviation of 10 whose sample mean is 55 or more is $p = .0571$, or 5.71%. Notice that we are talking about means that are 55 *or more*. That is because, strictly speaking, it's impossible to calculate the probability of a score taking on exactly 1 value since the “shaded region” would just be a line with no area to calculate.

Now let's do the same thing, but assume that instead of only having a sample of 10 people we took a sample of 50 people. First, we find z :

$$Z = (55-50)/10\sqrt{50} = 5/1.41 = 3.55$$

Then we shade the appropriate region of the normal distribution:

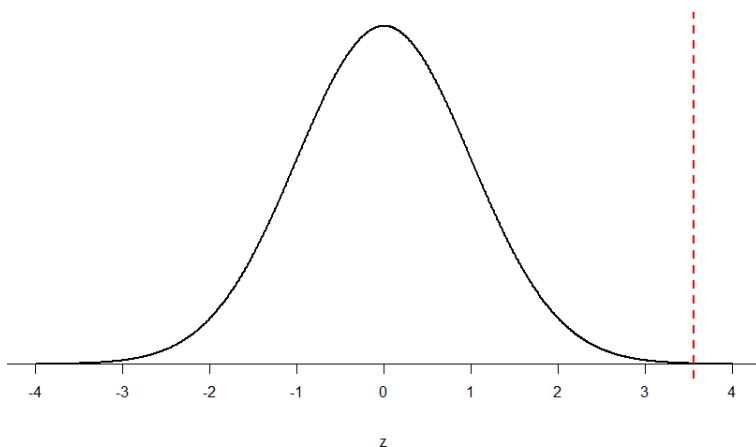


Figure 9: Area under the curve greater than $z = 3.55$

Notice that no region of Figure 5 appears to be shaded. That is because the area under the curve that far out into the tail is so small that it can't even be seen (the red line has been added to show exactly where the region starts). Thus, we already know that the probability must be smaller for $N = 50$ than $N = 10$ because the size of the area (the proportion) is much smaller.

We run into a similar issue when we try to find $z = 3.55$ on our Standard Normal Distribution Table. The table only goes up to 3.09 because everything beyond that is almost 0 and changes so little that it's not worth printing values. The closest we can get is subtracting the largest value, 0.9990, from 1 to get 0.001. We know that, technically, the actual probability is smaller than this (since 3.55 is farther into the tail than 3.09), so we say that the probability is $p < 0.001$, or less than 0.1%.

This example shows what an impact sample size can have. From the same population, looking for exactly the same thing, changing only the sample size took us from roughly a 5% chance (or about 1/20 odds) to a less than 0.1% chance (or less than 1 in 1000). As the sample size n increased, the standard error decreased, which in turn caused the value of z to increase, which finally caused the p-value (a term for

probability we will use a lot in Unit 2) to decrease.

You can think of this relation like gears: turning the first gear (sample size) clockwise causes the next gear (standard error) to turn counterclockwise, which causes the third gear (z) to turn clockwise, which finally causes the last gear (probability) to turn counterclockwise.



[Photo credit](#)

All of these pieces fit together, and the relations will always be the same: $n \uparrow \sigma_M \downarrow z \uparrow p \downarrow$

Let's look at this one more way. For the same population of sample size 50 and standard deviation 10, what proportion of sample means fall between 47 and 53 if they are of sample size 10 and sample size 50?

We'll start again with $n = 10$. Converting 47 and 53 into z-scores,

we get $z = -0.95$ and $z = 0.95$, respectively. From our z-table, we find that the proportion between these two scores is 0.6578 (the process here is left off for the student to practice converting X to z and z to proportions). So, 65.78% of sample means of sample size 10 will fall between 47 and 53. For $n = 50$, our z-scores for 47 and 53 are ± 2.13 , which gives us a proportion of the area as 0.9668, almost 97%! Shaded regions for each of these sampling distributions is displayed in Figure 9. The sampling distributions are shown on the original scale, rather than as z-scores, so you can see the effect of the shading and how much of the body falls into the range, which is marked off with dotted line.

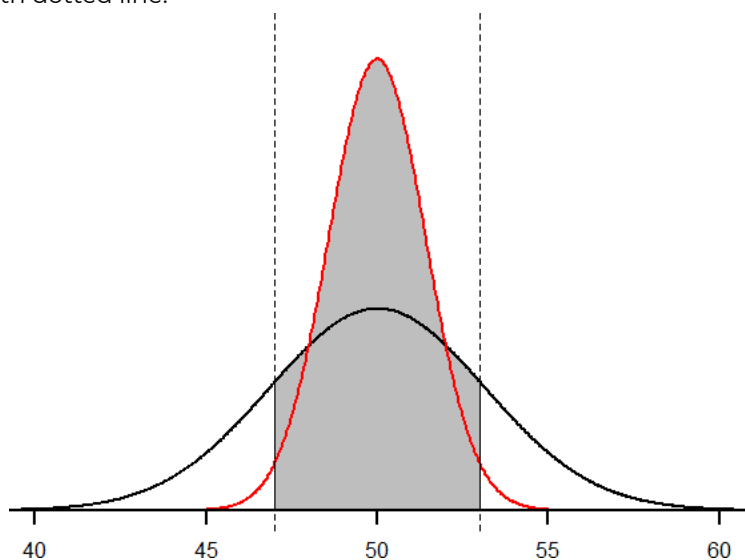


Figure 9. Areas between 47 and 53 for sampling distributions of $n = 10$ and $n = 50$

Sampling Distribution, Probability and Inference

We've seen how we can use the standard error to determine

probability based on our normal curve. We can think of the standard error as how much we would naturally expect our statistic – be it a mean or some other statistic – to vary. In our formula for z based on a sample mean, the numerator ($M - \mu$) is what we call an observed effect. That is, it is what we observe in our sample mean versus what we expected based on the population from which that sample mean was calculated.

Because the sample mean will naturally move around due to sampling error, our observed effect will also change naturally. In the context of our formula for z , then, our standard error is how much we would naturally expect the observed effect to change. Changing by a little is completely normal, but changing by a lot might indicate something is going on. This is the basis of inferential statistics and the logic behind hypothesis testing, the subject of Unit 2.

Recap

Earlier we learned that probability forms the direct link between samples and the population that they come from. This link serves as the foundation for inferential statistics.

As we learned earlier, the concept of a sampling distribution is perhaps the most basic concept in inferential statistics but it is also one of the most challenging because we have to accept hypothetical concepts and theories about how samples and normal distributions work. A few key things to remember are that the sample distribution of the means is based on sample statistics (sample means) not on individual scores. Second, the distribution of sample means is formed by statistics obtained by selecting all possible samples of a specific size from a population. The standard error of the mean is used to determine how close a sample mean is to a population mean.

When we are describing the parameters of a DSM, there are a few differences and a few similarities compared to the other distributions we have already learned about.

| | measure of central tendency | measure of variability |
|--|-----------------------------|------------------------|
| distribution of individual scores (sample) | M | S |
| distribution of individual scores (population) | μ | σ |
| distribution of sample means | M | σ_M |

Learning Objectives

Having read this chapter, you should be able to:

- Distinguish between a population and a sample, and between population parameters and sample statistics
- Describe the concepts of sampling error and sampling distribution
- Compute the z-score for distribution of sample means
- Compute the standard error of the mean
- Describe how the Central Limit Theorem determines the nature of the sampling distribution of the mean
- Use the distribution of sample means, z-scores, and unit normal table to determine probabilities corresponding to sample means.

We have come to the final chapter in this unit. We will now take the logic, ideas, and techniques we have developed and put them together to see how we can take a sample of data and use it to make inferences about what's truly happening in the broader population. This is the final piece of the puzzle that we need to understand in order to have the groundwork necessary

for formal hypothesis testing. Though some of the concepts in this chapter seem strange, they are all simple extensions of what we have already learned in previous chapters.

Exercises – Ch. 8

1. What is a sampling distribution?
2. What are the two mathematical facts that describe how sampling distributions work?
3. What is the difference between a sampling distribution and a regular distribution?
4. What effect does sample size have on the shape of a sampling distribution?
5. What is standard error?
6. For a population with a mean of 75 and a standard deviation of 12, what proportion of sample means of size $n = 16$ fall above 82?
7. For a population with a mean of 100 and standard deviation of 16, what is the probability that a random sample of size 4 will have a mean between 110 and 130?
8. Find the z-score for the following means taken from a population with mean 10 and standard deviation 2:
 1. $\bar{X} = 8, n = 12$
 2. $\bar{X} = 8, n = 30$
 3. $\bar{X} = 20, n = 4$
 4. $\bar{X} = 20, n = 16$
9. As the sample size increases, what happens to the p-value associated with a given sample mean?
10. For a population with a mean of 35 and standard deviation of 7, find the sample mean of size $n = 20$ that cuts off the top 5% of the sampling distribution.

Answers to Odd-Numbered Exercises – Ch. 8

1. The sampling distribution (or sampling distribution of the sample means) is the distribution formed by combining many sample means taken from the same population and of a single, consistent sample size.
3. A sampling distribution is made of statistics (e.g. the mean) whereas a regular distribution is made of individual scores.
5. Standard error is the spread of the sampling distribution and is the quantification of sampling error. It is how much we expect the sample mean to naturally change based on random chance.
7. 10.46% or 0.1046
9. As sample size increases, the p-value will decrease

9. Chapter 9

Hypothesis testing

The first unit was designed to prepare you for hypothesis testing. In the first chapter we discussed the three major goals of statistics:

- Describe: connects to unit 1 with descriptive statistics and graphing
- Decide: connects to unit 1 knowing your data and hypothesis testing
- Predict: connects to hypothesis testing and unit 3

The remaining chapters will cover many different kinds of hypothesis tests connected to different inferential statistics. Needless to say, hypothesis testing is the central topic of this course. This lesson is important but that does not mean the same thing as difficult. There is a lot of new language we will learn about when conducting a hypothesis test. Some of the components of a hypothesis test are the topics we are already familiar with:

- Test statistics
- Probability
- Distribution of sample means

Hypothesis testing is an inferential procedure that uses data from a sample to draw a general conclusion about a population. It is a formal approach and a statistical method that uses sample data to evaluate hypotheses about a population. When interpreting a research question and statistical results, a natural question arises as to whether the

finding could have occurred by chance. Hypothesis testing is a statistical procedure for testing whether chance (random events) is a reasonable explanation of an experimental finding. Once you have mastered the material in this lesson you will be used to solving hypothesis testing problems and the rest of the course will seem much easier. In this chapter, we will introduce the ideas behind the use of statistics to make decisions – in particular, decisions about whether a particular hypothesis is supported by the data.

Logic and Purpose of Hypothesis Testing

The statistician Ronald Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Fisher was an Australian statistician and is noted as the first person to formalize the process of hypothesis testing. His elegantly simple “Lady Tasting Tea” experiment demonstrated the logic of the hypothesis test.



Figure 1. A depiction of the lady tasting tea [Photo Credit](#)

Fisher would often have afternoon tea during his studies. He usually took tea with a woman who claimed to be a tea expert. In particular, she told Fisher that she could tell which was poured first in the tea cup, the milk or the tea, simply by tasting the cup. Fisher, being a scientist, decided to put this rather bizarre claim to the test. The lady accepted his challenge. Fisher brought her 8 cups of tea in succession; 4 cups would be prepared with the milk added first, and 4 with the tea added first. The cups would be presented in a random order unknown to the lady.

The lady would take a sip of each cup as it was presented and report which ingredient she believed was poured first. Using the laws of probability, Fisher determined the chances of her guessing all 8 cups correctly was $1/70$, or about 1.4%. In other words, if the lady was indeed guessing there was a 1.4% chance of her getting all 8 cups correct. On the day of the experiment, Fisher had 8 cups prepared just as he had requested. The lady drank each cup and made her decisions for each one.

After the experiment, it was revealed that the lady got all 8 cups correct! Remember, had she been truly guessing, the chance of getting this result was 1.4%. Since this probability was so low, Fisher instead concluded that the lady could indeed differentiate between the milk or the tea being poured first. Fisher's original hypothesis that she was just guessing was demonstrated to be false and was therefore rejected. The alternative hypothesis, that the lady could truly tell the cups apart, was then accepted as true.

This story demonstrates many components of hypothesis testing in a very simple way. For example, Fisher started with a hypothesis that the lady was guessing. He then determined that if she was indeed guessing, the probability of guessing all 8 right was very small, just 1.4%. Since that probability was so tiny, when she did get all 8 cups right, Fisher determined it

was extremely unlikely she was guessing. A more reasonable conclusion was that the lady had the skill to tell the cups apart.

In hypothesis testing, we will *always* set up a particular hypothesis that we want to demonstrate to be true. We then use probability to determine the likelihood of our hypothesis is correct. If it appears our original hypothesis was wrong, we reject it and accept the alternative hypothesis. The alternative hypothesis is usually the opposite of our original hypothesis. In Fisher's case, his original hypothesis was that the lady was guessing. His alternative hypothesis was the lady was not guessing.

Let's consider this with a James Bond twist. James Bond insisted that martinis should be shaken rather than stirred.

Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini (i.e., random assignment). Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result *does not* prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed to be 0.0106. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. A low probability gives us more confidence there is evidence Bond can tell whether the drink was shaken or stirred. There is also still a chance that Mr. Bond was very lucky (more on this

later!). The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

You may notice some patterns here:

- We have 2 hypotheses: the original (researcher prediction) and the alternative
- We collect data
- We determine how likely or unlikely the original hypothesis is to occur based on probability.
- We determine if we have enough evidence to support the original hypothesis and draw conclusions.

Now let's bring in some specific terminology:

Null hypothesis: In general, the null hypothesis, written H_0 ("H-naught"), is the idea that nothing is going on: there is no effect of our treatment, no relation between our variables, and no difference in our sample mean from what we expected about the population mean. The null hypothesis indicates that an apparent effect is due to chance. *This is always our baseline starting assumption, and it is what we (typically) seek to reject.* For mathematical notation, one uses $=$).

Alternative hypothesis: If the null hypothesis is rejected, then we will need some other explanation, which we call the alternative hypothesis, H_A or H_1 . The alternative hypothesis is simply the reverse of the null hypothesis. Thus, our alternative hypothesis is the mathematical way of stating our research question. In general, the alternative hypothesis (also called the research hypothesis) is there is an effect of treatment, the relation between variables, or differences in a sample mean compared to a population mean. The alternative hypothesis essentially shows evidence the findings are *not* due to chance. It is also called the research hypothesis as this is the most common outcome a researcher is looking for: evidence of

change, differences, or relationships. There are three options for setting up the alternative hypothesis, depending on where we expect the difference to lie. The alternative hypothesis always involves some kind of inequality (\neq not equal, $>$, or $<$).

- If we expect a specific direction of change/differences/relationships, which we call a **directional hypothesis**, then our alternative hypothesis takes the form based on the research question itself. One would expect a decrease in depression from taking an anti-depressant as a specific directional hypothesis. Or the direction could be larger, where for example, one might expect an increase in exam scores after completing a student success exam preparation module. The directional hypothesis (2 directions) makes up 2 of the 3 alternative hypothesis options. The other alternative is to state there are differences/changes, or a relationship but not predict the direction. We use a **non-directional alternative hypothesis** (typically see \neq for mathematical notation).

Probability value (p-value): the probability of a certain outcome assuming a certain state of the world. In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. *Using this terminology, the probability value is the probability of an outcome given the hypothesis.* It is *not* the probability of the hypothesis given the outcome. It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing. It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. *This is not at all what it means.* The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing).

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05 ($p < .05$). More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01 ($p < .01$). When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α level or simply α ("alpha"). It is also called the *significance level*. If α is not explicitly specified, assume that $\alpha = 0.05$.

Decision-making is part of the process and we have some language that goes along with that. *Importantly, null hypothesis testing operates under the assumption that the null hypothesis is true unless the evidence shows otherwise.* We (typically) seek to **reject the null hypothesis, giving us evidence to support the alternative hypothesis**. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the null hypothesis is false. Note that all probability calculations for all hypothesis tests center on the null hypothesis. In the James Bond example, the *null hypothesis* is that he *cannot* tell the difference between shaken and stirred martinis. The probability value is low that one is able to identify 13 of 16 martinis as shaken or stirred (0.0106), thus providing evidence that he can tell the difference. Note that we have *not* computed the probability that he can tell the difference.

The specific type of hypothesis testing reviewed is specifically known as *null hypothesis statistical testing* (NHST). We can break the process of null hypothesis testing down into a number of steps a researcher would use.

- Formulate a hypothesis that embodies our prediction

(before seeing the data)

- Specify null and alternative hypotheses
- Collect some data relevant to the hypothesis
- Compute a test statistic
- Identify the criteria probability (or compute the probability of the observed value of that statistic) assuming that the null hypothesis is true
- Drawing conclusions. Assess the “statistical significance” of the result

Steps in hypothesis testing

Let's consider another example as we set up some more formal steps for hypothesis testing that we will use throughout the semester. An experiment was done to determine whether physicians spend less time with obese patients, we will refer to the study as Physicians' Reactions. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

Identify some key aspects of the study design: it is an experiment with 2 levels or groups in the independent variable (IV) and random assignment was used to place physicians into only 1 level/group/condition. IV has 2 categories and is classified as a nominal scale of measurement. The outcome variable being examined is the dependent variable (DV), which is estimated minutes to examine the medical chart. Minutes can be classified as a ratio scale of

measurement.

Step 1: Formulate a hypothesis of interest

The researchers hypothesized that physicians spend less time with obese patients. The researchers hypothesis derived from an identified population. In creating a research hypothesis, we also have to decide whether we want to test a *directional* or *non-directional* hypotheses. Researchers typically will select a non-directional hypothesis for a more conservative approach, particularly when the outcome is unknown (more about why this is later).

Step 2: Specify the null and alternative hypotheses

Can you set up the null and alternative hypotheses for the Physician's Reaction Experiment?

In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as: $H_0: \mu_{\text{obese}} - \mu_{\text{average}} = 0$.

In the Physicians' Reactions example, the alternative hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as: $H_1: \mu_{\text{obese}} - \mu_{\text{average}} \neq 0$. Again, it is more common to take a non-directional approach. However, if there was previous research or evidence for physicians spending less time with obese patients, the alternative hypothesis could be written as: $H_1: \mu_{\text{obese}} - \mu_{\text{average}} < 0$.

A reminder that in setting up hypotheses, you will see parameters (μ for mean) used in hypotheses as we are interested in understanding the population, not just our sample.

Step 3: Determine the alpha level.

For this course, alpha will be given to you as .05 or .01. Researchers will decide on alpha and then determine the associated test statistic based from the sample. Researchers in the Physician Reaction study might set the alpha at .05 and identify the test statistics associated with the .05 for the sample size. Researchers might take extra precautions to be more confident in their findings (more on this later).

Step 4: Collect some data

For this course, the data will be given to you. Researchers collect the data and then start to summarize it using descriptive statistics. The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for normal-weight patients.

Step 5: Compute a test statistic

We next want to use the data to compute a statistic that will ultimately let us decide whether the null hypothesis is rejected or not. We can think of the test statistic as providing a measure of the size of the effect compared to the variability in the data. In general, this test statistic will have a probability distribution associated with it, because that allows us to determine how

likely our observed value of the statistic is under the null hypothesis.

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference ($31.4 - 24.7 = 6.7$ minutes) if the difference were, in fact, due solely to chance.

Step 6: Determine the probability of the observed result under the null hypothesis

Using methods presented in later chapters, this probability associated with the observed differences between the two groups for the Physician's Reaction was computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight (obese or not) (and is not due to chance). We can then reject the null hypothesis (there are no differences or differences seen are due to chance).

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically as part of the researcher's control of other variables. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

This is the step where NHST starts to violate our intuition. Rather than determining the likelihood that the null hypothesis is true given the data, we instead determine the likelihood under the null hypothesis of observing a statistic at

least as extreme as one that we have observed — because we started out by assuming that the null hypothesis is true! To do this, we need to know the expected probability distribution for the statistic under the null hypothesis, so that we can ask how likely the result would be under that distribution. This will be determined from a table we use for reference or calculated in a statistical analysis program. Note that when I say “how likely the result would be”, what I really mean is “how likely the observed result or one more extreme would be”. We need to add this caveat as we are trying to determine how weird our result would be if the null hypothesis were true, and any result that is more extreme will be even more weird, so we want to count all of those weirder possibilities when we compute the probability of our result under the null hypothesis.

Let's review some considerations for *Null hypothesis statistical testing* (NHST)!

Null hypothesis statistical testing (NHST) is commonly used in many fields. If you pick up almost any scientific or biomedical research publication, you will see NHST being used to test hypotheses, and in their introductory psychology textbook, Gerrig & Zimbardo (2002) referred to NHST as the “backbone of psychological research”. Thus, learning how to use and interpret the results from hypothesis testing is essential to understand the results from many fields of research.

It is also important for you to know, however, that NHST is flawed, and that many statisticians and

researchers think that it has been the cause of serious problems in science, which we will discuss in further in this unit. NHST is also widely misunderstood, largely because it violates our intuitions about how statistical hypothesis testing should work. Let's look at an example to see this.

There is great interest in the use of body-worn cameras by police officers, which are thought to reduce the use of force and improve officer behavior. However, in order to establish this we need experimental evidence, and it has become increasingly common for governments to use randomized controlled trials to test such ideas. A randomized controlled trial of the effectiveness of body-worn cameras was performed by the Washington, DC government and DC Metropolitan Police Department in 2015-2016. Officers were randomly assigned to wear a body-worn camera or not, and their behavior was then tracked over time to determine whether the cameras resulted in less use of force and fewer civilian complaints about officer behavior.

Before we get to the results, let's ask how you would think the statistical analysis might work. Let's say we want to specifically test the hypothesis of whether the use of force is decreased by the wearing of cameras. The randomized controlled trial provides us with the data to test the hypothesis – namely, the rates of use of force by officers assigned to either the camera or control groups. The next

obvious step is to look at the data and determine whether they provide convincing evidence for or against this hypothesis. That is: What is the likelihood that body-worn cameras reduce the use of force, given the data and everything else we know?

It turns out that this is *not* how null hypothesis testing works. Instead, we first take our hypothesis of interest (i.e. that body-worn cameras reduce use of force), and flip it on its head, creating a *null hypothesis* – in this case, the null hypothesis would be that cameras do not reduce use of force. Importantly, we then assume that the null hypothesis is true. We then look at the data, and determine how likely the data would be if the null hypothesis were true. If the data are sufficiently unlikely under the null hypothesis that we can reject the null in favor of the *alternative hypothesis* which is our hypothesis of interest. If there is not sufficient evidence to reject the null, then we say that we retain (or “fail to reject”) the null, sticking with our initial assumption that the null is true.

Understanding some of the concepts of NHST, particularly the notorious “p-value”, is invariably challenging the first time one encounters them, because they are so counter-intuitive. As we will see later, there are other approaches that provide a much more intuitive way to address hypothesis testing (but have their own complexities).

Step 7: Assess the “statistical significance” of the result. Draw conclusions.

The next step is to determine whether the p-value that results from the previous step is small enough that we are willing to reject the null hypothesis and conclude instead that the alternative is true. In the Physicians Reactions study, the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what “significant” usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

How much evidence do we require and what considerations are needed to better understand the significance of the findings? This is one of the most controversial questions in statistics, in part because it requires a subjective judgment – there is no “correct” answer.

What does a statistically significant result mean?

There is a great deal of confusion about what p-values actually mean (Gigerenzer, 2004). Let’s say that we do an experiment comparing the means between conditions, and we find a difference with a p-value of .01. There are a number of possible interpretations that one might entertain.

Does it mean that the probability of the null hypothesis being true is .01? No. Remember that in null hypothesis testing, the p-value is the probability of the data given the null hypothesis. It does not warrant conclusions about the probability of the null hypothesis given the data.

Does it mean that the probability that you are making the wrong decision is .01? No. Remember as above that p-values are probabilities of data under the null, not probabilities of hypotheses.

Does it mean that if you ran the study again, you would obtain the same result 99% of the time? No. The p-value is a statement about the likelihood of a particular dataset under the null; it does not allow us to make inferences about the likelihood of future events such as replication.

Does it mean that you have found a practically important effect? No. There is an essential distinction between *statistical significance* and *practical significance*. As an example, let's say that we performed a randomized controlled trial to examine the effect of a particular diet on body weight, and we find a statistically significant effect at $p < .05$. What this doesn't tell us is how much weight was actually lost, which we refer to as the *effect size* (to be discussed in more detail). If we think about a study of weight loss, then we probably don't think that the loss of one ounce (i.e. the weight of a few potato chips) is practically significant. Let's look at our ability to detect a significant difference of 1 ounce as the sample size increases.

A statistically significant result is not necessarily a strong one. Even a very weak result can be statistically significant if it is based on a large enough sample. This is why it is important to distinguish between the statistical significance of a result and the practical significance of that result. Practical significance refers to the importance or usefulness of the result in some real-world context and is often referred to as the **effect size**.

Many differences are statistically significant—and may even be interesting for purely scientific reasons—but they are not practically significant. In clinical practice, this same concept is often referred to as “clinical significance.” For example, a study on a new treatment for social phobia might show that it

produces a statistically significant positive effect. Yet this effect still might not be strong enough to justify the time, effort, and other costs of putting it into practice—especially if easier and cheaper treatments that work almost as well already exist. Although statistically significant, this result would be said to lack practical or clinical significance.

Be aware that the term effect size can be misleading because it suggests a causal relationship—that the difference between the two means is an “effect” of being in one group or condition as opposed to another. In other words, simply calling the difference an “effect size” does not make the relationship a causal one.

Figure 1 shows how the proportion of significant results increases as the sample size increases, such that with a very large sample size (about 262,000 total subjects), we will find a significant result in more than 90% of studies when there is a 1 ounce difference in weight loss between the diets. *While these are statistically significant, most physicians would not consider a weight loss of one ounce to be practically or clinically significant.* We will explore this relationship in more detail when we return to the concept of *statistical power* in Chapter X, but it should already be clear from this example that statistical significance is not necessarily indicative of practical significance.

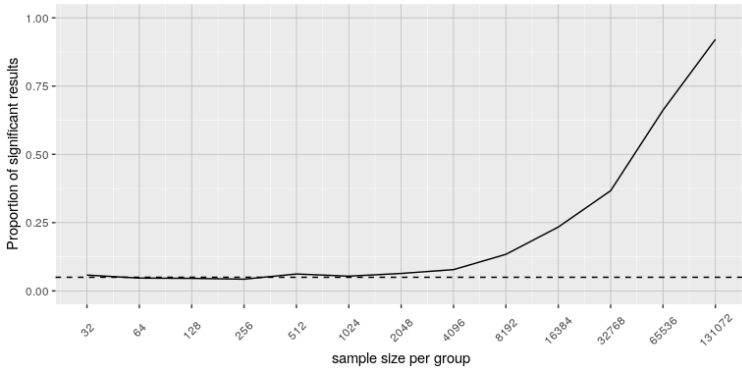


Figure 1: The proportion of significant results for a very small change (1 ounce, which is about .001 standard deviations) as a function of sample size.

Challenges with using p-values

Historically, the most common answer to this question has been that we should reject the null hypothesis if the p-value is less than 0.05. This comes from the writings of Ronald Fisher, who has been referred to as “the single most important figure in 20th century statistics” (Efron, [1998](#)):

“If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 ... it is convenient to draw the line at about the level at which we can say: Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials” (Fisher, [1925](#))

Fisher never intended $p < 0.05$ to be a fixed rule:

“no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas” (Fisher, [1956](#))

Instead, it is likely that $p < .05$ became a ritual due to the reliance upon tables of p-values that were used before computing made it easy to compute p values for arbitrary values of a statistic. All of the tables had an entry for 0.05, making it easy to determine whether one's statistic exceeded the value needed to reach that level of significance. Although we use tables in this class, statistical software examines the specific probability value for the calculated statistic.

Assessing Error Rate: Type I and Type II Error

Although there are challenges with p-values for decision making, we will examine a way we can think about hypothesis testing in terms of its error rate. This was proposed by Jerzy Neyman and Egon Pearson:

“no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong” (Neyman & Pearson, [1933](#))

That is: We can't know which specific decisions are right or wrong, but if we follow the rules, we can at least know how often our decisions will be wrong in the long run.

To understand the decision-making framework that Neyman and Pearson developed, we first need to discuss statistical decision-making in terms of the kinds of outcomes that can occur. There are two possible states of reality (H_0 is true, or H_0 is false), and two possible decisions (reject H_0 , or retain H_0). There are two ways in which we can make a correct decision:

- We can reject H_0 when it is false (in the language of signal detection theory, we call this a *hit*)
- We can retain H_0 when it is true (somewhat confusingly in this context, this is called a *correct rejection*)

There are also two kinds of errors we can make:

- We can reject H_0 when it is actually true (we call this a *false alarm*, or *Type I error*), **Type I error** means that we have concluded that there is a relationship in the population when in fact there is not. Type I errors occur because even when there is no relationship in the population, sampling error alone will occasionally produce an extreme result.
- We can retain H_0 when it is actually false (we call this a *miss*, or *Type II error*). **Type II error** means that we have concluded that there is no relationship in the population when in fact there is.

Summing up, when you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

| ACTION | H_0 IS ACTUALLY | |
|---------------------|---------------------|----------------------|
| | True | False |
| Do not reject H_0 | Correct Outcome | Type II error |
| Reject H_0 | Type I Error | Correct Outcome |

Table 1. The four possible outcomes in hypothesis testing.

1. The decision is **not to reject H_0** when **H_0 is true (correct decision)**.
2. The decision is to **reject H_0** when **H_0 is true** (incorrect decision known as a **Type I error**).
3. The decision is **not to reject H_0** when, in fact, **H_0 is false** (incorrect decision known as a **Type II error**).
4. The decision is to **reject H_0** when **H_0 is false (correct decision)**.

Neyman and Pearson coined two terms to describe the probability of these two types of errors in the long run:

- $P(\text{Type I error}) = \alpha$
- $P(\text{Type II error}) = \beta$

That is, if we set α to .05, then in the long run we should make a Type I error 5% of the time. The α (alpha), is associated with the p-value for the level of significance. Again it's common to set α as .05. In fact, when the null hypothesis is true and α is .05, we will mistakenly reject the null hypothesis 5% of the time. (This is why α is sometimes referred to as the "Type I error rate.") In principle, it is possible to reduce the chance of a Type I error by setting α to something less than .05. Setting it to .01, for example, would mean that if the null hypothesis is true, then there is only a 1% chance of mistakenly rejecting it. But making it harder to reject true null hypotheses also makes it harder to reject false ones and therefore increases the chance of a Type II error.

In practice, Type II errors occur primarily because the research design lacks adequate statistical power to detect the relationship (e.g., the sample is too small). Statistical power is the complement of Type II error. We will have more to say about statistical power shortly. The standard value for an

acceptable level of β (beta) is .2 – that is, we are willing to accept that 20% of the time we will fail to detect a true effect when it truly exists. It is possible to reduce the chance of a Type II error by setting α to something greater than .05 (e.g., .10). But making it easier to reject false null hypotheses also makes it easier to reject true ones and therefore increases the chance of a Type I error. This provides some insight into why the convention is to set α to .05. There is some agreement among researchers that level of α keeps the rates of both Type I and Type II errors at acceptable levels.

The possibility of committing Type I and Type II errors has several important implications for interpreting the results of our own and others' research. One is that we should be cautious about interpreting the results of any individual study because there is a chance that it reflects a Type I or Type II error. This is why researchers consider it important to replicate their studies. Each time researchers replicate a study and find a similar result, they rightly become more confident that the result represents a real phenomenon and not just a Type I or Type II error.

Test Statistic Assumptions

Last consideration we will revisit with each test statistic (e.g., t-test, z-test and ANOVA) in the coming chapters. There are four main assumptions. These assumptions are often taken for granted in using prescribed data for the course. In the real world, these assumptions would need to be examined, often tested using statistical software.

1. Assumption of random sampling. A sample is random when each person (or animal) point in your population has an equal chance of being included in the sample;

therefore selection of any individual happens by chance, rather than by choice. This reduces the chance that differences in materials, characteristics or conditions may bias results. Remember that random samples are more likely to be representative of the population so researchers can be more confident interpreting the results. Note: there is no test that statistical software can perform which assures random sampling has occurred but following good sampling techniques helps to ensure your samples are random.

2. Assumption of Independence. Statistical independence is a critical assumption for many statistical tests including the 2-sample t-test and ANOVA. It is assumed that observations are independent of each other often but often this assumption is not met. Independence means the value of one observation does not influence or affect the value of other observations. Independent data items are not connected with one another in any way (unless you account for it in your study). Even the smallest dependence in your data can turn into heavily biased results (which may be undetectable) if you violate this assumption. Note: there is no test statistical software can perform that assures independence of the data because this should be addressed during the research planning phase. Using a non-parametric test is often recommended if a researcher is concerned this assumption has been violated.
3. Assumption of Normality. Normality assumes that the continuous variables (dependent variable) used in the analysis are normally distributed. Normal distributions are symmetric around the center (the mean) and form a bell-shaped distribution. Normality is violated when sample data are skewed. With **large enough sample sizes** ($n > 30$) the violation of the normality assumption should not cause major problems (remember the central limit

theorem) but there is a feature in most statistical software that can alert researchers to an assumption violation.

4. Assumption of Equal Variance. Variance refers to the spread or of scores from the mean. Many statistical tests assume that although different samples can come from populations with different means, they have the same variance. Equality of variance (i.e., homogeneity of variance) is violated when variances across different groups or samples are significantly different. Note: there is a feature in most statistical software to test for this.

Recap

We will use 4 main steps for hypothesis testing:

1. Begin with two hypotheses. Write a null hypothesis and alternative hypothesis about the populations.
 1. Usually the hypotheses concern population parameters and predict the characteristics that a sample should have
 2. The hypotheses are contradictory
 1. Null: Null hypothesis (H_0) states that there is no difference, no effect or no change between population means and sample means. There is no difference.
 2. Alternative: Alternative hypothesis (H_1 or H_A) states that there is a difference or a change between the population and sample. It is the opposite of the null hypothesis.
2. Set criteria for a decision. In this step we must determine the boundary of our distribution at which the null hypothesis will be rejected. Researchers usually use either a 5% (.05) cutoff or 1% (.01) critical boundary. Recall from our earlier story about Ronald Fisher that the *lower the probability* the more confident he was that the Tea Lady

was not guessing. We will apply this to z in the next chapter.





3. Sample data are collected and analyzed by performing statistics (calculations)
 1. Compare sample and population to decide if the hypothesis has support
4. Make a decision and provide an explanation
 1. When a researcher uses hypothesis testing, the individual is making a decision about whether the data collected is sufficient to state that the population parameters are significantly different.

Further considerations

1. The probability value is the probability of a result as extreme or more extreme given that the null hypothesis is true. It is the probability of the data given the null hypothesis. It is not the probability that the null hypothesis is false.

2. A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true. We will learn more about assessing effect size later in this unit.

3. A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false. There is always a chance of error and 4 outcomes associated with hypothesis testing.

| | | THE TRUTH: | |
|---------------------|-----------------------|--|--|
| | | H_0 is True | H_0 is False |
| The decision maker: | Rejects H_0 | Type 1 Error (α)  | Correct Decision  |
| | Fail to Rejects H_0 | Correct Decision  | Type 2 Error  |

4. It is important to take into account the assumptions for each test statistic.

Learning objectives

Having read the chapter, you should be able to:

- Identify the components of a hypothesis test, including the parameter of interest, the null and alternative hypotheses, and the test statistic.
- State the hypotheses and identify appropriate critical areas depending on how hypotheses are set up.
- Describe the proper interpretations of a p-value as well as common misinterpretations.
- Distinguish between the two types of error in hypothesis testing, and the factors that determine them.
- Describe the main criticisms of null hypothesis statistical testing
- Identify the purpose of effect size and power.

Exercises – Ch. 9

1. In your own words, explain what the null hypothesis is.
2. What are Type I and Type II Errors?
3. What is α ?
4. Why do we phrase null and alternative hypotheses with population parameters and not sample means?
5. If our null hypothesis is " $H_0: \mu = 40$ ", what are the three possible alternative hypotheses?
6. Why do we state our hypotheses and decision criteria before we collect our data?
7. When and why do you calculate an effect size?

Answers to Odd- Numbered Exercises – Ch. 9

1. Your answer should include mention of the baseline assumption of no difference between the sample and the population.

3. Alpha is the significance level. It is the criteria we use when decided to reject or fail to reject the null hypothesis, corresponding to a given proportion of the area under the normal distribution and a probability of finding extreme scores assuming the null hypothesis is true.

5. $\mu > 40$; $\mu < 40$; $\mu \neq 40$

7. We calculate effect size to determine the strength of the finding. Effect size should always be calculated when we have rejected the null hypothesis. Effect size can be calculated for non-significant findings as a possible indicator of Type II error.

10. Chapter 10: Hypothesis Testing with Z

This chapter lays out the basic logic and process of hypothesis testing using a z. We will perform a test statistics using z, we use the z formula from chapter 8 and data from a sample mean to make an inference about a population.

Setting up the hypotheses

When setting up the hypotheses with z, the parameter is associated with a sample mean (in the previous chapter examples the parameters for the null used 0). Using z is an occasion in which the null hypothesis is a value other than 0. For example, if we are working with mothers in the U.S. whose children are at risk of low birth weight, we can use 7.47 pounds, the average birth weight in the US, as our null value and test for differences against that. For now, we will focus on testing a value of a single mean against what we expect from the population.

Using birthweight as an example, our null hypothesis takes the form: $H_0: \mu = 7.47$ Notice that we are testing the value for μ , the population parameter, NOT the sample statistic \bar{X} (or M). We are referring to the data right now in raw form (we have not standardized it using z yet). Again, using inferential statistics, we are interested in understanding the population, drawing from our sample observations. For the research question, we

have a *mean value from the sample* to use, we have specific data is – it is observed and used as a comparison for a set point.

As mentioned earlier, the alternative hypothesis is simply the reverse of the null hypothesis, and there are three options, depending on where we expect the difference to lie. We will set the criteria for rejecting the null hypothesis based on the directionality (greater than, less than, or not equal to) of the alternative.

If we expect our obtained sample mean to be above or below the null hypothesis value (knowing which direction), we set a **directional hypothesis**. Our alternative hypothesis takes the form based on the research question itself. In our example with birthweight, this could be presented as $H_A: \mu > 7.47$ or $H_A: \mu < 7.47$.

Note that we should only use a directional hypothesis if we have a good reason, based on prior observations or research, to suspect a particular direction. When we do not know the direction, such as when we are entering a new area of research, we use a **non-directional alternative hypothesis**. In our birthweight example, this could be set as $H_A: \mu \neq 7.47$

In working with data for this course we will need to set a critical value of the test statistic for alpha (α) for use of test statistic tables in the back of the book. This is determining the critical rejection region that has a set critical value based on α .

Determining Critical Value from α

We set alpha (α) before collecting data in order to determine whether or not we should reject the null hypothesis. *We set this value beforehand to avoid biasing ourselves by viewing our results and then determining what criteria we should use.*

When a research hypothesis predicts an effect but does not predict a direction for the effect, it is called a *non-directional hypothesis*. To test the significance of a non-directional hypothesis, we have to consider the possibility that the sample could be extreme at either tail of the comparison distribution. We call this a **two-tailed test**.

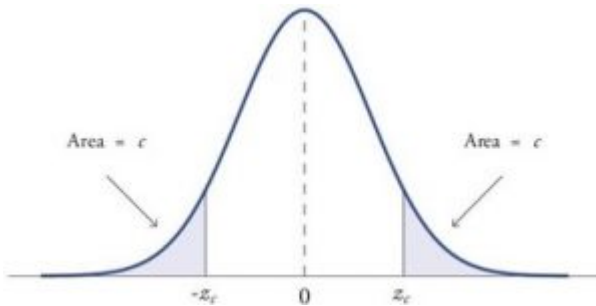


Figure 1. showing a 2-tail test for non-directional hypothesis for z for area C is the critical rejection region.

When a research hypothesis predicts a direction for the effect, it is called a *directional hypothesis*. To test the significance of a directional hypothesis, we have to consider the possibility that the sample could be extreme at one-tail of the comparison distribution. We call this a **one-tailed test**.

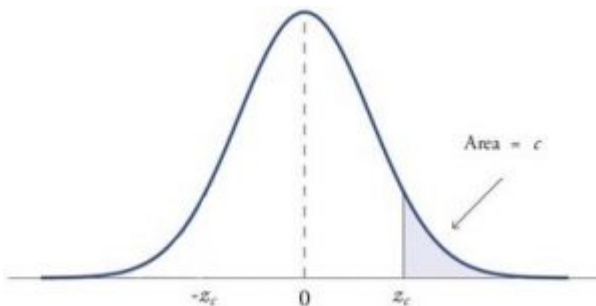


Figure 2. showing a 1-tail test for a directional hypothesis (predicting an increase) for z for area C is the critical rejection region.

Determining Cutoff Scores with Two-Tailed Tests

Typically we specify an α level before analyzing the data. If the data analysis results in a probability value below the α level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected. In other words, if our data produce values that meet or exceed this threshold, then we have sufficient evidence to *reject the null hypothesis*; if not, we *fail to reject the null* (we never “accept” the null). According to this perspective, if a result is significant, then it does not matter how significant it is. Moreover, if it is not significant, then it does not matter how close to being significant it is. Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically. Similarly, probability values of 0.06 and 0.34 are treated identically. Note we will discuss ways to address effect size (which is related to this challenge of NHST).

When setting the probability value, there is a special complication in a two-tailed test. We have to divide the significance percentage between the two tails. For example, with a 5% significance level, we reject the null hypothesis only if the sample is so extreme that it is in either the top 2.5% or the bottom 2.5% of the comparison distribution. This keeps the overall level of significance at a total of 5%. A one-tailed test does have such an extreme value but with a one-tailed test only one side of the distribution is considered.

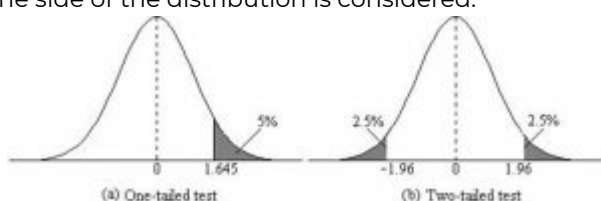


Figure 3. Critical value differences in one and two-tail tests.

[Photo Credit](#)

Let's review the set critical values for Z.

We discussed z-scores and probability in chapter 8. If we revisit the z-score for 5% and 1%, we can identify the critical regions for the critical rejection areas from the unit standard normal table.

- A two-tailed test at the 5% level has a critical boundary Z score of +1.96 and -1.96
- A one-tailed test at the 5% level has a critical boundary Z score of +1.64 or -1.64
- A two-tailed test at the 1% level has a critical boundary Z score of +2.58 and -2.58
- A one-tailed test at the 1% level has a critical boundary Z score of +2.33 or -2.33.

Review: Critical values, p-values, and significance level

There are two criteria we use to assess whether our data meet the thresholds established by our chosen significance level, and they both have to do with our discussions of probability and distributions. Recall that probability refers to the likelihood of an event, given some situation or set of conditions. In hypothesis testing, that situation is the assumption that the null hypothesis value is the correct value, or that there is *no effect*. The value laid out in H_0 is our condition under which we interpret our results. To reject this assumption, and thereby reject the null hypothesis, we need results that would be very unlikely if the null was true.

Now recall that values of z which fall in the tails of the standard normal distribution represent unlikely values. That is, the proportion of the area under the curve as or more extreme than z is very small as we get into the tails of the distribution. Our significance level corresponds to the area under the tail that is exactly equal to α : if we use our normal criterion of $\alpha = .05$, then 5% of the area under the curve becomes what we call the rejection region (also called the critical region) of the distribution. This is illustrated in Figure 4.

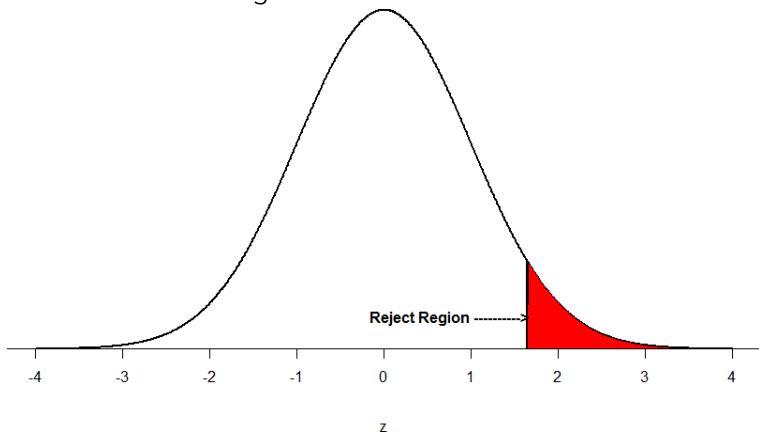


Figure 4: The rejection region for a one-tailed test

The shaded rejection region takes us 5% of the area under the curve. Any result which falls in that region is sufficient evidence to reject the null hypothesis.

The rejection region is bounded by a specific z -value, as is any area under the curve. In hypothesis testing, the value corresponding to a specific rejection region is called the critical value, z_{crit} ("z-crit") or z^* (hence the other name "critical region"). Finding the critical value works exactly the same as finding the z -score corresponding to any area under the curve like we did in Unit 1. If we go to

the normal table, we will find that the z-score corresponding to 5% of the area under the curve is equal to 1.645 ($z = 1.64$ corresponds to 0.0405 and $z = 1.65$ corresponds to 0.0495, so .05 is exactly in between them) if we go to the right and -1.645 if we go to the left. The direction must be determined by your alternative hypothesis, and drawing then shading the distribution is helpful for keeping directionality straight.

Suppose, however, that we want to do a non-directional test. We need to put the critical region in both tails, but we don't want to increase the overall size of the rejection region (for reasons we will see later). To do this, we simply split it in half so that an equal proportion of the area under the curve falls in each tail's rejection region. For $\alpha = .05$, this means 2.5% of the area is in each tail, which, based on the z-table, corresponds to critical values of $z^* = \pm 1.96$. This is shown in Figure 5.

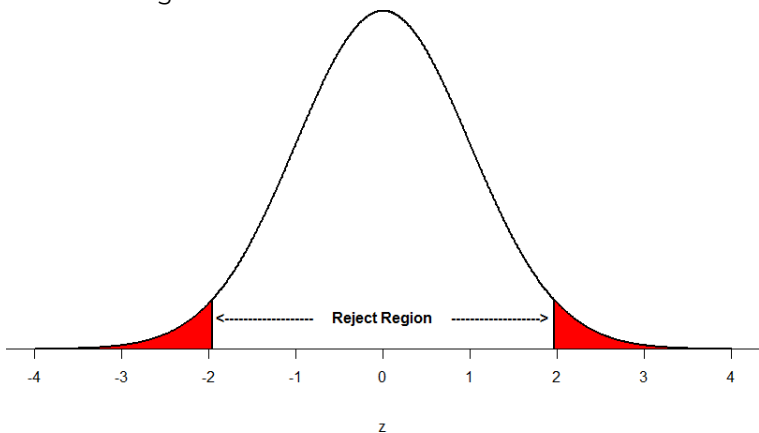


Figure 5: Two-tailed rejection region

Thus, any z-score falling outside ± 1.96 (greater than 1.96 in absolute value) falls in the rejection region. When we use z-scores in this way, the obtained value of z (sometimes called z-obtained) is something

known as a test statistic, which is simply an inferential statistic used to test a null hypothesis.

Calculate the test statistic: Z

Now that we understand setting up the hypothesis and determining the outcome, let's examine hypothesis testing with z! The next step is to carry out the study and get the actual results for our sample. Central to hypothesis test is comparison of the population and sample means. To make our calculation and determine where the sample is in the hypothesized distribution we calculate the Z for the sample data.

Make a decision

To decide whether to reject the null hypothesis, we compare our sample's Z score to the Z score that marks our critical boundary. If our sample Z score falls inside the rejection region of the comparison distribution (is greater than the z-score critical boundary) we reject the null hypothesis.

The formula for our z- statistic has not changed:

$$Z = \frac{M - \mu}{\sigma_M} \quad \text{where}$$
$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

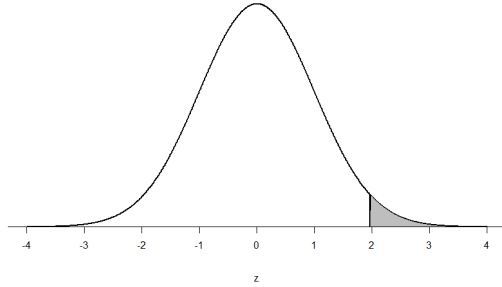
To formally test our hypothesis, we compare our obtained z-statistic to our critical z-value. If $Z_{obt} > Z_{crit}$, that means it falls

in the rejection region (to see why, draw a line for $z = 2.5$ on Figure 1 or Figure 2) and so we reject H_0 . If $z_{obt} < z_{crit}$, we fail to reject. Remember that as z gets larger, the corresponding area under the curve beyond z gets smaller. Thus, the proportion, or p-value, will be smaller than the area for α , and if the area is smaller, the probability gets smaller. Specifically, the probability of obtaining that result, or a more extreme result, under the condition that the null hypothesis is true gets smaller.

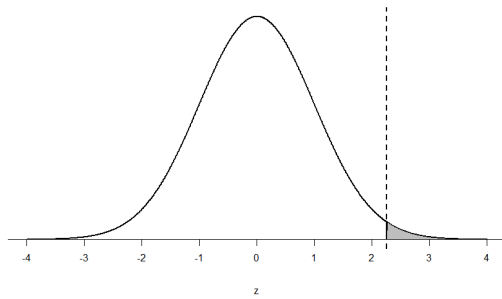
The z-statistic is very useful when we are doing our calculations by hand. However, when we use computer software, it will report to us a p-value, which is simply the proportion of the area under the curve in the tails beyond our obtained z-statistic. We can directly compare this p-value to α to test our null hypothesis: if $p < \alpha$, we reject H_0 , but if $p > \alpha$, we fail to reject. Note also that the reverse is always true: if we use critical values to test our hypothesis, we will always know if p is greater than or less than α . If we reject, we know that $p < \alpha$ because the obtained z-statistic falls farther out into the tail than the critical z-value that corresponds to α , so the proportion (p-value) for that z-statistic will be smaller.

Conversely, if we fail to reject, we know that the proportion will be larger than α because the z-statistic will not be as far into the tail. This is illustrated for a one- tailed test in Figure 6.

Rejection Region for $\alpha = 0.05$, $z^* = 1.96$



Shaded p-value for $z_{\text{obt}} = 2.25$, Reject H_0



Shaded p-value for $z_{\text{obt}} = 1.25$, Fail to Reject H_0

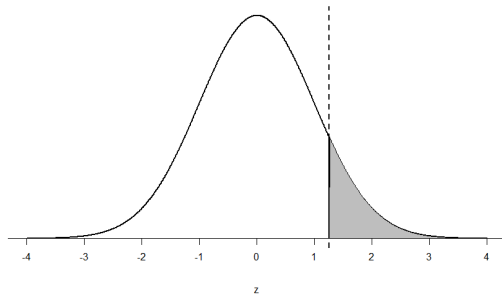


Figure 6. Relation between α , z_{obt} , and p

When the null hypothesis is rejected, the effect is said to be *statistically significant*. Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough.

Why does the word “significant” in the phrase “statistically

significant” mean something so different from other uses of the word? Interestingly, this is because the meaning of “significant” in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was “significant” if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of “significant” changed, leading to the potential misinterpretation.

Review: Steps of the Hypothesis Testing Process

The process of testing hypotheses follows a simple four-step procedure. This process will be what we use for the remainder of the textbook and course, and though the hypothesis and statistics we use will change, this process will not.

Step 1: State the Hypotheses

Your hypotheses are the first thing you need to lay out. Otherwise, there is nothing to test! You have to state the null hypothesis (which is what we test) and the alternative hypothesis (which is what we expect). These should be stated mathematically as they were presented above AND in words, explaining in normal English what each one means in terms of the research question.

Step 2: Find the Critical Values

Next, we formally lay out the criteria we will use to test our hypotheses. There are two pieces of information that inform

our critical values: α , which determines how much of the area under the curve composes our rejection region, and the directionality of the test, which determines where the region will be.

Step 3: Compute the Test Statistic

Once we have our hypotheses and the standards we use to test them, we can collect data and calculate our test statistic, in this case z . This step is where the vast majority of differences in future chapters will arise: different tests used for different data are calculated in different ways, but the way we use and interpret them remains the same.

Step 4: Make the Decision

Finally, once we have our obtained test statistic, we can compare it to our critical value and decide whether we should reject or fail to reject the null hypothesis. When we do this, we must interpret the decision in relation to our research question, stating what we concluded, what we based our conclusion on, and the specific statistics we obtained.

Example: Movie Popcorn

Let's see how hypothesis testing works in action by working through an example. Say that a movie theater owner likes to keep a very close eye on how much popcorn goes into each bag sold, so he knows that the average bag has 8 cups of popcorn and that this varies a little bit, about half a cup. That is, the known population mean is $\mu = 8.00$ and

the known population standard deviation is $\sigma = 0.50$. The owner wants to make sure that the newest employee is filling bags correctly, so over the course of a week he randomly assesses 25 bags filled by the employee to test for a difference ($n = 25$). He doesn't want bags overfilled or under filled, so he looks for differences in both directions. This scenario has all of the information we need to begin our hypothesis testing procedure.

Step 1: State the Hypotheses

Our manager is looking for a difference in the mean weight of popcorn bags compared to the population mean of 8. We will need both a null and an alternative hypothesis written both mathematically and in words. We'll always start with the null hypothesis:

H_0 : There is no difference in the weight of popcorn bags from this employee $H_0: \mu = 8.00$

Notice that we phrase the hypothesis in terms of the population parameter μ , which in this case would be the true average weight of bags filled by the new employee.

Our assumption of no difference, the null hypothesis, is that this mean is exactly

the same as the known population mean value we want it to match, 8.00. Now let's do the alternative:

H_A : There is a difference in the weight of popcorn bags from this employee $H_A: \mu \neq 8.00$

In this case, we don't know if the bags will be too full or not full enough, so we do a two-tailed alternative hypothesis that there is a difference.

Step 2: Find the Critical Values

Our critical values are based on two things: the directionality of the test and the level of significance. We decided in step 1 that a two-tailed test is the appropriate directionality. We were given no information about the level of significance, so we assume that $\alpha = 0.05$ is what we will use. As stated earlier in the chapter, the critical values for a two-tailed z-test at $\alpha = 0.05$ are $z^* = \pm 1.96$. This will be the criteria we use to test our hypothesis. We can now draw out our distribution so we can visualize the rejection region and make sure it makes sense

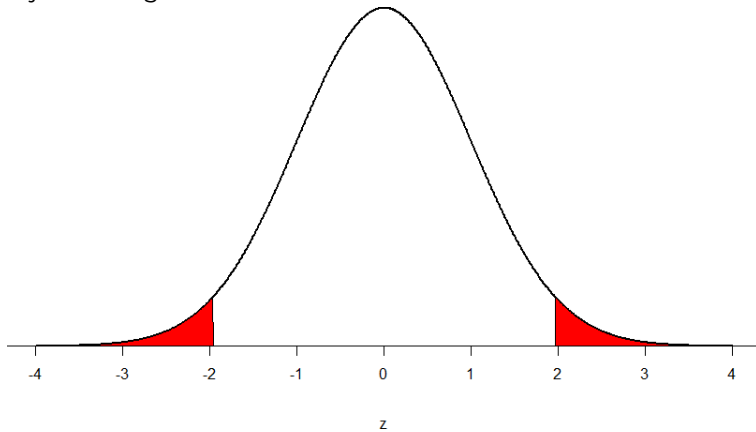


Figure 7: Rejection region for $z^* = \pm 1.96$

Step 3: Calculate the Test Statistic

Now we come to our formal calculations. Let's say that the manager collects data and finds that the average weight of this employee's popcorn bags is $\bar{X} = 7.75$ cups. We can now plug this value, along with the values presented in the original problem, into our equation for z:

$$z = \frac{7.75 - 8.00}{0.50/\sqrt{25}} = \frac{-0.25}{0.10} = -2.50$$

So our test statistic is $z = -2.50$, which we can draw onto our rejection region distribution:

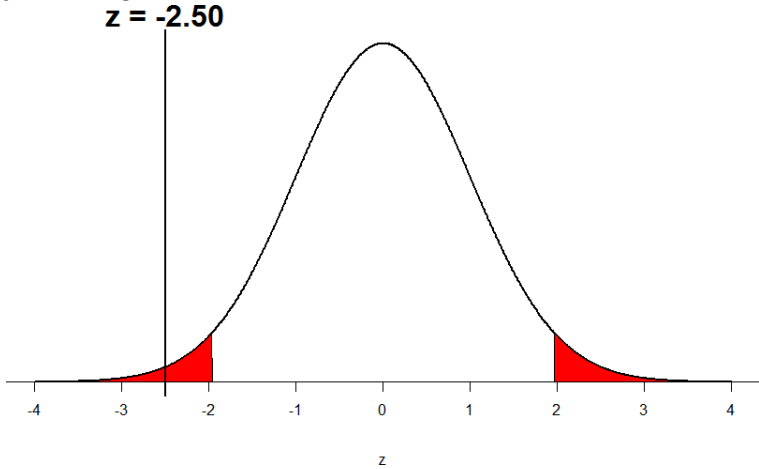


Figure 8: Test statistic location

Step 4: Make the Decision

Looking at Figure 5, we can see that our obtained z -statistic falls in the rejection region. We can also directly compare it to our critical value: in terms of absolute value, $-2.50 > -1.96$, so we reject the null hypothesis. We can now write our conclusion:

Reject H_0 . Based on the sample of 25 bags, we can conclude that the average popcorn bag from this employee is smaller ($\bar{X} = 7.75$ cups) than the average weight of popcorn bags at this movie theater, $z = -2.50$, $p < 0.05$.

When we write our conclusion, we write out the words to communicate what it actually means, but we also include the average sample size we calculated (the exact location doesn't matter, just somewhere that flows naturally and makes sense) and the z -statistic and p -value. We don't know the exact p -

value, but we do know that because we rejected the null, it must be less than α .

Effect Size

When we reject the null hypothesis, we are stating that the difference we found was statistically significant, but we have mentioned several times that this tells us nothing about practical significance. To get an idea of the actual size of what we found, we can compute a new statistic called an effect size. Effect sizes give us an idea of how large, important, or meaningful a statistically significant effect is.

For mean differences like we calculated here, our effect size is Cohen's d :

$$d = \frac{M - \mu}{\sigma}$$

This is very similar to our formula for z , but we no longer take into account the sample size (since overly large samples can make it too easy to reject the null). Cohen's d is interpreted in units of standard deviations, just like z .

For our example:

$$d = \frac{7.75 - 8.00}{0.50} = \frac{-0.25}{0.50} = -.50$$

Cohen's d is interpreted as small, moderate, or large. Specifically, $d = 0.20$ is small, $d = 0.50$ is moderate, and $d = 0.80$ is large. Obviously values can fall in between these guidelines, so we should use our best judgment and the context of the problem to make our final interpretation of size. Our effect size happened to be exactly equal to one of these, so we say that there was a moderate effect.

Effect sizes are incredibly useful and provide important information and clarification that overcomes some of the weakness of hypothesis testing. Whenever you find a significant result, you should always calculate an effect size

| d | Interpretation |
|-----------|-----------------------|
| 0.0 – 0.2 | negligible |
| 0.2 – 0.5 | small |
| 0.5 – 0.8 | medium |
| 0.8 – | large |

Table 1. Interpretation of Cohen's d

Example: Office Temperature

Let's do another example to solidify our understanding. Let's say that the office building you work in is supposed to be kept at 74 degree Fahrenheit but is allowed

to vary by 1 degree in either direction. You suspect that, as a cost saving measure, the temperature was secretly set higher. You set up a formal way to test your hypothesis.

Step 1: State the Hypotheses

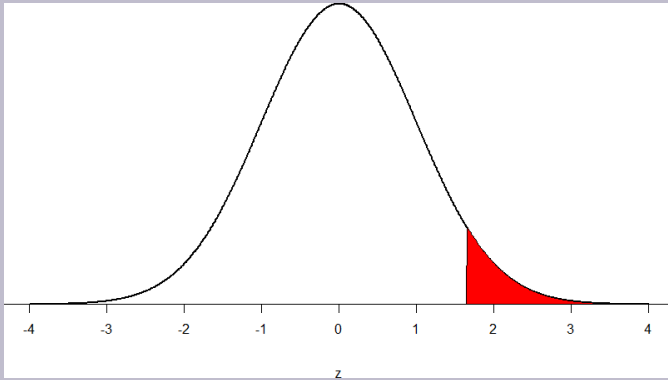
You start by laying out the null hypothesis:

H_0 : There is no difference in the average building temperature $H_0: \mu = 74$

Next you state the alternative hypothesis. You have reason to suspect a specific direction of change, so you make a one-tailed test:

H_A : The average building temperature is higher than claimed $H_A: \mu > 74$

Step 2: Find the Critical Values



You know that the most common level of significance is $\alpha = 0.05$, so you keep that the same and know that the critical value for a one-tailed z-test is $z_{\text{crit}}^* = 1.645$. To keep track of the directionality of the test and rejection region, you draw out your distribution:

Step 3: Calculate the Test Statistic

Now that you have everything set up, you spend one week collecting temperature data:

| Day | Temp |
|-----------|------|
| Monday | 77 |
| Tuesday | 76 |
| Wednesday | 74 |
| Thursday | 78 |
| Friday | 78 |

You calculate the average of these scores to be $\bar{x} = 76.6$ degrees. You use this to calculate the test statistic, using $\mu = 74$ (the supposed average temperature), $\sigma = 1.00$ (how much the temperature should vary), and $n = 5$ (how many data points you collected):

$$z = \frac{76.60 - 74.00}{1.00/\sqrt{5}} = \frac{2.60}{0.45} = 5.78$$

$$1.00/\sqrt{5} \quad 0.45$$

This value falls so far into the tail that it cannot even be plotted on the distribution!

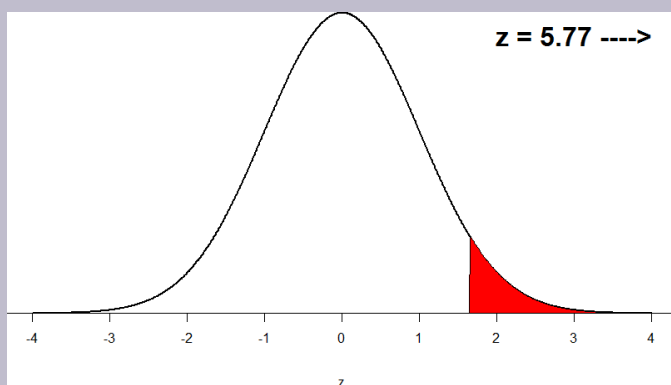


Figure 7: Obtained z-statistic

Step 4: Make the Decision

You compare your obtained z-statistic, $z = 5.77$, to the critical value, $z^* = 1.645$, and find that $z > z^*$. Therefore you reject the null hypothesis, concluding: Based on 5 observations, the average temperature ($\bar{\diamond} = 76.6$ degrees) is statistically significantly higher than it is supposed to be, $z = 5.77$, $p < .05$.

A significant result gives you more confidence to also calculate an effect size:

$$d = (76.60 - 74.00) / 1 = 2.60$$

The effect size you calculate is definitely large, meaning someone has some explaining to do!

Example: Different Significance Level

First, let's take a look at an example phrased in generic terms,

rather than in the context of a specific research question, to see the individual pieces one more time. This time, however, we will use a stricter significance level, $\alpha = 0.01$, to test the hypothesis.

Step 1: State the Hypotheses

We will use 60 as an arbitrary null hypothesis value: H_0 : The average score does not differ from the population $H_0: \mu = 50$

We will assume a two-tailed test: H_A : The average score does differ $H_A: \mu \neq 50$

Step 2: Find the Critical Values

We have seen the critical values for z-tests at $\alpha = 0.05$ levels of significance several times. To find the values for $\alpha = 0.01$, we will go to the standard normal table and find the z-score cutting off 0.005 (0.01 divided by 2 for a two-tailed test) of the area in the tail, which is $z_{crit}^* = \pm 2.575$. Notice that this cutoff is much higher than it was for $\alpha = 0.05$. This is because we need much less of the area in the tail, so we need to go very far out to find the cutoff. As a result, this will require a much larger effect or much larger sample size in order to reject the null hypothesis.

Step 3: Calculate the Test Statistic

We can now calculate our test statistic. The average of 10 scores is $M = 60.40$ with a $\mu = 60$. We will use $\sigma = 10$ as our known population standard deviation. From this information, we calculate our z-statistic as:

$$z = \frac{60.40 - 60.00}{10.00/\sqrt{10}} = \frac{0.40}{3.16} = 0.13$$

Step 4: Make the Decision

Our obtained z-statistic, $z = 0.13$, is very small. It is much less than our critical value of 2.575. Thus, this time, we fail to reject the null hypothesis. Our conclusion would look something like:

Based on the sample of 10 scores, we cannot conclude that there is no effect causing the mean ($M = 60.40$) to be statistically significantly different from 60.00, $z = 0.13$, $p > 0.01$.

Notice two things about the end of the conclusion. First, we wrote that p is greater than instead of p is less than, like we did in the previous two examples. This is because we failed to reject the null hypothesis. We don't know exactly what the p -value is, but we know it must be larger than the α level we used to test our hypothesis. Second, we used 0.01 instead of the usual 0.05, because this time we tested at a different level. The number you compare to the p -value should always be the significance level you test at. Because we did not detect a statistically significant effect, we do not need to calculate an effect size. *Note:* some statisticians will suggest to always calculate effects size as a possibility of Type II error. Although insignificant, calculating $d = (60.4 - 60)/10 = .04$ which suggests no effect (and not a possibility of Type II error).

Review Considerations in Hypothesis Testing

Errors in Hypothesis Testing

Keep in mind that rejecting the null hypothesis is not an all-or-nothing decision. The Type I error rate is affected by the α level: the lower the α level the lower the Type I error rate. It might

seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error. The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a Type II error. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true. A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called β (beta). The probability of correctly rejecting a false null hypothesis equals $1 - \beta$ and is called power.

Statistical power is simply our ability to correctly detect an effect that exists. It is influenced by the size of the effect (larger effects are easier to detect), the significance level we set (making it easier to reject the null makes it easier to detect an effect, but increases the likelihood of a Type I Error), and the sample size used (larger samples make it easier to reject the null).

Statistical Power

The statistical **power** of a research design is the probability of rejecting the null hypothesis given the sample size and expected relationship strength. Statistical power is the

complement of the probability of committing a Type II error. Clearly, researchers should be interested in the power of their research designs if they want to avoid making Type II errors. In particular, they should make sure their research design has adequate power before collecting data. A common guideline is that a power of .80 is adequate. This means that there is an 80% chance of rejecting the null hypothesis for the expected relationship strength.

Given that statistical power depends primarily on relationship strength and sample size, there are essentially two steps you can take to increase statistical power: increase the strength of the relationship or increase the sample size. Increasing the strength of the relationship can sometimes be accomplished by using a stronger manipulation or by more carefully controlling extraneous variables to reduce the amount of noise in the data (e.g., by using a within-subjects design rather than a between-subjects design). The usual strategy, however, is to increase the sample size. For any expected relationship strength, there will always be some sample large enough to achieve adequate power.

Inferential statistics uses data from a sample of individuals to reach conclusions about the whole population. The degree to which our inferences are valid depends upon how we selected the sample (sampling technique) and the characteristics (parameters) of population data. Statistical analyses assume that sample(s) and population(s) meet certain conditions called statistical assumptions.

It is easy to check assumptions when using statistical software and it is important as a researcher to check for violations; if violations of statistical assumptions are not appropriately addressed then results may be interpreted incorrectly.

Learning Objectives

Having read the chapter, students should be able to:

- Conduct a hypothesis test using a z-score statistics, locating critical region, and make a statistical decision including.
- Explain the purpose of measuring effect size and power, and be able to compute Cohen's d.

Exercises – Ch. 10

1. List the main steps for hypothesis testing with the z-statistic. When and why do you calculate an effect size?
2. Determine whether you would reject or fail to reject the null hypothesis in the following situations:
 1. $z = 1.99$, two-tailed test at $\alpha = 0.05$
 2. $z = 1.99$, two-tailed test at $\alpha = 0.01$
 3. $z = 1.99$, one-tailed test at $\alpha = 0.05$
 4. $z = 1.99$, one-tailed test at $\alpha = 0.05$
3. You are part of a trivia team and have tracked your team's performance since you started playing, so you know that your scores are normally distributed with $\mu = 78$ and $\sigma = 12$. Recently, a new person joined the team, and you think the scores have gotten better. Use hypothesis testing to see if the average score has improved based on the following 8 weeks' worth of score data: 82, 74, 62, 68, 79, 94, 90, 81, 80.
4. A study examines self-esteem and depression in teenagers. A sample of 25 teens with a low self-esteem are given the Beck Depression Inventory. The average score for the group is 20.9. For the general population, the average score is 18.3 with $\sigma = 12$. Use a two-tail test with $\alpha = 0.05$ to examine whether teenagers with low self-esteem

show significant differences in depression.

5. You get hired as a server at a local restaurant, and the manager tells you that servers' tips are \$42 on average but vary about \$12 ($\mu = 42$, $\sigma = 12$). You decide to track your tips to see if you make a different amount, but because this is your first job as a server, you don't know if you will make more or less in tips. After working 16 shifts, you find that your average nightly amount is \$44.50 from tips. Test for a difference between this value and the population mean at the $\alpha = 0.05$ level of significance.

Answers to Odd- Numbered Exercises – Ch. 10

1. List hypotheses. Determine critical region. Calculate z . Compare z to critical region. Draw Conclusion. We calculate an effect size when we find a statistically significant result to see if our result is practically meaningful or important

3. Step 1: $H_0: \mu = 78$ "The average score is not different after the new person joined", $H_A: \mu > 78$ "The average score has gone up since the new person joined."

Step 2: One-tailed test to the right, assuming $\alpha = 0.05$, $z^* = 1.645$.

Step 3: $\bar{X} = 78.889$, $s = 12/\sqrt{9} = 4$, $z = .889/4 = .22$

Step 4: $z < z^*$, Fail to Reject H_0 . Based on 8 weeks of games, we can conclude that our average score ($\bar{X} = 78.889$) is not different than the new person is on the team, $z = 0.22$, $p > .05$.

Since the result is not significant, we can still run an effect size: Cohen's $d = .889/12 = .07$, which is no effect.

5. Step 1: $H_0: \mu = 42$ "My average tips does not differ from other servers", $H_A: \mu \neq 42$ "My average tips do differ from others"

Step 2: Two-tailed test to the right, assuming $\alpha = 0.05$, $z^* = \pm 1.96$.

Step 3: $\bar{X} = 44.50$, $\sigma_{\bar{X}} = 12/\sqrt{16} = 3$, $z = .833$.

Step 4: $z > z^*$, Fail to reject H_0 . Based on 16 shifts, we can conclude that average tip ($\bar{X} = 44.50$) is not different from rest of servers, $z = 0.833$, $p > .05$. Since the result is not significant, we don't necessarily need an effect size – but we will check: Cohen's $d = 2.5/12 = .2$, which is actually a small effect. This may indicate that we have a Type II error (or we need more shifts examine for potential differences).

11. Chapter 11:

Introduction to t-tests

In this unit, we made a big leap from basic descriptive statistics into full hypothesis testing and inferential statistics. For the rest of the unit, we will be learning new tests, each of which is just a small adjustment on the test before it. In this chapter, we will learn about the first of three *t*-tests, and we will learn a new method of testing the null hypothesis: confidence intervals.

At this point, we may think we know all about hypothesis testing. Here's a surprise – what we know will not help you much as a researcher. Why? The procedures for testing hypotheses described up to this point were, of course, absolutely necessary for what comes next, but these procedures involved comparing a group of scores to a known population. In real research practice, we often compare two or more groups of scores to each other, without any direct information about populations. For example:

- Comparing the intelligence scores (IQ) of one sample to standardized IQ norms and population values.
- Comparing pre and post-test anxiety scores for a group of patients before and after psychotherapy or number of familiar versus unfamiliar words recalled in a memory experiment.
- Comparing scores on a cognitive test for a group of participants experiencing sleep deprivation and a group of participants who slept normally.
- Comparing scores on self-esteem test scores for a group of 10-year-old girls to a group of 10-year-old boys.

These kinds of research situations are among the most

common in psychology, where the only information available is usually from samples. Nothing is known about the populations that the samples are supposed to come from. In particular, the researcher does not know the variance of the populations involved. In this chapter, we will learn the solution to the problem of the unknown population variance.

The hypothesis-testing procedures we learn in this chapter (and a few others) are examples of t -tests. Its main principles were originally developed by William S. Gosset who published his research articles anonymously using the name “Student”.

William S. Gosset graduated from Oxford University in 1899 with degrees in mathematics and chemistry. It happened that in the same year the Guinness brewers in Dublin, Ireland, were seeking a few young scientists to take a first-ever scientific look at beer making. Gosset took one of these jobs and soon had immersed himself in barley, hops, and vats of brew.



Photo in the public domain.

The problem was how to make a beer of consistently high quality. Scientists want to make a quality beer less variable, and were especially interested in finding the cause of bad batches. But a business such as a brewery could not afford to waste money on experiments involving large numbers of vats. So, Gosset was forced to contemplate the probability of a certain strain of barley producing terrible beer when the experiment could consist of only a few batches of each strain. So, from this Gosset discovered the t distribution and invented the t -test.

Most of his work was done on the back of envelopes, with plenty of minor errors in arithmetic

that had to be worked out later. He published his paper on his “brewery methods” only when editors of scientific journals demanded. At that time, the Guinness brewery did not allow a scientist to publish papers, because more than one Guinness scientist has revealed brewery secrets. To this day, most scientists call the t -distribution “Student’s t ” because Gosset wrote under the anonymous name “Student” so that the brewery would not know about his writing or be identified through his being known to be its employee. Supposedly, the brewery learned of his scientific fame only at his death, when colleagues wanted to honor him.

The t -statistic for one-sample (compared to population mean)

Last chapter, we were introduced to hypothesis testing using the z -statistic for sample means that we learned in Unit 1. This was a useful way to link the material and ease us into the new way to looking at data, but it isn’t a very common test because it relies on knowing the population standard deviation, σ , which is rarely going to be the case. Instead, we will estimate that parameter σ using the sample statistics in the same way that we estimate μ using \bar{X} (μ will still appear in our formulas because we suspect something about its value and that is what we are testing). Our new statistic is called t , and for testing one population mean using a single sample (called a 1-sample t -test) it takes the form:

1-sample t -test:

t is mean differences over the estimated standard error

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

in other words

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Notice that t looks almost identical to z ; this is because they test the exact same thing: the value of a sample mean compared to what we expect of the population. The only difference is that the *standard error* is now denoted $s_{\bar{x}}$ or s_M to indicate that we use the sample statistic for standard deviation, s , instead of the population parameter σ . We call the standard error using s , the estimated standard error of the mean. The process of using and interpreting the standard error and the full test statistic remain exactly the same.

Estimated standard error of the mean $\sigma_{\bar{X}}$ or s_M

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

or equivalent

$$s_{\bar{X}} = \sqrt{\frac{s_X^2}{n}}$$

Note: Formulas can notate as $\sigma_{\bar{X}}$ or s_M . You can calculate estimated standard error using the sample standard deviation (s) or sample variance (s^2).

Setting up for step 2

In order to find the critical boundary for a t -test we must use degrees of freedom. In chapter 4 we learned that the formulae for sample standard deviation and population standard deviation differ by one key factor: the denominator for the parameter is N but the denominator for the statistic is $n - 1$, also known as degrees of freedom, df . As we learned earlier, degrees of freedom gets its name because it is the number of scores in a sample that are “free to vary”. The idea is that when finding the variance we must first know the mean. If we know the mean and all but one of the scores in a sample, we can figure out the one we do not know with a little math. In

this situation the degrees of freedom is the number of scores minus 1.

Because we are using a new measure of spread, we can no longer use the standard normal distribution and the z-table to find our critical values. For *t*-tests, we will use the *t*-distribution and *t*-table to find these values.

The *t*-distribution, like the standard normal distribution, is symmetric and normally distributed with a mean of 0 and standard error (as the measure of standard deviation for sampling distributions) of 1. However, because the calculation of standard error uses degrees of freedom, there will be a different *t*-distribution for every degree of freedom. Luckily, they all work exactly the same, so in practice this difference is minor.

Figure 1 shows four curves: a normal distribution curve labeled *z*, and three *t*-distribution curves for 2, 10, and 30 degrees of freedom. Remember degrees of freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample. Two things should stand out: First, for lower degrees of freedom (e.g., 2), the tails of the distribution are much fatter, meaning a larger proportion of the area under the curve falls in the tail. This means that we will have to go farther out into the tail to cut off the portion corresponding to 5% or $\alpha = 0.05$, which will, in turn, lead to higher critical (rejection) values. Second, as the degrees of freedom increase, we get closer and closer to the *z* curve. Even the distribution with $df = 30$, corresponding to a sample size of just 31 people, is nearly indistinguishable from *z*. In fact, a *t*-distribution with infinite degrees of freedom (theoretically, of course) is exactly the standard normal distribution. Because of this, the bottom row of the *t*-table also includes the critical values for *z*-tests at the specific significance levels. Even though these curves are very close,

it is still important to use the correct table and critical values, because small differences can add up quickly.

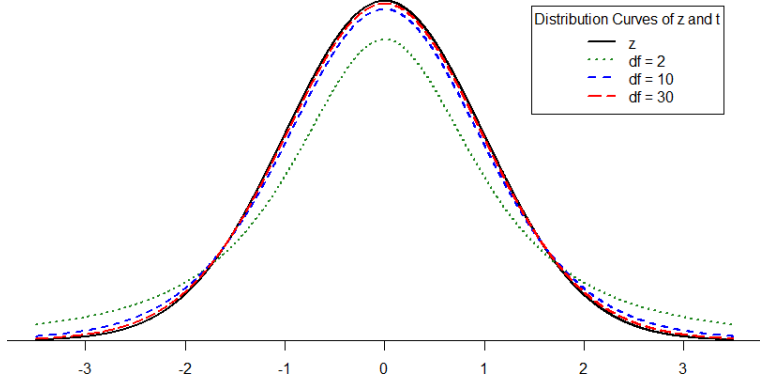


Figure 1. Distributions comparing effects of degrees of freedom

The *t*-distribution table lists critical values for one- and two-tailed tests at several levels of significance arranged into columns. The rows of the *t*-table list degrees of freedom up to $df = 100$ in order to use the appropriate distribution curve. It does not, however, list all possible degrees of freedom in this range, because that would take too many rows. Above $df = 40$, the rows jump in increments of 10. If a problem requires you to find critical values and the exact degrees of freedom is not listed, you always round down to the next smallest number. For example, if you have 48 people in your sample, the degrees of freedom are $n - 1 = 48 - 1 = 47$; *however, 47 doesn't appear on our table, so we round down and use the critical values for $df = 40$, even though 50 is closer.* We do this because it avoids inflating Type I Error (false positives, see chapter 9) by using criteria that are too lax.

Hypothesis Testing with *t*

Hypothesis testing with the *t*-statistic works exactly the same

way as z-tests did, following the four-step process of 1) Stating the Hypothesis, 2) Finding the Critical Values, 3) Computing the Test Statistic, and 4) Making the Decision. Just like the z-statistic, our ultimate goal is to decide whether to reject or fail to reject the null hypothesis.

We will work through an example: let's say that you move to a new city and find an auto shop to change your oil. Your old mechanic did the job in about 30 minutes (though you never paid close enough attention to know how much that varied), and you suspect that your new shop takes much longer. After 4 oil changes, you think you have enough evidence to demonstrate this.

Step 1: State the Hypotheses

Our hypotheses for 1-sample t-tests are identical to those we used for z-tests. We still state the null and alternative hypotheses mathematically in terms of the population parameter and written out in readable English. In the example, the individual hypothesized that the shop took longer (increased time) which is phrased as a directional research hypothesis, corresponding to a 1-tail test.

For our example:

H_0 : There is no difference in the average time to change a car's oil – mathematically can be phrased as, $H_0: \mu = 30$

H_A : This shop takes longer to change oil than your old mechanic – mathematically can be phrased as, $H_A: \mu > 30$

It is important to set up the hypotheses giving context of the problem/study (using words) and mathematically connecting to the population mean (μ) that is reported in the problem/study (or in real life you would look up or calculate from a given data set).

Step 2: Find the Critical Values

As noted above, our critical values still delineate the area in the tails under the curve corresponding to our chosen level of significance. Because we have no reason to change significance levels, we will use $\alpha = 0.05$, and because we suspect a direction of effect, we have a one-tailed test. To find our critical values for t , we need to add one more piece of information: the degrees of freedom.

For this example: $df = n - 1 = 4 - 1 = 3$

Going to our t -table, we find the column corresponding to our one-tailed significance level and find where it intersects with the row for 3 degrees of freedom. As shown in Figure 2: our critical value is $t^* = 2.353$

t-distribution Table

| df | 0.05 0.10 | 0.025 0.05 | 0.01 0.02 | 0.005 0.01 | 1-tailed α 2-tailed α |
|----|--------------|---------------|--------------|---------------|--|
| 1 | 6.314 | 12.706 | 31.821 | 63.657 | |
| 2 | 2.920 | 4.303 | 6.965 | 9.925 | |
| 3 | 2.353 | 3.182 | 4.541 | 5.841 | |
| 4 | 2.132 | 2.776 | 3.747 | 4.604 | |
| 5 | 2.015 | 2.571 | 3.365 | 4.032 | |
| 6 | 1.943 | 2.447 | 3.143 | 3.707 | |

Figure 2. t -table

We can then shade this region on our t -distribution to visualize our critical rejection region

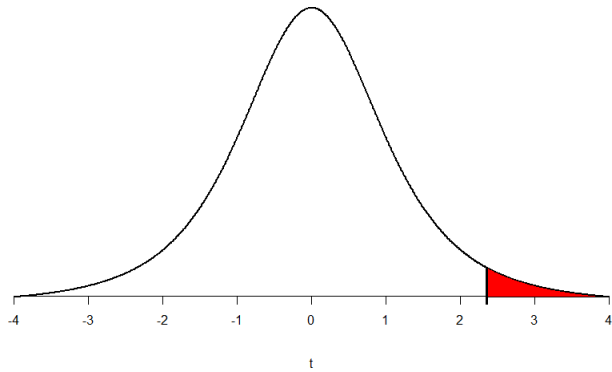


Figure 3. Critical Rejection Region calculated from df and identified using the t-distribution Table.

Step 3: Compute the Test Statistic

The four wait times you experienced for your oil changes at the new shop were 46 minutes, 58 minutes, 40 minutes, and 71 minutes. We will use these to calculate \bar{X} and s by first filling in the sum of squares table in Table 1:

| X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|----------------|---------------|-------------------|
| 46 | -7.75 | 60.06 |
| 58 | 4.25 | 18.06 |
| 40 | -13.75 | 189.06 |
| 71 | 17.25 | 297.56 |
| $\Sigma = 215$ | $\Sigma = 0$ | $\Sigma = 564.74$ |

Table 1. Sum of Squares Table

After filling in the first row to get $\Sigma X = 215$, we find that the mean is $\bar{X} = 53.75$ (215 divided by sample size 4), which allows us to fill in the rest of the table

to get our sum of squares $SS = 564.74$, which we then plug in to the formula for standard deviation from chapter 3:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} = \sqrt{\frac{SS}{df}}$$

Plugging in $SS = 564.74$ and $df = 3$, we get $s = 13.72$. Next, we take this value and plug it in to the formula for standard error:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Plugging in $s = 13.72$ and $\sqrt{n} = \sqrt{4} = 2$, we get $s_{\bar{x}} = 6.86$. And, finally, we put the standard error, sample mean, and null hypothesis value into the formula for our test statistic t :

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

Plugging in $\bar{x} = 53.75$, $\mu = 30$, and $s_{\bar{x}} = 6.86$, we get $23.75 / 6.86 = 3.46$. This may seem like a lot of steps, but it is really just taking our raw data to calculate one value at a time and carrying that value forward into the next equation: data → sample size/degrees of freedom → mean → sum of squares → standard deviation → standard error → test statistic. At each step, we simply match the symbols of what we just calculated to where they appear in the next formula to make sure we are plugging in everything correctly. Also, for this class, you may directly given the standard deviation and mean to work from,

so be sure to identify where you are starting from if not given the data to work through the entire problem.

Step 4: Make the Decision

Now that we have our critical value and test statistic, we can make our decision using the same criteria we used for a z-test. Our obtained t -statistic was $t = 3.46$ and our critical value was $t^* = 2.353$: $t > t^*$, so we reject the null hypothesis and conclude: Based on our four oil changes, the new mechanic takes longer on average ($\bar{X} = 53.75$) to change oil than our old mechanic, $t(3) = 3.46, p < .05$.

Notice that we also include the degrees of freedom in parentheses next to t . We need to calculate an effect size, which is still Cohen's d , but now we use s in place of σ :

Cohen's d

$$d = (M - \mu) / s$$

Example: $d = (53.75 - 30) / 13.72 = 1.73$

For this example, $d = (53.75 - 30) / 13.72 = 1.73$. This is a large effect. It should also be noted that for some things, like the minutes in our current example, we can also interpret the magnitude of the difference we observed (23 minutes and 45 seconds) as an indicator of importance since time is a familiar metric.

Confidence Intervals

Up to this point, we have learned how to estimate the population parameter for the mean using sample data and a sample statistic. From one point of view, this makes sense: we have one value for our parameter so we use a single value (called a point estimate) to estimate it. However, we have seen that all statistics have sampling error and that the value we find for the sample mean will bounce around based on the people in our sample, simply due to random chance. Thinking about estimation from this perspective, it would make more sense to take that error into account rather than relying just on our point estimate. To do this, we calculate what is known as a *confidence interval*.

A confidence interval starts with our point estimate then creates a range of scores considered plausible based on our standard deviation, our sample size, and the level of confidence with which we would like to estimate the parameter. This range, which extends equally in both directions away from the point estimate, is called the margin of error. We calculate the margin of error by multiplying our two-tailed critical value by our standard error:

Margin of Error (MOE)

$$t \frac{s}{\sqrt{n}}$$

Margin of Error = $t \frac{s}{\sqrt{n}}$ where
s = standard deviation and *t* is the critical value

t. The MOE is the t-critical value times the estimated standard error, $S \sqrt{\frac{1}{n}}$.

One important consideration when calculating the margin of error is that it can only be calculated using the critical value for a two-tailed test. This is because the margin of error moves away from the point estimate in both directions, so a one-tailed value does not make sense.

The critical value we use will be based on a chosen level of confidence, which is equal to $1 - \alpha$. Thus, a 95% level of confidence corresponds to $\alpha = 0.05$. Thus, at the 0.05 level of significance, we create a 95% Confidence Interval. How to interpret that is discussed further on.

Once we have our margin of error calculated, we add it to our point estimate for the mean to get an upper bound to the confidence interval and subtract it from the point estimate for the mean to get a lower bound for the confidence interval:

Confidence Intervals (CI)

$$\bar{X} \pm t \frac{s}{\sqrt{n}}$$

CI =

where

$$t \frac{s}{\sqrt{n}}$$

is the Margin of Error (MOE)

You will calculate an upper bound value and a lower bound value.

Upper bound (UB)= use the + in the CI formula = $\bar{X} + \text{MOE}$

Lower bound (LB) = use the - in the CI formula = $\bar{X} - \text{MOE}$

To write out a confidence interval, we always use soft brackets and put the lower bound, a comma, and the upper bound:

Confidence Interval = (LB value, UB value)

Let's see what this looks like with some actual numbers by taking our oil change data and using it to create a 95% confidence interval estimating the average length of time it takes at the new mechanic. We already found that our average was $\bar{X} = 53.75$ and our standard error was $s_{\bar{X}} = 6.86$.

We also found a critical value to test our hypothesis, but remember that we were testing a one-tailed hypothesis, so that critical value won't work. To see why that is, look at the column headers on the t -table. The column for one-tailed $\alpha = 0.05$ is the same as a two-tailed $\alpha = 0.10$. If we used the old critical value, we'd actually be creating a 90% confidence interval ($1.00 - 0.10 = 0.90$, or 90%). To find the correct value, we use the column for two-tailed $\alpha = 0.05$ and, again, the row for 3 degrees of freedom, to find $t^* = 3.182$.

Now we have all the pieces we need to construct our confidence interval:

$$95\% \text{ CI} = 53.75 \pm 3.182(6.86)$$

$$\text{upper bound} = 53.75 + 3.182(6.86)$$

$$\text{UB} = 53.75 + 21.83 = 75.58$$

$$\text{lower bound} = 53.75 - 3.182(6.86)$$

$$\text{LB} = 53.75 - 21.83 = 31.92$$

$$95\% \text{ CI} = (31.92, 75.58)$$

So we find that our 95% confidence interval runs from 31.92 minutes to 75.58 minutes, but what does that actually mean? The range (31.92, 75.58) represents values of the mean that we consider reasonable or plausible based on our observed data. It includes our point estimate of the mean, $\bar{X} = 53.75$, in the center, but it also has a range of values that could also have been the case based on what we know about how much these scores vary (i.e. our standard error).

It is very tempting to also interpret this interval by saying that we are 95% confident that the true population mean falls within the range (31.92, 75.58), *but this is not true*. The reason it is not true is that phrasing our interpretation this way suggests that we have firmly established an interval and the population mean does or does not fall into it, suggesting that our interval is firm and the population mean will move around. However, the population mean is an absolute that does not change; it is our interval that will vary from data collection to

data collection, even taking into account our standard error. The correct interpretation, then, is that we are 95% confident that the range (31.92, 75.58) brackets the true population mean. This is a very subtle difference, but it is an important one.

Interpreting Confidence Intervals

Confidence intervals are notoriously confusing, primarily because they don't mean what we might intuitively think they mean. If I tell you that I have computed a "95% confidence interval" for my statistic, then it would seem natural to think that we can have 95% confidence that the true parameter value falls within this interval. However, as we will see throughout the course, concepts in statistics often don't mean what we think they should mean. In the case of confidence intervals, we can't interpret them in this way because the population parameter has a fixed value – it either is or isn't in the interval, so it doesn't make sense to talk about the probability of that occurring. Jerzy Neyman, the inventor of the confidence interval, said:

"The parameter is an unknown constant and no probability statement concerning its value may be made."(Neyman [1937](#))

Instead, we have to view the confidence interval procedure from the same standpoint that we viewed hypothesis testing: As a procedure that in the long run will allow us to make correct statements with a particular probability. Thus, the

proper interpretation of the 95% confidence interval is that it is an interval that will contain the true population mean 95% of the time.

Hypothesis Testing with Confidence Intervals

As a function of how they are constructed, we can also use confidence intervals to test hypotheses. However, we are limited to testing two-tailed hypotheses only for this, because of how the intervals work, as discussed above.

Once a confidence interval has been constructed, using it to test a hypothesis is simple.

- The range of the confidence interval brackets (or contains, or is around) the null hypothesis value, we fail to reject the null hypothesis.
- If it does not bracket the null hypothesis value (i.e. if the entire range is above the null hypothesis value or below it), we reject the null hypothesis.

The reason for this is clear if we think about what a confidence interval represents. Remember: a confidence interval is a range of values that we consider reasonable or plausible based on our data. Thus, if the null hypothesis value is in that range, then it is a value that is plausible based on our observations. If the null hypothesis is plausible, then we have no reason to reject it. Thus, if our confidence interval brackets the null hypothesis value, thereby making it a reasonable or plausible value based on our observed data, then we have no evidence against the null hypothesis and fail to reject it. However, if we build a confidence interval of reasonable values based on our

observations and it does not contain the null hypothesis value, then we have no empirical (observed) reason to believe the null hypothesis value and therefore reject the null hypothesis.

Let's see an example of hypothesis testing using Confidence Intervals

You hear that the national average on a measure of friendliness is 38 points. You want to know if people in your community are more or less friendly than people nationwide, so you collect data from 30 random people in town to look for a difference. *We'll follow the same four-step hypothesis testing procedure as before.*

Step 1: State the Hypotheses

We will start by laying out our null and alternative hypotheses:

H_0 : There is no difference in how friendly the local community is compared to the national average; $H_0: \mu = 38$

H_A : There is a difference in how friendly the local community is compared to the national average; $H_A: \mu \neq 38$

Remember we must use a 2-tail test for this, so your hypotheses would be non-directional.

Step 2: Find the Critical Values

We need our critical values in order to determine the width of our margin of error.

We will assume a significance level of $\alpha = 0.05$ (which will give us a 95% CI).

From the *t*-table, a two-tailed critical value at $\alpha = 0.05$ with 29 degrees of freedom ($n - 1 = 30 - 1 = 29$) is $t^* = 2.045$.

Step 3: Calculations

Now we can construct our confidence interval. After we collect our data, we find that the average person in our community scored 39.85, or $\bar{X} = 39.85$, and our standard deviation was $s = 5.61$. First, we need to use this standard deviation, plus our sample size of $n = 30$, to calculate our standard error:

$$s_{\bar{X}} = s/\sqrt{n} = 5.61/\sqrt{30} = 1.02$$

Now we can put that value (1.02 as $S_{\bar{X}}$), our point estimate for the sample mean (39.85), and our critical *t*-value from step 2 (2.045) into the formula for a confidence interval:

$$95\% \text{ CI} = 39.85 \pm 2.045(1.02)$$

$$\text{UB} = 39.85 + 2.045(1.02) = 39.85 + 2.09 = 41.94$$

$$\text{LB} = 39.85 - 2.045(1.02) = 39.85 - 2.09 = 37.76$$

$$95\% \text{ CI} = (37.76, 41.94)$$

Step 4: Make the Decision

Finally, we can compare our confidence interval to our null hypothesis value. The null value of 38 is higher than our lower bound of 37.76 and lower than our upper bound of 41.94. Thus, the confidence

interval brackets our null hypothesis value, and we fail to reject the null hypothesis:

Fail to Reject H_0 . Based on our sample of 30 people, our community not different in average friendliness ($\bar{X} = 39.85$) than the nation as a whole, 95% CI = (37.76, 41.94).

Note that we don't report a test statistic or p-value because that is not how we tested the hypothesis, but we do report the value we found for our confidence interval.

An important characteristic of hypothesis testing is that both methods will always give you the same result. That is because both are based on the standard error and critical values in their calculations. To check this, we can calculate a t-statistic for the example above and find it to be $t = 1.81$, which is smaller than our critical value of 2.045 and fails to reject the null hypothesis.

Although we failed to reject the null, we can calculate Cohen's $d = (\bar{X} - \mu)/s = (39.85 - 38)/5.61 = .33$. $d = .33$ is a small effect size suggesting that we may have a Type II error.

Note: we can also calculate confidence intervals for z .

Confidence intervals can also be constructed using z-score criteria, if one knows the population standard deviation. The format, calculations, and interpretation are all exactly the same, only replacing t^* with z^* and $s\sqrt{n}$ with $\sigma\sqrt{n}$.

Exercises – Ch. 11

1. What is the difference between a z-test and a 1-sample t-test?
2. What does a confidence interval represent?
3. What is the relationship between a chosen level of confidence for a confidence interval and how wide that interval is? For instance, if you move from a 95% CI to a 90% CI, what happens? Hint: look at the t-table to see how critical values change when you change levels of significance.
4. Construct a confidence interval around the sample mean $\bar{X} = 25$ for the following conditions:
 1. $n = 25, s = 15, 95\%$ confidence level
 2. $n = 25, s = 15, 90\%$ confidence level
 3. $S_{\bar{X}} = 4.5, \alpha = 0.05, df = 20$
 4. $s = 12, df = 16$ (yes, that is all the information you need)
5. True or False: a confidence interval represents the most likely location of the true population mean.
6. You hear that college campuses may differ from the general population in terms of political affiliation, and you want to use hypothesis testing to see if this is true and, if so, how big the difference is. You know that the average political affiliation in the nation is $\mu = 4.00$ on a scale of 1.00 to 7.00, so you gather data from 150 college students across the nation to see if there is a difference. You find that the average score is 3.76 with a standard deviation of 1.52. Use a 1-sample t-test to see if there is a difference at

the $\alpha = 0.05$ level.

7. You hear a lot of talk about increasing global temperature, so you decide to see for yourself if there has been an actual change in recent years. You know that the average land temperature from 1951-1980 was 8.79 degrees Celsius. You find annual average temperature data from 1981-2017 and decide to construct a 99% confidence interval (because you want to be as sure as possible and look for differences in both directions, not just one) using this data to test for a difference from the previous average.
8. Determine whether you would reject or fail to reject the null hypothesis in the following situations:
 1. $t = 2.58$, $N = 21$, two-tailed test at $\alpha = 0.05$
 2. $t = 1.99$, $N = 49$, one-tailed test at $\alpha = 0.01$
 3. $\mu = 47.82$, 99% CI = (48.71, 49.28)
 4. $\mu = 0$, 95% CI = (-0.15, 0.20)
9. You are curious about how people feel about craft beer, so you gather data from 55 people in the city on whether or not they like it. You code your data so that 0 is neutral, positive scores indicate liking craft beer, and negative scores indicate disliking craft beer. You find that the average opinion was $\bar{X} = 1.10$ and the spread was $s = 0.40$, and you test for a difference from 0 at the $\alpha = 0.05$ level.
10. You want to know if college students have more stress in their daily lives than the general population ($\mu = 12$), so you gather data from 25 people to test your hypothesis. Your sample has an average stress score of $\bar{X} = 13.11$ and a standard deviation of $s = 3.89$. Use a 1-sample t -test to see if there is a difference.

Answers to Odd- Numbered Exercises – Ch. 11

1. A z-test uses population standard deviation for calculating standard error and gets critical values based on the standard normal distribution. A *t*-test uses sample standard deviation as an estimate when calculating standard error and gets critical values from the *t*-distribution based on degrees of freedom.

3. As the level of confidence gets higher, the interval gets wider. In order to speak with more confidence about having found the population mean, you need to cast a wider net. This happens because critical values for higher confidence levels are larger, which creates a wider margin of error.

5. False: a confidence interval is a range of plausible scores that may or may not bracket the true population mean.

7. $\bar{X} = 9.44$, $s = 0.35$, $s_{\bar{X}} = 0.06$, $df = 36$, $t^* = 2.719$, 99% CI = (9.28, 9.60); CI does not bracket μ , reject null hypothesis. $d = 1.83$

9. Step 1: $H_0: \mu = 0$ "The average person has a neutral opinion towards craft beer", $H_A: \mu \neq 0$ "Overall people will have an opinion about craft beer, either good or bad."

Step 2: Two-tailed test, $df = 54$, $t^* = 2.009$.

Step 3: $\bar{X} = 1.10$, $s_{\bar{X}} = 0.05$, $t = 22.00$.

Step 4: $t > t^*$, Reject H_0 . Based on opinions from 55 people, we can conclude that the average opinion of craft beer ($\bar{X} = 1.10$) is positive, $t(54) = 22.00$, $p < .05$. Since the result is significant, we need an effect size: Cohen's $d = 2.75$, which is a large effect.

12. Chapter 12:

Repeated Measures t-test

So far, we have dealt with data measured on a single variable at a single point in time, allowing us to gain an understanding of the logic and process behind statistics and hypothesis testing. Now, we will look at a slightly different type of data that has new information we couldn't get at before: change. Specifically, we will look at how the value of a variable, *within people*, changes across two timepoints. This is a very powerful thing to do, and, as we will see shortly, it involves only a very slight addition to our existing process and does not change the mechanics of hypothesis testing or formulas at all!

Change and Differences

Researchers are often interested in change over time. Sometimes we want to see if change occurs naturally, and other times we are hoping for change in response to some manipulation. In each of these cases, we measure a single variable at different times, and what we are looking for is whether or not we get the same score at time 2 as we did at time 1. This is a **repeated sample research design**, where a single group of individuals is obtained and each individual is measured in two treatment conditions that are then compared. Data consist of two scores for each individual. This means that all subjects participate in each treatment condition. Think about it like a pretest/posttest.

When we analyze data for a repeated research design, we calculate the difference between members of each pair of scores and then take the average of those differences. *The absolute value of our measurements does not matter – all that matters is the change.* If the average difference between scores in our sample is very large, compared to the difference between scores we would expect if the member was selected from the same population then we will conclude that the individuals were selected from different populations.

Let's look at an example:

| Before | After | Improvement |
|--------|-------|-------------|
| 6 | 9 | 3 |
| 7 | 7 | 0 |
| 4 | 10 | 6 |
| 1 | 3 | 2 |
| 8 | 10 | 2 |

Table 1. Raw and difference scores before and after training. Table 1 shows scores on a quiz that five employees received before they took a training course and after they took the course. The difference between these scores (i.e. the score after minus the score before) represents improvement in the employees' ability. This third column is what we look at when assessing whether or not our training was effective. We want to see positive scores, which indicate that the employees' performance went up. What we are not interested in is how good they were before they took the training or after the training. Notice that the lowest scoring employee before the training (with a score of 1) improved just as much as the highest scoring employee before the training (with a score of 8), regardless of how far apart they were to begin with. There's also one improvement score of 0, meaning that the training

did not help this employee. An important factor in this is that the participants received the same assessment at both time points. To calculate improvement or any other difference score, we must measure only a single variable.

When looking at change scores like the ones in Table 2, we calculate our difference scores by taking the time 2 score and subtracting the time 1 score. That is:

The difference score formula:

$$X_D = X_{T1} - X_{T2}$$

Note: T2 is the time 2 variable; T1 is the time 1 variable

Where X_D is the **difference score**, X_{T1} is the score on the variable at time 1, and X_{T2} is the score on the variable at time 2. The difference score, X_D (can also be noted as D for difference score), will be the data we use to test for improvement or change. Whether a difference score is positive or negative depends on the direction of change; it does not denote big or small, good or bad. The sign of the difference score (X_D or D) denotes the direction of the change.

We subtract time 2 minus time 1 for ease of interpretation; if scores get better, then the difference score will be positive. Similarly, if we're measuring something like reaction time or

depression symptoms that we are trying to reduce, then better outcomes (lower scores) will yield negative difference scores.

We can also test to see if people who are matched or paired in some way agree on a specific topic. We call this a **matched design**. For example, we can see if a parent and a child agree on the quality of home life, or we can see if two romantic partners agree on how serious and committed their relationship is. In these situations, we also subtract one score from the other to get a difference score. This time, however, it doesn't matter which score we subtract from the other because what we are concerned with is the agreement.

In both of these types of data, what we have are multiple scores on a single variable. That is, a single observation or data point is comprised of two measurements that are put together into one difference score. This is what makes the analysis of change unique – our ability to link these measurements in a meaningful way. This type of analysis would not work if we had two separate samples of people that weren't related at the individual level, such as samples of people from different states that we gathered independently. Such datasets and analyses are the subject of the following chapter.

A rose by any other name...

It is important to point out that this form of *t*-test has been called many different things by many different people over the years: “matched pairs”, “paired samples”, “repeated measures”, “dependent measures”, “dependent samples”, and many others. What all of these names have in common is that they describe the analysis of two scores that are related in a systematic way within people or within pairs, which is what each of the datasets usable in this analysis have in common. As

such, all of these names are equally appropriate, and the choice of which one to use comes down to preference. In this text, we will refer to *paired samples*, though the appearance of any of the other names throughout this chapter should not be taken to refer to a different analysis: they are all the same thing.

We are still working with t-tests. In chapter 11, we compared a sample to a population mean. For t-tests in this chapter, we are comparing 2 groups of scores, yet both are from the same individuals. We call this a dependent t-test or a paired t-test. Think of it like you are having 2 cups of tea.



2 cups of tea for me: for a repeated measures design the same individuals are in both conditions for a t-test. [Photo credit](#)

Now that we have an understanding of what difference scores are and know how to calculate them, we can use them to test hypotheses. As we will see, this works exactly the same way as testing hypotheses about one sample mean with a t -statistic. The only difference is in the format of the null and alternative hypotheses, where for focus on the difference score.

Hypotheses of Change and Differences for step 1

When we work with difference scores, our research questions have to do with change. Did scores improve? Did symptoms get better? Did prevalence go up or down? Our hypotheses will reflect this. Remember that the null hypothesis is the idea that there is nothing interesting, notable, or impactful represented in our dataset. In a paired samples t -test, that takes the form of 'no change'. There is no improvement in scores or decrease in symptoms.

Thus, our null hypothesis is: H_0 : There is no change or difference $H_0: \mu_D = 0$

Let's be clear, $H_0: \mu_D = 0$ does not say that everyone in the population will stay the same it only says that on average, the entire population will show a mean difference of 0. As with our other null hypotheses, we express the null hypothesis for paired samples t -tests in both words and mathematical notation. The exact wording of the written-out version should be changed to match whatever research question we are addressing (e.g. "There is no change in ability scores after training"). However, the mathematical version of the null hypothesis is always exactly the same: the average change score is equal to zero. Our population parameter for the average is still μ , but it now has a subscript D to denote the fact that it is the average change score and not the average raw observation before or after our manipulation. Obviously individual difference scores can go up or down, but the null

hypothesis states that these positive or negative change values are just random chance and that the true average change score across all people is 0.

Our alternative hypotheses will also follow the same format that they did before: they can be directional if we suspect a change or difference in a specific direction, or we can use an inequality sign to test for any change:

H_A : There is a change or difference $H_A: \mu_D \neq 0$

H_A : The average score increases $H_A: \mu_D > 0$

H_A : The average score decreases $H_A: \mu_D < 0$

Just as before, your choice of which alternative hypothesis to use should be specified before you collect data based on your research question and any evidence you might have that would indicate a specific directional (or non-directional) change. Additionally, it should be noted that a non-directional research/alternative hypothesis is a more conservative approach when you have an expected direction for change.

Choosing 1-tail vs 2-tail test

How do you choose whether to use a one-tailed versus a two-tailed test? The two-tailed test is always going to be more conservative, so it's always a good bet to use that one, unless you had a very strong prior reason for using a one-tailed test. In that case, you should have written down the hypothesis before you ever looked at the data. In Chapter 19, we will discuss the idea of pre-registration of hypotheses, which formalizes the idea of writing down your hypotheses before you ever see the actual data. You should *never* make a

decision about how to perform a hypothesis test once you have looked at the data, as this can introduce serious bias into the results.

We do have to make one main assumption when we use the randomization test, which we refer to as *exchangeability*. This means that all of the observations are distributed in the same way, such that we can interchange them without changing the overall distribution. The main place where this can break down is when there are related observations in the data; for example, if we had data from individuals in 4 different families, then we couldn't assume that individuals were exchangeable, because siblings would be closer to each other than they are to individuals from other families. In general, if the data were obtained by random sampling, then the assumption of exchangeability should hold.

Critical Values and Decision Criteria for step 2

As with before, once we have our hypotheses laid out, we need to find our critical values that will serve as our decision criteria. This step has not changed at all from the last chapter. Our critical values are based on our level of significance (still usually $\alpha = 0.05$), the directionality of our test (one-tailed or two-tailed), and the degrees of freedom, which are still calculated as $df = n - 1$. Because this is a *t*-test like the last chapter, we will find our critical values on the same *t*-table using the same process of

identifying the correct column based on our significance level and directionality and the correct row based on our degrees of freedom or the next lowest value if our exact degrees of freedom are not presented. After we calculate our test statistic, our decision criteria are the same as well: $p < \alpha$ or $t_{\text{obt}} > t_{\text{crit}}^*$.

Test Statistic for step 3

Our test statistic for our change scores follows exactly the same format as it did for our 1-sample t -test. In fact, the only difference is in the data that we use. For our change test, we first calculate a difference score as shown above. Then, we use those scores as the raw data in the same mean calculation, standard error formula, and t -statistic. Let's look at each of these.

Mean Difference (top of t -formula):

$$\bar{D} = \frac{\sum D}{n}$$

which can also be

$$M_D = \frac{\sum D}{n}$$

noted as

The mean difference score is calculated in the same way as any

other mean: sum each of the individual difference scores and divide by the sample size.

Here we are using the subscript D to keep track of that fact that these are difference scores instead of raw scores; it has no actual effect on our calculation.

Using this, we calculate the standard deviation of the difference scores the same way as well:

Standard deviation for D (S_D) and variance for D is s_D^2 :

$$S_D = \sqrt{\frac{\sum (XD - MD)^2}{n-1}} = \sqrt{\frac{SS}{df}}$$

$$S_D^2 = \frac{\sum (XD - MD)^2}{n-1} = \frac{SS}{df}$$

or may see S_D noted as

$$SD = \sqrt{\frac{\sum (D - \bar{D})^2}{n-1}}$$

where $x_D = D$ &

$$D^- = M_D$$

$$\sqrt{s_D^2}$$

Note: $s_D^2 = s_D * s_D$ and $s_D =$

We will find the numerator, the Sum of Squares, using the same table format that we learned in chapter 3. Once we have our standard deviation, we can find the standard error:

Standard Error

Standard error of the mean differences (S_{MD}) (bottom of t-formula):

$$S_{MD} = \sqrt{\frac{s^2}{n}}$$

which can also

$$s\bar{D} = \frac{s_D}{\sqrt{n}}$$

be noted as

Note: the formula can also be noted as S_{MD} or s_{D^-} and you can calculate it from the variance ($\sqrt{s^2/n}$) or standard deviation (s/\sqrt{n})

Finally, our test statistic t has the same structure as well:

t-test for paired samples:

$$t = \frac{\bar{D} - \mu(\text{hyp})}{s\bar{D}}$$

where

$\mu(\text{hyp})$ is expected to be 0 and is dropped from the

$$t = \frac{\bar{D}}{s\bar{D}}$$

calculation formula leaving

or

$$t = \frac{M_D}{S_{M_D}}$$

Note:

Both formulas are the same with the mean noted as M_D or \bar{D} and the estimated standard error notes as S_{M_D} or $s\bar{D}$

Effect size:

Cohen's d

There are several different ways that the effect size can be quantified, which depend on the nature of the data. One of the most common measures of effect size is known as *Cohen's d*

$$d = \frac{M_D}{\sqrt{s^2}} = \frac{M_D}{s}$$

Note: M_D is the mean of the difference scores.

Another way to examine effect size is to report the explained variance for the treatment effect, in other words the percentage of variance accounted for the treatment. This is known as r^2 .

$$r^2 = \frac{t^2}{t^2 + df}$$

Note: r^2 is calculated when there is a reported effect (in other words, null is rejected). df is the same df from step 2.

As we can see, once we calculate our difference scores from our raw measurements, everything else is exactly the same. Let's see an example.

Example: Increasing Satisfaction at Work

Workers at a local company have been complaining that working conditions have gotten very poor, hours are too long, and they don't feel supported by the management. The company hires a consultant to come in and help fix the situation before it gets so bad that the employees start to quit. The consultant first assesses 49 of the employee's level of job satisfaction as part of focus groups used to identify specific changes that might help. The company institutes some of these changes, and six months later the consultant returns to measure job satisfaction again. Knowing that some interventions miss the mark and can actually make things worse, the consultant tests for a difference in either direction (i.e. and increase or a decreased in average job satisfaction) at the $\alpha = 0.05$ level of significance.

Step 1: State the Hypotheses

In this case, we are hoping that the changes we made will improve employee satisfaction, and, because we based the changes on employee recommendations, we have good reason to believe that they will. However we will take a conservative approach and will use a two-tail alternative hypothesis.

Thus, we state our null and alternative hypotheses as

H_0 : There is no change in average job satisfaction $H_0: \mu_D = 0$

H_A : There is a change in average job satisfaction $H_A: \mu_D \neq 0$

Step 2: Find critical value

Our critical values will once again be based on our level of significance, which we know is $\alpha = 0.05$, the directionality of our test, which is two-tailed, and our degrees of freedom. For our dependent-samples t -test, the degrees of freedom are still given as $df = n - 1$. For this problem, we have 49 people, so our degrees of freedom are 48. Our table does not have 48, so we go with the closest lower value (40). Going to our t -table, we find that the critical value is $t^* = 2.021$. As shown in Figure 1, the cut off or critical value helps with decision making in step 4.

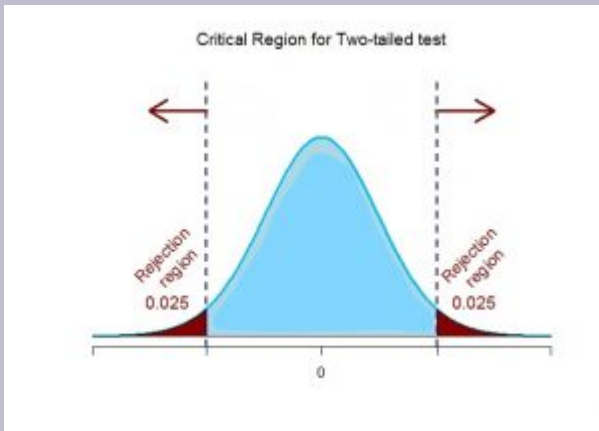


Figure 1. Critical region for two-tailed t -test at $\alpha = 0.05$

Step 3: Calculate test statistic

Now that the criteria are set, it is time to calculate the test statistic. The data obtained by the consultant found that the difference scores from time 1 to time 2 had a mean of M_D or $\bar{D} = 2.96$ and a standard deviation of $s_D = 2.85$. Using this information, plus the size of the sample ($n = 49$), we first calculate the standard error:

$$s_{\bar{D}} = \frac{SD}{\sqrt{n}}$$

Plugging in the values we get
 $2.85/(\sqrt{49}) = 2.85/7 = 0.41$

Now, we can put that value $s_{\bar{D}} = 0.41$, along with our sample mean (2.96), into the formula for t and calculate the test statistic:

$$t = \frac{\bar{D} - 0}{s_{\bar{D}}} = 2.96/0.41 = 7.22$$

Notice that, because the null hypothesis value of a dependent samples t -test is always 0, we can simply divide our obtained sample mean by the standard error.

Step 4: Make a decision

We have obtained a test statistic of $t = 7.22$ that we can compare to our previously established critical value of $t^* = 2.021$. 7.22 is larger than 2.021, so $t > t^*$ and we reject the null hypothesis:

Reject H_0 . Based on the sample data from 49 workers, we can say that the intervention statistically significantly improved job satisfaction ($\bar{D} = 2.96$) among the workers, $t(48) = 7.22, p < 0.05$.

Because this result was statistically significant, we will want to calculate Cohen's d as an effect size using the same format as we did for the last t -test:

$$d = \frac{M_D}{\sqrt{s^2}} = \frac{M_D}{s} \quad \text{where the } M_D \text{ or } \bar{D} = 2.96$$

and a standard deviation of $s = 2.85$. Plugging in the values we get $d = 2.96/2.85 = 1.04$ which is a large effect size. We could also calculate r^2 for effect size.

$$r^2 = \frac{t^2}{t^2 + df}$$

where $t^2 = 7.22 * 7.22 = 52.13$ and $df = 48$. Then plugging in, $r^2 = 52.13 / (52.13 + 48) = .52$. This can be interpreted as 52% of the variance in worker job satisfaction is due to changes the company made.

Hopefully the above example made it clear that running a dependent samples t -test to look for differences before and after some treatment works exactly the same way as a regular 1-sample t -test does from chapter 11 (which was just a small change in how z -tests were performed in chapter 10). At this point, this process should feel familiar, and we will continue to make small adjustments to this familiar process as we encounter new types of data to test new types of research questions.

Confidence Intervals

$$\bar{X} \pm t \frac{s}{\sqrt{n}}$$

Last chapter, $CI = \bar{X} \pm t \frac{s}{\sqrt{n}}$ but now the mean is the mean difference (\bar{D} or M_D) and s/\sqrt{n} becomes $s_{\bar{D}}$

Our adjusted

CI formula for paired or dependent t-test:

$$\bar{D} \pm t(s_{\bar{D}})$$

CI =

Note: We still calculate an upper bound and lower bound value and t is still the critical value t . CI formula is very similar using the notations for standard error. CI still notated as CI = (LB, UB).

Example with Confidence Interval Hypothesis Testing: Bad Press

Let's say that a bank wants to make sure that their new commercial will make them look good to the public, so they recruit 7 people to view the commercial as a focus group. The focus group members fill out a short questionnaire about how they view the company, then watch the commercial and fill out the same questionnaire a second time. The bank really wants to find significant results, so they test for a change at $\alpha = 0.05$. However, they use a 2-tailed test since they know that past commercials have not gone over well with the public, and they want to make sure the new one does not backfire. They decide to test their hypothesis using a confidence interval to see just how spread out the opinions are. As we will see, confidence intervals work the same way as they did before, just like with the test statistic.

Step 1: State the Hypotheses

As always, we start with hypotheses, and with confidence interval hypothesis test, we must use a 2-tail test.

H₀: There is no change in how people view the bank

H₀: $\mu_D = 0$

H_A: There is a change in how people view the bank

H_A: $\mu_D \neq 0$

Step 2: Find the Critical Values

Just like with our regular hypothesis testing procedure, we will need critical values from the appropriate level of significance and degrees of freedom in order to form our confidence interval. Because we have 7 participants, our degrees of freedom are $df = 6$. From our t -table, we find that the critical value corresponding to this df at this level of significance is $t^* = 2.447$.

Step 3: Calculate the Confidence Interval

The data collected before (time 1) and after (time 2) the participants viewed the commercial is presented in Table 1. In order to build our confidence interval, we will first have to calculate the mean and standard deviation of the difference scores, which are also in Table 1. As a reminder, the difference scores (D^- or M_D) are calculated as Time 2 – Time 1.

| Time 1 | Time 2 | D^- |
|-----------|-----------|-------|
| 3 | 2 | -1 |
| 3 | 6 | 3 |
| 5 | 3 | -2 |
| 8 | 4 | -4 |
| 3 | 9 | 6 |
| 1 | 2 | 1 |
| 4 | 5 | 1 |

Table 1. Opinions of the bank

The mean of the difference scores is: $D^- = 4/7 = .57$

The standard deviation will be solved by first using the Sum of Squares Table:

| D | $D - D^-$ | $(D - D^-)^2$ |
|--------------|--------------|-----------------------|
| -1 | -1.57 | 2.46 |
| 3 | 2.43 | 5.90 |
| -2 | -2.57 | 6.60 |
| -4 | -4.57 | 20.88 |
| 6 | 5.43 | 29.48 |
| 1 | 0.43 | 0.18 |
| 1 | 0.43 | 0.18 |
| $\Sigma = 4$ | $\Sigma = 0$ | $\Sigma = 65.68$ (our |

$s = \sqrt{SS/df}$ where $SS = 65.68$ and $df = n-1 = 7-1 = 6$

$s = \sqrt{65.68/6} = \sqrt{10.94} = 3.308$

Finally, we find the standard error (s_{D^-}) taking $s = 3.308$ and $n = 7$.

$s_{D^-} = 3.308/\sqrt{7} = 1.25$

We now have all the pieces needed to compute our confidence interval:

95% CI = $D^- \pm t(s_{D^-})$

Upper Bound (UB) = $0.57 + 1.943(1.25) = 0.57 + 2.43 = 3.00$

Lower Bound (LB) = $0.57 - 1.943(1.25) = 0.57 - 2.43 = -1.86$

95% CI = (LB, UB) = (-1.86, 3.00)

Step 4: Make the Decision

Remember that the confidence interval represents a range of values that seem plausible or reasonable based on our observed data. The interval spans -1.86 to 3.00, which includes 0, our null hypothesis value. Because the null hypothesis value is in the interval, it is considered a reasonable value, and because it is a reasonable value, we have no evidence against it. We fail to reject the null hypothesis.

Fail to Reject H_0 . Based on our focus group of 7 people, we cannot say that the average change in opinion ($D^- = 0.57$) was any better or worse after viewing the commercial, CI: (-1.86, 3.00).

It is optional to calculate effect size. Performing Cohen's $d = D^-/s = .57/3.308 = .17$ which indicates a possible Type II error (very small sample size). As with before, we only report the confidence interval to indicate how we performed the test.

Assumptions of paired t-test

Assumptions are conditions that must be met in order for our hypothesis testing conclusion to be valid. [Important: If the assumptions are *not* met then our hypothesis testing

conclusion is not likely to be valid. Testing errors can still occur even if the assumptions for the test are met.]

Recall that inferential statistics allow us to make inferences (decisions, estimates, predictions) about a population based on data collected from a sample. Recall also that an inference about a population is true only if the sample studied is representative of the population. A statement about a population based on a biased sample is not likely to be true.

Assumption 1: Individuals in the sample were selected randomly and independently, so the sample is highly likely to be representative of the larger population.

- Random sampling ensures that each member of the population is equally likely to be selected.
- An independent sample is one which the selection of one member has no effect on the selection of any other.

Assumption 2: The distribution of sample differences (DSD) is a normal, because we drew the samples from a population that was normally distributed.

- This assumption is very important because we are estimating probabilities using the t- table – which provide accurate estimates of probabilities for events distributed normally.

Assumption 3: Sampled populations have equal variances or have homogeneity of variance.

Advantages & Disadvantages of using a repeated measures design

Advantages. Repeated measure designs reduce the probability of Type I errors when compared with independent sample designs because repeated measure t-tests reduce the probability that we will get a statistically significant difference that is due to an extraneous variable that differed between groups by chance (due to some other factor than the one in which we are interested).

Repeated measure designs are also more powerful (sensitive) than independent sample designs because two scores from each person are compared so each person serves as his or her own control group (we analyze the difference between scores). A special type of repeated measures design is known as the matched pairs design. If we are designing a study and suspect that there are important factors that could differ between our groups even if we randomly select and assign subjects, then we may use this type of design.

Because members of a matched-pair are similar to each other there is greater likelihood of our statistical test finding an “effect” when one person is present (power) in a repeated sample design as compared to a two-repeated sample design (in which subjects for two groups are picked randomly and independently – not matched on any traits).

Disadvantages. Repeated measure t-tests are very sensitive to outside influences and treatment influences. Outside Influences refers to factors outside of the experiment that may interfere with testing an individual across treatment/trials. Examples include mood or health or motivation of the individual participants. Think about it, if a participant tries really hard during the pretest but does not try very hard during the posttest, these differences can create problems later when analyzing the data.

Treatment Influences refers to the events that happen within the testing experience that interferes with how the data are collected. Three of the most common treatment influences are: 1. Practice effects, 2. Fatigue effects, and 3. Order effects.

Practice effect is present where participants perform a task better in later conditions because they have had a chance to practice it. Another type is a **fatigue effect**, where participants perform a task worse in later conditions because they become tired or bored. **Order effects** refer to differences in research participants' responses that result from the order (e.g., first, second, third) in which the experimental materials are presented to them.

Imagine, for example, that participants judge the guilt of an attractive defendant and then judge the guilt of an unattractive defendant. If they judge the unattractive defendant more harshly, this might be

because of his unattractiveness. But it could be instead that they judge him more harshly because they are becoming bored or tired. In other words, the order of the conditions is a confounding variable. The attractive condition is always the first condition and the unattractive condition the second. Thus any difference between the conditions in terms of the dependent variable could be caused by the order of the conditions and not the independent variable itself.

There is a solution to the problem of order effects, however, that can be used in many situations. It is counterbalancing, which means testing different participants in different orders. For example, some participants would be tested in the attractive defendant condition followed by the unattractive defendant condition, and others would be tested in the unattractive condition followed by the attractive condition. With three conditions, there would be six different orders (ABC, ACB, BAC, BCA, CAB, and CBA), so some participants would be tested in each of the six orders. With counterbalancing, participants are assigned to orders randomly, using the techniques we have already discussed. Thus random assignment plays an important role in within-subjects designs just as in between-subjects designs. Here, instead of randomly assigning to conditions, they are randomly assigned to different orders of conditions. In fact, it can safely be said that if a study does not involve random assignment in one form or another, it is not an experiment.

Because the repeated-measures design requires that each individual participate in more than one treatment, there is always the risk that exposure to the first treatment will cause a change in the participants that influences their scores in the second treatment that have nothing to do with the intervention. For example, if students are given the same test before and after the intervention the change in the posttest might be because the student got practice taking the test, not because the intervention was successful.

Learning Objectives

Having read this chapter, a student should be able to:

- identify when appropriate to calculate a paired or dependent t-test
- perform a hypothesis test using the paired or dependent t-test
- compute and interpret effect size for dependent or paired t-test
- list the assumptions for running a paired or dependent t-test
- list the advantages and disadvantages for a repeated measures design

Exercises – Ch. 12

1. What is the difference between a 1-sample t -test and a dependent-samples t -test? How are they alike?
2. Name 3 research questions that could be addressed using a dependent-samples t -test.
3. What are difference scores and why do we calculate them?
4. Why is the null hypothesis for a dependent-samples t -test always $\mu_D = 0$?
5. A researcher is interested in testing whether explaining the processes of statistics helps increase trust in computer algorithms. He wants to test for a difference at the $\alpha = 0.05$ level and knows that some people may trust the algorithms less after the training, so he uses a two-tailed test. He gathers pre- post data from 35 people and finds that the average difference score is 12.10 with a standard deviation (s) is 17.39. Conduct a hypothesis test to answer the research question.
6. Decide whether you would reject or fail to reject the null hypothesis in the following situations:
 1. $M_{\text{diff}} = 3.50$, $s = 1.10$, $n = 12$, $\alpha = 0.05$, two-tailed test
 2. 95% CI = (0.20, 1.85)
 3. $t = 2.98$, $t^* = -2.36$, one-tailed test to the left
 4. 90% CI = (-1.12, 4.36)
7. Calculate difference scores for the following data:

| Time 1 | Time 2 | X_D or D |
|--------|--------|------------|
| 61 | 83 | |
| 75 | 89 | |
| 91 | 98 | |
| 83 | 92 | |
| 74 | 80 | |
| 82 | 88 | |
| 98 | 98 | |
| 82 | 77 | |
| 69 | 88 | |
| 76 | 79 | |
| 91 | 91 | |
| 70 | 80 | |

8. You want to know if an employee's opinion about an organization is the same as the opinion of that employee's boss. You collect data from 18 employee-supervisor pairs and code the difference scores so that positive scores indicate that the employee has a higher opinion and negative scores indicate that the boss has a higher opinion (meaning that difference scores of 0 indicate no difference and complete agreement). You find that the mean difference score is $\bar{D} = -3.15$ with a standard deviation of $s_D = 1.97$. Test this hypothesis at the $\alpha = 0.01$ level.

9. Construct confidence intervals from a mean = 1.25, standard error of 0.45, and $df = 10$ at the 90%, 95%, and 99% confidence level. Describe what happens as confidence changes and whether to reject H_0 .

10. A professor wants to see how much students learn over the course of a semester. A pre-test is given before the class begins

to see what students know ahead of time, and the same test is given at the end of the semester to see what students know at the end. The data are below. Test for an improvement at the $\alpha = 0.05$ level. Did scores increase? How much did scores increase?

| Pretest | Posttest | X_D |
|---------|----------|-------|
| 90 | 8 | |
| 60 | 66 | |
| 95 | 99 | |
| 93 | 91 | |
| 95 | 100 | |
| 67 | 64 | |
| 89 | 91 | |
| 90 | 95 | |
| 94 | 95 | |
| 83 | 89 | |
| 75 | 82 | |
| 87 | 92 | |
| 82 | 83 | |
| 82 | 85 | |
| 88 | 93 | |
| 66 | 69 | |
| 90 | 90 | |
| 93 | 100 | |
| 86 | 95 | |
| 91 | 96 | |

Answers to Odd- Numbered Exercises – Ch. 12

1. A 1-sample t -test uses raw scores to compare an average to a specific value. A dependent samples t -test uses two raw scores from each person to calculate difference scores and test for an average difference score that is equal to zero. The calculations, steps, and interpretation is exactly the same for each.

3. Difference scores indicate change or discrepancy relative to a single person or pair of people. We calculate them to eliminate individual differences in our study of change or agreement.

5. Step 1: $H_0: \mu = 0$ "The average change in trust of algorithms is 0", $H_A: \mu \neq 0$ "People's opinions of how much they trust algorithms changes."

Step 2: Two-tailed test, $df = 34$, $t^* = 2.032$.

Step 3: $\bar{D} = 12.10$, $s_{\bar{D}} = 2.94$, $t = 4.12$.

Step 4: $t > t^*$, Reject H_0 . Based on opinions from 35 people, we can conclude that people trust algorithms more ($\bar{D} = 12.10$) after learning statistics, $t(34) = 4.12$, $p < .05$. Since the result is significant, we need an effect size: Cohen's $d = 0.70$, which is a moderate to large effect.

7. See table last column.

| Time 1 | Time 2 | D or X_D |
|--------|--------|------------|
| 61 | 83 | 22 |
| 75 | 89 | 14 |
| 91 | 98 | 7 |
| 83 | 92 | 9 |
| 74 | 80 | 6 |
| 82 | 88 | 6 |
| 98 | 98 | 0 |
| 82 | 77 | -5 |
| 69 | 88 | 19 |
| 76 | 79 | 3 |
| 91 | 91 | 0 |
| 70 | 80 | 10 |

9. At the 90% confidence level, $t^* = 1.812$ and $CI = (0.43, 2.07)$ so we reject H_0 . At the 95% confidence level, $t^* = 2.228$ and $CI = (0.25, 2.25)$ so we reject H_0 . At the 99% confidence level, $t^* = 3.169$ and $CI = (-0.18, 2.68)$ so we fail to reject H_0 . As the confidence level goes up, our interval gets wider (which is why we have higher confidence), and eventually we do not reject the null hypothesis because the interval is so wide that it contains 0.

13. Chapter 13:

Independent Samples

We have seen how to compare a single mean against a given value and how to utilize difference scores to look for meaningful, consistent change via a single mean difference using a repeated measures design. Now, we will learn how to compare two separate means from separate groups that do not overlap to see if there is a difference between them. The process of testing hypotheses about two means is exactly the same as it is for testing hypotheses about a single mean, and the logical structure of the formulae is the same as well. However, we will be adding a few extra steps this time to account for the fact that our data are coming from different sources.

Difference of Means

Last chapter, we learned about mean differences, that is, the average value of difference scores. Those difference scores came from ONE group and TWO time points (or two perspectives). Now, we will deal with the difference of the means, that is, the average values of separate groups that are represented by separate descriptive statistics. This analysis involves TWO groups and ONE time point. As with all of our other tests as well, both of these analyses are concerned with a single variable.

It is very important to keep these two tests separate and understand the distinctions between them because they assess very different questions and require different

approaches to the data. When in doubt, think about how the data were collected and where they came from. If they came from two time points with the same people (sometimes referred to as “longitudinal” data), you know you are working with repeated measures data (the measurement literally was repeated) and will use a paired/dependent samples *t*-test. If it came from a single time point that used separate groups, you need to look at the nature of those groups and if they are related. Can individuals in one group being meaningfully matched up with one and only one individual from the other group? For example, are they a romantic couple? If so, we call those data matched and we use a matched pairs/dependent samples *t*-test. However, if there’s no logical or meaningful way to link individuals across groups, or if there is no overlap between the groups, then we say the groups are independent and use the **independent samples *t*-test**, the subject of this chapter.

Research Questions about Independent Means

An **independent samples *t*-test** is also designed to compare populations. If we want to know if two populations differ and we do not know the mean of either population, we take a sample from each and then conduct an independent sample *t*-test. Many research ideas in the behavioral sciences and other areas of research are concerned with whether or not two means are the same or different. Logically, we can then say that these research questions are concerned with group mean differences. That is, on average, do we expect a person from Group A to be higher or lower on some variable than a person from Group B. In any time of research design looking at group mean differences, there are some key criteria we must

consider: the groups must be mutually exclusive (i.e. you can only be part of one group at any given time) and the groups have to be measured on the same variable (i.e. you can't compare personality in one group to reaction time in another group since those values would not be the same anyway).

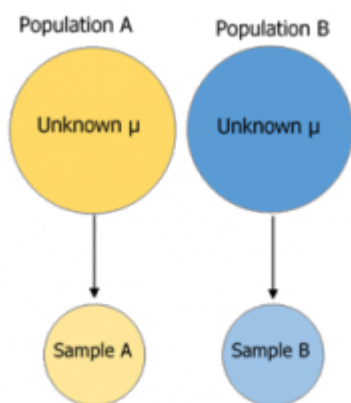


Figure 1. Collecting data from 2 different groups.

If the difference between the sample means given in the problem is very large compared to the differences we would expect to see between samples drawn from the same population, then we will conclude that the two samples must be from different populations. The language of the two independent samples t-tests involves probability statements because we know that there is variability in the samples that we draw from populations. If we were to draw two samples from a particular population we would expect a difference between the means of the samples by chance alone. In an independent sample situation we are given two sample means and understand that they are probably not equal BUT this does NOT provide evidence that the samples are from different populations. In independent samples t-tests we must estimate the probability of drawing particular differences between sample means from a population before deciding whether the

difference between the sample means given in the problem is sufficiently large to lead us to conclude that the samples must be from different populations.

Let's look at one of the most common and logical examples: testing a new medication. When a new medication is developed, the researchers who created it need to demonstrate that it effectively treats the symptoms they are trying to alleviate. The simplest design that will answer this question involves two groups: one group that receives the new medication (the "treatment" group) and one group that receives a placebo (the "control" group). Participants are randomly assigned to one of the two groups (remember that random assignment is the hallmark of a true experiment), and the researchers test the symptoms in each person in each group after they received either the medication or the placebo. They then calculate the average symptoms in each group and compare them to see if the treatment group did better (i.e. had fewer or less severe symptoms) than the control group.

In this example, we had two groups: treatment and control. Membership in these two groups was mutually exclusive: each individual participant received either the experimental medication or the placebo. No one in the experiment received both, so there was no overlap between the two groups. Additionally, each group could be measured on the same variable: symptoms related to the disease or ailment being treated. Because each group was measured on the same variable, the average scores in each group could be meaningfully compared. If the treatment was ineffective, we would expect that the average symptoms of someone receiving the treatment would be the same as the average symptoms of someone receiving the placebo (i.e. there is no difference between the groups). However, if the treatment WAS effective, we would expect fewer symptoms from the treatment group, leading to a lower group average. Now let's

look at an example using groups that already exist. A common, and perhaps salient, question is how students feel about their job prospects after graduation. Suppose that we have narrowed our potential choice of college down to two universities and, in the course of trying to decide between the two, we come across a survey that has data from each university on how students at those universities feel about their future job prospects. As with our last example, we have two groups: University A and University B, and each participant is in only one of the two groups (assuming there are no transfer students who were somehow able to rate both universities). Because students at each university completed the same survey, they are measuring the same thing, so we can use a t -test to compare the average perceptions of students at each university to see if they are the same. If they are the same, then we should continue looking for other things about each university to help us decide on where to go. But, if they are different, we can use that information in favor of the university with higher job prospects.

As we can see, the grouping variable we use for an independent samples t -test can be a set of groups we create (as in the experimental medication example) or groups that already exist naturally (as in the university example). There are countless other examples of research questions relating to two group means, making the independent samples t -test one of the most widely used analyses around.

Independent samples t -test



[Photo credit](#)

We know can say tea for two, as the saying goes, one for me and one for you, is for two different people having tea at the same time. This follows the independent t-test having two separate groups at the same time point.

Setting up step 1

In chapter 12, it is the same participants tracked over 2 timepoints for the paired samples *t*-test. This is an example of a **repeated measures** or **within-group design**. This chapter focuses on two *different* samples compared, using independent samples *t*-test, known as a **between-group design**. In setting up step 1, there will be a few variations to how to set up the null and alternative (H1 or HA) hypotheses. You can see these are all ways to set up the two types of research hypotheses (three hypotheses for the 2 directional and 1 for the non-directional). You can set it up by comparing both groups (first 2 columns) or examining differences (which is similar to how we set up the top of your *t*-formula).

| Research Question | Hypotheses in 3 ways | | |
|--|---|--|--|
| Are male scores higher than female scores? (between-group design) | $H_0: \mu_M - \mu_F \leq 0$ $H_1: \mu_M - \mu_F > 0$ | in other words: $H_0: \mu_M \leq \mu_F$ $H_1: \mu_M > \mu_F$ | in other words: $H_0: \mu_M \leq \mu_F$ $H_1: \mu_M > \mu_F$ |
| Are male scores lower than female scores? (between-group design) | $H_0: \mu_M - \mu_F \geq 0$ $H_1: \mu_M - \mu_F < 0$ | in other words: $H_0: \mu_M \geq \mu_F$ $H_1: \mu_M < \mu_F$ | in other words: $H_0: \mu_M \geq \mu_F$ $H_1: \mu_M < \mu_F$ |
| Are male scores different from female scores? (between-group design) | $H_0: \mu_M - \mu_F = 0$ $H_1: \mu_M - \mu_F \neq 0$ | in other words: $H_0: \mu_M = \mu_F$ $H_1: \mu_M \neq \mu_F$ | in other words: $H_0: \mu_M = \mu_F$ $H_1: \mu_M \neq \mu_F$ |
| Do athlete performance improve after training? (within-group design) | $H_0: \mu_D \leq 0$ $H_1: \mu_D > 0$ | in other words: $H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$ | in other words: $H_0: \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 - \mu_2 > 0$ |
| Do athlete reaction times decrease after training? (within-group design) | $H_0: \mu_D \geq 0$ $H_1: \mu_D < 0$ | in other words: $H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$ | in other words: $H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$ |
| Does training have an effect on athlete reaction times? (within-group design) | $H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$ | in other words: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ | in other words: $H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$ |

Table 1. Examples of hypotheses set up for 2 samples based on between-group and within-group design.

Hypotheses and Decision Criteria

The process of testing hypotheses using an independent samples t -test is the same as it was in the last three chapters, and it starts with stating our hypotheses and laying out the criteria we will use to test them.

Our null hypothesis for an independent samples t -test is the same as all others: there is no difference. The means of the two groups are the same under the null hypothesis, no matter how those groups were formed. Mathematically, this takes on two equivalent forms:

$$H_0: \mu_1 = \mu_2 \text{ or } H_0: \mu_1 - \mu_2 = 0$$

Both of these formulations of the null hypothesis tell us exactly the same thing: that the numerical value of the means is the same in both groups. This is more clear in the first formulation, but the second formulation also makes sense (any number minus itself is always zero) and helps us out a little when we get to the math of the test statistic. Either one is acceptable and you only need to report one. The English interpretation of both of them is also the same: H_0 : There is no difference between the means of the two groups.

Our alternative hypotheses are also unchanged: we simply replace the equal sign ($=$) with one of the three inequalities ($>$, $<$, \neq):

$H_A: \mu_1 > \mu_2$ or $\mu_1 - \mu_2 > 0$; H_A : Group 1 is more than/
stronger than group 2.

$H_A: \mu_1 < \mu_2$ or $\mu_1 - \mu_2 < 0$; H_A : Group 1 is less than/
weaker than group 2.

$H_A: \mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$; H_A : There is no difference
between the two groups.

Whichever formulation you chose for the null hypothesis should be the one that connects to the research or alternative hypothesis. Notice that we are now dealing with two means instead of just one, so it will be very important to keep track of

which mean goes with which population and, by extension, which dataset and sample data. We use subscripts to differentiate between the populations, so make sure to keep track of which is which. If it is helpful, you can also use more descriptive subscripts.

To use the experimental medication example:

H_0 : There is no difference between the means of the treatment (T) and control (C) groups. $H_0: \mu_T = \mu_C$

H_A : There is a difference between the means of the treatment (T) and control (C) group. $H_A: \mu_T \neq \mu_C$

Step 2

Once we have our hypotheses laid out, we can set our criteria to test them using the same three pieces of information as before: significance level (α), directionality (left, right, or two-tailed), and degrees of freedom. We will use the same critical value t table, but a new degrees of freedom for the independent samples t -test.

degrees of freedom for independent samples t -test

$$df = (n_1 - 1) + (n_2 - 1)$$

where n_1 represents the sample size for group 1 and n_2 represents the sample size for group 2. We have 2 separate groups, each with a calculated degrees of freedom.

This looks different than before, but it is just adding

the individual degrees of freedom from each group $(n - 1)$ together. Notice that the sample sizes, n , also get subscripts so we can tell them apart. For an independent samples t -test, it is often the case that our two groups will have slightly different sample sizes, either due to chance or some characteristic of the groups themselves. Generally, this is not as issue, so long as one group is not massively larger than the other group, and there are not large differences in the variance of each group (more on this later).

Independent Samples t -statistic

The test statistic for our independent samples t -test takes on the same logical structure and format as our other t -tests: our observed effect minus our null hypothesis value, all divided by the standard error:

Independent samples t -test

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_{(M1 - M2)}}$$

M_1 is the sample mean for group 1 and M_2 is the sample mean for group 2. This looks like more work to calculate, but remember that our null hypothesis

states that the quantity $\mu_1 - \mu_2 = 0$, so we can drop that out of the equation and are left with:

$$t = \frac{M_1 - M_2}{S_{m_1 - m_2}}$$

Our standard error in the denominator (bottom of the formula) is denoting what it is the standard error of (derived from 2 samples). Because we are dealing with the difference between two separate means, rather than a single mean or single mean of difference scores, we put both means in the subscript.

Calculating our standard error, as we will see next, is where the biggest differences between this t -test and other t -tests appears. However, once we do calculate it and use it in our test statistic, everything else goes back to normal. Our decision criteria is still comparing our obtained test statistic to our critical value, and our interpretation based on whether or not we reject the null hypothesis is unchanged as well.

Standard Error and Pooled Variance

Recall that the standard error is the average distance between any given sample mean and the center of its corresponding sampling distribution, and it is a function of the standard deviation of the population (either given or estimated) and the sample size. This definition and interpretation hold true for our independent samples t -test as well, but because we are working with two samples drawn from two populations, we have to first combine their estimates of standard deviation – or, more accurately, their estimates of variance – into a single value that we can then use to calculate our standard error.

The combined estimate of variance using the information from each sample is called the pooled variance and is denoted S_p^2 ; the subscript p serves as a reminder indicating that it is the pooled variance. The term “pooled variance” is a literal name because we are simply pooling or combining the information on variance – the Sum of Squares and Degrees of Freedom – from both of our samples into a single number. The result is a weighted average of the observed sample variances, the weight for each being determined by the sample size, and will always fall between the two observed variances. The computational formula for the pooled variance is:

Pooled variance used to get to our new
standard error formula

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

This formula can look daunting at first, but it is in fact just a weighted average. Even more conveniently, some simple algebra can be employed to greatly reduce the complexity of the calculation. The simpler and more appropriate formula to use when calculating pooled variance is:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

Both formula will give you the same pooled variance. The most common formula used is the SS/df (which is how we typically think about variance). Using the SS/df formula, it's very simple to see that we are just adding together the same pieces of information we have been calculating since chapter 3.

Once we have our pooled variance calculated, we can drop it into the equation for our standard error:

Standard error for independent t-test

$$s_{M-M} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Once again, although this formula may seem different than it was before, in reality it is just a different way of writing the same thing. Think back to the standard error options presented in chapter 9, when our standard error was

$$s_{\bar{X}} = \sqrt{\frac{s_X^2}{n}}$$

Looking at that, we can now see that, once again, we are simply adding together two pieces of information: no new logic or interpretation required. Once the standard error is calculated, it goes in the denominator of our test statistic, as shown above and as was the case in all previous chapters. Thus, the only additional step to calculating an independent samples *t*-statistic is computing the pooled variance. Let's see an example in action.

Example: Movies and Mood

We are interested in whether the type of movie someone sees at the theater affects their mood when they leave. We decide to ask people about their mood as they leave one of two movies: a comedy (group 1, $n = 35$) or a horror film (group 2, $n = 29$). Our data are coded so that higher scores indicate a more positive mood. We have good reason to believe that people leaving the comedy will be in a better mood, so we use a one-tailed test at $\alpha = 0.05$ to test our hypothesis.

Step 1: State the Hypotheses

As always, we start with hypotheses:

H_0 : There is no difference in
average mood between the two
movie types

$$H_0: \mu_1 - \mu_2 = 0 \text{ or } H_0: \mu_1 = \mu_2$$

H_A : The comedy film will give a
better average mood than the
horror film

$$H_A: \mu_1 - \mu_2 > 0 \text{ or } H_A: \mu_1 > \mu_2$$

Notice that in the first formulation of the alternative hypothesis we say that the first mean minus the second mean will be greater than zero. This is based on how we code the data (higher is better), so we suspect

that the mean of the first group will be higher. Thus, we will have a larger number minus a smaller number, which will be greater than zero. Be sure to pay attention to which group is which and how your data are coded (higher is almost always used as better outcomes) to make sure your hypothesis makes sense!

Step 2: Find the Critical Values

Just like before, we will need critical values, which come from our t -table. In this example, we have a one-tailed test at $\alpha = 0.05$ and expect a positive answer (because we expect the difference between the means to be greater than zero). Our degrees of freedom for our independent samples t -test is just the degrees of freedom from each group added together: $35 + 29 - 2 = 62$. From our t -table, we find that our critical value is $t^* = 1.671$. Note: because 62 does not appear on the critical values table, we use the next lowest value, which in this case is 60.

Step 3: Compute the Test Statistic

The data from our two groups are presented in the tables below. Table 1 shows the summary values for the Comedy group, and Table 2 shows the summary values for the Horror group (the end of the chapter contains the raw data).

| Group 1: Comedy Film | | |
|----------------------|-----------------------|-----------------------------------|
| n | M | SS |
| 35 | $\Sigma X/n = 840/35$ | $\Sigma(X - \bar{X})^2 = 5061.60$ |

Table 1. Raw scores and Sum of Squares for Group 1

| Group 2: Horror Film | | |
|----------------------|--------------------------------|-----------------------------------|
| n | M | SS |
| 29 | $\Sigma X/n = 478.6/29 = 16.5$ | $\Sigma(X - \bar{X})^2 = 3896.45$ |

Table 2. Raw scores and Sum of Squares for Group 1.

the Sum of Squares for each group: $SS_1 = 5061.60$ and $SS_2 = 3896.45$. These values have all been calculated and take on the same interpretation as they have since chapter 3. Before we move on to the pooled variance that will allow us to calculate standard error, let's compute our standard deviation for each group; even though we will not use them in our calculation of the test statistic, they are still important descriptors of our data:

$$s_1 = \sqrt{\frac{5061.60}{34}} = 12.20$$

$$s_2 = \sqrt{\frac{3896.45}{28}} = 11.80$$

Now we can move on to our new calculation, the pooled variance, which is just the Sums of Squares that we calculated from our table and the degrees of freedom, which is just $n - 1$ for each group:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{5061.60 + 3896.45}{34 + 28} = \frac{8958.05}{62} = 144.48$$

As you can see, if you follow the regular process of calculating standard deviation using the Sum of Squares table, finding the pooled variance is very easy. Now we can use that value to calculate our standard error, the last step before we can find our test statistic:

$$\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{144.48}{35} + \frac{144.48}{29}} = \sqrt{4.13 + 4.98} = \sqrt{9.11} = 3.02$$

Finally, we can use our standard error and the means we calculated earlier to compute our test statistic.

$$t = \frac{M_1 - M_2}{S_{m_1 - m_2}}$$

$$= \frac{24.00 - 16.50}{3.02} = 2.48$$

Now we can move on to the final step of the hypothesis testing procedure.

Step 4: Make the Decision

Our test statistic has a value of $t = 2.48$, and in step 2 we found that the critical value is $t^* = 1.671$. $2.48 > 1.671$, so we reject the null hypothesis:

Reject H_0 . Based on our sample data from people who watched different kinds of movies, we can say that the average mood after a

comedy movie ($M_2 = 24.00$) is better than the average mood after a horror movie ($M_2 = 16.50$), $t(62) = 2.48$, $p < .05$.

Effect Sizes and Confidence Intervals

We have seen in previous chapters that even a statistically significant effect needs to be interpreted along with an effect size to see if it is practically meaningful. We have also seen that our sample means, as a point estimate, are not perfect and would be better represented by a range of values that we call a confidence interval. As with all other topics, this is also true of our independent samples t -tests. Our effect size for the independent samples t -test is still Cohen's d , and it is still just our observed effect divided by the standard deviation. Remember that standard deviation is just the square root of the variance, and because we work with pooled variance in our test statistic, we will use the square root of the pooled variance as our denominator in the formula for Cohen's d . We also can still calculate r^2 , the percentage of variance accounted for by the independent variable/treatment effect.

Effect size options

Cohen's d for independent t -test:

$$d = \frac{M_1 - M_2}{\sqrt{S_p^2}}$$

Just like chapter 12, you also have r^2 as an option.

$$r^2 = \frac{t^2}{t^2 + df}$$

where df is the df calculated in step 2. The r^2 is interpreted as percent of variance in the dependent variable accounted for by the independent variable. Remember that r^2 is calculated when you have an effect.

For our example above, $M_1 = 24$, $M_2 = 16.5$, $s_p^2 = 144.48$, we can calculate Cohen's d to be:

$$\begin{aligned} d &= 24 - 16.50 / \sqrt{144.48} = 7.5 / \\ &12.02 = 0.62 \end{aligned}$$

We interpret this using the same guidelines as before, so we would consider this a moderate or moderately large effect.

For our example above, $t = 2.48$ (thus $t^2 = 2.48 \times 2.48 = 6.15$), $df = 62$, we can calculate r^2 to be:

$$r^2 = 6.15 / (6.15 + 62) = 6.15 / 68.15 = 0.09$$

We interpret this using the same guidelines as before, so 9% of the variance in mood (our DV) is from type of movie (our IV).

Our confidence intervals also take on the same form and interpretation as they have in the past. The value we are interested in is the difference between the two means, so our point estimate is the value of one mean minus the other, or $M_1 - M_2$. Just like before, this is our observed effect and is the same value as the one we place in the numerator of our test statistic. We calculate this value then place the margin of error – still our critical value times our standard error – above and below it.

Confidence Intervals

CI =

$$(M_1 - M_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where you would still calculate upper bound (UB) and lower bound (LB) values with t being the critical value t -value using a 2-tail test.

Because our hypothesis testing example used a one-tailed test, it would be inappropriate to calculate a

confidence interval on those data (remember that we can only calculate a confidence interval for a two-tailed test because the interval extends in both directions).

Example: Confidence Interval

Let's say we find summary statistics on the average life satisfaction of people from two different towns and want to create a confidence interval to see if the difference between the two might actually be zero.

Our sample data are $M_1 = 28.65$, $s_1 = 12.40$, $n_1 = 40$ and $M_2 = 25.40$, $s_2 = 15.68$, $n_2 = 42$. At face value, it looks like the people from the first town have higher life satisfaction (28.65 vs. 25.40), but it will take a confidence interval (or complete hypothesis testing process) to see if that is true or just due to random chance.

First, we want to calculate the difference between our sample means, which is $28.65 - 25.40 = 3.25$.

Next, we need a critical value from our t -table. If we want to test at the normal 95% level of confidence, then our sample sizes will yield degrees of freedom equal to $40 + 42 - 2 = 80$. From our table, that gives us a critical value of $t^* = 1.990$.

Finally, we need our standard error. Recall

that our standard error for an independent samples t-test uses pooled variance, which requires the Sum of Squares and degrees of freedom. Up to this point, we have calculated the Sum of Squares using raw data, but in this situation, we do not have access to it. So, what are we to do?

If we have summary data like standard deviation and sample size, it is very easy to calculate the pooled variance using the other formula presented

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

If $s_1 = 12.40$, then $s_1^2 = 12.40 \times 12.40 = 153.76$, and if $s_2 = 15.68$, then $s_2^2 = 15.68 \times 15.68 = 245.86$. With $n_1 = 40$ and $n_2 = 42$, we are all set to calculate the pooled variance.

$$s_p^2 = [(40-1)(153.76) + (42-1)(245.86)] / (40+42-2) = [(39)(153.76) + (41)(245.86)] / 80 = (5996.64 + 10080.36) / 80 = 16077 / 80 = 200.96$$

Plugging in $s_p^2 = 200.96$, $n_1 = 40$, and $n_2 = 42$ nd, our standard error equals:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{200.96}{40} + \frac{200.96}{42}} = \sqrt{5.02 + 4.78} = \sqrt{9.89} = 3.14$$

All of these steps are just slightly different ways of using the same formulae, numbers, and ideas we have worked with up to this point. Once we get our standard error, it's time to build our confidence interval.

$$95\%CI = 3.25 \pm 1.990(3.14)$$

$$\text{Upper Bound} = 3.25 + 1.990(3.14)$$

$$UB = 3.25 + 6.25$$

$$UB = 9.50$$

$$\text{Lower Bound} = 3.25 - 1.990(3.14)$$

$$LB = 3.25 - 6.25$$

$$LB = -3.00$$

$$95\%CI = (-3.00, 9.50)$$

Our confidence interval, as always, represents a range of values that would be considered reasonable or plausible based on our observed data. In this instance, our interval $(-3.00, 9.50)$ does contain zero. Thus, even though the means look a little bit different, it may very well be the case that the life satisfaction in both of these towns is the same. Proving otherwise would require more data.

Hypothesis testing and confidence intervals

As mentioned when confidence intervals were introduced, there is a close relationship between confidence intervals and hypothesis tests. In particular, if the confidence interval does not

include the null hypothesis, then the associated statistical test would be statistically significant. However, things get trickier if we want to compare the means of two conditions (Schenker and Gentleman, [2001](#)). There are a couple of situations that are clear. First, if each mean is contained within the confidence interval for the other mean, then there is definitely no significant difference at the chosen confidence level. Second, if there is no overlap between the confidence intervals, then there is certainly a significant difference at the chosen level; in fact, this test is substantially *conservative*, such that the actual error rate will be lower than the chosen level. But what about the case where the confidence intervals overlap one another but don't contain the means for the other group? In this case the answer depends on the relative variability of the two variables, and there is no general answer. However, one should in general avoid using the “eyeball test” for overlapping confidence intervals.

Homogeneity of Variance

Before wrapping up the coverage of independent samples *t*-tests, there is one other important topic to cover. Using the pooled variance to calculate the test statistic relies on an assumption known as homogeneity of variance. In statistics, an assumption is some characteristic that we assume is true about our data, and our ability to use our inferential statistics

accurately and correctly relies on these assumptions being true. If these assumptions are not true, then our analyses are at best ineffective (e.g. low power to detect effects) and at worst inappropriate (e.g. too many Type I errors). A detailed coverage of assumptions is beyond the scope of this course, but it is important to know that they exist for all analyses.

For the current analysis, one important assumption is homogeneity of variance. This is fancy statistical talk for the idea that the true population variance for each group is the same and any difference in the observed sample variances is due to random chance (if this sounds eerily similar to the idea of testing the null hypothesis that the true population means are equal, that's because it is exactly the same!) This notion allows us to compute a single pooled variance that uses our easily calculated degrees of freedom. If the assumption is shown to not be true, then we have to use a very complicated formula to estimate the proper degrees of freedom. There are formal tests to assess whether or not this assumption is met, but we will not discuss them here.

Many statistical programs incorporate the test of homogeneity of variance automatically and can report the results of the analysis assuming it is true or assuming it has been violated. You can easily tell which is which by the degrees of freedom: the corrected degrees of freedom (which is used when the assumption of homogeneity of variance is violated) will have decimal places. Fortunately, the independent samples t-test is very robust to violations of this assumption (an analysis is "robust" if it works well even when its assumptions are not met), which is why we do not bother going through the tedious work of testing and estimating new degrees of freedom by hand.

Review: Statistical Power

There are three factors that can affect statistical power:

- Sample size: Larger samples provide greater statistical power
- Effect size: A given design will always have greater power to find a large effect than a small effect (because finding large effects is easier)
- Type I error rate: There is a relationship between Type I error and power such that (all else being equal) decreasing Type I error will also decrease power.

We can see this through simulation. First let's simulate a single experiment, in which we compare the means of two groups using a standard t-test. We will vary the size of the effect (specified in terms of Cohen's d), the Type I error rate, and the sample size, and for each of these we will examine how the proportion of significant results (i.e. power) is affected. Figure 1 shows an example of how power changes as a function of these factors.

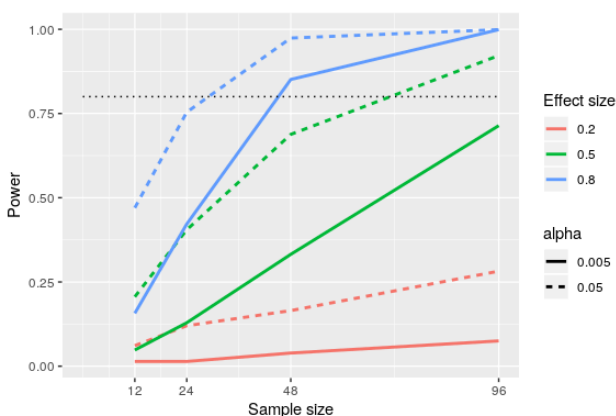


Figure 1: Results from power simulation, showing power as a function of sample size, with effect sizes shown as different colors, and alpha shown as line type. The standard criterion of 80 percent power is shown by the dotted black line.

This simulation shows us that even with a sample size of 96, we will have relatively little power to find a small effect ($d=0.2$) with $\alpha=0.005$. This means that a study designed to do this would be *futile* – that is, it is almost guaranteed to find nothing even if a true effect of that size exists.

There are at least two important reasons to care about statistical power. First, if you are a researcher, you probably don't want to spend your time doing futile experiments. Running an underpowered study is essentially futile, because it means that there is a very low likelihood that one will find an effect, even if it exists. Second, it turns out that any positive

findings that come from an underpowered study are more likely to be false compared to a well-powered study, a point we discuss in more detail in Chapter [19](#). Fortunately, there are tools available that allow us to determine the statistical power of an experiment. The most common use of these tools is in planning an experiment, when we would like to determine how large our sample needs to be in order to have sufficient power to find our effect of interest.

Assumptions of independent t-test

Assumptions are conditions that must be met in order for our hypothesis testing conclusion to be valid. [Important: If the assumptions are *not* met then our hypothesis testing conclusion is not likely to be valid. Testing errors can still occur even if the assumptions for the test are met.]

Recall that inferential statistics allow us to make inferences (decisions, estimates, predictions) about a population based on data collected from a sample. Recall also that an inference about a population is true only if the sample studied is representative of the population. A statement about a population based on a biased sample is not likely to be true.

Assumption 1: Individuals in the sample were selected randomly and independently, so the sample is highly likely to be representative of the larger population.

- Random sampling ensures that each member of the population is equally likely to be selected.
- An independent sample is one which the selection of one member has no effect on the selection of any other.

Assumption 2: The distribution of sample differences (DSD) is a normal, because we drew the samples from a population that was normally distributed.

- This assumption is very important because we are estimating probabilities using the *t*-table – which provide accurate estimates of probabilities for events distributed normally.

Assumption 3: Sampled populations have equal variances or have homogeneity of variance.

Learning Objectives

Having read this chapter, a student should be able to:

- Identify under what study design to use an independent *t*-test
- Use the independent *t*-test to test hypotheses about mean differences between two populations or treatment conditions
- Calculate and evaluate effect size options, Cohen's *d* and r^2

Exercises – Ch. 13

1. What is meant by “the difference of the means” when talking about an independent samples *t*-test? How does it differ from the “mean of the differences” in a repeated measures *t*-test?
2. Describe three research questions that could be tested using an independent samples *t*-test.

3. Calculate pooled variance from the following raw data:

| Group 1 | Group 2 |
|---------|---------|
| 16 | 4 |
| 11 | 10 |
| 9 | 15 |
| 7 | 13 |
| 5 | 12 |
| 4 | 9 |
| 12 | 8 |

4. Calculate the standard error from the following descriptive statistics

1. $s_1 = 24$, $s_2 = 21$, $n_1 = 36$, $n_2 = 49$
2. $s_1 = 15.40$, $s_2 = 14.80$, $n_1 = 20$, $n_2 = 23$
3. $s_1 = 12$, $s_2 = 10$, $n_1 = 25$, $n_2 = 25$

5. Determine whether to reject or fail to reject the null hypothesis in the following situations:

1. $t(40) = 2.49$, $\alpha = 0.01$, one-tailed test to the right
2. $\overline{X} = 64$, $\overline{X} = 54$, $n_1 = 14$, $n_2 = 12$, $\diamond \text{—————} = 9.75$, $\alpha = 0.05$, two-tailed test
3. 95% Confidence Interval: (0.50, 2.10)

6. A professor is interest in whether or not the type of software program used in a statistics lab affects how well students learn the material. The professor teaches the same lecture material to two classes but has one class use a point-and-click software program in lab and has the other class use a basic programming language. The professor tests for a difference between the two classes on their final exam scores.

| Point-and-Click | Programming |
|-----------------|-------------|
| 83 | 86 |
| 83 | 79 |
| 63 | 100 |
| 77 | 74 |
| 86 | 70 |
| 84 | 67 |
| 78 | 83 |
| 61 | 85 |
| 65 | 74 |
| 75 | 86 |
| 100 | 87 |
| 60 | 61 |
| 90 | 76 |
| 66 | 100 |
| 54 | |

7. A researcher wants to know if there is a difference in how busy someone is based on whether that person identifies as an early bird or a night owl. The researcher gathers data from people in each group, coding the data so that higher scores represent higher levels of being busy, and tests for a difference between the two at the .05 level of significance.

| Early Bird | Night Owl |
|------------|-----------|
| 23 | 26 |
| 28 | 10 |
| 27 | 20 |
| 33 | 19 |
| 26 | 26 |
| 30 | 18 |
| 22 | 12 |
| 25 | 25 |
| 26 | |

8. Lots of people claim that having a pet helps lower their stress level. Use the following summary data to test the claim that there is a lower average stress level among pet owners (group 1) than among non-owners (group 2) at the .05 level of significance. $M_1 = 16.25$, $M_2 = 20.95$, $s_1 = 4.00$, $s_2 = 5.10$, $n_1 = 29$, $n_2 = 25$

9. Administrators at a university want to know if students in different majors are more or less extroverted than others. They provide you with descriptive statistics they have for English majors (coded as 1) and History majors (coded as 2) and ask you to create a confidence interval of the difference between them. Does this confidence interval suggest that the students from the majors differ? $M_1 = 3.78$, $M_2 = 2.23$, $s_1 = 2.60$, $s_2 = 1.15$, $n_1 = 45$, $n_2 = 40$

10. Researchers want to know if people's awareness of environmental issues varies as a function of where they live. The researchers have the following summary data from two states, Alaska and Hawaii, that they want to use to test for a

difference. $M_H = 47.50$, $M_A = 45.70$, $s_H = 14.65$, $s_A = 13.20$, $n_H = 139$, $n_A = 150$

Answers to Odd- Numbered Exercises – Ch. 13

1. The difference of the means is one mean, calculated from a set of scores, compared to another mean which is calculated from a different set of scores; the independent samples t-test looks for whether the two separate values are different from one another. This is different than the “mean of the differences” because the latter is a single mean computed on a single set of difference scores that come from one data collection of matched pairs. So, the difference of the means deals with two numbers but the mean of the differences is only one number.

3. $SS_1 = 106.86$, $SS_2 = 78.86$, $s^2 = 15.48$

5. 1) Reject 2) Fail to Reject 3) Reject

7. Step 1: $H_0: \mu_1 - \mu_2 = 0$ “There is not difference in the average business of early birds versus night owls”, $H_A: \mu_1 - \mu_2 \neq 0$ “There is a difference in the average business of early birds versus night owls.”

Step 2: Two-tailed test, $df = 15$, $t^* = 2.131$.

Step 3: $M_1 = 26.67$, $M_2 = 19.50$, $s^2 = 27.73$, $s_{M_1-M_2} = 2.37$, $t = 3.03$.

Step 4: $t > t^*$, Reject H_0 . Based on our data of early birds and night owls, we can conclude that early birds are busier ($M_1 = 26.67$) than night owls ($M_2 = 19.50$), $t(15) = 3.03$, $p < .05$. Since the result is significant, we need an effect size: Cohen's $d = 1.47$, which is a large effect.

9. $M_1 - M_2 = 1.55$, $t^* = 1.990$, $s_{M_1-M_2} = 0.45$, $CI = (0.66, 2.44)$. This confidence interval does not contain zero, so it does suggest that there is a difference between the extroversion of English majors and History majors.

*Raw scores from the mood and movies
example*

| Group 1: Comedy Film | | |
|----------------------|-----------------|-------------------|
| X | $(X - \bar{X})$ | $(X - \bar{X})^2$ |
| 39.10 | 15.10 | 228.01 |
| 38.00 | 14.00 | 196.00 |
| 14.90 | -9.10 | 82.81 |
| 20.70 | -3.30 | 10.89 |
| 19.50 | -4.50 | 20.25 |
| 32.20 | 8.20 | 67.24 |
| 11.00 | -13.00 | 169.00 |
| 20.70 | -3.30 | 10.89 |
| 26.40 | 2.40 | 5.76 |
| 35.70 | 11.70 | 136.89 |
| 26.40 | 2.40 | 5.76 |
| 28.80 | 4.80 | 23.04 |
| 33.40 | 9.40 | 88.36 |
| 13.70 | -10.30 | 106.09 |
| 46.10 | 22.10 | 488.41 |
| 13.70 | -10.30 | 106.09 |
| 23.00 | -1.00 | 1.00 |
| 20.70 | -3.30 | 10.89 |
| 19.50 | -4.50 | 20.25 |
| 11.40 | -12.60 | 158.76 |

| | | |
|----------------|--------------|------------------|
| 24.10 | 0.10 | 0.01 |
| 17.20 | -6.80 | 46.24 |
| 38.00 | 14.00 | 196.00 |
| 10.30 | -13.70 | 187.69 |
| 35.70 | 11.70 | 136.89 |
| 41.50 | 17.50 | 306.25 |
| 18.40 | -5.60 | 31.36 |
| 36.80 | 12.80 | 163.84 |
| 54.10 | 30.10 | 906.01 |
| 11.40 | -12.60 | 158.76 |
| 8.70 | -15.30 | 234.09 |
| 23.00 | -1.00 | 1.00 |
| 14.30 | -9.70 | 94.09 |
| 5.30 | -18.70 | 349.69 |
| 6.30 | -17.70 | 313.29 |
| $\Sigma = 840$ | $\Sigma = 0$ | $\Sigma=5061.60$ |

| Group 2: Horror Film | | |
|----------------------|-----------------|-------------------|
| X | $(X - \bar{X})$ | $(X - \bar{X})^2$ |
| 24.00 | 7.50 | 56.25 |
| 17.00 | 0.50 | 0.25 |
| 35.80 | 19.30 | 372.49 |
| 18.00 | 1.50 | 2.25 |
| -1.70 | -18.20 | 331.24 |
| 11.10 | -5.40 | 29.16 |
| 10.10 | -6.40 | 40.96 |
| 16.10 | -0.40 | 0.16 |
| -0.70 | -17.20 | 295.84 |
| 14.10 | -2.40 | 5.76 |
| 25.90 | 9.40 | 88.36 |
| 23.00 | 6.50 | 42.25 |
| 20.00 | 3.50 | 12.25 |
| 14.10 | -2.40 | 5.76 |
| -1.70 | -18.20 | 331.24 |
| 19.00 | 2.50 | 6.25 |
| 20.00 | 3.50 | 12.25 |
| 30.90 | 14.40 | 207.36 |
| 30.90 | 14.40 | 207.36 |
| 22.00 | 5.50 | 30.25 |
| 6.20 | -10.30 | 106.09 |
| 27.90 | 11.40 | 129.96 |
| 14.10 | -2.40 | 5.76 |
| 33.80 | 17.30 | 299.29 |

| | | |
|------------------|-----------------|------------------|
| 26.90 | 10.40 | 108.16 |
| 5.20 | -11.30 | 127.69 |
| 13.10 | -3.40 | 11.56 |
| 19.00 | 2.50 | 6.25 |
| -15.50 | -32.00 | 1024.00 |
| $\Sigma = 478.6$ | $\Sigma = 0.10$ | $\Sigma=3896.45$ |

Table 1. Raw scores and Sum of Squares for Group 1

Table 2. Raw scores and Sum of Squares for Group 2.

14. Chapter 14: Analysis of Variance

Additional Hypothesis Tests

In unit 1, we learned the basics of statistics – what they are, how they work, and the mathematical and conceptual principles that guide them. In unit 2, we put applied these principles to the process and ideas of hypothesis testing – how we take observed sample data and use it to make inferences about our populations of interest – using one continuous variable and one categorical variable. We will now continue to use this same hypothesis testing logic and procedure on new types of data. We will focus on group mean differences on more than two groups, using Analysis of Variance.

Analysis of variance, often abbreviated to **ANOVA** for short, serves the same purpose as the t -tests we learned earlier in unit 2: it tests for differences in group means. ANOVA is more flexible in that it can handle any number of groups, unlike t -tests which are limited to two groups (independent samples) or two time points (paired samples). Thus, the purpose and interpretation of ANOVA will be the same as it was for t -tests, as will the hypothesis testing procedure. However, ANOVA will, at first glance, look much different from a mathematical perspective, though as we will see, the basic logic behind the test statistic for ANOVA is actually the same.

ANOVA basics

An Analysis of Variance (ANOVA) is an inferential statistical tool that we use to find statistically significant differences among the *means* of two or more populations.

We calculate variance but the goal is still to compare population mean differences. The test statistic for the ANOVA is called F. It is a ratio of two estimates of the population variance based on the sample data.

Experiments are designed to determine if there is a cause and effect relationship between two variables. In the language of the ANOVA, the factor is the variable hypothesized to cause some change (effect) in the response variable (dependent variable).

An ANOVA conducted on a design in which there is only one factor is called a **one-way ANOVA**. If an experiment has two factors, then the ANOVA is called a *two-way ANOVA*. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels. ANOVAs can also be used for within-group/repeated and between subjects designs. For this chapter we will focus on *between subject one-way ANOVA*.

In a One-Way ANOVA we compare two types of variance: the variance between groups and the variance within groups, which we will discuss in the next section.

Observing and Interpreting Variability

We have seen time and again that scores, be they individual data or group means, will differ naturally. Sometimes this is

due to random chance, and other times it is due to actual differences. Our job as scientists, researchers, and data analysts is to determine if the observed differences are systematic and meaningful (via a hypothesis test) and, if so, what is causing those differences. Through this, it becomes clear that, although we are usually interested in the mean or average score, it is the variability in the scores that is key.

Take a look at figure 1, which shows scores for many people on a test of skill used as part of a job application. The x-axis has each individual person, in no particular order, and the y-axis contains the score each person received on the test. As we can see, the job applicants differed quite a bit in their performance, and understanding why that is the case would be extremely useful information. However, there's no interpretable pattern in the data, especially because we only have information on the test, not on any other variable (remember that the x-axis here only shows individual people and is not ordered or interpretable).

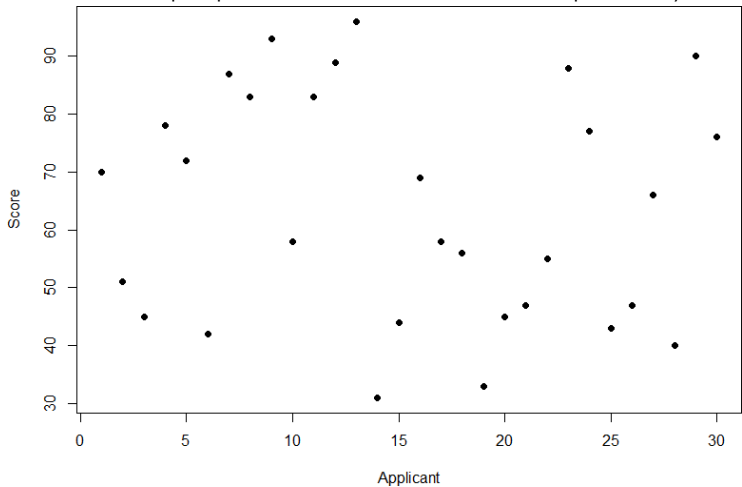


Figure 1. Scores on a job test

Our goal is to explain this variability that we are seeing in the dataset. Let's assume that as part of the job application procedure we also collected data on the highest degree each

applicant earned. With knowledge of what the job requires, we could sort our applicants into three groups: those applicants who have a college degree related to the job, those applicants who have a college degree that is not related to the job, and those applicants who did not earn a college degree. This is a common way that job applicants are sorted, and we can use ANOVA to test if these groups are actually different. Figure 2 presents the same job applicant scores, but now they are color coded by group membership (i.e. which group they belong in). Now that we can differentiate between applicants this way, a pattern starts to emerge: those applicants with a relevant degree (coded red) tend to be near the top, those applicants with no college degree (coded black) tend to be near the bottom, and the applicants with an unrelated degree (coded green) tend to fall into the middle. However, even within these groups, there is still some variability, as shown in Figure 2.

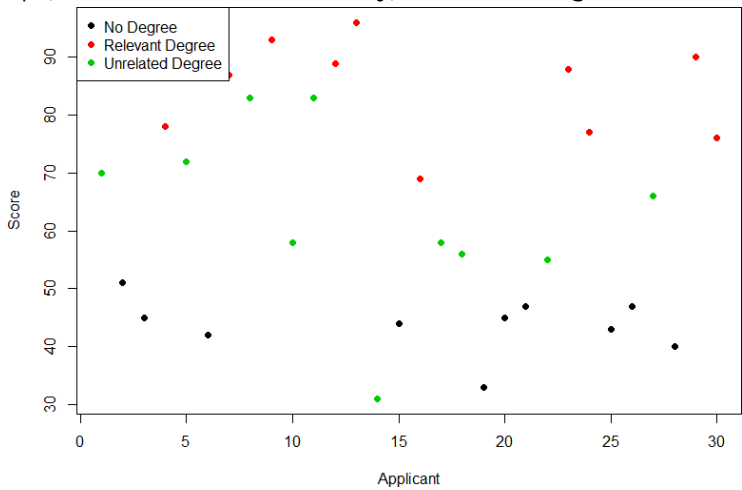


Figure 2. Applicant scores coded by degree earned

This pattern is even easier to see when the applicants are sorted and organized into their respective groups, as shown in Figure 3.

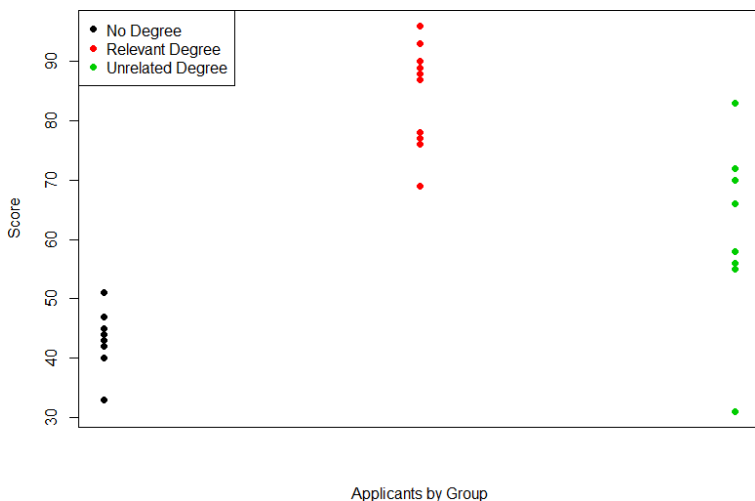


Figure 3. Applicant scores by group

Now that we have our data visualized into an easily interpretable format, we can clearly see that our applicants' scores differ largely along group lines. Those applicants who do not have a college degree received the lowest scores, those who had a degree relevant to the job received the highest scores, and those who did have a degree but one that is not related to the job tended to fall somewhere in the middle. Thus, we have systematic variance *between* our groups.

We can also clearly see that *within* each group, our applicants' scores differed from one another. Those applicants without a degree tended to score very similarly, since the scores are clustered close together. Our group of applicants with relevant degrees varied a little but more than that, and our group of applicants with unrelated degrees varied quite a bit. It may be that there are other factors that cause the observed score differences within each group, or they could just be due to random chance. Because we do not have any other explanatory data in our dataset, the variability we observe within our groups is considered random error, with any deviations between a person and that person's group mean caused only by chance. Thus, we have unsystematic (random)

variance *within* our groups.

The process and analyses used in ANOVA will take these two sources of variance (systematic variance between groups and random error within groups, or how much groups differ from each other and how much people differ within each group) and compare them to one another to determine if the groups have any explanatory value in our outcome variable. By doing this, we will test for statistically significant differences between the group means, just like we did for t -tests. We will go step by step to break down the math to see how ANOVA actually works.

ANOVA (analysis of variance) breaks down to:

$$F = \frac{\text{between group variance}}{\text{within group variance}}$$

where F is the new statistic reported for ANOVAs

Sources of Variance

ANOVA is all about looking at the different sources of variance (i.e. the reasons that scores differ from one another) in a dataset. Fortunately, the way we calculate these sources of variance takes a very familiar form: the Sum of Squares. Before we get into the calculations themselves, we must first lay out some important terminology and notation.

In ANOVA, we are working with two variables, a *grouping* or

explanatory variable and a *continuous outcome variable*. The **grouping variable** is our predictor (it predicts or explains the values in the outcome variable) or, in experimental terms, our *independent variable*, and it made up of k groups, with k being any whole number 2 or greater. That is, ANOVA requires two or more groups to work, and it is usually conducted with three or more. In ANOVA, we refer to groups as “levels”, so the number of levels is just the number of groups, which again is k . In the above example, our grouping variable was education, which had 3 levels, so $k = 3$. When we report any descriptive value (e.g. mean, sample size, standard deviation) for a specific group, we will use a subscript 1... k to denote which group it refers to. For example, if we have three groups and want to report the standard deviation s for each group, we would report them as s_1 , s_2 , and s_3 .

Our second variable is our *outcome variable*. This is the variable on which people differ, and we are trying to explain or account for those differences based on group membership. In the example above, our outcome was the score each person earned on the test. Our outcome variable will still use X for scores as before. When describing the outcome variable using means, we will use subscripts to refer to specific group means. So if we have $k = 3$ groups, our means will be \bar{X}_1 , \bar{X}_2 , and \bar{X}_3 . We will also have a single mean representing the average of all participants across all groups. This is known as the **grand mean**, and we use the symbol \bar{X}_G . These different means – the individual group means and the overall grand mean – will be how we calculate our sums of squares.

Finally, we now have to differentiate between several different sample sizes. Our data will now have sample sizes for each group, and we will denote these with a lower case “ n ” and a subscript, just like with our other descriptive statistics: n_1 , n_2 , and n_3 . We also have the overall sample size in our dataset, and

we will denote this with a capital N . The total sample size (N) is just the group sample sizes added together.

Between Groups Sum of Squares

One source of variability we can identify in Figure 3 of the above example was differences or variability between the groups. That is, the groups clearly had different average levels. The variability arising from these differences is known as the between groups variability, and it is quantified using Between Groups Sum of Squares.

Our calculations for sums of squares in ANOVA will take on the same form as it did for regular calculations of variance. Each observation, in this case the group means, is compared to the overall mean, in this case the grand mean, to calculate a deviation score. These deviation scores are squared so that they do not cancel each other out and sum to zero. The squared deviations are then added up, or summed. There is, however, one small difference. Because each group mean represents a group composed of multiple people, before we sum the deviation scores we must multiply them by the number of people within that group. Incorporating this, we find our equation for Between Groups Sum of Squares.

Between Groups Sum of Squares

$$SS_B = \sum n_j (\bar{X}_J - \bar{X}_G)^2$$

The subscript j refers to the “ j^{th} ” group where $j = 1 \dots k$ to keep track of which group mean and sample size we are working with. As you can see, the only difference between this equation and the familiar sum of squares for variance is that we are adding in the sample size. Everything else logically fits together in the same way.

Within Groups Sum of Squares

The other source of variability in the figures comes from differences that occur within each group. That is, each individual deviates a little bit from their respective group mean, just like the group means differed from the grand mean. We therefore label this source the Within Groups Sum of Squares. Because we are trying to account for variance based on group-level means, any deviation from the group means indicates an inaccuracy or error. Thus, our within groups variability represents our error in ANOVA.

The formula for this sum of squares is again going to take on the same form and logic. What we are looking for is the distance between each individual person and the mean of the group to which they belong. We calculate this deviation score, square it so that they can be added together, then sum all of

them into one overall value.

Sum of Squares within-group

$$SS_W = \sum (X_{ij} - \bar{X}_j)^2$$

In this instance, because we are calculating this deviation score for each individual person, there is no need to multiple by how many people we have. The subscript j again represents a group and the subscript i refers to a specific person. So, X_{ij} is read as “the i^{th} person of the j^{th} group.” It is important to remember that the deviation score for each person is only calculated relative to their group mean: do not calculate these scores relative to the other group means.

Total Sum of Squares

The Between Groups and Within Groups Sums of Squares represent all variability in our dataset. We also refer to the total variability as the Total Sum of Squares, representing the overall variability with a single number. The calculation for this score is exactly the same as it would be if we were calculating the overall variance in the dataset (because that’s what we are interested in explaining) without worrying about or even knowing about the groups into which our scores fall:

Total Sum of Squares

$$SS_T = \sum (X_i - \overline{X_G})^2$$

We can see that our Total Sum of Squares is just each individual score minus the grand mean. As with our Within Groups Sum of Squares, we are calculating a deviation score for each individual person, so we do not need to multiply anything by the sample size; that is only done for Between Groups Sum of Squares.

An important feature of the sums of squares in ANOVA is that they all fit together. We could work through the algebra to demonstrate that if we added together the formulas for SS_B and SS_W , we would end up with the formula for SS_T . That is:

$$SS_T = SS_B + SS_W$$

This will prove to be very convenient, because if we know the values of any two of our sums of squares, it is very quick and easy to find the value of

the third. It is also a good way to check calculations: if you calculate each SS by hand, you can make sure that they all fit together as shown above, and if not, you know that you made a math mistake somewhere.

We can see from the above formulas that calculating an ANOVA by hand from raw data can take a very, very long time. For this reason, you will not be required to calculate the SS values by hand, but you should still take the time to understand how they fit together and what each one represents to ensure you understand the analysis itself.

ANOVA Table

All of our sources of variability fit together in meaningful, interpretable ways as we saw above, and the easiest way to do this is to organize them into a table. The ANOVA table, shown in Table 1, is how we calculate our test statistic.

| Source | SS | df | MS |
|---------|--------|-------|------------|
| Between | SS_B | $k-1$ | MS_B |
| Within | SS_W | $N-k$ | MS_W |
| Total | SS_T | $N-1$ | (MS is var |

Table 1. ANOVA table.

The first column of the ANOVA table, labeled “Source”, indicates which of our sources of variability we are using: between groups, within groups, or total. The second column, labeled “SS”, contains our values for the sums of squares that we learned to calculate above. As noted previously, calculating these by hand takes too long, and so the formulas are not presented in Table 1. However, remember that the Total is the sum of the other two, in case you are only given two SS values and need to calculate the third.

The next column in Table 1, labeled “df”, is our degrees of freedom. As with the sums of squares, there is a different df for each group, and the formulas are presented in the table. Notice that the total degrees of freedom, $N - 1$, is the same as it was

for our regular variance. This matches the SS_T formulation to again indicate that we are simply taking our familiar variance term and breaking it up into difference sources. Also remember that the capital N in the df calculations refers to the overall sample size, not a specific group sample size. Notice that the total row for degrees of freedom, just like for sums of squares, is just the Between and Within rows added together. If you take $N - k + k - 1$, then the “ $- k$ ” and “ $+ k$ ” portions will cancel out, and you are left with $N - 1$. This is a convenient way to quickly check your calculations.

The third column, labeled “MS”, is our Mean Squares for each source of variance. A “mean square” is just another way to say variability. Each mean square is calculated by dividing the sum of squares by its corresponding degrees of freedom. Notice that we do this for the Between row and the Within row, but not for the Total row. There are two reasons for this. First, our Total Mean Square would just be the variance in the full dataset (put together the formulas to see this for yourself), so it would not be new information. Second, the Mean Square values for Between and Within would not add up to equal the Mean Square Total because they are divided by different denominators. This is in contrast to the first two columns, where the Total row

was both the conceptual total (i.e. the overall variance and degrees of freedom) and the literal total of the other two rows.

The final column in the ANOVA table (Table 1), labeled “F”, is our test statistic for ANOVA. The F statistic, just like a t - or z -statistic, is compared to a critical value to see whether we can reject for fail to reject a null hypothesis. Thus, although the calculations look different for ANOVA, we are still doing the same thing that we did in all of Unit 2. We are simply using a new type of data to test our hypotheses. We will see what these hypotheses look like shortly, but first, we must take a moment to address why we are doing our calculations this way.

ANOVA

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

We will typically work from having Sum of Squares calculated, but here are the basic formulas for the 3 types of Sum of Squares for the ANOVA:

1. Total sum of squares (SS_T): $\sum x^2 - (\sum x)^2/n$
2. Within sum of squares (SS_W): add up the sum of squares for each treatment condition
3. Between sum of squares (SS_B): $SST - SSW = SSB$

While there are other ways to calculate the SS s, these are the formulas we can use for this class if needed.

ANOVA and Type I Error

You may be wondering why we do not just use another t -test to test our hypotheses about three or more groups the way we did in Unit 2. After all, we are still just looking at group mean differences. The reason is that our t -statistic formula can only handle up to two groups, one minus the other. With only two groups, we can move our population parameters for the group means around in our null hypothesis and still get the same interpretation: the means are equal, which can also be concluded if one mean minus the other mean is equal to zero. However, if we tried adding a third mean, we would no longer be able to do this. So, in order to use t -tests to compare three or more means, we would have to run a series of individual group comparisons.

For only three groups, we would have three t -tests: group 1 vs group 2, group 1 vs group 3, and group 2 vs group 3. This may not sound like a lot, especially with the advances in technology that have made running an analysis very fast, but it quickly

scales up. With just one additional group, bringing our total to four, we would have six comparisons: group 1 vs group 2, group 1 vs group 3, group 1 vs group 4, group 2 vs group 3, group 2 vs group 4, and group 3 vs group 4. This makes for a logistical and computation nightmare for five or more groups. When we reject the null hypothesis in a one-way ANOVA, we conclude that the group means are not all the same in the population. But this can indicate different things. With three groups, it can indicate that all three means are significantly different from each other. Or it can indicate that one of the means is significantly different from the other two, but the other two are not significantly different from each other. For this reason, statistically significant one-way ANOVA results are typically followed up with a series of post hoc comparisons of selected pairs of group means to determine which are different from which others.

A bigger issue, however, is our probability of committing a Type I Error. Remember that a Type I error is a false positive, and the chance of committing a Type I error is equal to our significance level, α . This is true if we are only running a single analysis (such as a *t*-test with only two groups) on a single dataset.

However, when we start running multiple analyses on the same dataset, our Type I error rate increases, raising the probability that we are capitalizing on random chance and rejecting a null hypothesis when we should not. ANOVA, by comparing all groups simultaneously with a single analysis, averts this issue and keeps our error rate at the α we set.

Hypotheses in ANOVA

So far we have seen what ANOVA is used for, why we use it, and how we use it. Now we can turn to the formal hypotheses we will be testing. As with before, we have a null and an alternative

hypothesis to lay out. Our null hypothesis is still the idea of “no difference” in our data. Because we have multiple group means, we simply list them out as equal to each other:

H_0 : There is
no difference
in the group
means. H_0 :
 $\mu_1 = \mu_2 = \mu_3$

We list as many μ parameters as groups we have. In the example above, we have three groups to test ($k = 3$), so we have three parameters in our null hypothesis. If we had more groups, say, four, we would simply add another μ to the list and give it the appropriate subscript, giving us: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. Notice that we *do not* say that the means are all equal to zero, we only say that they are equal to one another; it does not matter what the actual value is, so long as it holds for all groups equally.

Our alternative hypothesis for ANOVA is a little bit different. Let’s take a look at it and then dive deeper into what it means:

H_A : At least 1 mean is different

The first difference is obvious: there is no mathematical statement of the alternative hypothesis in ANOVA. This is due to the second difference: we are not saying *which* group is going to be different, only that *at least one* will be. Because we do not hypothesize about which mean will be different, there is no way to write it mathematically. Related to this, we do not have directional hypotheses (greater than or less than) like we did with the z-statistic and t- statistics. Due to this, our alternative hypothesis is always exactly the same: at least one mean is different.

With t-tests, we saw that, if we reject the null

hypothesis, we can adopt the alternative, and this made it easy to understand what the differences looked like. In ANOVA, we will still adopt the alternative hypothesis as the best explanation of our data if we reject the null hypothesis. However, when we look at the alternative hypothesis, we can see that it does not give us much information. We will know that a difference exists somewhere, but we will not know where that difference is. The ANOVA is an **ominous test** meaning you just know there are differences. More specifically, at least 1 group is different from the rest. Is only group 1 different but groups 2 and 3 the same? Is it only group 2? Are all three of them different? Based on just our alternative hypothesis, there is no way to be sure. We will come back to this issue later and see how to find out specific differences. For now, just remember that we are testing for *any* difference in group means, and it does not matter where that difference occurs. Now that we have our hypotheses for ANOVA, let's work through an example. We will continue to use the data from Figures 1 through 3 for continuity.

Example: Scores on Job Application Tests

Our data come from three groups of 10 people each, all of whom applied for a single job opening: those with no college degree, those with a college degree that is

not related to the job opening, and those with a college degree from a relevant field. We want to know if we can use this group membership to account for our observed variability and, by doing so, test if there is a difference between our three group means ($k = 3$). We will follow the same steps for hypothesis testing as we did in previous chapters. Let's start, as always, with our hypotheses.

Step 1: State the Hypotheses

Our hypotheses are concerned with the means of groups based on education level, so:

H_0 : There is no difference between educational levels. $H_0: \mu_1 = \mu_2 = \mu_3$

H_A : At least 1 educational level is different.

Again, we phrase our null hypothesis in terms of what we are actually looking for, and we use a number of population parameters equal to our number of groups. Our alternative hypothesis is always exactly the same.

Step 2: Find the Critical Value

Our test statistic for ANOVA, as we saw above, is F . Because we are using a new test statistic, we will get a new table: the F distribution table, the top of which is shown in Figure 4:

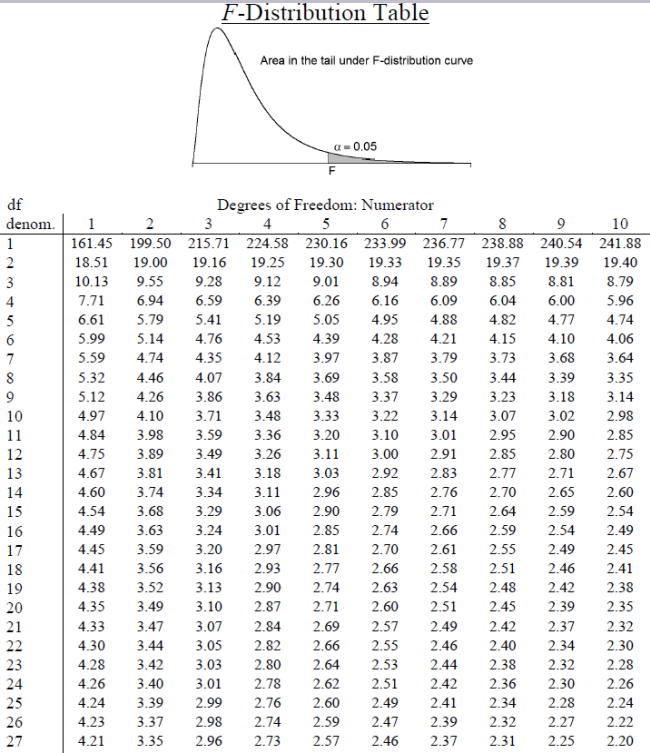


Figure 4. *F* distribution table.

The *F* table only displays critical values for $\alpha = 0.05$. This is because other significance levels are uncommon and so it is not worth it to use up the space to present them. There are now two degrees of freedom we must use to find our critical value: Numerator and Denominator. These correspond to the numerator and denominator of our test statistic, which, if you look at the ANOVA table presented earlier, are our Between Groups and Within Groups rows, respectively. The df_B

is the “Degrees of Freedom: Numerator” because it is the degrees of freedom value used to calculate the Mean Square Between, which in turn was the numerator of our F statistic. Likewise, the df_W is the “df denom.” (short for denominator) because it is the degrees of freedom value used to calculate the Mean Square Within, which was our denominator for F .

The formula for df_B is $k - 1$, and remember that k is the number of groups we are assessing. In this example, $k = 3$ so our $df_B = 2$. This tells us that we will use the second column, the one labeled 2, to find our critical value. To find the proper row, we simply calculate the df_W , which was $N - k$. The original prompt told us that we have “three groups of 10 people each,” so our total sample size is 30. This makes our value for $df_W = 27$. If we follow the second column down to the row for 27, we find that our critical value is 3.35. We use this critical value the same way as we did before: it is our criterion against which we will compare our obtained test statistic to determine statistical significance.

Step 3: Calculate the Test Statistic

Now that we have our hypotheses and the criterion we will use to test them, we can calculate our test statistic. To do this, we will fill in the ANOVA table. When we do so, we will work our way from left to right, filling in

each cell to get our final answer. Here will be are basic steps for calculating ANOVA:

1. 3 Sum of Square calculations
2. 3 degrees of freedom calculations
3. 2 variance calculations
4. 1 F – score

We will assume that we are given the SS values as shown below:

| Source | SS | d_f | S_M | F |
|---------|-----|-------|-------|---|
| Between | 824 | | | |
| n | 6 | | | |
| Within | 302 | | | |
| | 0 | | | |
| Total | | | | |

These may seem like random numbers, but remember that they are based on the distances between the groups themselves and within each group. Figure 5 shows the plot of the data with the group means and grand mean included. If we wanted to, we could use this information, combined with our earlier information that each group has 10 people, to calculate the Between Groups Sum of Squares by hand.

However, doing so would take some time, and without the specific values of the data points, we would not be able to calculate our Within Groups Sum of Squares, so we will trust that these values are the correct ones.

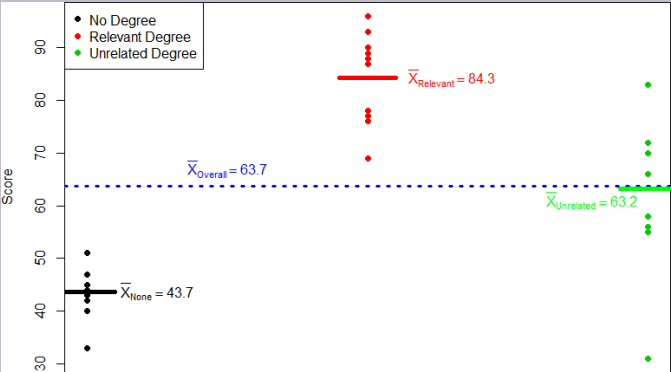


Figure 5. Means

We were given the sums of squares values for our first two rows, so we can use those to calculate the Total Sum of Squares.

| Source | SS | f^d | S^M | F |
|--------------|-----------------|-------|-------|---|
| Between n | 8246 | | | |
| Within | 3020 | | | |
| Total | 8246+3020=11266 | | | |

We also calculated our degrees of freedom earlier, so we can fill in those values. Additionally, we know that the total degrees of freedom is $N - 1$, which is 29. This value of 29 is also the sum of the other two degrees of freedom, so everything checks out.

| Source | SS | df | M_S | F |
|---------|-------|-----------|-------|---|
| Between | 8246 | $3-1=2$ | | |
| Within | 3020 | $29-2=27$ | | |
| Total | 11266 | $30-1=29$ | | |

Now we have everything we need to calculate our mean squares. Our MS values for each row are just the SS divided by the df for that row, giving us:

| Source | SS | df | MS |
|---------|-------|----|-------------|
| Between | 8246 | 2 | $8246/2 =$ |
| Within | 3020 | 27 | $3020/27 =$ |
| Total | 11266 | 29 | |

Remember that we do not calculate a Total Mean Square, so we leave that cell blank. Finally, we have the information we need to calculate our test statistic. F is our MS_B divided by MS_W .

| Source | SS | df | MS | F |
|---------|--------|----|--------|-------|
| Between | 82.46 | 2 | 41.23 | 36.86 |
| Within | 30.20 | 27 | 1.1185 | |
| Total | 112.66 | 29 | | |

Remember that we do not calculate a Total Mean Square, so we leave that cell blank. Finally, we have the information we need to calculate our test statistic. F is our MS_B divided by MS_W .

So, working our way through the table given only two SS values and the sample size and group size given before, we calculate our test statistic to be $F_{obt} = 36.86$, which we will compare to the critical value in step 4.

Step 4: Make a decision

Our obtained test statistic was calculated to be $F_{obt} = 36.86$ and our critical value was found to be $F^* = 3.35$. Our obtained statistic is larger than our critical value, so we can reject the null hypothesis.

Reject H_0 . Based on our 3 groups of 10 people, we can conclude that job test scores are statistically significantly

different based on education level, $F(2,27)$
 $= 36.86, p < .05$.

Notice that when we report F , we include both degrees of freedom. We always report the numerator then the denominator, separated by a comma. We must also note that, because we were only testing for any difference, we cannot yet conclude which groups are different from the others. We will do so shortly, but first, because we found a statistically significant result, we need to calculate an effect size to see how big of an effect we found.

Effect Size: Variance Explained

Recall that the purpose of ANOVA is to take observed variability and see if we can explain those differences based on group membership. To that end, our effect size will be just that: the variance explained. You can think of variance explained as the proportion or percent of the differences we are able to account for based on our groups. We know that the overall observed differences are quantified as the Total Sum of Squares, and that our observed effect of group membership is the Between Groups Sum of Squares. Our effect size, therefore, is the ratio of these to sums of squares.

Effect size, η^2 (eta-square) also known as R^2

The effect size η^2 or R^2 is called “eta-squared” and represents variance explained.

$$\eta^2 = \frac{SS_B}{SS_T}$$

or stated as

$$R^2 = \frac{SS_{between}}{SS_{total}}$$

Eta-square is reported as percentage of variance of the outcome/dependent variable explained by the predictor/independent variable.

Although you report variance explained by the predictor/independent variable, you can also use the η^2 guidelines for effect size:

| | | |
|---|--|--|
| | | |
| ◇2 | | |
| 0.01 | | |
| 0.09 | | |
| 0.25 | | |
| Note: if less than .01, no effect is reported | | |

Example continued adding on effect size for scores on job application tests

For our example, $SS_B = 8246$ and $SS_T = 11266$, our values give an effect size, η^2 , of:

$$\eta^2 = \frac{8246}{11266} = 0.73$$

So, we are able to explain 73% of the variance in job test scores based on education. This is, in fact, a huge effect size, and most of the time we will not explain nearly that much variance.

So, we found that not only do we have a statistically significant result, but that our observed effect was very large! However, we still do not know specifically which groups are different

from each other. It could be that they are all different, or that only those who have a relevant degree are different from the others, or that only those who have no degree are different from the others. To find out which is true, we need to do a special analysis called a post hoc test.

Post Hoc Tests

A post hoc test is used only after we find a statistically significant result and need to determine where our differences truly came from. The term “post hoc” comes from the Latin for “after the event”. There are many different post hoc tests that have been developed, and most of them will give us similar answers.

Post hoc testing is NOT running a series of independent-samples t tests comparing each group mean to each of the other group means. As discussed earlier, if we conduct several t -tests when the null hypothesis is true, the chance of mistakenly rejecting at least one null hypothesis increases with each test we conduct. This is a similar issue as explained with ANOVA and Type I Error. This referred to **experiment-wise error**. Instead we have a few options to determine significant differences between the groups. We will only focus here on the most commonly used ones. Further we will only discuss the concepts behind each and will not worry about calculations. (Note: these all would be run in statistical analysis software — and so would the ANOVA!)

Bonferroni Test

A Bonferroni test is perhaps the simplest post hoc analysis. A Bonferroni test is a series of t -tests performed on each pair of groups. As we discussed earlier, the number of groups quickly

grows the number of comparisons, which inflates Type I error rates. To avoid this, a Bonferroni test divides our significance level α by the number of comparisons we are making so that when they are all run, they sum back up to our original Type I error rate. Once we have our new significance level, we simply run independent samples t -tests to look for difference between our pairs of groups. This adjustment is sometimes called a Bonferroni Correction, and it is easy to do by hand if we want to compare obtained p -values to our new corrected α level, but it is more difficult to do when using critical values like we do for our analyses so we will leave our discussion of it to that.

Tukey's Honest Significant Difference

Tukey's Honest Significant Difference (HSD) is a very popular post hoc analysis. This analysis, like Bonferroni's, makes adjustments based on the number of comparisons, but it makes adjustments to the test statistic when running the comparisons of two groups. These comparisons give us an estimate of the difference between the groups and a confidence interval for the estimate. We use this confidence interval in the same way that we use a confidence interval for a regular independent samples t -test: if it contains 0.00, the groups are not different, but if it does not contain 0.00 then the groups are different.

Example continued adding on post hoc for scores on job application tests: Tukey

Remember we are comparing scores from those whom applied for a single job opening: those with no college degree (none), those with a college degree that is not related to the job opening (unrelated), and those with a college degree from a relevant field (relevant).

Tukey

Below are the differences between the group means and the Tukey's HSD confidence intervals for the differences:

| Comparison | Difference | Tukey's HSD CI |
|-----------------------|------------|----------------|
| None vs Relevant | 40.60 | (28.87, 52.33) |
| None vs Unrelated | 19.50 | (7.77, 31.23) |
| Relevant vs Unrelated | 21.10 | (9.37, 32.83) |

As we can see, none of these intervals contain 0.00, so we can conclude that all three groups are different from one another.

Scheffe's Test

Another common post hoc test is Scheffe's Test. Like Tukey's HSD, Scheffe's test adjusts the test statistic for how many comparisons are made, but it does so in a slightly different way. The result is a test that is "conservative," which means that it is less likely to commit a Type I Error, but this comes at the cost of less power to detect effects. We can see this by looking

at the confidence intervals that Scheffe's test gives us for our example.

Example continued adding on post hoc for scores on job application tests: Scheffe

Scheffe

Below are the differences between the group means and the Scheffe confidence intervals for the differences:

| Comparison | Difference | Scheffe's CI |
|-----------------------|------------|----------------|
| None vs Relevant | 40.60 | (28.35, 52.85) |
| None vs Unrelated | 19.50 | (7.25, 31.75) |
| Relevant vs Unrelated | 21.10 | (8.85, 33.35) |

As we can see, these are slightly wider than the intervals we got from Tukey's HSD. This means that, all other things being equal, they are more likely to contain zero. In our case, however, the results are the same, and we again conclude that all three groups differ from one another.

There are many more post hoc tests than just these three, and they all approach the task in different ways, with some being more conservative and others being more powerful. In general, though,

they will give highly similar answers. What is important here is to be able to interpret a post hoc analysis. If you are given post hoc analysis confidence intervals, like the ones seen above, read them the same way we read confidence intervals previously comparing two groups: if they contain zero, there is no difference; if they do not contain zero, there is a difference.

Other ANOVA Designs

We have only just scratched the surface on ANOVA in this chapter. There are many other variations available for the one-way ANOVA presented here. There are also other types of ANOVAs that you are likely to encounter. The first is called a **factorial ANOVA**. Factorial ANOVAs use multiple grouping variables, not just one, to look for group mean differences. Just as there is no limit to the number of groups in a one-way ANOVA, there is no limit to the number of grouping variables in a Factorial ANOVA, but it becomes very difficult to find and interpret significant results with many factors, so usually they are limited to two or three grouping variables with only a small number of groups in each. Another ANOVA is called a **Repeated Measures ANOVA**. This is an extension of a repeated measures or matched pairs t-test, but in this case we are measuring each person three or more times to look for a change. We can even combine both of these advanced ANOVAs into mixed designs to test very specific and valuable questions. These topics are far beyond the scope of this text, but you should know about their existence.

Our treatment of ANOVA here is a small first step into a much larger world!

Learning Objectives

Having read the chapter, students should be able to:

- understand the basic purpose for analysis of variance (ANOVA) and the general logic that underlies the statistical procedure
- perform an ANOVA to evaluate data from a single factor, between subjects research design
- understand when post hoc tests are necessary and purpose that they serve
- calculate and interpret effect size

Exercises – Ch. 14

1. What are the three pieces of variance analyzed in ANOVA?
2. What does rejecting the null hypothesis in ANOVA tell us? What does it not tell us?
3. What is the purpose of post hoc tests?
4. Based on the ANOVA table below, do you reject or fail to reject the null hypothesis? What is the effect size?

| Source | SS | df | MS | F |
|---------|--------|----|-------|------|
| Between | 60.72 | 3 | 20.24 | 3.88 |
| Within | 213.61 | 41 | 5.21 | |
| Total | 274.33 | 44 | | |

5. Finish filling out the following ANOVA tables:

Problem 1: N = 14

| Source | SS | df | MS | F |
|---------|-------|----|-------|---|
| Between | | 2 | 14.10 | |
| Within | | | | |
| Total | 64.65 | | | |

Problem 2:

| Source | SS | df | MS | F |
|---------|----|----|------|-------|
| Between | | 2 | | 42.36 |
| Within | | 54 | 2.48 | |
| Total | | | | |

6. You know that stores tend to charge different prices for similar or identical products, and you want to test whether or not these differences are, on average, statistically significantly different. You go online and collect data from 3 different stores, gathering information on 15 products at each store. You find that the average prices at each store are: Store 1 M = \$27.82, Store 2 M= \$38.96, and Store 3 M = \$24.53. Based on the overall variability in the products and the variability within each store, you find the following values for the Sums of Squares: SST = 683.22, SSW = 441.19. Complete the ANOVA table and use the 4 step hypothesis testing procedure to see if there are systematic price differences between the stores.

7. You and your friend are debating which type of candy is the best. You find data on the average rating for hard candy (e.g. jolly ranchers, $\bar{X} = 3.60$), chewable candy (e.g. starburst, $\bar{X} = 4.20$), and chocolate (e.g. snickers, $\bar{X} = 4.40$); each type of candy was rated by 30 people. Test for differences in average candy rating using SSB = 16.18 and SSW = 28.74.

8. Administrators at a university want to know if students in

different majors are more or less extroverted than others. They provide you with data they have for English majors ($\bar{X} = 3.78$, $n = 45$), History majors ($\bar{X} = 2.23$, $n = 40$), Psychology majors ($\bar{X} = 4.41$, $n = 51$), and Math majors ($\bar{X} = 1.15$, $n = 28$). You find the $SSB = 75.80$ and $SSW = 47.40$ and test at $\alpha = 0.05$.

9. You are assigned to run a study comparing a new medication ($\bar{X} = 17.47$, $n = 19$), an existing medication ($\bar{X} = 17.94$, $n = 18$), and a placebo ($\bar{X} = 13.70$, $n = 20$), with higher scores reflecting better outcomes. Use $SSB = 210.10$ and $SSW = 133.90$ to test for differences.

10. You are in charge of assessing different training methods for effectiveness. You have data on 4 methods: Method 1 ($\bar{X} = 87$, $n = 12$), Method 2 ($\bar{X} = 92$, $n = 14$), Method 3 ($\bar{X} = 88$, $n = 15$), and Method 4 ($\bar{X} = 75$, $n = 11$). Test for differences among these means, assuming $SSB = 64.81$ and $SST = 399.45$.

Answers to Odd- Numbered Exercises – Ch. 14

1. Variance between groups (SSB), variance within groups (SSW) and total variance (SST).

3. Post hoc tests are run if we reject the null hypothesis in ANOVA; they tell us which specific group differences are significant.5. Finish filling out the following ANOVA tables:

Problem 1:

| Source | SS | df | MS | F |
|---------|-------|----|-------|------|
| Between | 28.20 | 2 | 14.10 | 4.26 |
| Within | 36.45 | 11 | 3.31 | |
| Total | 64.65 | 13 | | |

Problem 2:

| Source | SS | df | MS |
|---------|--------|----|--------|
| Between | 210.10 | 2 | 105.05 |
| Within | 133.92 | 54 | 2.48 |
| Total | 344.02 | | |

7. Step 1: $H_0: \mu_1 = \mu_2 = \mu_3$ "There is no difference in average rating of candy quality", H_A : "At least one mean is different."

Step 2: 3 groups and 90 total observations yields $df_{num} = 2$ and $df_{den} = 87$, $\alpha = 0.05$, $F^* = 3.11$.

Step 3: based on the given SSB and SSW and the computed df from step 2, is:

| Source | SS | df | MS | F |
|---------|-------|----|------|-------|
| Between | 16.18 | 2 | 8.09 | 24.52 |
| Within | 28.74 | 87 | 0.33 | |
| Total | 44.92 | 89 | | |

Step 4: $F > F^*$, reject H_0 . Based on the data in our 3 groups, we can say that there is a statistically significant difference in the quality of different types of candy, $F(2,87) = 24.52$, $p < .05$. Since the result is significant, we need an effect size: $\eta^2 = 16.18/44.92 = .36$, which is a large effect.

9. Step 1: $H_0: \mu_1 = \mu_2 = \mu_3$ "There is no difference in average outcome based on treatment", H_A : "At least one mean is different."

Step 2: 3 groups and 57 total participants yields $df_{num} = 2$ and $df_{den} = 54$, $\alpha = 0.05$, $F^* = 3.18$.

Step 3: based on the given SSB and SSW and the computed df from step 2, is:

| Source | SS | df | MS | F |
|---------|--------|----|--------|-------|
| Between | 210.10 | 2 | 105.02 | 42.36 |
| Within | 133.90 | 54 | 2.48 | |
| Total | 344.00 | 56 | | |

Step 4: $F > F^*$, reject H_0 . Based on the data in our 3 groups, we can say that there is a statistically significant difference in the effectiveness of the treatments, $F(2,54) = 42.36$, $p < .05$. Since the result is significant, we need an effect size: $\eta^2 = 210.10/344.00 = .61$, which is a large effect.

15. Chapter 15: 2 - Factor ANOVA

A single factor ANOVA is the statistical analysis appropriate when we are analyzing the results of an experiment in which we have one factor and are looking for differences in the response variable among three or more groups, each of which is receiving different levels or amounts of the factor. In chapter 14, we learned about the single factor ANOVA, also known as the one-way. We will now conceptually review a multi-factor ANOVA. We will keep it on the simpler side and use 2-factors (two independent/predictor variables) using a between-subjects design.

Logic of a 2 Factor ANOVA

A two factor ANOVA is used when we believe that more than one factor may affect a particular response (dependent) variable. For example, believe that the age of an adolescent will have an impact on number of phone calls made to the opposite sex and I also suspect that gender of the adolescent will have an impact on the number of phone calls made to the opposite sex.

To test my hypothesis that Age and Gender of adolescent will impact the number of phone calls made to the opposite sex in the past week. In this case, we have a between-subjects design for both age and gender. I have 2 conditions/levels/groups for each factor/variable. I will have to collect data for these for 4 samples of subjects:

| Age | Gender | |
|-----|-------------|-------|
| | Teen Males | Teen |
| | Older Males | Older |

Table 1. Example of 2×2 ANOVA

A 2×2 ANOVA gives you 4 conditions. *Note:* one way to identify the total conditions in a factorial study is to multiply the conditions for each factor. Thus, a 2×2 design is 2 times 2 giving us 4 total conditions for the study. We will discuss this more in a moment.

Remember that there are different types of ANOVAs based on design. In this case, we have a between-subjects design. An individual can only be in 1 condition for gender and 1 condition for age. So among the 4 total conditions/levels/groups between the 2 factors, an individual is only in 1 of the samples. For a between-subjects design, there are 4 different samples. Two Factor ANOVA data is commonly organized like the table above and is referred to a **matrix**. When the data is organized in a matrix it is very easy to see the factors, as well as the separate levels of the factors.

- **Factorial designs** like the 2-Factor ANOVA allow a researcher to examine more than one independent variable on the dependent variable
 - Individually for each factor, reporting out a F for each
 - Collectively where the collective influence of the factors is referred to as an **interaction**. An interaction is the result of the two independent variables combining to produce a result that is different from a result that is produced by either variable alone.
- A 2-Factor ANOVA allows a researcher to assess the main

effects (the independent variables) and the interaction yielding three outcomes (3 Fs), a F for factor 1, a F for factor 2 and an interaction between factor 1 and 2.

Let's go back to our example:

- Main Effect of Factor A
 - Is there a significant effect of age of teen (Factor A) on number of phone calls made to the opposite sex (response variable).
- Main Effect of Factor B
 - Is there a significant effect of sex of the teen (Factor B) on number of phone calls made to the opposite sex (response variable).
- Interaction of AxB
 - Does the effect of age of teen (Factor A) on the number of phone calls made to the opposite sex (response variable) depend on the sex of the teen (Factor B)?

Conducting a Two Factor ANOVA

Before we begin the process of calculating a 2-Factor ANOVA we need to review several key elements of the study:

- **Factors:** the independent variables/predictors
- **Levels of each factor:** how many conditions/groups/treatments a factor has
- **Response variable:** this is the dependent variable/outcome variable/measurement taken

- **Total number of condition in the experiment:** this is identified by multiplying out the number of levels for each factor
- **Number of subjects per condition, n:** how many participants are in each level/group/treatment
- **Total number of experiment participants, N:** this will be determined by type of factor for each. In a between-group design, there will be four different conditions of participants. In a complete repeated measures design, all participants are in all conditions. In mixed design, it will vary by the study design for each factor. For this chapter, we are focused on a between subjects design.

Remember that in experiments that are designed to test for a cause and effect relationship between two variables (experimental designs) the factor is the variable hypothesized to cause something to happen. The **response variable** is the variable we believe will be affected (changed) by the factor.

Level of each factor refers to the categories of a factor represented in the experiment. In our example of age and gender the number of levels was 2×2 – we refer to design by its *levels* (can also call them conditions/groups/treatments).

| Age | Gender | |
|-----|-------------|---------------|
| | Teen Males | Teen Females |
| | Older Males | Older Females |

Our example from Table 1 was a 2×2 design because there were two levels of the age variable (i.e., younger and older) and two levels of gender (i.e., male and female).

Total number of groups in the experiment equals the number of levels in Factor A multiplied by the number of levels for Factor B. For our example, there were four conditions. Another way to think about the number of groups or conditions is the number of cells in the matrix.

In a factorial design like the 2-Factor ANOVA, the number of subjects per condition is denoted by n and the total number of experiment participants is denoted by N . For example, if each condition has 10 participants, then $n = 10$. The experiment would have $N = 40$. In other words, 4 conditions with 10 participations ($n = 10$) (4×10) = 40 participants in the study.

Hypothesis Testing

We use the same steps for 2- Factor ANOVA that we have used for all other test statistics.

Write the alternative and null hypotheses

- 3 separate set of hypotheses: one set for each F
 - A effect (factor 1)
 - B effect (factor 2)
 - Interaction (A x B or factor 1 x 2)

These are three separate ANOVA tests yielding 3 Fs that are independent and the results are unrelated to the outcome for either of the other two. The hypotheses are set up in the same way as chapter 14. We will see an example for an interaction later in the chapter.

Set criteria for decision making

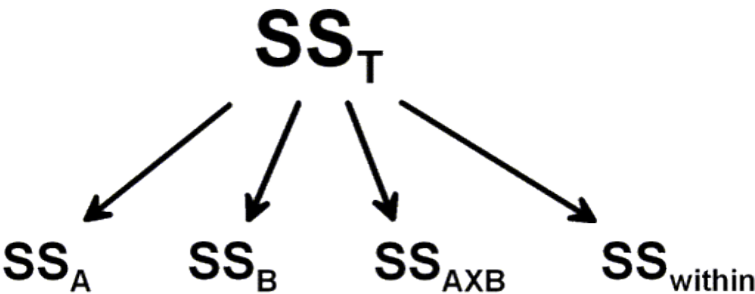
There are three hypotheses and three F scores so there will be three critical boundaries. The critical boundary of F comes from the F distribution table.

We need to know:

- Alpha (α)
- degrees of freedom Factor A = $df_A = (k_A - 1)$ where k_A is number of levels
- degrees of freedom Factor B = $df_B = (k_B - 1)$ where k_B is number of levels
- degrees of freedom Interaction (A x B) = $df_{A*B} = k_A * k_B$
- degrees of freedom for within treatment = $df_{total} - (df_A + df_B + df_{A*B})$ [within treatment is also called error]
- degrees of freedom total = $df_{total} = N - 1$ where N is the total number of scores

Note: We would still use the critical value ANOVA table for the critical F-values. The critical values may not be the same for each hypothesis; it will depend on the number of rows and columns used in the study! We will see this in an example later in the chapter.

Sample data are collected and analyzed by performing statistics (calculations for our adjusted step 3)



In the first stage of calculations Sum of Squares (SS) Total is calculated and then separated into the two components SS Between Treatments and SS Within Treatments.

In the second stage the SS Between Treatments is separated into the three factors: Factor A, Factor B & Factor A X B (interaction factor)

| Source | SS | df |
|-------------------------|---------------------------------|---|
| Between Treatment (b/t) | $SS_A + SS_B + SS_{A \times B}$ | $(k_A - 1) + (k_B - 1)$ |
| Factor A | (identify from info. given) | $(k_A - 1)$ |
| Factor B | (identify from info. given) | $(k_B - 1)$ |
| Interaction | (identify from info. given) | $(k_A)(k_B - 1)$ |
| Within Treatment (w/i) | $SS_{total} - SS_{b/t}$ | $df_{total} - df_{b/t}$ or $N - df_{b/t}$ |
| Total | $SS_{Between} + SS_{within}$ | $N - 1$ |

Table 2. ANOVA summary table with calculations

Note: In real life, we would run this through a statistical

program with the raw data to calculate the Fs! We are focusing conceptually on calculating the 3 Fs for a two-way factorial ANOVA. Notice that in Table 2, the Sum of Squares Between is adding up the Sum of Squares from each of the factors. You also see that to get to our F-ratios, we need the Mean Squares (just like chapter 14). We have an F for each: Factor A, Factor B and the Interaction Factor. The calculations for Sum of Square for the factors can be found by knowing the df and MS, or knowing the Sum of Squares Between.

You would also be most likely given the means and standard deviations for the 4 study conditions. Here is an example from Table 1 (made up data). You will see the main value as the mean and the standard deviation in parentheses.

| Age | Gender | |
|-----|----------------------------|--------------------------------|
| | Teen Males M = 3.5 (.3) | Teen Females M = 4.5 (.25) |
| | Older Males M = 8 (.5) | Older Females M = 12.5 (.8) |

Table 2. Means and Standard Deviations example from Table 1 study design.

Possible outcomes for 2-way factorial ANOVA

In a 2 X 2 design, there are eight possibilities:

1. A main effect for factor A only
2. A main effect for factor B only
3. Main effects for A and B only
4. A main effect for A, plus an interaction
5. A main effect for B, plus an interaction
6. Main effects for both A and B, plus an interaction
7. An interaction only, no main effects

8. No main effects, no interaction

Figure 1 shows examples of how these findings might look graphing the means of each of the 4 study conditions in a 2x2 design.

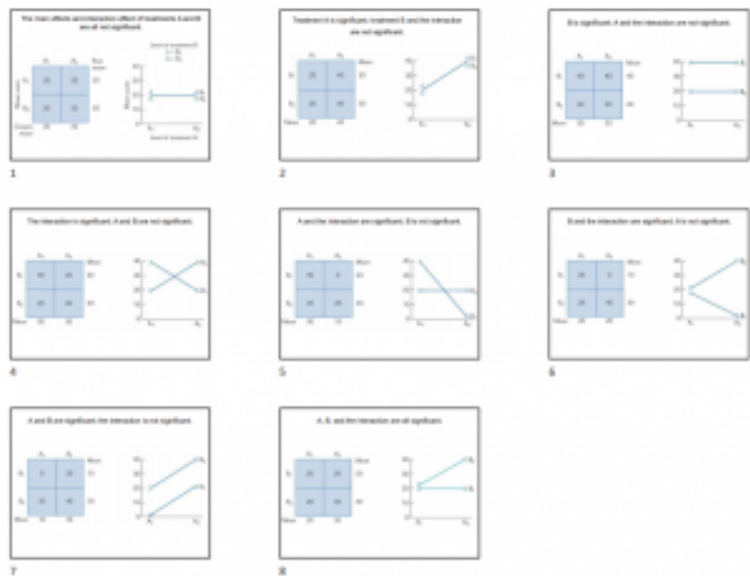


Figure 1. Examples of eight outcomes of a 2x2 ANOVA

Make your decision and explain the results (adjusted step 4).

- When making a statistical decision you should begin by looking for patterns in the means from each of the total conditions rather than focusing on the main effects or the interaction. After identifying patterns begin interpreting with the interaction effects first.
- Interaction means that the effect of one factor depends on the level of a second factor – so then there is no consistent main effect. If you get a significant interaction, emphasize

that finding over any significant main effects. In other words, if there is an interaction effect, then the main effect cannot be discussed without a qualifier.

Calculate effect size

- Effect size is calculated for each F that is statistically significant.
- Effect size reported is typically eta-square. Remember that from chapter 14, eta-square is the percentage of total variance explained variance by the factor. Again, just as you have a F for factor A, a F for factor B, and an F for the interaction, you would have eta-squares for each.

Graphing the Results of Factorial Experiments

The results of factorial experiments with two independent variables can be graphed by representing one independent variable on the x-axis and representing the other by using different kinds of bars or lines. The y-axis is always reserved for the dependent variable.

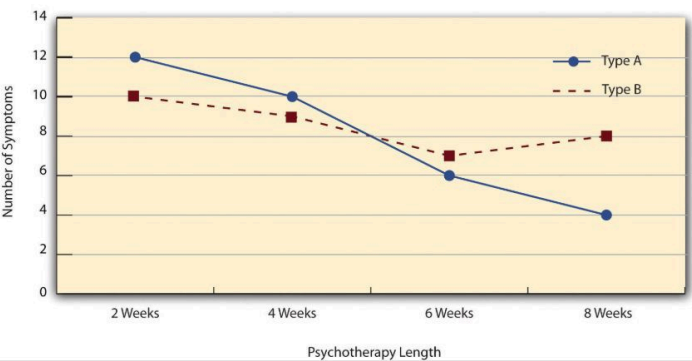


Figure 2. A 4 (Psychotherapy Length) x 2 (Type) ANOVA.
The figure above is a line graph that shows results for

a hypothetical 4 x 2 factorial experiment Psychotherapy length, is represented along the x-axis and has four levels (e.g., 2 weeks, 4 weeks, 6 weeks and 8 weeks) and the other variable (psychotherapy type) is represented by differently formatted lines.

Advantages & Disadvantages

Considerations

A 2-Factor ANOVA design is relatively easy to carry out and requires fewer subjects than other types of designs. There is no pre-testing necessary because one group could serve as the control. Although identifying sample sizes and study design for power is an important consideration using a factorial ANOVA.

Disadvantages

A 2-Factor ANOVA using a between-subjects design provides little information about the effect of the independent variable. The statistic provides information about whether the two groups differed (on average) and in which direction but it is not sensitive to individual differences. Other considerations for 2-Factor ANOVAs include using a *repeated measures ANOVA*. In this case for a 2-factor ANOVA, each person would be in every condition. So if you had a 2x2 an individual would be in all 4 study conditions. Another considerations is having a *mixed design*. For a mixed design, one factor would be between-subjects and the other would be within-subjects (repeated measures). For example, you might wish to conduct a 2x2 study on drug therapy. You can examine gender differences as one factor and type of drug as the other factor. Participants are only in 1 gender category but would receive both types of drug. A mixed design would give you individual differences in how each participant responded to the drug, but also has some of the challenges of using a within-subjects design (see short

discussion in chapter 12 on advantages and disadvantages of using a repeated measures design).

Learning Objectives

Having read the chapter, students should be able to:

- Explain the concept of a two-factor research design and recognize a matrix with levels of one factor being rows and levels of the second factor being columns
- Explain main effects and interactions in a two-factor ANOVA including patterns of findings
- Complete a ANOVA table given some information from the study
- Interpret effect size

Exercises – Ch 15

1. True or false. The bigger the differences between the sample means, the more likely it is that at least one of the Fs will be significant.
2. True or false. The advantage of combining two factors into a single research study is that the two factor study provides information about the interaction of the two factors and the main effects of each factor.
3. Complete the ANOVA table given this is a 2×3 ANOVA (two-way ANOVA; factor A = 2 levels with $n = 5$; factor B = 3 levels with $n = 5$; $N = 30$)

| Source | SS | df | MS | F |
|-------------------|-----|----|----|---|
| Between Treatment | 60 | | | |
| Factor A | | | | 5 |
| Factor B | | | | |
| Interaction | 30 | | | |
| Within Treatment | | | 2 | |
| Total | 108 | | | |

4. What is the df for factor A, B and AxB for the following?
What are the corresponding F-critical values?
- factor A $n=14$; factor B $n = 18$; $N = 32$

Answers to Exercises – Ch 15

- true
- 3.

| Source | SS | df | MS | F |
|-------------------|-----|----|----|-----|
| Between Treatment | 60 | 5 | | |
| Factor A | 10 | 1 | 10 | 5 |
| Factor B | 20 | 2 | 10 | 5 |
| Interaction | 30 | 2 | 15 | 7.5 |
| Within Treatment | 48 | 24 | 2 | |
| Total | 108 | 29 | | |

16. Chapter 16:

Correlations

Hypothesis testing beyond t-tests and ANOVAs

All of our analyses thus far have focused on comparing the value of a continuous variable across different groups via mean differences (t-tests and ANOVAs). These next few chapters will take you beyond having the predictor variable as categorical (nominal) with a continuous (interval/ratio) outcome variable. We will continue to use the same hypotheses testing logic and procedures with new types of data.

The type of data we have used in most chapters (except chapter 15) is **bivariate** data — “bi” for two variables. In reality, statisticians use **multivariate** data, meaning many variables. In this lesson, you will be studying correlation which is the relationship between two variables. We will also be covering the simplest form of regression – linear regression – with one independent variable (x). This chapter is focused on how to assess the relation between two continuous variables in the form of correlations. As we will see, the logic behind correlations is the same as it was group means (focus on previous chapters with hypothesis testing), but we will now have the ability to assess an entirely new data structure.

There are several different types of correlation coefficients. A correlation coefficient is a measure that varies from -1 to 1, where a value of 1 represents a perfect positive relationship

between the variables, 0 represents no relationship, and -1 represents a perfect negative relationship.

In this chapter we will focus on **Pearson's r** , which is a measure of the strength of the linear relationship between two *continuous variables*. r was developed by Karl Pearson in the early 1900s. We will see r as a way to quantify the relation between two variables to describe a linear relationship.



Karl Pearson at his desk [Source](#)

Figure 1 shows examples of various levels of correlation using randomly generated data for two continuous variables (reporting Pearson's r s). We will learn more about interpreting a correlation coefficient when we discuss *direction* and *magnitude* later in the chapter.

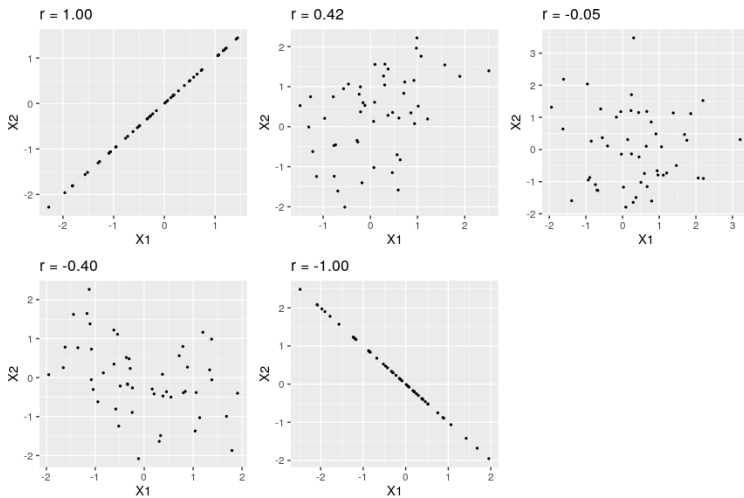


Figure 1: Examples of various levels of Pearson's r .

Variability and Covariance

A common theme throughout statistics is the notion that individuals will differ on different characteristics and traits, which we call variance. In inferential statistics and hypothesis testing, our goal is to find systematic reasons for differences and rule out random chance as the cause. By doing this, we are using information on a different variable – which so far has been group membership like in ANOVA – to explain this variance. In correlations, we will instead use a continuous variable to account for the variance. Because we have two continuous variables, we will have two characteristics or score on which people will vary. What we want to know is do people vary on the scores together. That is, as one score changes, does the other score also change in a predictable or consistent way? This notion of variables differing together is called covariance (the prefix “co” meaning “together”).

Let's look at our formula for sample variance on a single

variable (learned in chapter 4):

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$

We use X to represent a person's score on the variable at hand, and \bar{X} to represent the mean of that variable. The numerator of this formula is the *Sum of Squares*, which we have seen several times for various uses. Recall that squaring a value is just multiplying that value by itself. Thus, we can write the same equation but use $\Sigma(X - \bar{X})(X - \bar{X})$ on top. *This is the same formula* and works the same way as before, where we multiply the deviation score by itself (we square it) and then sum across squared deviations.

Now, let's look at the formula for *covariance*. In this formula, we will still use X to represent the score on one variable, and we will now use Y to represent the score on the second variable. We will still use bars to represent averages of the scores.

The formula for covariance (cov_{XY} with the subscript XY to indicate covariance across the X and Y variables) is:

Covariance sample formula:

$$Cov(X, Y) = \frac{\Sigma(X_i - \bar{X}) * (Y_i - \bar{Y})}{n - 1}$$

As we can see, this is the exact same structure as the previous

formula. Now, instead of multiplying the deviation score by itself on one variable, we take the deviation scores from a single person on each variable and multiply them together. We do this for each person (exactly the same as we did for variance) and then sum them to get our numerator. The numerator in this is called the **Sum of Products**.

Sum of Products formula:

$$SP = \sum (X_i - \bar{X}) * (Y_i - \bar{Y})$$

We will calculate the sum of products using the same table we used to calculate the sum of squares. In fact, the table for sum of products is simply a sum of squares table for X, plus a sum of squares table for Y, with a final column of products, as shown below.

| X | (X - \bar{X}) | (X - \bar{X}) ² | Y | (Y - \bar{Y}) | (Y - \bar{Y}) ² | (X - \bar{X})(Y - \bar{Y}) |
|-----|------------------|-------------------------------|-----|------------------|-------------------------------|----------------------------------|
| | | (if need s ²) | | | (if need s ²) | |
| ... | ... | ... | ... | ... | ... | ... |
| | | | | | | Σ (total up) |

Table 1. Example for calculating Sum of Products

This table works the same way that it did before (remember that the column headers tell you exactly what to do in that column). We list our raw data for the X and Y variables in the X and Y columns, respectively, then add them up so we can calculate the mean of each variable. We then take those

means and subtract them from the appropriate raw score to get our deviation scores for each person on each variable, and the columns of deviation scores will both add up to zero. We will square our deviation scores for each variable to get the sum of squares for X and Y so that we can compute the variance and standard deviation of each (we will use the standard deviation in our equation below). Finally, we take the deviation score from each variable and multiply them together to get our product score. Summing this column will give us our sum of products. It is very important that you multiply the raw deviation scores from each variable, NOT the squared deviation scores. The squared deviation scores are included in case standard deviation (s) or variance (s^2) are needed).

Our sum of products will go into the numerator of our formula for covariance, and then we only have to divide by $n - 1$ to get our covariance. *Unlike the sum of squares, both our sum of products and our covariance can be positive, negative, or zero, and they will always match (e.g. if our sum of products is positive, our covariance will always be positive).* A positive sum of products and covariance indicates that the two variables are related and move in the same direction. That is, as one variable goes up, the other will also go up, and vice versa. A negative sum of products and covariance means that the variables are related but move in opposite directions when they change, which is called an inverse relation. In an inverse relation, as one variable goes up, the other variable goes down. If the sum of products and covariance are zero, then that means that the variables are not related. As one variable goes up or down, the other variable does not change in a consistent or predictable way.

The previous paragraph brings us to an important definition about relations between variables. What we are looking for in a relation is a consistent or predictable pattern. That is, the

variables change together, either in the same direction or opposite directions, in the same way each time. It doesn't matter if this relation is positive or negative, only that it is not zero. If there is no consistency in how the variables change within a person, then the relation is zero and does not exist. We will revisit this notion of direction vs zero relation later on.

Visualizing Relations

Chapter 3 covered many different forms of data visualization, and visualizing data remains an important first step in understanding and describing out data before we move into inferential statistics. Nowhere is this more important than in correlation. Correlations are visualized by a **scatterplot**, where our X variable values are plotted on the X-axis, the Y variable values are plotted on the Y-axis, and each point or marker in the plot represents a single person's score on X and Y. Figure 2 shows a scatterplot for hypothetical scores on job satisfaction (X) and worker well-being (Y). We can see from the axes that each of these variables is measured on a 10- point scale, with 10 being the highest on both variables (high satisfaction and good health and well-being) and 1 being the lowest (dissatisfaction and poor health).When we look at this plot, we can see that the variables do seem to be related. The higher scores on job satisfaction tend to also be the higher scores on well-being, and the same is true of the lower scores.

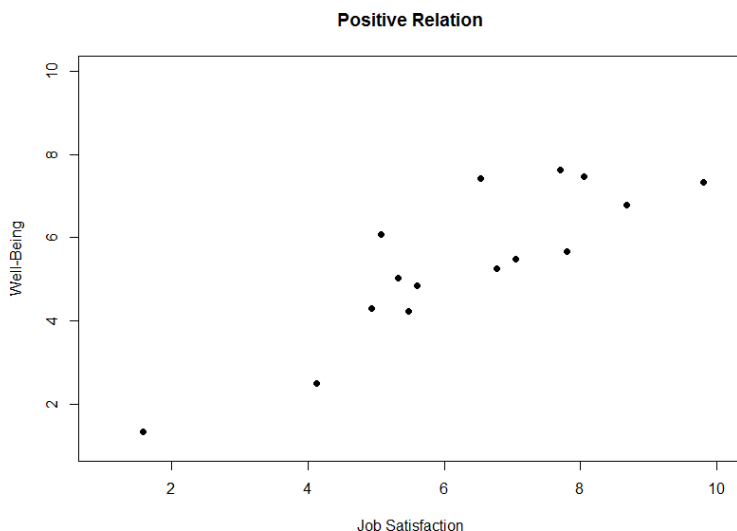


Figure 2. Plotting satisfaction and well-being scores.

Figure 2 demonstrates a positive relation. As scores on X increase, scores on Y also tend to increase. Although this is not a perfect relation (if it were, the points would form a single straight line), it is nonetheless very clearly positive. This is one of the key benefits to scatterplots: they make it very easy to see the direction of the relation. As another example, figure 3 shows a negative relation between job satisfaction (X) and burnout (Y). As we can see from this plot, higher scores on job satisfaction tend to correspond to lower scores on burnout, which is how stressed, unenergetic, and unhappy someone is at their job. As with figure 2, this is not a perfect relation, but it is still a clear one. As these figures show, points in a positive relation moves from the bottom left of the plot to the top right, and points in a negative relation move from the top left to the bottom right.

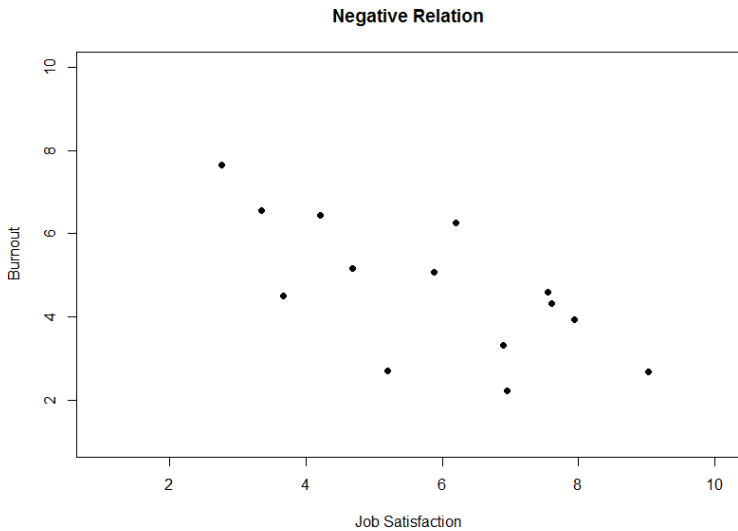


Figure 3. Plotting satisfaction and burnout scores.

Scatterplots can also indicate that there is no relation between the two variables. In these scatterplots (an example is shown below in figure 4 plotting job satisfaction and job performance) there is no interpretable shape or line in the scatterplot. The points appear randomly throughout the plot. If we tried to draw a straight line through these points, it would basically be flat. The low scores on job satisfaction have roughly the same scores on job performance as do the high scores on job satisfaction. Scores in the middle or average range of job satisfaction have some scores on job performance that are about equal to the high and low levels and some scores on job performance that are a little higher, but the overall picture is one of inconsistency.

As we can see, scatterplots are very useful for giving us an approximate idea of whether or not there is a relation between the two variables and, if there is, if that relation is positive or negative. They are also useful for another reason: they are the only way to determine one of the characteristics of correlations that are discussed next: form.

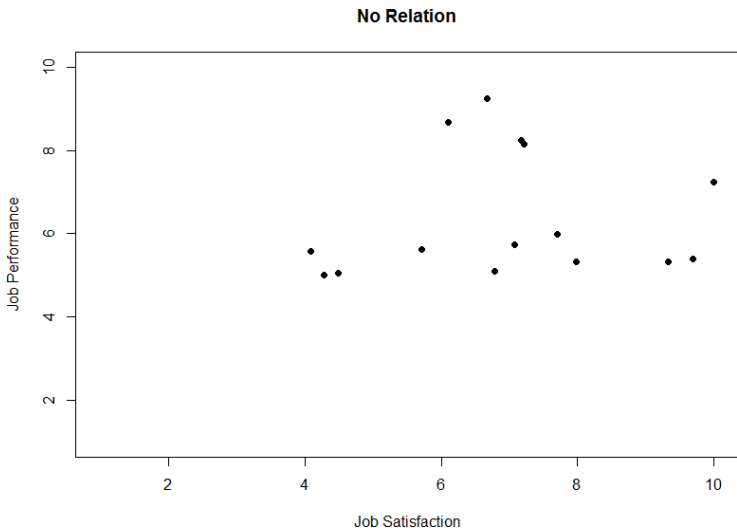


Figure 4. Plotting no relation between satisfaction and job performance.

Three Characteristics

When we talk about correlations, there are three traits that we need to know in order to truly understand the relation (or lack of relation) between X and Y: **form**, **direction**, and **magnitude**. We will discuss each of them in turn.

Form

The first characteristic of relations between variables is their form. The form of a relation is the shape it takes in a scatterplot, and a scatterplot is the only way it is possible to assess the form of a relation. there are three forms we look for: *linear*, *curvilinear*, or *no relation*. A **linear relation** is what we saw in figures 1, 2, and 3. If we drew a line through the middle points in the any of the scatterplots, we would be best suited

with a straight line. The term “linear” comes from the word “line”. A *linear relation* is what we will always assume when we calculate correlations. All of the correlations presented here are only valid for linear relations. Thus, it is important to plot our data to make sure we meet this assumption.

The relation between two variables can also be **curvilinear**. As the name suggests, a curvilinear relation is one in which a line through the middle of the points in a scatterplot will be curved rather than straight. Two examples are presented in figures 5 and 6.

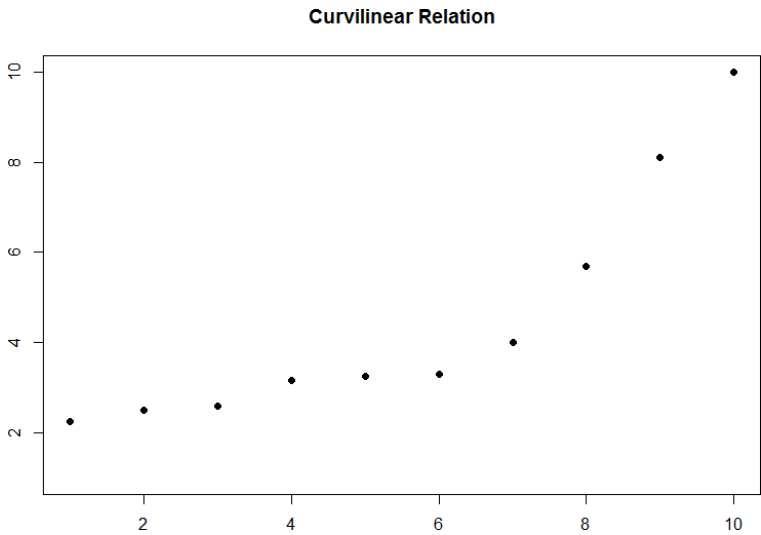


Figure 5. Exponentially increasing curvilinear relation

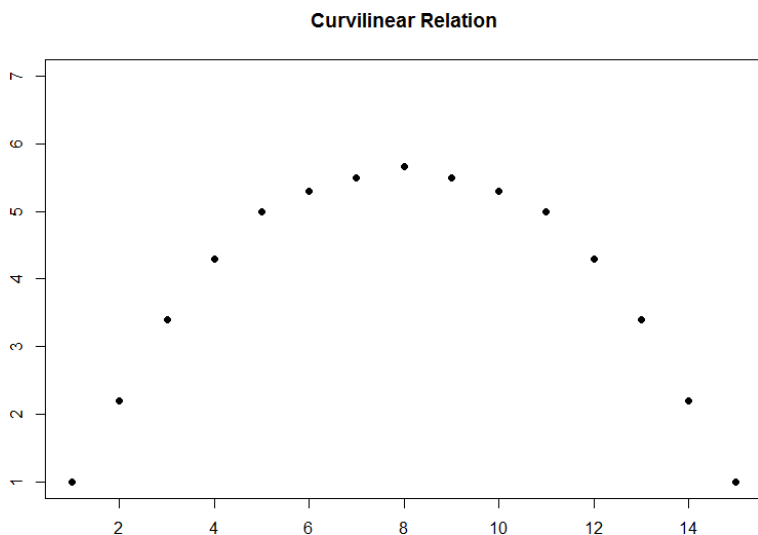


Figure 6. Inverted-U curvilinear relation.

Curvilinear relations can take many shapes, and the two examples above are only a small sample of the possibilities. What they have in common is that they both have a very clear pattern but that pattern is not a straight line. If we try to draw a straight line through them, we would get a result similar to what is shown in figure 7.

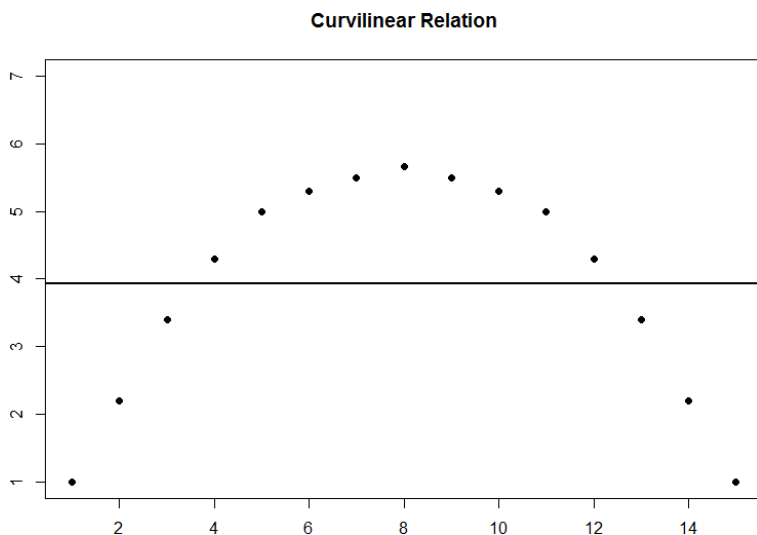


Figure 7. Overlaying a straight line on a curvilinear relation.

Although that line is the closest it can be to all points at the same time, it clearly does a very poor job of representing the relation we see. Additionally, the line itself is flat, suggesting there is no relation between the two variables even though the data show that there is one. This is important to keep in mind, because the math behind our calculations of correlation coefficients will only ever produce a straight line – we cannot create a curved line with the techniques discussed here.

Finally, sometimes when we create a scatterplot, we end up with **no interpretable relation** at all. An example of this is shown below in figure 8. The points in this plot show no consistency in relation, and a line through the middle would once again be a straight, flat line.

Sometimes when we look at scatterplots, it is tempting to get biased by a few points that fall far away from the rest of the points and seem to imply that there may be some sort of relation. These points are called outliers, and we will discuss them in more detail later in the chapter. These can be common, so it is important to formally test for a relation between our variables, not just rely on visualization. This is the

point of hypothesis testing with correlations, and we will go in depth on it soon. First, however, we need to describe the other two characteristics of relations: *direction* and *magnitude*.

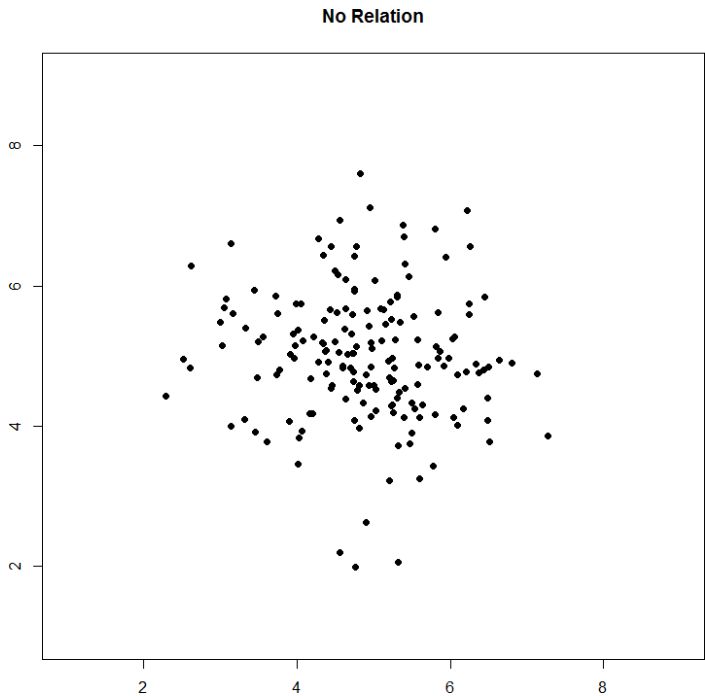


Figure 8. No relation

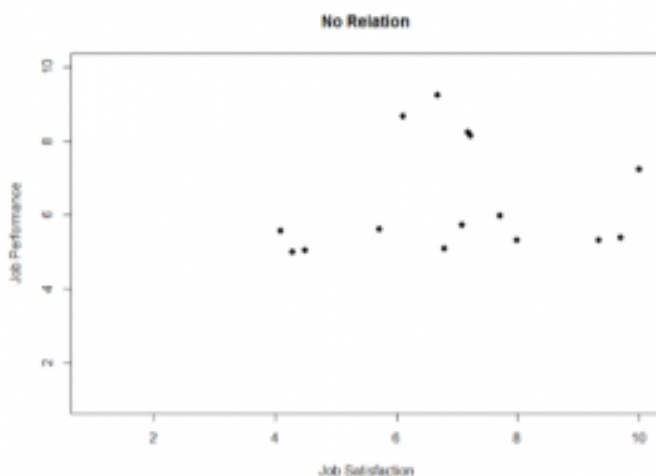


Figure 9. No relations fictional data scatterplot between job satisfaction and job performance

Direction

The **direction** of the relation between two variables tells us whether the variables change in the same way at the same time or in opposite ways at the same time. We saw this concept earlier when first discussing scatterplots, and we used the terms positive and negative. A **positive relation** is one in which X and Y change in the same direction: as X goes up, Y goes up, and as X goes down, Y also goes down. A **negative relation** is just the opposite: X and Y change together in opposite directions: as X goes up, Y goes down, and vice versa.

As we will see soon, when we calculate a correlation coefficient, we are quantifying the relation demonstrated in a scatterplot. That is, we are putting a number to it. That number will be either positive, negative, or zero, and we interpret the sign of the number as our direction. If the number is positive, it is a positive relation, and if it is negative, it is a negative relation. If

it is zero, then there is no relation. The direction of the relation corresponds directly to the slope of the hypothetical line we draw through scatterplots when assessing the form of the relation. If the line has a positive slope that moves from bottom left to top right, it is positive, and vice versa for negative. If the line is flat, that means it has no slope, and there is no relation, which will in turn yield a zero for our correlation coefficient.

Magnitude

The number we calculate for our correlation coefficient, which we will describe in detail below, corresponds to the magnitude of the relation between the two variables. The **magnitude** is how strong or how consistent the relation between the variables is. Higher numbers mean greater magnitude, which means a stronger relation. Our correlation coefficients will take on any value between -1.00 and 1.00, with 0.00 in the middle, which again represents no relation. A correlation of -1.00 is a perfect negative relation; as X goes up by some amount, Y goes down by the same amount, consistently. Likewise, a correlation of 1.00 indicates a perfect positive relation; as X goes up by some amount, Y also goes up by the same amount. Finally, a correlation of 0.00, which indicates no relation, means that as X goes up by some amount, Y may or may not change by any amount, and it does so inconsistently.

The vast majority of correlations do not reach -1.00 or positive 1.00. Instead, they fall in between, and we use rough cut offs for how strong the relation is based on this number.

Importantly, the sign of the number (the direction of the relation) has no bearing on how strong the relation is. The only thing that matters is the magnitude, or the absolute value of the correlation coefficient. A correlation of -1 is just as strong as a correlation of 1. We generally use values of 0.10, 0.30, and 0.50 as indicating weak, moderate, and strong relations,

respectively.

The strength of a relation, just like the form and direction, can also be inferred from a scatterplot, though this is much more difficult to do. Some examples of weak and strong relations are shown in figures 9 and 10, respectively. Weak correlations still have an interpretable form and direction, but it is much harder to see. Strong correlations have a very clear pattern, and the points tend to form a line. The examples show two different directions, but remember that the direction does not matter for the strength, only the consistency of the relation and the size of the number, which we will see next.

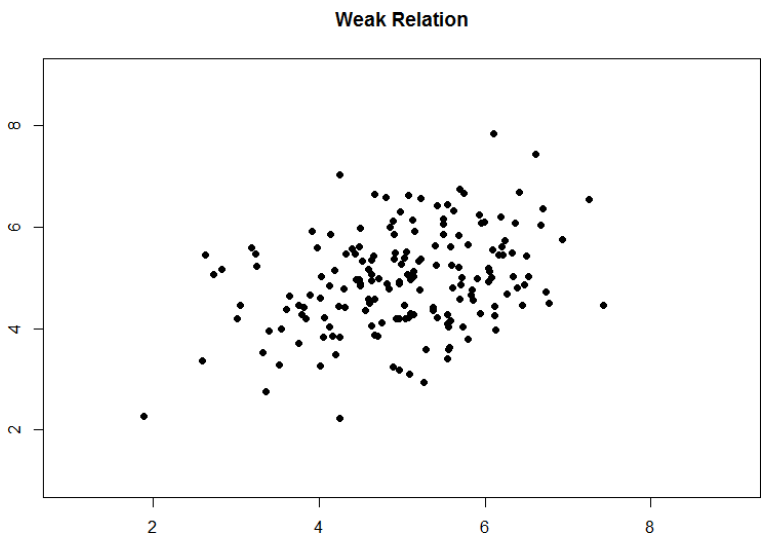


Figure 10. Weak positive correlation.

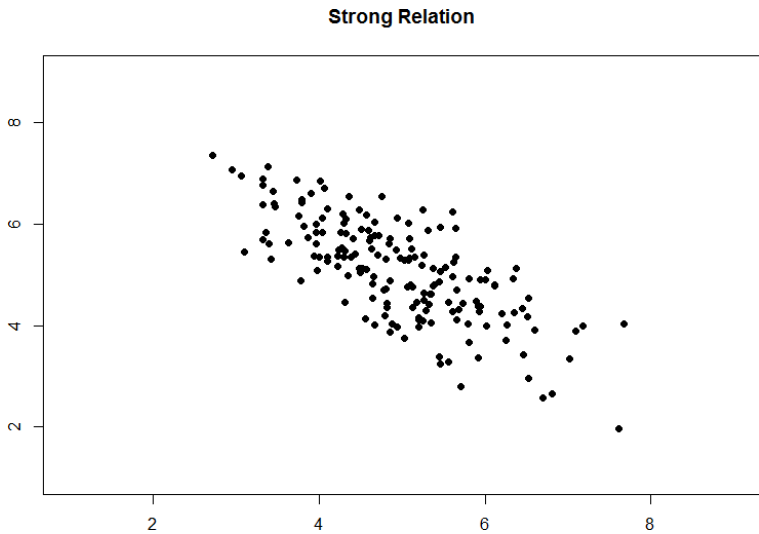


Figure 11. Strong negative correlation.

Pearson's r

There are several different types of correlation coefficients, but we will only focus on the most common: **Pearson's r** . r is a very popular correlation coefficient for assessing linear relations, and it serves as both a descriptive statistic (like \bar{X} aka M) and as a test statistic (like t). It is descriptive because it describes what is happening in the scatterplot; r will have both a sign (+/-) for the direction and a number (0 – 1 in absolute value) for the magnitude. As noted above, assumes a linear relation, so nothing about r will suggest what the form is – it will only tell what the direction and magnitude would be if the form is linear (Remember: always make a scatterplot first!). r also works as a test statistic because the magnitude of r will correspond directly to a t value as the specific degrees of freedom, which can then be compared to a critical value. Luckily, we do not need to do this conversion by hand. Instead, we will have a

table of r critical values that looks very similar to our t table, and we can compare our r directly to those.

The **conceptual formula for r** is very simple: it is just the covariance (defined above) divided by the standard deviations of X and Y:

$$r = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{SP}{\sqrt{SSX * SSY}}$$

Note: This formula gives a direct sense of what a correlation is: a covariance standardized onto the scale of X and Y.

We can also compute Pearson another way. The second formula is computationally simpler and faster. Both of these equations will give the same value. When we do this calculation, we will find that our answer is always between -1.00 and 1.00 (if it's not, check the math again), which gives us a standard, interpretable metric, similar to what z-scores did.

Computation r formula:

The diagram shows the formula for the sample correlation coefficient r . The formula is $r = \frac{\sum (z_x z_y)}{n - 1}$. Red boxes with arrows point to specific parts of the formula:

- A box labeled "Correlation coefficient" points to r .
- A box labeled "The z-score for the X value" points to z_x .
- A box labeled "The z-score for the Y value" points to z_y .
- A box labeled "The number of pairs of scores" points to $n - 1$.

If data

$$r = \frac{\sum z_X z_Y}{N}$$

population:

Correlation as a descriptive and as a test statistic

It was stated earlier that r is a descriptive statistic like \bar{X} (or M), and just like \bar{X} (or M), it corresponds to a population parameter. For correlations, the population parameter is the lowercase Greek letter ρ ("rho"); be careful not to confuse ρ with a p -value – they look quite similar. r is an estimate of ρ just like \bar{X} is an estimate of μ . Thus, we will test our observed value of r that we calculate from the data and compare it to a value of ρ specified by our null hypothesis to see if the relation between our variables is significant, as we will see in our example next.

Correlation & z-scores

| Pair of Raw Scores | | Pair of Z Scores | | Cross-Product of Z Scores | Effect on Correlation |
|--------------------|--------|------------------|----------------|-------------------------------|-------------------------------------|
| X | Y | Z _x | Z _y | Z _x Z _y | |
| High | High | + | + | + | Contributes to positive correlation |
| Low | Low | - | - | + | Contributes to positive correlation |
| High | Low | + | - | - | Contributes to negative correlation |
| Low | High | - | + | - | Contributes to negative correlation |
| Middle | Any | 0 | +, -, or 0 | 0 | Makes correlation near 0 |
| Any | Middle | +, -, or 0 | 0 | 0 | Makes correlation near 0 |

Note: + indicates a positive number, - indicates a negative number.

Table. 2.

The table 2 shows that positive products of Z scores contribute toward making a positive correlation, negative products of Z scores contribute toward making a negative correlation, and products of Z scores that are zero (or close to zero) contribute toward making a correlation of zero.

We still need to determine the strength of a positive or negative correlation on some standard scale. You cannot judge the strength of the correlation from the sum of the cross-products alone, because it gets bigger just by adding the cross-products of more people together.

The solution is to divide this sum of the cross-products by the number of people in the study. That is, you figure the average of the cross-products of Z scores.

It turns out that because of the nature of Z scores, this average can never be more than +1, which would be a positive linear perfect correlation. It can

never be less than -1, which would be a negative linear perfect correlation.

In the situation of no linear correlation, the average of the cross-products of Z scores is 0.

Example: Anxiety and Depression

Anxiety and depression are often reported to be highly linked (or “comorbid”). Our hypothesis testing procedure follows the same four-step process as before, starting with our null and alternative hypotheses. We will look for a positive relation between our variables among a group of 10 people because that is what we would expect based on them being comorbid.

Step 1: State the Hypotheses

Our hypotheses for correlations start with a baseline assumption of no relation, and our alternative will be directional if we expect to find a specific type of relation. For this example, we expect a positive relation:

H_0 : There is no relation
between anxiety and
depression, $H_0: \rho = 0$

H_A : There is a positive relation between anxiety and
depression, $H_A: \rho > 0$

Remember that ρ (“rho”) is our population parameter for the correlation that we estimate with r , just like \bar{X} and μ for means. Remember also that if there is no relation between

variables, the magnitude will be 0, which is where we get the null and alternative hypothesis values.

Step 2: Find the Critical Values

The critical values for correlations come from the correlation table, which looks very similar to the t -table (see figure 12). Just like our t -table, the column of critical values is based on our significance level (α) and the directionality of our test. The row is determined by our degrees of freedom. For correlations, we have $n - 2$ degrees of freedom, rather than $n - 1$ (why this is the case is not important at the moment). For our example, we have 10 people, so our degrees of freedom = $10 - 2 = 8$.

Critical Values for Pearson's r

| df | 0.05 | 0.025 | 0.01 | 0.005 | 1-tailed α |
|----|-------|-------|-------|-------|-------------------|
| | 0.10 | 0.05 | 0.02 | 0.01 | 2-tailed α |
| 1 | 0.988 | 0.997 | 1.000 | 1.000 | |
| 2 | 0.900 | 0.950 | 0.980 | 0.990 | |
| 3 | 0.805 | 0.878 | 0.934 | 0.959 | |
| 4 | 0.729 | 0.811 | 0.882 | 0.917 | |
| 5 | 0.669 | 0.754 | 0.833 | 0.874 | |
| 6 | 0.622 | 0.707 | 0.789 | 0.834 | |
| 7 | 0.582 | 0.666 | 0.750 | 0.798 | |
| 8 | 0.549 | 0.632 | 0.716 | 0.765 | |
| 9 | 0.521 | 0.602 | 0.685 | 0.735 | |
| 10 | 0.497 | 0.576 | 0.658 | 0.708 | |
| 11 | 0.476 | 0.553 | 0.634 | 0.684 | |
| 12 | 0.458 | 0.532 | 0.612 | 0.661 | |
| 13 | 0.441 | 0.514 | 0.592 | 0.641 | |
| 14 | 0.426 | 0.497 | 0.574 | 0.623 | |
| 15 | 0.412 | 0.482 | 0.558 | 0.606 | |

Figure 12. Correlation table

We were not given any information about the level of significance at which we should test our hypothesis, so we will assume $\alpha = 0.05$ as always. From our table, we can see that a 1-tailed test (because we expect only a positive relation) at the $\alpha = 0.05$ level has a critical value of $r^* = 0.549$. Thus, if our observed correlation is greater than 0.549, it will be statistically significant. This is a rather high bar (remember, the guideline for a strong relation is $r = 0.50$); this is because we have so few people. Larger samples make it easier to find significant relations.

Step 3: Calculate the Test Statistic

We have laid out our hypotheses and the criteria we will use to assess them, so now we can move on to our test statistic. Before we do that, we must first create a scatterplot of the data to make sure that the most likely form of our relation is in fact linear. Figure 13 below shows our data plotted out, and it looks like they are, in fact, linearly related, so Pearson’s r is appropriate.

Checking for Linear Form

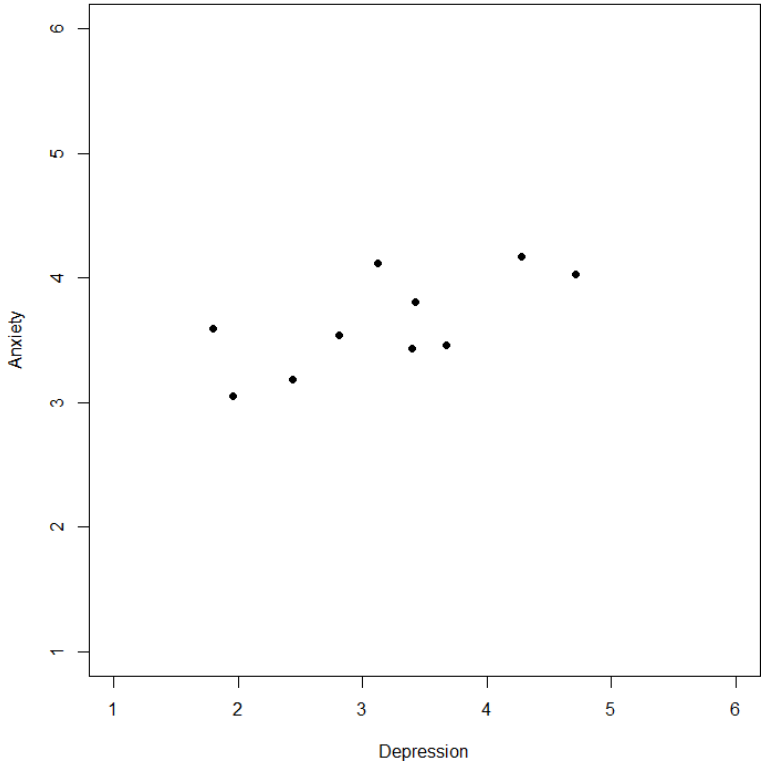


Figure 13. Scatterplot of anxiety and depression
The data we gather from our participants ($n=10$) is as follows:

| | | | | | | | | | | |
|-----|------|------|------|------|------|------|------|------|------|------|
| Dep | 2.81 | 1.96 | 3.43 | 3.40 | 4.71 | 1.80 | 4.27 | 3.68 | 2.44 | 3.13 |
| Anx | 3.54 | 3.05 | 3.81 | 3.43 | 4.03 | 3.59 | 4.17 | 3.46 | 3.19 | 4.12 |

Table 3. Data for step 3 to calculate r.

Steps for Calculating r using the *computational formula*:

1. Change all scores to Z scores.
1. This requires using the mean and the standard deviation of each variable, then changing each raw score to a Z score. This step is converting the raw scores to z-scores using the computational formula. See Table 4.

| | | | | | | | | | | | | |
|-----|------|------|------|------|------|------|------|------|------|------|-----------|----------|
| Dep | 2.81 | 1.96 | 3.43 | 3.40 | 4.71 | 1.80 | 4.27 | 3.68 | 2.44 | 3.13 | M = 3.163 | s = 0.94 |
| Anx | 3.54 | 3.05 | 3.81 | 3.43 | 4.03 | 3.59 | 4.17 | 3.46 | 3.19 | 4.12 | M = 3.639 | s = 0.38 |

Table 4. z-scores for anxiety and depression

2. Figure the cross-product of the Z scores for each person.
- That is, for each person, multiply the person's Z score on one variable by the person's Z score on the other variable.

| Zx | Zy | Zx*Zy |
|-------|-------|-----------------|
| -.38 | -.26 | 0.10 |
| -1.28 | -1.55 | 1.98 |
| 0.28 | 0.45 | 0.13 |
| 0.25 | -0.55 | -0.14 |
| 1.65 | 1.03 | 1.69 |
| -1.45 | -0.13 | 0.19 |
| 1.78 | 1.0 | 1.65 |
| 0.55 | -0.47 | -0.26 |
| -0.77 | -1.18 | 0.91 |
| -0.04 | 1.27 | -0.04 |
| | | $\Sigma = 6.20$ |

3. Add up the cross-products of the Z scores for Sum of Products.

- Adding up the third column we get $\Sigma = 6.20$.

4. Divide by the $n-1$ using sample in the study.

- There were 10 participants in the study.
- $10-1 = 9$
- $6.2/9 = .69$

5. Describe the relationship in words.

So our observed correlation between anxiety and depression is $r = 0.69$, which, based on sign and magnitude, is a strong, positive correlation. Now we need to compare it to our critical value to see if it is also statistically significant.

Step 4: Make a Decision

Our critical value was $r^* = 0.549$ and our obtained value was $r = 0.69$. Our obtained value was larger than our critical value, so we can reject the null hypothesis.

Reject H_0 . Based on our sample of 10 people, there is a statistically significant, strong, positive relation between anxiety and depression, $r(8) = 0.70$, $p < .05$.

Notice in our interpretation that, because we already know the magnitude and direction of our correlation, we can interpret that. We also report the degrees of freedom, just like with t , and we know that $p < \alpha$ because we rejected the null hypothesis. As we can see, even though we are dealing with a very different type of data, our process of hypothesis testing has remained unchanged.

Effect Size

Pearson's r is an incredibly flexible and useful statistic. Not only is it both descriptive and inferential, as we saw above, but because it is on a standardized metric (always between -1.00 and 1.00), it can also serve as its own effect size. In general, we use $r = 0.10$, $r = 0.30$, and $r = 0.50$ as our guidelines for small, medium, and large effects. Just like with Cohen's d , these guidelines are not absolutes, but they do serve as useful indicators in most situations. Notice as well that these are the same guidelines we used earlier to interpret the magnitude of the relation based on the correlation coefficient.

In addition to r being its own effect size, there is an additional effect size we can calculate for our results. This effect size is r^2 , and it is exactly what it looks like – it is the squared value of our correlation coefficient. Just like η^2 in ANOVA, r^2 is interpreted as the amount of variance explained in the outcome variance,

and the cut scores are the same as well: 0.01, 0.09, and 0.25 for small, medium, and large, respectively. Notice here that these are the same cutoffs we used for regular r effect sizes, but squared ($0.10^2 = 0.01$, $0.30^2 = 0.09$, $0.50^2 = 0.25$) because, again, the r^2 effect size is just the squared correlation, so its interpretation should be, and is, the same. The reason we use r^2 as an effect size is because our ability to explain variance is often important to us.

The similarities between η^2 and r^2 in interpretation and magnitude should clue you in to the fact that they are similar analyses, even if they look nothing alike. That is because, behind the scenes, they actually are! In the next chapter, we will learn a technique called **Linear Regression**, which will formally link the two analyses together.

Correlation versus Causation

We cover a great deal of material in introductory statistics and, as mentioned chapter 1, many of the principles underlying what we do in statistics can be used in your day to day life to help you interpret information objectively and make better decisions. We now come to what may be the most important lesson in introductory statistics: the difference between correlation and causation.

It is very, very tempting to look at variables that are correlated and assume that this means they are causally related; that is, it gives the impression that X is causing Y. However, in reality, correlation do not – and cannot – do this. *Correlations DO NOT prove causation*. No matter how logical or how obvious or how convenient it may seem, no correlational analysis can demonstrate causality. The ONLY way to demonstrate a causal relation is with a properly designed and controlled experiment.

Many times, we have good reason for assessing the correlation between two variables, and often that reason will be that we suspect that one causes the other. Thus, when we run our analyses and find strong, statistically significant results, it is very tempting to say that we found the causal relation that we are looking for. The reason we cannot do this is that, without an experimental design that includes random assignment and control variables, the relation we observe between the two variables may be caused by something else that we failed to measure. These “third variables” are lurking variables or confound variables, and they are impossible to detect and control for without an experiment.

Confound variables, which we will represent with Z , can cause two variables X and Y to appear related when in fact they are not. They do this by being the hidden– or lurking – cause of each variable independently. That is, if Z causes X and Z causes Y , the X and Y will appear to be related. However, if we control for the effect of Z (the method for doing this is beyond the scope of this text), then the relation between X and Y will disappear.

A popular example for this effect is the correlation between ice cream sales and deaths by drowning. These variables are known to correlate very strongly over time. However, this does not prove that one causes the other. The lurking variable in this case is the weather – people enjoy swimming and enjoy eating ice cream more during hot weather as a way to cool off. As another example, consider shoe size and spelling ability in elementary school children. Although there should clearly be no causal relation here, the variables are nonetheless consistently correlated. The confound in this case? Age. Older children spell better than younger children and are also bigger, so they have larger shoes.

When there is the possibility of confounding variables being the hidden cause of our observed correlation, we will often collect data on Z as well and control for it in our analysis. This is

good practice and a wise thing for researchers to do. Thus, it would seem that it is easy to demonstrate causation with a correlation that controls for Z. However, the number of variables that could potentially cause a correlation between X and Y is functionally limitless, so it would be impossible to control for everything. That is why we use experimental designs; by randomly assigning people to groups and manipulating variables in those groups, we can balance out individual differences in any variable that may be our cause. It is not always possible to do an experiment, however, so there are certain situations in which we will have to be satisfied with our observed relation and do the best we can to control for known confounds. However, in these situations, even if we do an excellent job of controlling for many extraneous (a statistical and research term for “outside”) variables, we must be very careful not to use causal language. That is because, even after controls, sometimes variables are related just by chance.

Sometimes, variables will end up being related simply due to random chance, and we call these correlation spurious. Spurious just means random, so what we are seeing is random correlations because, given enough time, enough variables, and enough data, sampling error will eventually cause some variables to be related when they should not. Sometimes, this even results in incredibly strong, but completely nonsensical, correlations. This becomes more and more of a problem as our ability to collect massive datasets and dig through them improves, so it is very important to think critically about any relation you encounter.

A Reminder about Experimental Design

When we say that one thing *causes* another, what do we mean? There is a long history in philosophy of discussion about the meaning of causality, but in statistics one way that we commonly think of causation is in terms of experimental control. That is, if we think that factor X causes factor Y, then manipulating the value of X should also change the value of Y.

Often we would like to test causal hypotheses but we can't actually do an experiment, either because it's impossible ("What is the relationship between human carbon emissions and the earth's climate?") or unethical ("What are the effects of severe abuse on child brain development?"). However, we can still collect data that might be relevant to those questions. For example, we can potentially collect data from children who have been abused as well as those who have not, and we can then ask whether their brain development differs.

Let's say that we did such an analysis, and we found that abused children had poorer brain development than non-abused children. Would this demonstrate that abuse *causes* poorer brain development? No. Whenever we observe a statistical association between two variables, it is certainly possible that one of those two variables causes the other. However, it is also possible that

both of the variables are being influenced by a third variable; in this example, it could be that child abuse is associated with family stress, which could also cause poorer brain development through less intellectual engagement, food stress, or many other possible avenues. The point is that a correlation between two variables generally tells us that something is *probably* causing something else, but it doesn't tell us what is causing what.

Final Considerations

Correlations, although simple to calculate, and be very complex, and there are many additional issues we should consider. We will look at two of the most common issues that affect our correlations, as well as discuss some other correlations and reporting methods you may encounter.

Range Restriction

The strength of a correlation depends on how much variability is in each of the variables X and Y. This is evident in the formula for Pearson's r , which uses both covariance (based on the sum of products, which comes from deviation scores) and the standard deviation of both variables (which are based on the sums of squares, which also come from deviation scores). Thus, if we reduce the amount of variability in one or both variables, our correlation will go down. Failure to capture the full variability of a variability is called range restriction.

Take a look at figures 14 and 15 below. The first shows a strong relation ($r = 0.67$) between two variables. An oval is overlain on top of it to make the relation even more distinct. The second shows the same data, but the bottom half of the X variable (all scores below 5) have been removed, which causes our relation (again represented by a red oval) to become much weaker ($r = 0.38$). Thus range restriction has truncated (made smaller) our observed correlation.

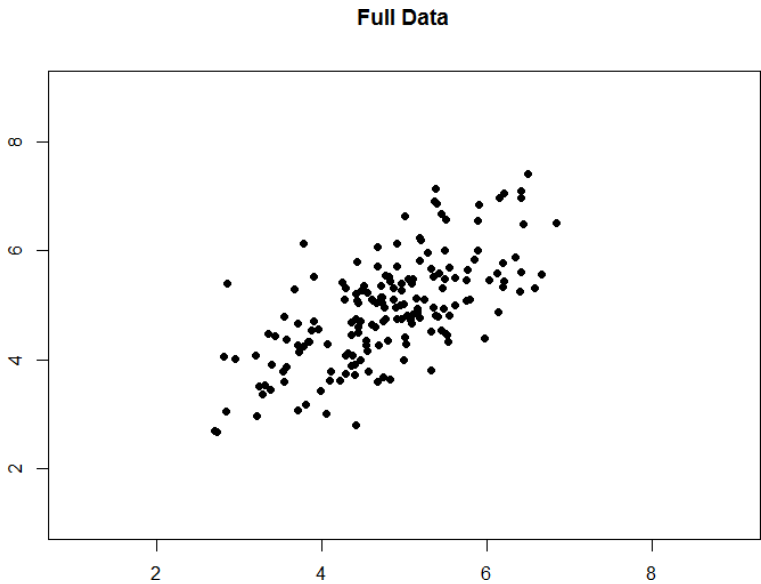


Figure 14. Strong, positive correlation.

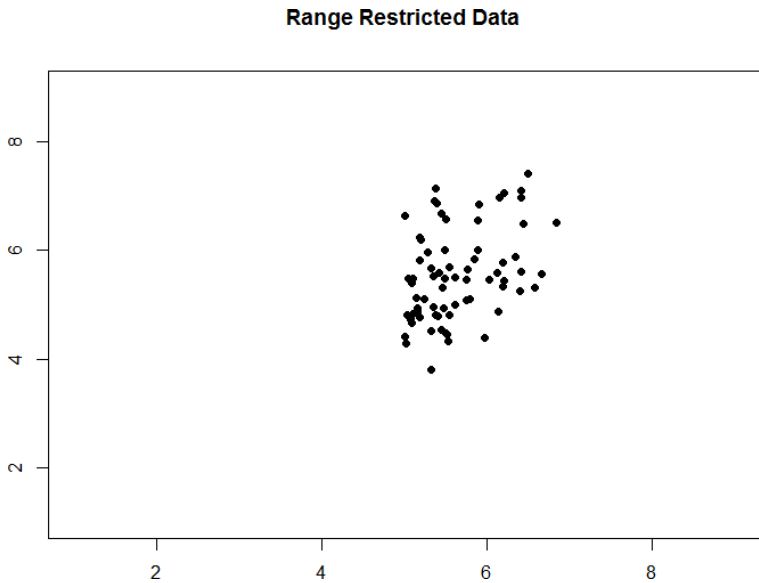


Figure 15. Effect of range restriction.

Sometimes range restriction happens by design. For example, we rarely hire people who do poorly on job applications, so we would not have the lower range of those predictor variables. Other times, we inadvertently cause range restriction by not properly sampling our population. Although there are ways to correct for range restriction, they are complicated and require much information that may not be known, so it is best to be very careful during the data collection process to avoid it.

Outliers

Another issue that can cause the observed size of our correlation to be inappropriately large or small is the presence of outliers. An outlier is a data point that falls far away from the rest of the observations in the dataset. Sometimes outliers are the result of incorrect data entry, poor or intentionally misleading responses, or simple random chance. Other times,

however, they represent real people with meaningful values on our variables. The distinction between meaningful and accidental outliers is a difficult one that is based on the expert judgment of the researcher. Sometimes, we will remove the outlier (if we think it is an accident) or we may decide to keep it (if we find the scores to still be meaningful even though they are different).

Pearson's r is sensitive to outliers. For example, in Figure 16 we can see how a single outlying data point can cause a very high positive correlation value, even when the actual relationship between the other data points is perfectly negative.

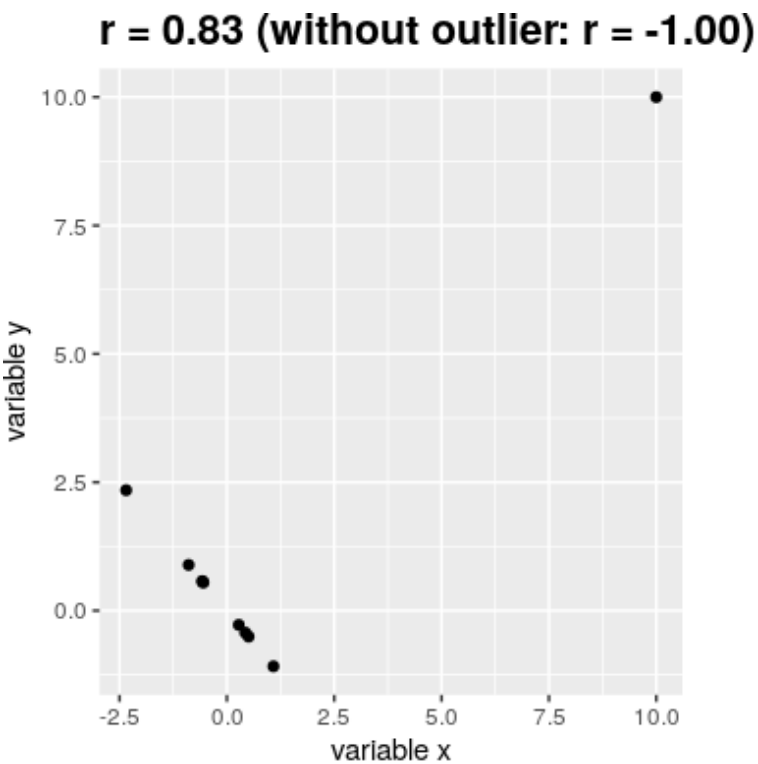


Figure 16. An simulated example of the effects of outliers on

correlation. Without the outlier the remainder of the data points have a perfect negative correlation, but the single outlier changes the correlation value to highly positive.

One way to address outliers is to compute the correlation on the ranks of the data after ordering them, rather than on the data themselves; this is known as the *Spearman correlation*. Whereas the Pearson correlation for the example in Figure 15 was 0.83, the Spearman correlation is -0.45, showing that the rank correlation reduces the effect of the outlier and reflects the negative relationship between the majority of the data points.

Here are some more examples. The plots below in figure 16 show the effects that an outlier can have on data. In the first, we have our raw dataset. You can see in the upper right corner that there is an outlier observation that is very far from the rest of our observations on both the X and Y variables. In the middle, we see the correlation computed when we include the outlier, along with a straight line representing the relation; here, it is a positive relation. In the third image, we see the correlation after removing the outlier, along with a line showing the direction once again. Not only did the correlation get stronger, it completely changed direction!

In general, there are three effects that an outlier can have on a correlation: it can change the magnitude (make it stronger or weaker), it can change the significance (make a non-significant correlation significant or vice versa), and/or it can change the direction (make a positive relation negative or vice versa). Outliers are a big issue in small datasets where a single observation can have a strong weight compared to the rest. However, as our samples sizes get very large (into the hundreds), the effects of outliers diminishes because they are outweighed by the rest of the data. Nevertheless, no matter how large a dataset you have, it is always a good idea to screen for outliers, both statistically (using analyses that we do not

cover here) and/or visually (using scatterplots). Also, one way to address outliers is to compute the correlation on the ranks of the data after ordering them, rather than on the data themselves; this is known as the *Spearman correlation*.

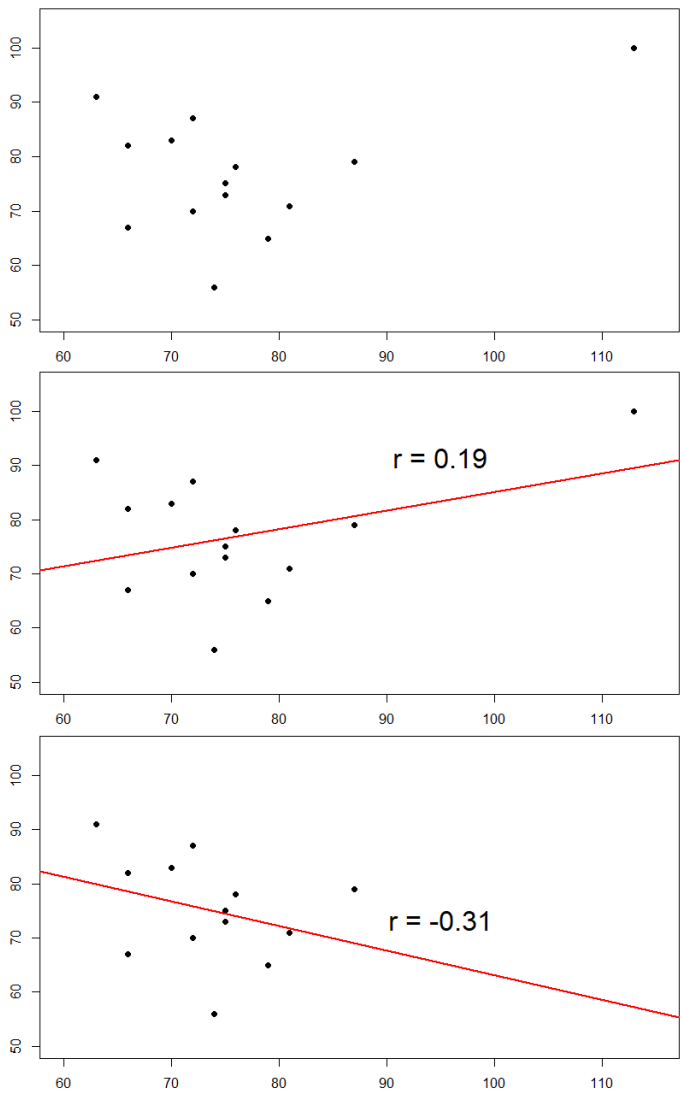


Figure 17. Three plots showing correlations with and without outliers.

An misreported media example: Hate crimes and income inequality

In 2017, the web site Fivethirtyeight.com published a story titled [*Higher Rates Of Hate Crimes Are Tied To Income Inequality*](#) which discussed the relationship between the prevalence of hate crimes and income inequality in the wake of the 2016 Presidential election. The story reported an analysis of hate crime data from the FBI and the Southern Poverty Law Center, on the basis of which they report:

“we found that income inequality was the most significant determinant of population-adjusted hate crimes and hate incidents across the United States”.

The analysis reported in the story focused on the relationship between income inequality (defined by a quantity called the *Gini index* — see Appendix for more details) and the prevalence of hate crimes in each state.

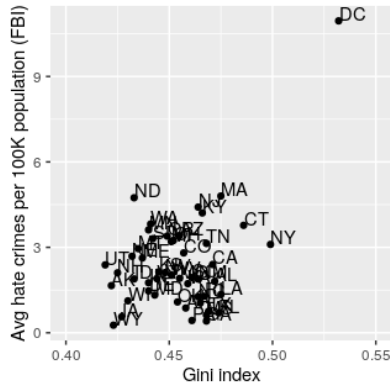


Figure 18: Plot of rates of hate crimes vs. Gini index.

The relationship between income inequality and rates of hate crimes is shown in Figure 18. Looking at the data, it seems that there may be a positive relationship between the two variables. The correlation value of 0.42 between hate crimes and income inequality seems to indicate a reasonably strong relationship between the two, but we can also imagine that this could occur by chance even if there is no relationship. We can test the null hypothesis that the correlation is zero using a statistical program (similar to our step 2). We get $r(48) = .42, [.16, .63], p = .002$. The numbers reported in the brackets are the 95% confidence interval for r . This test shows that the likelihood of an r value this

extreme or more is quite low under the null hypothesis, so we would reject the null hypothesis there is no relationship. Note that this test assumes that both variables are normally distributed. However, you may have noticed something a bit odd in Figure 17 – one of the datapoints (the one for the District of Columbia) seemed to be quite separate from the others. We refer to this as an *outlier*, and the standard correlation coefficient is very sensitive to outliers. When we calculate Spearman's rho (ρ) (less sensitive to outliers), $\rho = .033$, $p = .08$. Now we see that the correlation is no longer significant (and in fact is very near zero), suggesting that the claims of the FiveThirtyEight blog post may have been incorrect due to the effect of the outlier.

Other Correlation Coefficients

In this chapter we have focused on Pearson's r as our correlation coefficient because it very common and very useful. There are, however, many other correlations out there, each of which is designed for a different type of data. The most common of these is Spearman's rho (ρ), which is designed to be used on ordinal data rather than continuous data. This is a very useful analysis if we have ranked data or our data do not conform to the normal distribution. There are even more correlations for ordered categories, but they are much less common and beyond the scope of this chapter.

Additionally, the principles of correlations underlie many other advanced analyses. In the next chapter, we will learn about regression, which is a formal way of running and analyzing a correlation that can be extended to more than two variables. Regression is a very powerful technique that serves as the basis for even our most advanced statistical models, so what we have learned in this chapter will open the door to an entire world of possibilities in data analysis.

Correlation Matrices

Many research studies look at the relation between more than two continuous variables. In such situations, we could simply list out all of our correlations, but that would take up a lot of space and make it difficult to quickly find the relation we are looking for. Instead, we create correlation matrices so that we can quickly and simply display our results. A matrix is like a grid that contains our values. There is one row and one column for each of our variables, and the intersections of the rows and columns for different variables contain the correlation for those two variables. At the beginning of the chapter, we saw scatterplots presenting data for correlations between job satisfaction, well-being, burnout, and job performance. We can create a correlation matrix to quickly display the numerical values of each. Such a matrix is shown below.

| | Satisfaction | Well-Being | Burnout | Performance |
|--------------|--------------|------------|---------|-------------|
| Satisfaction | 1.00 | | | |
| Well-Being | 0.41 | 1.00 | | |
| Burnout | -0.54 | -0.87 | 1.00 | |
| Performance | 0.08 | 0.21 | -0.33 | 1.00 |

Table 5. Example Correlation Matrix

Notice that there are values of 1.00 where each row and

column of the same variable intersect. This is because a variable correlates perfectly with itself, so the value is always exactly 1.00. Also notice that the upper cells are left blank and only the cells below the diagonal of 1s are filled in. This is because correlation matrices are symmetrical: they have the same values above the diagonal as below it. Filling in both sides would provide redundant information and make it a bit harder to read the matrix, so we leave the upper triangle blank. Correlation matrices are a very condensed way of presenting many results quickly, so they appear in almost all research studies that use continuous variables. Many matrices also include columns that show the variable means and standard deviations, as well as asterisks showing whether or not each correlation is statistically significant.

Summary

Value of the correlation coefficient (r)

- The value of r is always between -1 and $+1$
- The size of the correlation r indicates the strength of the linear relationship between x and y and values close to -1 or to $+1$ indicate a stronger linear relationship between x and y .
 - $r = 1$ represents a perfect positive correlation. A correlation of 1 indicates a perfect linear relationship.
 - $r = -1$ represents a perfect negative correlation. A correlation of -1 indicates a perfect negative relationship.
 - If $r = 0$ there is absolutely no linear relationship between x and y . A correlation of zero indicates no linear relationship.

Direction of the correlation coefficient (r)

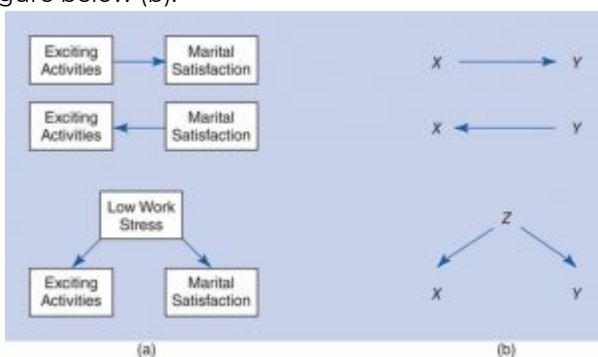
- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease (positive correlation).
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase (negative correlation).

If two variables have a significant linear correlation we normally might assume that there is something causing them to go together. However, we cannot know the direction of causality (what is causing what) just from the fact that the two variables are correlated.

Consider this example, the relationship between doing exciting activities with your significant other and satisfaction with the relationship. There are three possible directions of causality for these two variables:

- X could be causing Y
- Y could be causing X
- Some third factor could be causing both X and Y

These three possible directions of causality are shown in the figure below (b).



- Correlation is a relationship that has established that X and Y are related – if we know one then the other can be

predicted but we cannot conclude that one variable causes the other.

- Causation is a relationship for which we have to establish that X causes Y. To establish causation an experiment must demonstrate that Y can be controlled by presenting or removing X.
 - For example, when we apply heat (X) the temperature of water (Y) increases and when we remove heat (X) the temperature of water (Y) decreases.

Learning Objectives

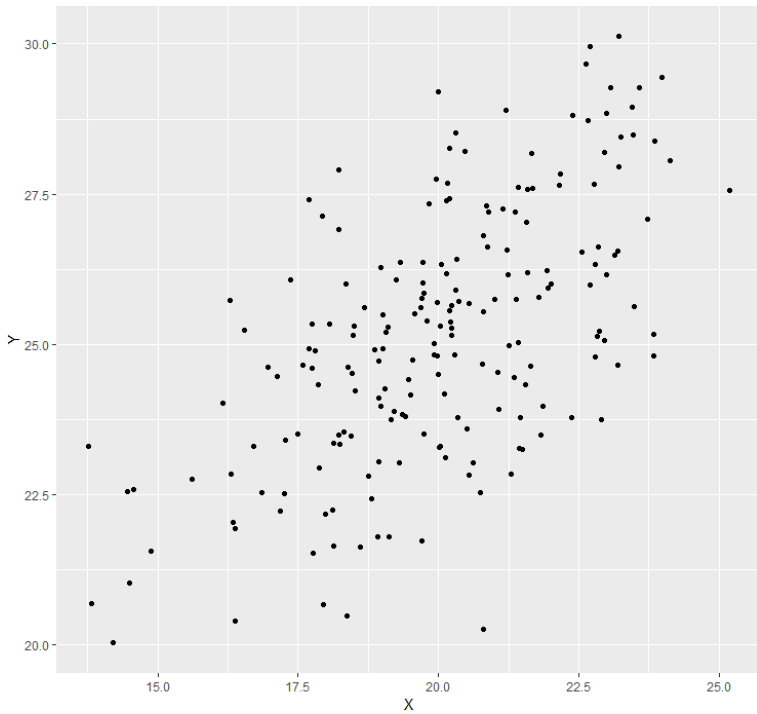
Having read this chapter, a student should be able to:

- Describe the concept of the correlation coefficient and its interpretation
- Understand Pearson correlation as a descriptive statistic and test statistic
- Compute the Pearson correlation
- Identify type of correlation based on the data (Pearson vs Spearman)
- Describe the effect of outlier data points and how to address them.
- Describe the potential causal influences that can give rise to an observed correlation.

Exercises – Ch. 16

1. What does a correlation assess?
2. What are the three characteristics of a correlation coefficient? Why is it important to visualize correlational data in a scatterplot before performing analyses?

3. What sort of relation is displayed in the scatterplot below?



4. What is the direction and magnitude of the following correlation coefficients?

1. -0.81
2. 0.40
3. 0.15
4. -0.08
5. 0.29

5. Create a scatterplot from the following data:

| Hours Studying | Overall Class Performance |
|----------------|---------------------------|
| 0.62 | 2.02 |
| 1.50 | 4.62 |
| 0.34 | 2.60 |
| 0.97 | 1.59 |
| 3.54 | 4.67 |
| 0.69 | 2.52 |
| 1.53 | 2.28 |
| 0.32 | 1.68 |
| 1.94 | 2.50 |
| 1.25 | 4.04 |
| 1.42 | 2.63 |
| 3.07 | 3.53 |
| 3.99 | 3.90 |
| 1.73 | 2.75 |
| 1.29 | 2.95 |

6. In the following correlation matrix, what is the relation (number, direction, and magnitude) between...

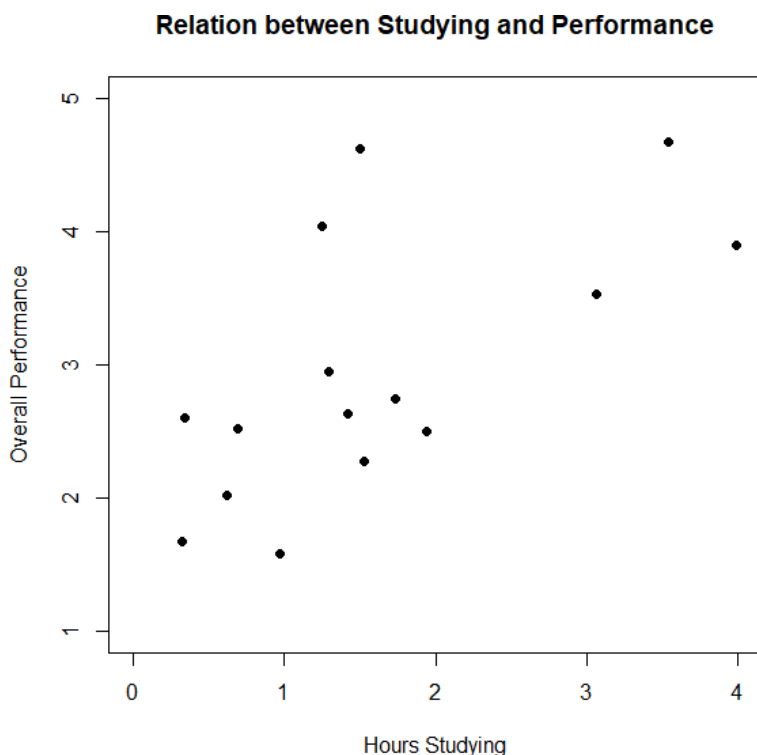
- Pay and Satisfaction
- Stress and Health

| Workplace | Pay | Satisfaction | Stress | Health |
|--------------|------|--------------|--------|--------|
| Pay | 1.00 | | | |
| Satisfaction | .68 | 1.00 | | |
| Stress | 0.02 | -0.23 | 1.00 | |
| Health | 0.05 | 0.15 | -0.48 | 1.00 |

7. A researcher collects data from 100 people to assess whether there is any relation between level of education and levels of civic engagement. The researcher finds the following descriptive values: $\bar{X} = 4.02$, $s_x = 1.15$, $\bar{Y} = 15.92$, $s_y = 5.01$, $SS_x = 130.93$, $SS_y = 2484.91$, $SP = 159.39$. Test for a significant relation using the four step hypothesis testing procedure.

Answers to Odd- Numbered Exercises – Ch. 16

1. Correlations assess the linear relation between two continuous variables
3. Strong, positive, linear relation
5. Your scatterplot should look similar to this:



7. Step 1: $H_0: \rho = 0$, "There is no relation between time spent studying and overall performance in class", $H_A: \rho > 0$, "There is a positive relation between time spent studying and overall performance in class."

Step 2: $df = 15 - 2 = 13$, $\alpha = 0.05$, 1-tailed test, $r^* = 0.441$.

Step 3: Using the Sum of Products table, you should find: $\bar{X} = 1.61$, $SS_X = 17.44$, $\bar{Y} = 2.95$, $SS_Y = 13.60$, $SP = 10.06$, $r = 0.65$.

Step 4: Obtained statistic is greater than critical value, reject H_0 . There is a statistically significant, strong, positive relation between time spent studying and performance in class, $r(13) = 0.65$, $p < .05$.

Appendix: conceptual calculations for anxiety and depression example.

We will use X for depression and Y for anxiety to keep track of our data, but be aware that this choice is arbitrary and the math will work out the same if we decided to do the opposite. Our table is thus:

| X | $(X - \bar{X})$ | $(X - \bar{X})^2$ | Y | $(Y - \bar{Y})$ | $(Y - \bar{Y})^2$ | $(X - \bar{X})(Y - \bar{Y})$ |
|-------|-----------------|-------------------|-------|-----------------|-------------------|------------------------------|
| 2.81 | -0.35 | 0.12 | 3.54 | -0.10 | 0.01 | 0.04 |
| 1.96 | -1.20 | 1.44 | 3.05 | -0.59 | 0.35 | 0.71 |
| 3.43 | 0.27 | 0.07 | 3.81 | 0.17 | 0.03 | 0.05 |
| 3.40 | 0.24 | 0.06 | 3.43 | -0.21 | 0.04 | -0.05 |
| 4.71 | 1.55 | 2.40 | 4.03 | 0.39 | 0.15 | 0.60 |
| 1.80 | -1.36 | 1.85 | 3.59 | -0.05 | 0.00 | 0.07 |
| 4.27 | 1.11 | 1.23 | 4.17 | 0.53 | 0.28 | 0.59 |
| 3.68 | 0.52 | 0.27 | 3.46 | -0.18 | 0.03 | -0.09 |
| 2.44 | -0.72 | 0.52 | 3.19 | -0.45 | 0.20 | 0.32 |
| 3.13 | -0.03 | 0.00 | 4.12 | 0.48 | 0.23 | -0.01 |
| total | total | total | total | total | total | total (SP) |
| 31.63 | 0.03 | 7.97 | 36.39 | -0.01 | 1.33 | 2.22 |

The bottom row is the sum of each column. We can see from this that the sum of the X observations is 31.63, which makes the mean of the X variable $\bar{X} = 3.16$. The deviation scores for X sum to 0.03, which is very close to 0, given rounding error, so everything looks right so far. The next column is the squared deviations for X, so we can see that the sum of squares for X is $SS_X = 7.97$. The same is true of the Y columns, with an average of $\bar{Y} = 3.64$, deviations that sum to zero within rounding error, and a sum of squares as $SS_Y = 1.33$. The final column is the

product of our deviation scores (NOT of our squared deviations), which gives us a sum of products of $SP = 2.22$.

$$r = \frac{SP}{\sqrt{SSX * SSY}} = \frac{2.22}{\sqrt{7.97 * 1.33}} = .70$$

Our calculation before was $r = .69$, difference due to rounding issues!

17. Chapter 17: Linear Regression

In chapter 14, we learned about ANOVA, which involves a new way of looking at how our data are structured and the inferences we can draw from that. In chapter 16, we learned about correlations, which analyze two continuous variables at the same time to see if they systematically relate in a linear fashion. In this chapter, we will combine these two techniques in an analysis called *simple linear regression*, or *regression* for short. **Regression** uses the technique of variance partitioning from ANOVA to more formally assess the types of relations looked at in correlations. Regression is the most general and most flexible analysis covered in this book, and we will only scratch the surface.

A major practical application of statistical methods is making predictions. Psychologists often call this kind of prediction regression. Regression literally means going back or returning. We use the term regression here because the predicted score on the criterion variable is closer (in terms of standard deviation units) to the mean of the criterion variable compared to the distance from the value of the predictor variable to the mean of the predictor variable. So we can think of this in terms of the predicted value of the criterion variable regressing, or going back, toward the mean of the criterion variable. Again, the concepts in this chapter are directly related to correlation. This is because if two variables are correlated it means that we can predict one from the other. So if sleep the night before is correlated with happiness the next day, this means that we should be able, to some extent, predict how happy a person will be the next day from knowing how much sleep the person got the night before. The concepts in the chapter are also related

to ANOVA as the goal of regression is the same as the goal of ANOVA: to take what we know about one variable (X) and use it to explain our observed differences in another variable (Y) – we are just two continuous variables.

Line of Best Fit

In correlations, we referred to a linear trend in the data. That is, we assumed that there was a straight line we could draw through the middle of our scatterplot that would represent the relation between our two variables, X and Y. Regression involves solving for the equation of that line, which is called the *Line of Best Fit*.

The line of best fit can be thought of as the central tendency of our scatterplot. The term “best fit” means that the line is as close to all points (with each point representing both variables for a single person) in the scatterplot as possible, with a balance of scores above and below the line. This is the same idea as the mean, which has an equal weighting of scores above and below it and is the best singular descriptor of all our data points for a single variable.

We have already seen many scatterplots in chapters 3 and 16, so we know by now that no scatterplot has points that form a perfectly straight line. Because of this, when we put a straight line through a scatterplot, it will not touch all of the points, and it may not even touch any! This will result in some distance between the line and each of the points it is supposed to represent, just like a mean has some distance between it and all of the individual scores in the dataset.

The distances between the line of best fit and each individual data point go by two different names that mean the same thing: errors and residuals. The term “error” in regression is closely aligned with the meaning of error in statistics (think

standard error or sampling error); it does not mean that we did anything wrong, it simply means that there was some discrepancy or difference between what our analysis produced and the true value we are trying to get at it. The term “residual” is new to our study of statistics, and it takes on a very similar meaning in regression to what it means in everyday parlance: there is something left over. In regression, what is “left over” – that is, what makes up the residual – is an imperfection in our ability to predict values of the Y variable using our line. This definition brings us to one of the primary purposes of regression and the line of best fit: predicting scores.

Prediction

The goal of regression is the same as the goal of ANOVA: to take what we know about one variable (X) and use it to explain our observed differences in another variable (Y). In ANOVA, we talked about – and tested for – group mean differences, but in regression we do not have groups for our explanatory variable; we have a continuous variable, like in correlation. Because of this, our vocabulary will be a little bit different, but the process, logic, and end result are all the same.

In regression, we most frequently talk about prediction, specifically predicting our outcome variable Y from our explanatory variable X, and we use the line of best fit to make our predictions. Let's take a look at the equation for the line, which is quite simple.

Regression equation

$$\hat{Y} = a + bX$$

The terms in the equation are defined as:

\hat{Y}

:

t

h

e

p

r

e

d

ic

t

e

d

v

a

l

u

e

o

f

Y

f

o

r

a

n

i

n

d

iv
i
d
u
a
l
p
e
r
s
o
n
a:
t
h
e
i
n
t
e
r
c
e
p
t
o
f
t
h
e
li

n
e

b: the slope of the line

X: the observed value of X for an
individual person

Additionally we have formulas for a and b:

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{SP}{SSX} = r \left(\frac{s_y}{s_x} \right)$$

What this shows us is that we will use our known value of X for each person to predict the value of Y for that person. The predicted value, \hat{Y} , is called “y-hat” and is our best guess for what a person’s score on the outcome is. Notice also that the form of the equation is very similar to very simple linear equations that you have likely encountered before and has only two parameter estimates: an intercept (where the line crosses the Y-axis) and a slope (how steep – and the direction, positive or negative – the line is). These are parameter estimates because, like everything else in statistics, we are interested in approximating the true value of the relation in the population but can only ever estimate it using sample data. We will soon see that one of these parameters, the slope, is the focus of our hypothesis tests (the intercept is only there to make the math work out properly and is rarely interpretable).

It is very important to point out that the Y values in the equations for a and b are our observed Y values in the dataset, NOT the predicted Y values (\hat{Y}) from our equation for the line

of best fit. Thus, we will have 3 values for each person: the observed value of X (X), the observed value of Y (Y), and the predicted value of Y (\hat{Y}). You may be asking why we would try to predict Y if we have an observed value of Y , and that is a very reasonable question. The answer has two explanations: first, we need to use known values of Y to calculate the parameter estimates in our equation, and we use the difference between our observed values and predicted values ($Y - \hat{Y}$) to see how accurate our equation is; second, we often use regression to create a predictive model that we can then use to predict values of Y for other people for whom we only have information on X .

Applied examples for using regression

Example 1: Businesses often have more applicants for a job than they have openings available, so they want to know who among the applicants is most likely to be the best employee. There are many criteria that can be used, but one is a personality test for conscientiousness, with the belief being that more conscientious (more responsible) employees are better than less conscientious employees. A business might give their employees a personality inventory to assess conscientiousness and existing performance data to look for a relation. In this example, we have known values of the predictor (X , conscientiousness) and outcome (Y , job performance), so we can estimate an equation for a line of best fit and see how accurately conscientious predicts job performance, then use this

equation to predict future job performance of applicants based only on their known values of conscientiousness from personality inventories given during the application process.

Example 2: Assume a researcher is interested in examining whether SAT scores can be an accurate predictor of college GPA. In this case, SAT scores would be the predictor variable or X and college GPA would be the criterion variable or Y .

The key assessing whether a linear regression works well is the difference between our observed and known Y values and our predicted \hat{Y} values. As mentioned in passing above, we use subtraction to find the difference between them ($Y - \hat{Y}$) in the same way we use subtraction for deviation scores and sums of squares. The value ($Y - \hat{Y}$) is our **residual**, which, as defined above, is how close our line of best fit is to our actual values. We can visualize residuals to get a better sense of what they are by creating a scatterplot and overlaying a line of best fit on it, as shown in Figure 1.

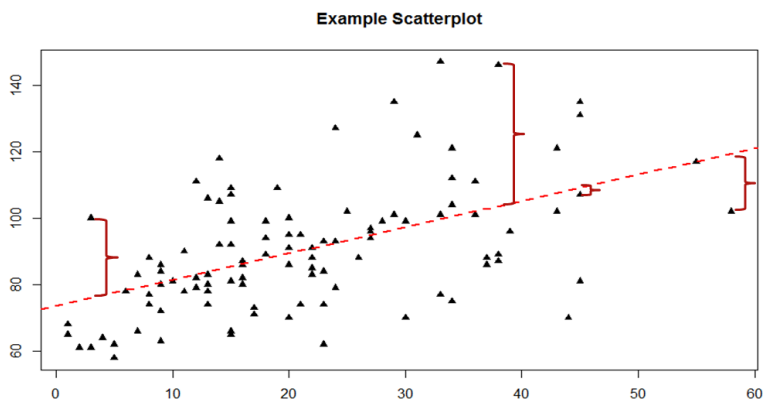


Figure 1. Scatterplot with residuals

In figure 1, the triangular dots represent observations from each person on both X and Y and the dashed bright red line is the line of best fit estimated by the equation $\hat{Y} = a + bX$. For every person in the dataset, the line represents their predicted score. The dark red bracket between the triangular dots and the predicted scores on the line of best fit are our residuals (they are only drawn for four observations for ease of viewing, but in reality there is one for every observation); you can see that some residuals are positive and some are negative, and that some are very large and some are very small. This means that some predictions are very accurate and some are very inaccurate, and the some predictions overestimated values and some underestimated values. Across the entire dataset, the line of best fit is the one that minimizes the total (sum) value of all residuals. That is, although predictions at an individual level might be somewhat inaccurate, across our full sample and (theoretically) in future samples our total amount of error is as small as possible.

We call this property of the line of best fit the **Least Squares Error Solution**. This term means that the solution – or equation – of the line is the one that provides the smallest possible value of the squared errors (squared so that they can be summed, just like in standard deviation) relative to any other straight line we could draw through the data.

Predicting Scores and Explaining Variance

We have now seen that the purpose of regression is twofold: we want to predict scores based on our line and, as stated earlier, explain variance in our observed Y variable just like in ANOVA. These two purposes go hand in hand, and our ability to predict scores is literally our ability to explain variance. That is, if we

cannot account for the variance in Y based on X, then we have no reason to use X to predict future values of Y.

We know that the overall variance in Y is a function of each score deviating from the mean of Y (as in our calculation of variance and standard deviation). So, just like the red brackets in figure 1 representing residuals, given as $(Y - \hat{Y})$, we can visualize the overall variance as each score's distance from the overall mean of Y, given as $(Y - \overline{Y})$, our normal deviation score. This is shown in figure 2.

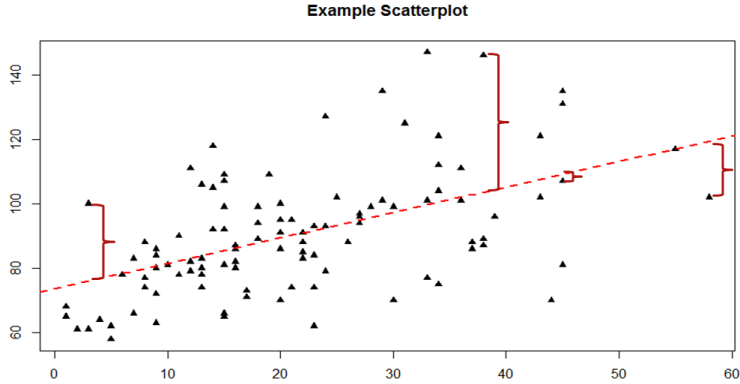


Figure 2. Scatterplot with residuals and deviation scores.

In figure 2, the solid blue line is the mean of Y, and the blue brackets are the deviation scores between our observed values of Y and the mean of Y. This represents the overall variance that we are trying to explain. Thus, the residuals and the deviation scores are the same type of idea: the distance between an observed score and a given line, either the line of best fit that gives predictions or the line representing the mean that serves as a baseline. The difference between these two values, which is the distance between the lines themselves, is our model's ability to predict scores above and beyond the baseline mean; that is, it is our models ability to explain the variance we observe in Y based on values of X. If we have no ability to explain variance, then our line will be flat (the slope will be 0.00) and will be the same as the line

representing the mean, and the distance between the lines will be 0.00 as well.

We now have three pieces of information: the distance from the observed score to the mean, the distance from the observed score to the prediction line, and the distance from the prediction line to the mean. These are our three pieces of information needed to test our hypotheses about regression and to calculate effect sizes. They are our three Sums of Squares, just like in ANOVA. Our distance from the observed score to the mean is the Sum of Squares Total, which we are trying to explain. Our distance from the observed score to the prediction line is our Sum of Squares Error, or residual, which we are trying to minimize. Our distance from the prediction line to the mean is our Sum of Squares Model, which is our observed effect and our ability to explain variance. Each of these will go into the ANOVA table to calculate our test statistic.

ANOVA Table

Our ANOVA table in regression follows the exact same format as it did for ANOVA (hence the name). Our top row is our observed effect, our middle row is our error, and our bottom row is our total. The columns take on the same interpretations as well: from left to right, we have our sums of squares, our degrees of freedom, our mean squares, and our *F* statistic.

| Source | SS | df | MS | F |
|--------|-----------------------|-------|-------------|-------------|
| Model | $\sum(\hat{Y} - Y)^2$ | 1 | SS_M/df_M | MS_M/MS_E |
| Error | $\sum(Y - \hat{Y})^2$ | $n-2$ | SS_E/df_E | |
| Total | $\sum(Y - \bar{Y})^2$ | $n-1$ | | |

As with ANOVA, getting the values for the SS column is a straightforward but somewhat arduous process. First, you take the raw scores of X and Y and calculate the means, variances, and covariance using the sum of products table introduced in our chapter on correlations. Next, you use the variance of X and the covariance of X and Y to calculate the slope of the line, b , the formula for which is given above. After that, you use the means and the slope to find the intercept, a , which is given alongside b . After that, you use the full prediction equation for the line of best fit to get predicted Y scores (\hat{Y}) for each person. Finally, you use the observed Y scores, predicted Y scores, and mean of Y to find the appropriate deviation scores for each person for each sum of squares source in the table and sum them to get the Sum of Squares Model, Sum of Squares Error, and Sum of Squares Total. As with ANOVA, you won't be required to compute the SS values by hand, but you will need to know what they represent and how they fit together. The other columns in the ANOVA table are all familiar. The degrees of freedom column still has $N - 1$ for our total, but now we have $N - 2$ for our error degrees of freedom and 1 for our model degrees of freedom; this is because simple linear regression only has one predictor, so our degrees of freedom for the model is always 1 and does not change. The total degrees of freedom must still be the sum of the other two, so our degrees of freedom error will always be $N - 2$ for simple linear regression. The mean square columns are still the SS column divided by the df column, and the test statistic F is still the ratio of the mean squares. Based on this, it is now explicitly clear that not only do regression and ANOVA have the same goal but they are, in fact, the same analysis entirely. The only difference is the type of data we feed into the predictor side of the equations: continuous for regression and categorical for ANOVA.

Hypothesis Testing in Regression

Regression, like all other analyses, will test a null hypothesis in our data. In regression, we are interested in predicting Y scores and explaining variance using a line, the slope of which is what allows us to get closer to our observed scores than the mean of Y can. Thus, our hypotheses concern the slope of the line, which is estimated in the prediction equation by b . Specifically, we want to test that the slope is not zero:

H_0 : There is no explanatory
relation between our variables,

$$H_0: \beta = 0$$

H_A : There is an explanatory
relation between our variables,

$$H_A: \beta \neq 0$$

or if directional – specify direction for relation (positive or negative), $H_A: \beta > 0$, $H_A: \beta < 0$

A non-zero slope indicates that we can explain values in Y based on X and therefore predict future values of Y based on X. Our alternative hypotheses are analogous to those in correlation: positive relations have values above zero, negative relations have values below zero, and two-tailed tests are possible. Just like ANOVA, we will test the significance of this relation using the F statistic calculated in our ANOVA table compared to a critical value from the F distribution table. Let's take a look at an example and regression in action.

Example: Happiness and Well-Being

Researchers are interested in explaining differences in how happy people are based on how healthy people are. They gather data on each of these variables from 18 people and fit a

linear regression model to explain the variance. We will follow the four-step hypothesis testing procedure to see if there is a relation between these variables that is statistically significant.

Step 1: State the Hypotheses

The null hypothesis in regression states that there is no relation between our variables. The alternative states that there is a relation, but because our research description did not explicitly state a direction of the relation, we will use a *non-directional hypothesis*.

H₀: There is no explanatory relation between health and happiness, H₀: $\beta = 0$

H_A: There is an explanatory relation between health and happiness, H_A: $\beta \neq 0$

Step 2: Find the Critical Value

Because regression and ANOVA are the same analysis, our critical value for regression will come from the same place: the *F* distribution table, which uses two types of degrees of freedom. We saw above that the degrees of freedom for our numerator – the Model line – is always 1 in simple linear regression, and that the denominator degrees of freedom – from the Error line – is $N - 2$. In this instance, we have 18 people so our degrees of freedom for the denominator is 16. Going to our *F* table, we find that the appropriate critical value for 1 and 16 degrees of freedom is $F^* = 4.49$, shown below in figure 3.

| df denom. | Degrees of Freedom: Numerator | | | | | | |
|--------------|-------------------------------|--------|--------|--------|--------|--------|--------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 |
| 10 | 4.97 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.10 | 3.01 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 |

Figure 3. Critical value from F distribution table

Step 3: Calculate the Test Statistic

The process of calculating the test statistic for regression first involves computing the parameter estimates for the line of best fit. To do this, we first calculate the means, standard deviations, and sum of products for our X and Y variables, as shown below.

| X | $(X - \bar{X})$ | $(X - \bar{X})^2$ | Y | $(Y - \bar{Y})$ | $(Y - \bar{Y})^2$ | $(X - \bar{X})(Y - \bar{Y})$ |
|-----------------------------------|-----------------|-------------------|--------|-----------------|-------------------|------------------------------|
| 17.65 | -2.13 | 4.53 | 10.36 | -7.10 | 50.37 | 15.10 |
| 16.99 | -2.79 | 7.80 | 16.38 | -1.08 | 1.16 | 3.01 |
| 18.30 | -1.48 | 2.18 | 15.23 | -2.23 | 4.97 | 3.29 |
| 18.28 | -1.50 | 2.25 | 14.26 | -3.19 | 10.18 | 4.79 |
| 21.89 | 2.11 | 4.47 | 17.71 | 0.26 | 0.07 | 0.55 |
| 22.61 | 2.83 | 8.01 | 16.47 | -0.98 | 0.97 | -2.79 |
| 17.42 | -2.36 | 5.57 | 16.89 | -0.56 | 0.32 | 1.33 |
| 20.35 | 0.57 | 0.32 | 18.74 | 1.29 | 1.66 | 0.73 |
| 18.89 | -0.89 | 0.79 | 21.96 | 4.50 | 20.26 | -4.00 |
| 18.63 | -1.15 | 1.32 | 17.57 | 0.11 | 0.01 | -0.13 |
| 19.67 | -0.11 | 0.01 | 18.12 | 0.66 | 0.44 | -0.08 |
| 18.39 | -1.39 | 1.94 | 12.08 | -5.37 | 28.87 | 7.48 |
| 22.48 | 2.71 | 7.32 | 17.11 | -0.34 | 0.12 | -0.93 |
| 23.25 | 3.47 | 12.07 | 21.66 | 4.21 | 17.73 | 14.63 |
| 19.91 | 0.13 | 0.02 | 17.86 | 0.40 | 0.16 | 0.05 |
| 18.21 | -1.57 | 2.45 | 18.49 | 1.03 | 1.07 | -1.62 |
| 23.65 | 3.87 | 14.99 | 22.13 | 4.67 | 21.82 | 18.08 |
| 19.45 | -0.33 | 0.11 | 21.17 | 3.72 | 13.82 | -1.22 |
| totals/Σ | | | | | | |
| 356.02 | 0.00 | 76.14 | 314.18 | 0.00 | 173.99 | 58.29 |

From the raw data in our X and Y columns, we find that the means are $\bar{X} = 19.78$ and $\bar{Y} = 17.45$. The deviation scores for each variable sum to zero, so all is well there. The sums of squares for X and Y ultimately lead us to standard deviations of $S_x = 2.12$ and $S_y = 3.20$. Finally, our sum of products is 58.29, which gives us a covariance of $\text{cov}_{XY} = 3.43$, so we know our

relation will be positive. This is all the information we need for our equations for the line of best.

First, we must calculate the slope of the line:

$$b = SS_x / SP = 58.29 / 76.14 = 0.77$$

This means that as X changes by 1 unit, Y will change by 0.77.

In terms of our problem, as health increases by 1, happiness goes up by 0.77, which is a positive relation. Next, we use the slope, along with the means of each variable, to compute the intercept:

$$a = \bar{Y} - b\bar{X} = 17.45 - (0.77 * 19.78) = 17.45 - 15.03 = 2.42$$

For this particular problem (and most regressions), the intercept is not an important or interpretable value, so we will not read into it further.

Now that we have all of our parameters estimated, we can give the full equation for our line of best fit:

$$\hat{Y} = 2.42 + 0.77X$$

We can plot this relation in a scatterplot and overlay our line onto it, as shown in figure 4.

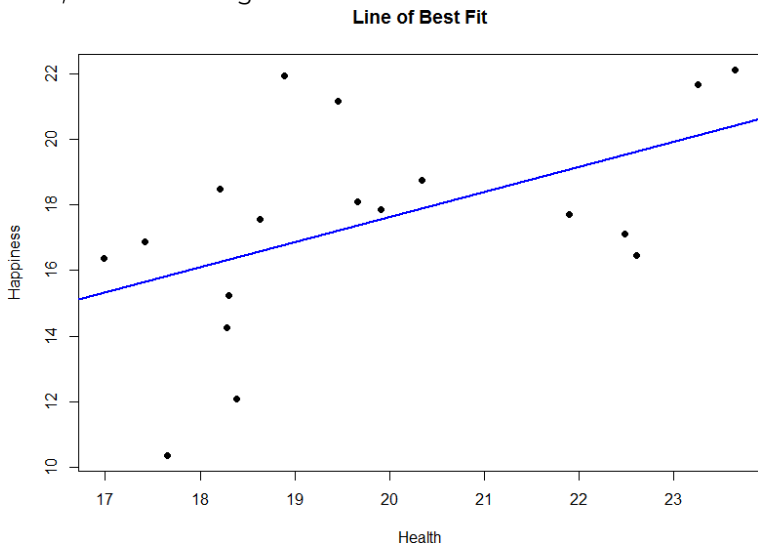


Figure 4. Health and happiness data and line.

We can use the line equation to find predicted values for each observation and use them to calculate our sums of squares model and error, but this is tedious to do by hand, so we will let the computer software do the heavy lifting in that column of our ANOVA table:

| Source | SS | df | MS | F |
|--------|--------|----|----|---|
| Model | 44.62 | | | |
| Error | 129.37 | | | |
| Total | | | | |

Now that we have these, we can fill in the rest of the ANOVA table. We already found our degrees of freedom in Step 2:

| Source | SS | df | MS | F |
|--------|--------|----|----|---|
| Model | 44.62 | 1 | | |
| Error | 129.37 | 16 | | |
| Total | | | | |

Our total line is always the sum of the other two lines, giving us:

| Source | SS | df | MS | F |
|--------|--------|----|----|---|
| Model | 44.62 | 1 | | |
| Error | 129.37 | 16 | | |
| Total | 173.99 | 17 | | |

Our mean squares column is only calculated for the model and error lines and is always our SS divided by our df, which is:

| Source | SS | df | MS | F |
|--------|--------|----|-------|---|
| Model | 44.62 | 1 | 44.62 | |
| Error | 129.37 | 16 | 8.09 | |
| Total | 173.99 | 17 | | |

Finally, our F statistic is the ratio of the mean squares:

| Source | SS | df | MS | F |
|--------|--------|----|-------|------|
| Model | 44.62 | 1 | 44.62 | 5.52 |
| Error | 129.37 | 16 | 8.09 | |
| Total | 173.99 | 17 | | |

This gives us an obtained F statistic of 5.52, which we will now use to test our hypothesis.

Step 4: Make the Decision

We now have everything we need to make our final decision. Our obtained test statistic was $F = 5.52$ and our critical value was $F^* = 4.49$. Since our obtained test statistic is greater than our critical value, we can reject the null hypothesis.

Reject H_0 . Based on our sample of 18 people, we can predict levels of happiness based on how healthy someone is, $F(1,16) = 5.52, p < .05$.

Effect Size

We know that, because we rejected the null hypothesis, we should calculate an effect size. In regression, our effect size is variance explained, just like it was in ANOVA. Instead of using η^2 to represent this, we instead use R^2 , as we saw in correlation

(yet more evidence that all of these are the same analysis).

From the example above, we get $R^2 = .26$. We are explaining 26% of the variance in happiness based on health, which is a large effect size (R^2 uses the same effect size cutoffs as η^2).

Accuracy in Prediction

We found a large, statistically significant relation between our variables, which is what we hoped for. However, if we want to use our estimated line of best fit for future prediction, we will also want to know how precise or accurate our predicted values are. What we want to know is the average distance from our predictions to our actual observed values, or the average size of the residual ($Y - \hat{Y}$). The average size of the residual is known by a specific name: the standard error of the estimate $s(Y - \hat{Y})$. The formula is almost identical to our standard deviation formula, and it follows the same logic. For our example, $s(Y - \hat{Y}) = 2.84$. So *on average, our predictions are just under 3 points away from our actual values*. There are no specific cutoffs or guidelines for how big our standard error of the estimate can or should be; it is highly dependent on both our sample size and the scale of our original Y variable, so expert judgment should be used. In this case, the estimate is not that far off and can be considered reasonably precise.

Quick recap of regression (without the math)

Two variables of regression

1. Predictor (X)
2. Criterion (Y)

With correlation it did not matter which variable was the

predictor variable or the criterion variable. But with prediction we have to decide which variable is being predicted from and which variable is being predicted. The variable being predicted from is called the **predictor variable**. The variable being predicted is called the **criterion variable**. In equations, the predictor variable is usually labeled X , and the criterion is labeled Y .

The Linear Prediction Rule: Ideally we want to make a prediction rule that is both simple and depends on every case for each prediction. In a linear prediction rule the formal name for the baseline number is the regression constant or just constant. It has the name constant because it is a fixed value that we always add in to the prediction.

The number we multiplied by the person's score on the predictor variable, b , is called the **regression coefficient** because a "coefficient" is a number we multiply by something.

Let's revisit example 2, predicting college GPA from SAT scores. For our SAT and GPA example, the rule might be "to predict a person's graduating GPA, start with .3 and at the result of multiplying .004 by the person's SAT scores". So, the baseline number (a) would be .3 and the predictor value (b) is .004. If a person had an SAT of 600 we would predict the person would graduate with a GPA of 2.7. This idea is known as the linear prediction rule. Lows go with lows and highs with highs, or lows with highs and highs with lows.

Criterion Variable (\hat{Y})

The variable we are predicting in a regression equation is called the criterion variable. It is labeled as \hat{Y} . The mark above Y indicates that this variable is a predicted variable and is dependent on the value of X .

Slope of the Regression Line (b)

The steepness of the angle of the regression line, called its slope, is the amount that the line moves up for every unit it is moved across. In our SAT example the line moves up .004 on the GPA scale for every additional point on the SAT. In fact, the slope of the line is exactly b , the regression coefficient.

Intercept of the Regression Line (a)

The point at which the regression line crosses or intersects the vertical axis is called the intercept.

- The intercept is the predicted score on the criterion variable when the score on the predictor variable is 0. It turns out that the intercept is the same as the regression constant.
- The reason this works is the regression constant is the number we always add in – a kind of baseline number, the number we start with.
- It is reasonable that the best baseline number would be the number we predict from a predictor score of 0.

In the SAT example the line crosses the vertical axis at .3. That is, when a person has an SAT score of zero, they are predicted to have a college GPA .3.

Linear regression standardized coefficient (β)

Standardized regression coefficient (β):

$$\beta = (b) \frac{\sqrt{SS_X}}{\sqrt{SS_Y}}$$

This formula has the effect of changing the regular (unstandardized) regression coefficient (b), to a standardized regression coefficient (β) that shows the relationship between the predictor and criterion variables in terms of standard deviation units.

Multiple Regression and Other Extensions

Simple linear regression as presented here is only a stepping stone towards an entire field of research and application. Regression is an incredibly flexible and powerful tool, and the extensions and variations on it are far beyond the scope of this chapter (indeed, even entire books struggle to accommodate all possible applications of the simple principles laid out here). The next step in regression is to study multiple regression, which uses multiple X variables as predictors for a single Y variable at the same time. The math of multiple regression is very complex but the logic is the same: we are trying to use variables that are statistically significantly related to our

outcome to explain the variance we observe in that outcome. Other forms of regression include curvilinear models that can explain curves in the data rather than the straight lines used here, as well as moderation models that change the relation between two variables based on levels of a third. The possibilities are truly endless and offer a lifetime of discovery.

Learning Objectives

Having read this chapter, a student should be able to:

- Explain the concept of a linear equation, including slope and intercept
- Explain how regression is related to correlation and ANOVA
- Understand the concept of least-square solution
- Understand the concept of multiple regression

Exercises – Ch. 17

1. How are ANOVA and linear regression similar? How are they different?
2. What is a residual?
3. How are correlation and regression similar? How are they different?
4. What are the two parameters of the line of best fit, and what do they represent?
5. What is our criteria for finding the line of best fit?
6. Fill out the rest of the ANOVA tables below for simple linear regressions: a.

| Source | SS | df | MS | F |
|--------|-------|----|-------|---|
| Model | 34.21 | 1 | 34.21 | |
| Error | | | | |
| Total | 66.12 | 54 | | |

7. In chapter 15, we found a statistically significant correlation between overall performance in class and how much time someone studied. Use the summary statistics calculated in that problem (provided here) to compute a line of best fit predicting success from study times: $\bar{X} = 1.61$, $s_X = 1.12$, $\bar{Y} = 2.95$, $s_Y = 0.99$, $r = 0.65$.

8. Using the line of best fit equation created in problem 7, predict the scores for how successful people will be based on how much they study:

- a. $X = 1.20$
- b. $X = 3.33$
- c. $X = 0.71$
- d. $X = 4.00$

9. You have become suspicious that the draft rankings of your fantasy football league have no predictive value for how teams place at the end of the season. You go back to historical league data and find rankings of teams after the draft and at the end of the season (below) to test for a statistically significant predictive relation. Assume $SSM = 2.65$ and $SSE = 337.35$

| Draft Projection | Final Rankings |
|------------------|----------------|
| 1 | 14 |
| 2 | 6 |
| 3 | 8 |
| 4 | 13 |
| 5 | 2 |
| 6 | 15 |
| 7 | 4 |
| 8 | 10 |
| 9 | 11 |
| 10 | 16 |
| 11 | 9 |
| 12 | 7 |
| 13 | 14 |
| 14 | 12 |
| 15 | 1 |
| 16 | 5 |

10. You have summary data for two variables: how extroverted some is (X) and how often someone volunteers (Y). Using these values, calculate the line of best fit predicting volunteering from extroversion then test for a statistically significant relation using the hypothesis testing procedure: $\bar{X} = 12.58$, $s_X = 4.65$, $\bar{Y} = 7.44$, $s_Y = 2.12$, $r = 0.34$, $N = 67$, $SSM = 19.79$, $SSE = 215.77$.

Answers to Odd- Numbered Exercises – Ch. 17

1. ANOVA and simple linear regression both take the total observed variance and partition it into pieces that we can explain and cannot explain and use the ratio of those pieces to test for significant relations. They are different in that ANOVA uses a categorical variable as a predictor whereas linear regression uses a continuous variable.

3. Correlation and regression both involve taking two continuous variables and finding a linear relation between them. Correlations find a standardized value describing the direction and magnitude of the relation whereas regression finds the line of best fit and uses it to partition and explain variance.

5. Least Squares Error Solution; the line that minimizes the total amount of residual error in the dataset.

7. $b = r(s_y/s_x) = 0.65*(0.99/1.12) = 0.72$; $a = \bar{Y} - b\bar{X} = 2.95 - (0.72*1.61) = 1.79$; $\hat{Y} = 1.79 + 0.72X$

9. Step 1: $H_0: \beta = 0$ "There is no predictive relation between draft rankings and final rankings in fantasy football," $H_A: \beta \neq 0$, "There is a predictive relation between draft rankings and final rankings in fantasy football."

Step 2: Our model will have 1 (based on the number of predictors) and 14 (based on how many observations we have) degrees of freedom, giving us a critical value of $F^* = 4.60$.

Step 3: Using the sum of products table, we find: $\bar{X} = 8.50$, $\bar{Y} = 8.50$, $SS_X = 339.86$, $SP = 29.99$, giving us a line of best fit of: $b = 29.99/339.86 = 0.09$; $a = 8.50 - 0.09*8.50 = 7.74$; $\hat{Y} = 7.74 + 0.09X$.

Our given SS values and our df from step 2 allow us to fill in the ANOVA table:

| Source | SS | df | MS | F |
|--------|--------|----|-------|------|
| Model | 2.65 | 1 | 2.65 | 0.11 |
| Error | 337.35 | 14 | 24.10 | |
| Total | 339.86 | 15 | | |

Step 4: Our obtained value was smaller than our critical value, so we fail to reject the null hypothesis. There is no evidence to suggest that draft rankings have any predictive value for final fantasy football rankings, $F(1,14) = 0.11$, $p > .05$

18. Chapter 18.

Chi-square

We come at last to our final statistic: chi-square (χ^2). This test is a special form of analysis called a non-parametric test, so the structure of it will look a little bit different from what we have done so far. However, the logic of hypothesis testing remains unchanged. The purpose of chi-square is to understand the frequency distribution of a single categorical variable or find a relation between two categorical variables, which is a frequently very useful way to look at our data.

Categories and Frequency Tables

Our data for the χ^2 test are categorical, specifically nominal, variables. Recall from unit 1 that nominal variables have no specified order and can only be described by their names and the frequencies with which they occur in the dataset. Thus, unlike our other variables that we have tested, we cannot describe our data for the χ^2 test using means and standard deviations. Instead, we will use frequencies tables.

| | Cat | Dog | Other | Total |
|----------|-----|-----|-------|-------|
| Observed | 14 | 17 | 5 | 36 |
| Expected | 12 | 12 | 12 | 36 |

Table 1. Pet Preferences

Table 1 gives an example of a frequency table used for a χ^2 test. The columns represent the different categories within our single variable, which in this example is pet preference. The

χ^2 test can assess as few as two categories, and there is no technical upper limit on how many categories can be included in our variable, although, as with ANOVA, having too many categories makes our computations long and our interpretation difficult. The final column in the table is the total number of observations, or N. The χ^2 test assumes that each observation comes from only one person and that each person will provide only one observation, so our total observations will always equal our sample size.

There are two rows in this table. The first row gives the observed frequencies of each category from our dataset; in this example, 14 people reported liking preferring cats as pets, 17 people reported preferring dogs, and 5 people reported a different animal. The second row gives expected values; expected values are what would be found if each category had equal representation.

The calculation for an expected value is:

$$E = N / C$$

Where N is the total number of people in our sample and C is the number of categories in our variable (also the number of columns in our table).

The expected values correspond to the null hypothesis for χ^2 tests: equal representation of categories. Our first of two χ^2 tests, the Goodness-of-Fit test, will assess how well our data lines up with, or deviates from, this assumption.

Goodness-of-Fit

The first of our two χ^2 tests assesses one categorical variable against a null hypothesis of equally sized frequencies. Equal frequency distributions are what we would expect to get if categorization was completely random. We could, in theory, also test against a specific distribution of category sizes if we have a good reason to (e.g. we have a solid foundation of how the regular population is distributed), but this is less common, so we will not deal with it in this text.

Hypotheses

All χ^2 tests, including the goodness-of-fit test, are **non-parametric**. This means that there is no population parameter we are estimating or testing against; we are working only with our sample data. Because of this, there are no mathematical statements for χ^2 hypotheses. This should make sense because the mathematical hypothesis statements were always about population parameters (e.g. μ), so if we are non-parametric, we have no parameters and therefore no mathematical statements.

We do, however, still state our hypotheses verbally. For goodness-of-fit χ^2 tests, our null hypothesis is that there is an equal number of observations in each category. That is, there is no difference between the categories in how prevalent they are. Our alternative hypothesis says that the categories do differ in their frequency. We do not have specific directions or one-tailed tests for χ^2 , matching our lack of mathematical statement.

Degrees of Freedom and the χ^2 table

Our degrees of freedom for the χ^2 test are based on the number of categories we have in our variable, not on the number of people or observations like it was for our other tests. Luckily, they are still as simple to calculate.

degrees of freedom for χ^2 Goodness of fit

$$df = C - 1$$

So for our pet preference example, we have 3 categories, so we have 2 degrees of freedom. Our degrees of freedom, along with our significance level (still defaulted to $\alpha = 0.05$) are used to find our critical values in the χ^2 table, which is shown in figure 1. Because we do not have directional hypotheses for χ^2 tests, we do not need to differentiate between critical values for 1- or 2-tailed tests. In fact, just like our F tests for regression and ANOVA, all χ^2 tests are 1-tailed tests.

| df | 1-tailed α | | | |
|----|-------------------|--------|--------|--------|
| | 0.10 | 0.05 | 0.02 | 0.01 |
| 1 | 2.706 | 3.841 | 5.024 | 6.635 |
| 2 | 4.605 | 5.991 | 7.378 | 9.210 |
| 3 | 6.251 | 7.815 | 9.348 | 11.345 |
| 4 | 7.779 | 9.488 | 11.143 | 13.277 |
| 5 | 9.236 | 11.070 | 12.833 | 15.086 |
| 6 | 10.645 | 12.592 | 14.449 | 16.812 |
| 7 | 12.017 | 14.067 | 16.013 | 18.475 |
| 8 | 13.362 | 15.507 | 17.535 | 20.090 |
| 9 | 14.684 | 16.919 | 19.023 | 21.666 |
| 10 | 15.987 | 18.307 | 20.483 | 23.209 |

Figure 1. First 10 rows of the χ^2 table

χ^2 Statistic

The calculations for our test statistic in χ^2 tests combine our information from our observed frequencies (O) and our expected frequencies (E) for each level of our categorical variable. For each cell (category) we find the difference between the observed and expected values, square them, and divide by the expected values. We then sum this value across cells for our test statistic.

$$\chi^2$$

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

For our pet preference data, we would have:

$$\chi^2 = \frac{(14 - 12)^2}{12} + \frac{(17 - 12)^2}{12} + \frac{(5 - 12)^2}{12} = 0.33 + 2.08 + 4.08 = 6.49$$

Notice that, for each cell's calculation, the expected value in the numerator and the expected value in the denominator are the same value. Let's now take a look at an example from start to finish.

Goodness-of-Fit Example: Pineapple on Pizza

There is a very passionate and on-going debate on whether or not pineapple should go on pizza. Being the objective, rational data analysts that we are, we will collect empirical data to see if we can settle this debate once and for all. We gather data from a group of adults asking for a simple Yes/No answer.

Step 1: State the Hypotheses

We start, as always, with our hypotheses. Our null hypothesis of no difference will state that an equal number of people will say they do or do not like pineapple on pizza, and our alternative will be that one side wins out over the other:

H_0 : An equal number of people do and do not like pizza.

H_A : A significant majority of people will agree one way or another

Step 2: Find the Critical Value

To avoid any potential bias in this crucial analysis, we will leave α at its typical level. We have two options in our data (Yes or No), which will give us two categories. Based on this, we will have 1 degree of freedom. From our χ^2 table, we find a critical value of 3.84.

Step 3: Calculate the Test Statistic

The results of the data collection are presented in table 2. We had data from 45 people in all and 2 categories, so our expected values are $E = 45/2 = 22.50$.

| | Yes | No | Total |
|----------|-------|-------|-------|
| Observed | 26 | 19 | 45 |
| Expected | 22.50 | 22.50 | 45 |

We can use these to calculate our χ^2 statistic:

$$\chi^2 = \frac{(26 - 22.50)^2}{22.50} + \frac{(19 - 22.50)^2}{22.50} = 0.54 + 0.54 = 1.08$$

Step 4: Make the Decision

Our observed test statistic had a value of 1.08 and our critical value was 3.84. Our test statistic was smaller than our critical value, so we fail to reject the null hypothesis, and the debate rages on.

Goodness-of-Fit Example 2: Favorite candy

We can also use goodness of fit to determine if we see differences in people's selection of favorite candy. To keep it simple, we had 3 categories to select from: chocolate, licorice, and bubblegum. We gathered data from a group of adults to select from the three options which was the favorite.

Step 1: State the Hypotheses

Our null hypothesis of no difference will state that an equal number of people select favorite candy, and our alternative will be that one type of candy is more popular:

H_0 : The proportion of each type of candy is equal. People have evenly distributed candy preference among our 3 choices.

HA: The proportion of each type of candy is not equal. There is an unequal distribution for candy preference.

Step 2: Find the Critical Value

To avoid any potential bias in this crucial analysis, we will leave α at its typical level. We have three options for favorite candy. Based on this, we will have 2 degree of freedom. From our χ^2 table, we find a critical value of 5.99.

Step 3: Calculate Statistic

The results of the data collection are presented in table 3. We had data from 100 people in all and 3 categories, so our expected values are $E = 100/3 = 33.333$.

| Candy Type | Count | Expected |
|------------|-------|----------|
| chocolate | 30 | 33.333 |
| licorice | 33 | 33.333 |
| gumball | 37 | 33.333 |

Table 3 Observed and expected counts for candy data

We can use these to calculate our χ^2 statistic: $\chi^2 = 11.11/33.33 + 0.11/33.33 + 13.44/33.33 = 0.333 + 0.003 + 0.403 = 0.739$

Step 4: Make the Decision

For the candy example, the observed counts of candies are not particularly surprising based on the proportions printed on the bag of candy, and we would not reject the null hypothesis of equal proportions.

Contingency Tables for Two Variables

The goodness-of-fit test is a useful tool for assessing a single categorical variable. However, what is more common is wanting to know if two categorical variables are related to one another. This type of analysis is similar to a correlation, the only difference being that we are working with nominal data, which violates the assumptions of traditional correlation coefficients. This is where the **χ^2 test for independence** comes in handy.

As noted above, our only description for nominal data is frequency, so we will again present our observations in a frequency table. When we have two categorical variables, our frequency table is crossed. That is, each combination of levels from each categorical variable are presented. This type of frequency table is called a **contingency table** because it shows the frequency of each category in one variable, contingent upon the specific level of the other variable. An example contingency table is shown in table 3, which displays whether or not 168 college students watched college sports growing up (Yes/No) and whether the students' final choice of which college to attend was influenced by the college's sports teams (Yes – Primary, Yes – Somewhat, No):

| College Sports | | Affected Decision | | | Total |
|----------------|-----|-------------------|----------|----|-------|
| | | Primary | Somewhat | No | |
| Watched | Yes | 47 | 26 | 14 | 87 |
| | No | 21 | 23 | 37 | 81 |
| Total | | 68 | 49 | 51 | 168 |

Table 3. Contingency table of college sports and decision making

In contrast to the frequency table for our goodness-of-fit test, our contingency table does not contain expected values, only observed data. Within our table, wherever our rows and columns cross, we have a cell. A cell contains the frequency of observing it's corresponding specific levels of each variable at the same time. The top left cell in table 3 shows us that 47 people in our study watched college sports as a child AND had college sports as their primary deciding factor in which college to attend.

Cells are numbered based on which row they are in (rows are numbered top to bottom) and which column they are in (columns are numbered left to right). We always name the cell using (R,C), with the row first and the column second. A quick and easy way to remember the order is that R/C Cola exists but C/R Cola does not. Based on this convention, the top left cell containing our 47 participants who watched college sports as a child and had sports as a primary criteria is cell (1,1). Next to it, which has 26 people who watched college sports as a child but had sports only somewhat affect their decision, is cell (1,2), and so on. We only number the cells where our categories cross. We do not number our total cells, which have their own special name: marginal values. Marginal values are the total values for a single category of one variable, added up across levels of the other variable. In table 3, these marginal values have been italicized for ease of explanation, though this is not normally the case. We can see that, in total, 87 of our participants (47+26+14) watched college sports growing up and 81 (21+23+37) did not. The total of these two marginal values is 168, the total number of people in our study. Likewise, 68 people used sports as a primary criteria for deciding which college to attend, 50 considered it somewhat, and 50 did not use it as criteria at all. The total of these marginal values is also

168, our total number of people. The marginal values for rows and columns will always both add up to the total number of participants, N , in the study. If they do not, then a calculation error was made and you must go back and check your work.

Expected Values of Contingency Tables

Our expected values for contingency tables are based on the same logic as they were for frequency tables, but now we must incorporate information about how frequently each row and column was observed (the marginal values) and how many people were in the sample overall (N) to find what random chance would have made the frequencies out to be.

Expected values formula

$$E(r, c) = \frac{n(r) \times c(r)}{n}$$

where:

r = row in question

c = column in question

n = corresponding total

The subscripts $n(r)$ is the count for the row and $n(c)$ the count for the column, respectively, correspond to the cell we are calculating the expected frequency for, and n is still the total sample size.

Example: Using the data from table 3, we can calculate the expected frequency for cell, $E(1,1)$, the college sport watchers who used sports at their primary criteria, is

$$E_{(1,1)} = (87)(68) / 168 = 35.21$$

| College Sports | | Affected Decision | | | Total |
|----------------|-------|-------------------|----------|----|-------|
| | | Primary | Somewhat | No | |
| Watched | Yes | 47 | 26 | 14 | 87 |
| | No | 21 | 23 | 7 | 51 |
| | Total | 68 | 49 | 51 | 168 |

We can follow the same math to find all the expected values for this table:

| Expected Values | | Affected Decision | | | Total |
|-----------------|-------|-------------------|----------|-------|-------|
| | | Primary | Somewhat | No | |
| Watched | Yes | 35.21 | 25.38 | 26.41 | 87 |
| | No | 32.79 | 23.62 | 24.59 | 81 |
| | Total | 68 | 49 | 51 | 168 |

Table 4. Expected Values derived from Table 3.

Notice that the marginal values still add up to the same totals as before. This is because the expected frequencies are just row

and column averages simultaneously. Our total N will also add up to the same value.

The observed and expected frequencies can be used to calculate the same χ^2 statistic as we did for the goodness-of-fit test. Before we get there, though, we should look at the hypotheses and degrees of freedom used for contingency tables.

Test for Independence

The χ^2 test performed on contingency tables is known as the test for independence. In this analysis, we are looking to see if the values of each categorical variable (that is, the frequency of their levels) is related to or independent of the values of the other categorical variable. Because we are still doing a χ^2 test, which is non-parametric, we still do not have mathematical versions of our hypotheses. The actual interpretations of the hypotheses are quite simple: the null hypothesis says that the variables are independent or not related, and alternative says that they are not independent or that they are related. For step 2, the only change is degrees of formula. Our critical value will come from the same table that we used for the goodness-of-fit test, but our degrees of freedom will change. Because we now have rows and columns (instead of just columns) our new degrees of freedom.

degrees of freedom for χ^2 independence test

$$df = (R - 1)(C - 1)$$

For step 3, we still use the χ^2 but we need to compute expected frequencies. Step 4 is the same process. Let's see an example.

Example: College Sports

Using this set up and the data provided in table 3, let's formally test for whether or not watching college sports as a child is related to using sports as a criteria for selecting a college to attend. We will follow the same 4 step procedure as we have since chapter 7.

Step 1: Hypotheses

Our null hypothesis of no difference will state that there is no relation between our variables, and our alternative will state that our variables are related (in other words there is a relationship): H_0 : College choice criteria is independent of college sports viewing as a child. H_A : College choice criteria is related of college sports viewing as a child.

Step 2: Criteria

In our example: $df = (2 - 1)(3 - 1) = 1 * 2 = 2$. Based on our 2 degrees of freedom, our critical value from using the table is 5.991. You use the same critical value table as goodness of fit as it is only the degrees of freedom calculation that has changed.

Step 3: Calculate the Test Statistic

The same formula for χ^2 is used once again. We are using the expected frequency values from table 4:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$\chi^2 = \frac{(47 - 35.21)^2}{35.21} + \frac{(26 - 25.38)^2}{25.38} + \frac{(14 - 26.41)^2}{26.41} + \frac{(21 - 32.79)^2}{32.79} + \frac{(23 - 23.62)^2}{23.62} + \frac{(37 - 24.59)^2}{24.59}$$

$$= 3.94 + 0.02 + 5.83 + 4.24 + 0.02 + 6.26 = 20.31$$

Step 4: Decision

The final decision for our test of independence is still based on our observed value (20.31) and our critical value (5.991). Because our observed value is greater than our critical value, we can reject the null hypothesis.

Reject H_0 . Based on our data from 168 people, we can say that there is a statistically significant relation between whether or not someone watches college sports growing up and how much a college's sports team factor in to that person's decision on which college to attend, $\chi^2(2) = 20.31, p < 0.05$.

Effect Size for χ^2

Like all other significance tests, χ^2 tests – *both* goodness-of-fit and tests for independence – have effect sizes that can and should be calculated for statistically significant results. There are many options for which effect size to use, and the ultimate decision is based on the type of data, the structure of your

frequency or contingency table, and the types of conclusions you would like to draw. For the purpose of our introductory course, we will focus only on a single effect size that is simple and flexible: Cramer's V .

Cramer's V is a type of correlation coefficient that can be computed on categorical data.

Cramer's V formula

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

For this calculation, k is the smaller value of either R (the number of rows) or C (the number of columns). The numerator is simply the test statistic (χ^2) we calculate during step 3 of the hypothesis testing procedure.

Example Continued: College Sports

Effect size

For our example, we had 2 rows and 3 columns, so $k = 2$:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{20.38}{168(2-1)}} = \sqrt{0.12} = 0.35$$

So the statistically significant relation between our variables was moderately strong examining the effect size table below.

Like other statistic effect sizes there are range cut offs of small, medium, and large. The effect size ranges of Cramer's V are in Table 6.

| | small | medium | large |
|--------|-------|--------|-------|
| df = 1 | 0.10 | 0.30 | 0.50 |
| df = 2 | 0.07 | 0.21 | 0.35 |
| df = 3 | 0.06 | 0.17 | 0.29 |

Beyond Pearson's Chi-Square Test: Standardized Residuals

For a more applicable example, let's take the question of whether a Black driver is more likely to be searched when they are pulled over by a police officer, compared to a white driver. The Stanford Open Policing Project (<https://openpolicing.stanford.edu/>) has studied this, and provides data that we can use to analyze the question. We will use the data from the State of Connecticut since they are fairly small and thus easier to analyze.

The standard way to represent data from a categorical analysis is through a *contingency table*, which presents the number or proportion of observations falling into each possible combination of values for each of the variables. Table 6 below shows the contingency table for the police search data. It can also be useful to look at the contingency table using proportions rather than raw numbers, since they are easier to compare visually, so we include both absolute and relative numbers here.

| searched | Black | White | Black (relative) | White (relative) |
|----------|-------|--------|------------------|------------------|
| FALSE | 36244 | 239241 | 0.13 | 0.86 |
| TRUE | 1219 | 3108 | 0.00 | 0.01 |

Table 6. Contingency Table for Police Search Data

The Pearson chi-squared test (discussed above) allows us to test whether observed frequencies are different from expected frequencies, so we need to determine what frequencies we would expect in each cell if searches and race were unrelated – which we can define as being *independent*. If we perform this test easily using our statistical software, $\chi^2(1) = 828, p < .001$. This shows that the observed data would be highly unlikely if there was truly no relationship between race and police searches, and thus we should reject the null hypothesis of independence.

When we find a significant effect with the chi-squared test, this tells us that the data are unlikely under the null hypothesis, but it doesn't tell us *how* the data differ. To get a deeper insight into how the data differ from what we would expect under the null hypothesis, we can examine the residuals from the model, which reflects the deviation of the data (i.e., the observed frequencies) from the model (i.e., the expected frequencies) in each cell. Rather than looking at the raw residuals (which will vary simply depending on the number of observations in the

data), it's more common to look at the **standardized residuals** (sometimes called *Pearson residuals*).

Table 7 shows these for the police stop data from χ^2 above. Remember that we examined the question of whether a Black driver is more likely to be searched when they are pulled over by a police officer, compared to a white driver. These standardized residuals can be interpreted as Z scores – in this case, we see that the number of searches for Black individuals are substantially higher than expected based on independence, and the number of searches for white individuals are substantially lower than expected. This provides us with the context that we need to interpret the significant chi-squared result.

| searched | driver_race | Standardized residuals |
|----------|-------------|------------------------|
| FALSE | Black | -3.3 |
| TRUE | Black | 26.6 |
| FALSE | White | 1.3 |
| TRUE | White | -10.4 |

Table 7. Summary of standardized residuals for police stop data

Beware of Simpson's paradox

The contingency tables that represent summaries of large numbers of observations, but summaries can sometimes be misleading. Let's take an example from baseball. The table below shows the

batting data (hits/at bats and batting average) for Derek Jeter and David Justice over the years 1995-1997:

| Player | 1995 | | 1996 | | 1997 | | Combined | |
|---------------|-------|------|------|------|------|------|----------|------|
| Derek Jeter | 12/48 | .250 | 3/58 | .052 | 0/65 | .000 | 385/1284 | .300 |
| David Justice | 10/41 | .244 | 5/14 | .357 | 3/49 | .061 | 312/1046 | .298 |

Table 9. Player Batting data for 2 baseball players

If you look closely, you will see that something odd is going on: In each individual year Justice had a higher batting average than Jeter, but when we combine the data across all three years, Jeter's average is actually higher than Justice's! This is an example of a phenomenon known as ***Simpson's paradox***, in which a pattern that is present in a combined dataset may not be present in any of the subsets of the data. This occurs when there is another variable that may be changing across the different subsets – in this case, the number of at-bats varies across years, with Justice batting many more times in 1995 (when batting averages were low). We refer to this as a *lurking variable*, and it's always important to be attentive to such variables whenever one examines categorical data.

Learning objectives

Having read the chapter, a student should be able to:

- Identify when appropriate to run a chi-square test of goodness-of-fit or independence.
- Describe the concept of a contingency table for categorical data.
- Compute it for a given contingency table.
- Complete hypothesis test for chi-square test of goodness-of-fit and independence.
- Compute and interpret effect size for chi-square chi-square test of goodness-of-fit or independence.
- Describe Simpson's paradox and why it is important for categorical data analysis.

Exercises – Ch. 18

1. What does a frequency table display? What does a contingency table display?
2. What does a goodness-of-fit test assess?
 3. How do expected frequencies relate to the null hypothesis?
 4. What does a test-for-independence assess?
 5. Compute the expected frequencies for the following contingency table:

| | Category A | Category B |
|------------|------------|------------|
| Category C | 22 | 38 |
| Category D | 16 | 14 |

6. Test significance and find effect sizes (if significant) for the following tests:

1. $N = 19, R = 3, C = 2, \chi^2(2) = 7.89, \alpha = .05$

2. $N = 12, R = 2, C = 2, \chi^2 (1) = 3.12, \alpha = .05$
3. $N = 74, R = 3, C = 3, \chi^2 (4) = 28.41, \alpha = .01$

7. You hear a lot of people claim that *The Empire Strikes Back* is the best movie in the original Star Wars trilogy, and you decide to collect some data to demonstrate this empirically (pun intended). You ask 48 people which of the original movies they liked best; 8 said *A New Hope* was their favorite, 23 said *The Empire Strikes Back* was their favorite, and 17 said *Return of the Jedi* was their favorite. Perform a chi-square test on these data at the .05 level of significance.

8. A pizza company wants to know if people order the same number of different toppings. They look at how many pepperoni, sausage, and cheese pizzas were ordered in the last week; fill out the rest of the frequency table and test for a difference.

| | Pepperoni | Sausage | Cheese | Total |
|----------|-----------|---------|--------|-------|
| Observed | 320 | 275 | 251 | |
| Expected | | | | |

9. A university administrator wants to know if there is a difference in proportions of students who go on to grad school across different majors. Use the data below to test whether there is a relation between college major and going to grad school.

| | | Major | | |
|-----------------|-----|------------|----------|------|
| | | Psychology | Business | Math |
| Graduate School | Yes | 32 | 8 | 36 |
| | No | 15 | 41 | 12 |

10.A company you work for wants to make sure that they are not discriminating against anyone in their promotion process. You have been asked to look across gender to see if there are differences in promotion rate (i.e. if gender and promotion rate are independent or not). The following data should be assessed at the normal level of significance:

| | | Promoted in last two years? | |
|--------|-------|-----------------------------|----|
| | | Yes | No |
| Gender | Women | 8 | 5 |
| | Men | 9 | 7 |

Answers to Odd- Numbered Exercises – Ch. 18

- Frequency tables display observed category frequencies and (sometimes) expected category frequencies for a single categorical variable. Contingency tables display the frequency of observing people in crossed category levels for two categorical variables, and (sometimes) the marginal totals of each variable level.
- Expected values are what we would observe if the proportion of categories was completely random (i.e. no consistent difference other than chance), which is the same

was what the null hypothesis predicts to be true.

5.

| Observed | Category A | Category B | Total |
|------------|------------|------------|-------|
| Category C | 22 | 38 | 60 |
| Category D | 16 | 14 | 30 |
| Total | 38 | 52 | 90 |

| Expected | Category A | Category B | Total |
|------------|---------------------------|---------------------------|-------|
| Category C | $((60*38)/90)$ = 25.33 | $((60*52)/90)$ = 34.67 | 60 |
| Category D | $((30*38)/90)$ = 12.67 | $((30*52)/90)$ = 17.33 | 30 |
| Total | 38 | 52 | 90 |

7. Step 1: H_0 : “There is no difference in preference for one movie”, H_A : “There is a difference in how many people prefer one movie over the others.” Step 2: 3 categories (columns) gives $df = 2$, $\chi^2_{crit} = 5.991$. Step 3: Based on the given frequencies:

| | New Hope | Empire | Jedi | Total |
|----------|----------|--------|------|-------|
| Observed | 8 | 23 | 17 | 48 |
| Expected | 16 | 16 | 16 | |

$\chi^2 = 7.13$. Step 4: Our obtained statistic is greater than our critical value, reject H_0 . Based on our sample of 48 people, there is a statistically significant difference in the proportion of people who prefer one Star Wars movie over the others, $\chi^2(2) = 7.13$, $p < .05$. Since this is a statistically significant result, we should calculate an effect size: Cramer’s $V = \sqrt{7.13/48(3-1)} = 0.27$, which is a moderate effect size.

9. Step 1: H_0 : “There is no relation between college major and

going to grad school”, H_A : “Going to grad school is related to college major.” Step 2: $df = 2$, $\chi^2_{crit} = 5.991$. Step 3: Based on the given frequencies:

| Expected Values | | Major | | |
|-----------------|-----|------------|----------|-------|
| | | Psychology | Business | Math |
| Graduate School | Yes | 24.81 | 25.86 | 25.33 |
| | No | 22.19 | 23.14 | 22.67 |

$\chi^2 = 2.09+12.34+4.49+2.33+13.79+5.02 = 40.05$. Step 4: Obtained statistic is greater than the critical value, reject H_0 . Based on our data, there is a statistically significant relation between college major and going to grad school, $\chi^2(2) = 40.05$, $p < .05$, Cramer’s $V = 0.53$, which is a large effect.

19. Chapter 19. Doing reproducible research

Most people think that science is a reliable way to answer questions about the world. When our physician prescribes a treatment we trust that it has been shown to be effective through research, and we have similar faith that the airplanes that we fly in aren't going to fall from the sky. However, since 2005 there has been an increasing concern that science may not always work as well as we have long thought that it does. In this chapter we will discuss these concerns about reproducibility of scientific research, and outline the steps that one can take to make sure that our statistical results are as reproducible as possible.

How we think science should work

Let's say that we are interested in a research project on how children choose what to eat. This is a question that was asked in a study by the well-known eating researcher Brian Wansink and his colleagues in 2012. The standard (and, as we will see, somewhat naive) view goes something like this:

- You start with a hypothesis
 - Branding with popular characters should cause children to choose “healthy” food more often
- You collect some data
 - Offer children the choice between a cookie and an apple with either an Elmo-branded sticker or a control sticker, and record what they choose

- You do statistics to test the null hypothesis
 - “The preplanned comparison shows Elmo-branded apples were associated with an increase in a child’s selection of an apple over a cookie, from 20.7% to 33.8% ($\chi^2_{\text{chi}^2}=5.158$; $P=.02$)” (Wansink, Just, and Payne [2012](#))
- You make a conclusion based on the data
 - “This study suggests that the use of branding or appealing branded characters may benefit healthier foods more than they benefit indulgent, more highly processed foods. Just as attractive names have been shown to increase the selection of healthier foods in school lunchrooms, brands and cartoon characters could do the same with young children.”(Wansink, Just, and Payne [2012](#))

How science (sometimes) actually works

Brian Wansink is well known for his books on “Mindless Eating”, and his fee for corporate speaking engagements was at one point in the tens of thousands of dollars. In 2017, a set of researchers began to scrutinize some of his published research, starting with a set of papers about how much pizza people ate at a buffet. The researchers asked Wansink to share the data from the studies but he refused, so they dug into his published papers and found a large number of inconsistencies and statistical problems in the papers. The publicity around this analysis led a number of others to dig into Wansink’s past, including obtaining emails between Wansink and his collaborators. As [reported by Stephanie Lee at Buzzfeed](#), these emails showed just how far Wansink’s actual research practices were from the naive model:

...back in September 2008, when Payne was looking over the data soon after it had been collected, he found no strong apples-and-Elmo link — at least not yet. ... “I have attached some initial results of the kid study to this message for your report,” Payne wrote to his collaborators. “Do not despair. It looks like stickers on fruit may work (with a bit more wizardry).” ... Wansink also acknowledged the paper was weak as he was preparing to submit it to journals. The p-value was 0.06, just shy of the gold standard cutoff of 0.05. It was a “sticking point,” as he put it in a Jan. 7, 2012, email. ... “It seems to me it should be lower,” he wrote, attaching a draft. “Do you want to take a look at it and see what you think. If you can get the data, and it needs some tweeking, it would be good to get that one value below .05.” ... Later in 2012, the study appeared in the prestigious JAMA Pediatrics, the 0.06 p-value intact. But in September 2017, it was retracted and replaced with a version that listed a p-value of 0.02. And a month later, it was retracted yet again for an entirely different reason: Wansink admitted that the experiment had not been done on 8- to 11-year-olds, as he’d originally claimed, but on preschoolers.

This kind of behavior finally caught up with Wansink; [fifteen of his research studies have been retracted](#) and in 2018 he resigned from his faculty position at Cornell University.

The reproducibility crisis in science

While we think that the kind of fraudulent behavior seen in Wansink’s case is relatively rare, it has become increasingly clear that problems with reproducibility are much more widespread in science than previously thought. This became

particularly evident in 2015, when a large group of researchers published a study in the journal *Science* titled “Estimating the reproducibility of psychological science”(Open Science Collaboration [2015](#)). In this paper, the researchers took 100 published studies in psychology and attempted to reproduce the results originally reported in the papers. Their findings were shocking: Whereas 97% of the original papers had reported statistically significant findings, only 37% of these effects were statistically significant in the replication study. Although these problems in psychology have received a great deal of attention, they seem to be present in nearly every area of science, from cancer biology (Errington et al. [2014](#)) and chemistry (Baker [2017](#)) to economics (Christensen and Miguel [2016](#)) and the social sciences (Camerer et al. [2018](#)).

The reproducibility crisis that emerged after 2010 was actually predicted by John Ioannidis, a physician from Stanford who wrote a paper in 2005 titled “Why most published research findings are false”(Ioannidis [2005](#)). In this article, Ioannidis argued that the use of null hypothesis statistical testing in the context of modern science will necessarily lead to high levels of false results. Additionally, statistical power remains low in many areas of science (Smaldino and McElreath, [2016](#)), suggesting that many published research findings are false. An amusing example of this was seen in a paper by Jonathan Schoenfeld and John Ioannidis, titled “Is everything we eat associated with cancer? A systematic cookbook review”(Schoenfeld and Ioannidis, [2013](#)). They examined a large number of papers that had assessed the relation between different foods and cancer risk, and found that 80% of ingredients had been associated with either increased or decreased cancer risk. In most of these cases, the statistical evidence was weak, and when the results were combined across studies, the result was null.

Low power

Another kind of error can also occur when statistical power is low: Our estimates of the effect size will be inflated. This phenomenon often goes by the term “winner’s curse”, which comes from economics, where it refers to the fact that for certain types of auctions (where the value is the same for everyone, like a jar of quarters, and the bids are private), the winner is guaranteed to pay more than the good is worth. In science, the winner’s curse refers to the fact that the effect size estimated from a significant result (i.e. a winner) is almost always an overestimate of the true effect size.

We can simulate this in order to see how the estimated effect size for significant results is related to the actual underlying effect size. Let’s generate data for which there is a true effect size of $d = 0.2$, and estimate the effect size for those results where there is a significant effect detected. The left panel of Figure 20.2 shows that when power is low, the estimated effect size for significant results can be highly inflated compared to the actual effect size.

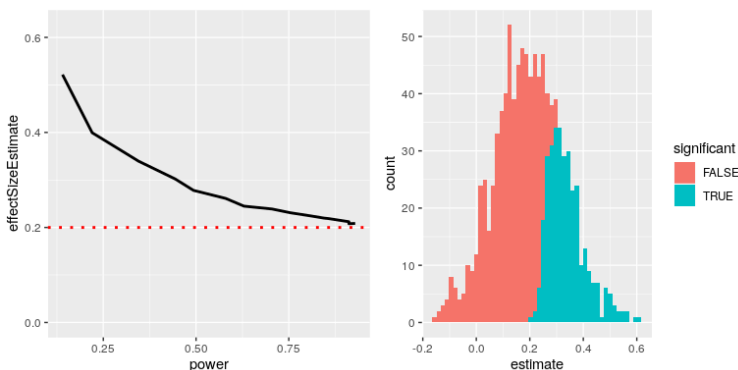


Figure 20.2: Left: A simulation of the winner’s curse as a function of statistical power (x axis). The solid line shows the estimated effect size, and the dotted line shows the actual

effect size. Right: A histogram showing effect size estimates for a number of samples from a dataset, with significant results shown in blue and non-significant results in red.

We can look at a single simulation to see why this is the case. In the right panel of Figure [20.2](#), you can see a histogram of the estimated effect sizes for 1000 samples, separated by whether the test was statistically significant. It should be clear from the figure that if we estimate the effect size only based on significant results, then our estimate will be inflated; only when most results are significant (i.e. power is high and the effect is relatively large) will our estimate come near the actual effect size.

Questionable research practices (QRPs)

A popular book entitled “The Compleat Academic: A Career Guide”, published by the American Psychological Association (Darley, Zanna, and Roediger, [2004](#)), aims to provide aspiring researchers with guidance on how to build a career. In a chapter by well-known social psychologist Daryl Bem titled “Writing the Empirical Journal Article”, Bem provides some suggestions about how to write a research paper. Unfortunately, the practices that he suggests are deeply problematic, and have come to be known as **questionable research practices** (QRPs).

Which article should you write? There are two possible articles you can write: (1) the article you planned to write when you designed your study or (2) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (2).

What Bem suggests here is known as **HARKing** (Hypothesizing

After the Results are Known)(Kerr, [1998](#)). This might seem innocuous, but is problematic because it allows the researcher to re-frame a post-hoc conclusion (which we should take with a grain of salt) as an a priori prediction (in which we would have stronger faith). In essence, it allows the researcher to rewrite their theory based on the facts, rather than using the theory to make predictions and then test them – akin to moving the goalpost so that it ends up wherever the ball goes. It thus becomes very difficult to disconfirm incorrect ideas, since the goalpost can always be moved to match the data. Bem continues:

Analyzing data Examine them from every angle. Analyze the sexes separately. Make up new composite indices. If a datum suggests a new hypothesis, try to find further evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something — anything — interesting. No, this is not immoral.

What Bem suggests here is known as *p-hacking*, which refers to trying many different analyses until one finds a significant result. Bem is correct that if one were to report every analysis done on the data then this approach would not be “immoral”. However, it is rare to see a paper discuss all of the analyses that were performed on a dataset; rather, papers often only present the analyses that *worked* – which usually means that they found a statistically significant result. There are many different ways that one might p-hack:

- Analyze data after every subject, and stop collecting data once $p < .05$

- Analyze many different variables, but only report those with $p < .05$
- Collect many different experimental conditions, but only report those with $p < .05$
- Exclude participants to get $p < .05$
- Transform the data to get $p < .05$

A well-known paper by Simmons, Nelson, and Simonsohn (2011) showed that the use of these kinds of p-hacking strategies could greatly increase the actual false positive rate, resulting in a high number of false positive results.

ESP or QRP?

In 2011, that same Daryl Bem published an article (Bem, 2011) that claimed to have found scientific evidence for extrasensory perception (ESP). The article states:

This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by “time-reversing” well-established psychological effects so that the individual’s responses are obtained before the putatively causal stimulus events occur. ...The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results.

As researchers began to examine Bem’s article, it became clear that he had engaged in all of the

QRPs that he had recommended in the chapter discussed above. As Tal Yarkoni pointed out in [a blog post that examined the article](#):

- Sample sizes varied across studies
- Different studies appear to have been lumped together or split apart
- The studies allow many different hypotheses, and it's not clear which were planned in advance
- Bem used one-tailed tests even when it's not clear that there was a directional prediction (so alpha is really 0.1)
- Most of the p-values are very close to 0.05
- It's not clear how many other studies were run but not reported

Doing reproducible research

In the years since the reproducibility crisis arose, there has been a robust movement to develop tools to help protect the reproducibility of scientific research.

Pre-registration

One of the ideas that has gained the greatest traction is *pre-registration*, in which one submits a detailed description of a study (including all data analyses) to a trusted repository (such as the [Open Science Framework](#) or [AsPredicted.org](#)). By

specifying one's plans in detail prior to analyzing the data, pre-registration provides greater faith that the analyses do not suffer from p-hacking or other questionable research practices.

The effects of pre-registration in clinical trials in medicine have been striking. In 2000, the National Heart, Lung, and Blood Institute (NHLBI) began requiring all clinical trials to be pre-registered using the system at [ClinicalTrials.gov](https://clinicaltrials.gov). This provides a natural experiment to observe the effects of study pre-registration. When Kaplan and Irvin (2015) examined clinical trial outcomes over time, they found that the number of positive outcomes in clinical trials was greatly reduced after 2000 compared to before. While there are many possible causes, it seems likely that prior to study registration researchers were able to change their methods or hypotheses in order to find a positive result, which became more difficult after registration was required.

Reproducible practices

The paper by Simmons, Nelson, and Simonsohn (2011) laid out a set of suggested practices for making research more reproducible, all of which should become standard for researchers:

- Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
- Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
- Authors must list all variables collected in a study.
- Authors must report all experimental conditions, including failed manipulations.
- If observations are eliminated, authors must also

report what the statistical results are if those observations are included.

- If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

Replication

One of the hallmarks of science is the idea of *replication* – that is, other researchers should be able to perform the same study and obtain the same result. Unfortunately, as we saw in the outcome of the Replication Project discussed earlier, many findings are not replicable. The best way to ensure replicability of one's research is to first replicate it on your own; for some studies this just won't be possible, but whenever it is possible one should make sure that one's finding holds up in a new sample. That new sample should be sufficiently powered to find the effect size of interest; in many cases, this will actually require a larger sample than the original.

It's important to keep a couple of things in mind with regard to replication. First, the fact that a replication attempt fails does not necessarily mean that the original finding was false; remember that with the standard level of 80% power, there is still a one in five chance that the result will be nonsignificant, even if there is a true effect. For this reason, we generally want to see multiple replications of any important finding before we decide whether or not to believe it. Unfortunately, many fields including psychology have failed to follow this advice in the past, leading to “textbook” findings that turn out to be likely false. With regard to Daryl Bem's studies of ESP, a large replication attempt involving 7 studies failed to replicate his findings (Galak et al. [2012](#)).

Second, remember that the p-value doesn't provide us with a measure of the likelihood of a finding to replicate. As we

discussed previously, the p-value is a statement about the likelihood of one's data under a specific null hypothesis; it doesn't tell us anything about the probability that the finding is actually true (as we learned in the chapter on Bayesian analysis). In order to know the likelihood of replication we need to know the probability that the finding is true, which we generally don't know.

Doing reproducible data analysis

So far we have focused on the ability to replicate other researchers' findings in new experiments, but another important aspect of reproducibility is to be able to reproduce someone's analyses on their own data, which we refer to a *computational reproducibility*. This requires that researchers share both their data and their analysis code, so that other researchers can both try to reproduce the result as well as potentially test different analysis methods on the same data. There is an increasing move in psychology towards open sharing of code and data; for example, the journal *Psychological Science* now provides “badges” to papers that share research materials, data, and code, as well as for pre-registration.

The ability to reproduce analyses is one reason that we strongly advocate for the use of scripted analyses (such as those using R) rather than using a “point-and-click” software package. It's also a reason that we advocate the use of free and open-source software (like R) as opposed to commercial software packages, which would require others to buy the software in order to reproduce any analyses.

There are many ways to share both code and data. A common way to share code is via web sites that support *version control* for software, such as [Github](https://github.com). Small datasets can also be shared via these same sites; larger datasets can be shared

through data sharing portals such as [Zenodo](#), or through specialized portals for specific types of data (such as [OpenNeuro](#) for neuroimaging data).

Conclusion: Doing better science

It is every scientist's responsibility to improve their research practices in order to increase the reproducibility of their research. It is essential to remember that the goal of research is not to find a significant result; rather, it is to ask and answer questions about nature in the most truthful way possible. Most of our hypotheses will be wrong, and we should be comfortable with that, so that when we find one that's right, we will be even more confident in its truth.

Learning objectives

- Describe the concept of P-hacking and its effects on scientific practice
- Describe the concept of positive predictive value and its relation to statistical power
- Describe the concept of pre-registration and how it can help protect against questionable research practices

Suggested Readings

- [Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions](#), by Richard Harris
- [Improving your statistical inferences](#) – an online course on how to do better statistical analysis, including many of the points raised in this chapter.

Appendix

More resources to be added. For no, there is no appendix.

References

References

<!--chapter:end:99-References.Rmd-->

Baker, Monya. 2017. "Reproducibility: Check Your Chemistry." *Nature* 548 (7668): 485–88. <https://doi.org/10.1038/548485a>.

Bem, Daryl J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *J Pers Soc Psychol* 100 (3): 407–25. <https://doi.org/10.1037/a0021524>.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statist. Sci.* 16 (3). The Institute of Mathematical Statistics: 199–231. <https://doi.org/10.1214/ss/1009213726>.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015." *Nature Human Behaviour* 2: 637–44.

Christensen, Garret S, and Edward Miguel. 2016. "Transparency, Reproducibility, and the Credibility of Economics Research." Working Paper 22989. Working Paper Series. National Bureau of Economic Research. <https://doi.org/10.3386/w22989>.

Copas, J. B. 1983. "Regression, Prediction and Shrinkage (with Discussion)." *Journal of the Royal Statistical Society, Series B: Methodological* 45: 311–54.

Darley, John M, Mark P Zanna, and Henry L Roediger. 2004. *The Compleat Academic: A Career Guide*. 2nd ed. Washington, DC: American Psychological Association. <http://www.loc.gov/catdir/toc/fy037/2003041830.html>.

Dehghan, Mahshid, Andrew Mente, Xiaohe Zhang, Sumathi Swaminathan, Wei Li, Viswanathan Mohan, Romaina Iqbal, et al. 2017. "Associations of Fats and Carbohydrate Intake with Cardiovascular Disease and Mortality in 18 Countries from Five Continents (Pure): A Prospective Cohort Study." *Lancet* 390 (10107): 2050–62. [https://doi.org/10.1016/S0140-6736\(17\)32252-3](https://doi.org/10.1016/S0140-6736(17)32252-3).

Efron, Bradley. 1998. "R. A. Fisher in the 21st Century (Invited Paper Presented at the 1996 R. A. Fisher Lecture)." *Statist. Sci.* 13 (2). The Institute of Mathematical Statistics: 95–122. <https://doi.org/10.1214/ss/1028905930>.

Errington, Timothy M, Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax, and Brian A Nosek. 2014. "An Open Investigation of the Reproducibility of Cancer Biology Research." *Elife* 3 (December). <https://doi.org/10.7554/eLife.04333>.

Fisher, R.A. 1925. *Statistical Methods for Research Workers*. Edinburgh Oliver & Boyd.

Fisher, Ronald Aylmer. 1956. *Statistical Methods and Scientific Inference*. New York: Hafner Pub. Co.

Galak, Jeff, Robyn A LeBoeuf, Leif D Nelson, and Joseph P Simmons. 2012. "Correcting the Past: Failures to Replicate Psi." *J Pers Soc Psychol* 103 (6): 933–48. <https://doi.org/10.1037/a0029709>.

Gardner, Christopher D, Alexandre Kiazand, Sofiya Alhassan, Soowon Kim, Randall S Stafford, Raymond R Balise, Helena C Kraemer, and Abby C King. 2007. "Comparison of the Atkins,

Zone, Ornish, and Learn Diets for Change in Weight and Related Risk Factors Among Overweight Premenopausal Women: The a to Z Weight Loss Study: A Randomized Trial." *JAMA* 297 (9): 969–77. <https://doi.org/10.1001/jama.297.9.969>.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Med* 2 (8): e124. <https://doi.org/10.1371/journal.pmed.0020124>.

Kaplan, Robert M, and Veronica L Irvin. 2015. "Likelihood of Null Effects of Large Nhlbi Clinical Trials Has Increased over Time." *PLoS One* 10 (8): e0132382. <https://doi.org/10.1371/journal.pone.0132382>.

Kerr, N. L. 1998. "HARKing: Hypothesizing After the Results Are Known." *Pers Soc Psychol Rev* 2 (3): 196–217. https://doi.org/10.1207/s15327957pspr0203_4.

Neyman, J. 1937. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 236 (767). The Royal Society: 333–80. <https://doi.org/10.1098/rsta.1937.0005>.

Neyman, J., and K. Pearson. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 231 (694-706). The Royal Society: 289–337. <https://doi.org/10.1098/rsta.1933.0009>.

Open Science Collaboration. 2015. "PSYCHOLOGY. Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716. <https://doi.org/10.1126/science.aac4716>.

Pesch, Beate, Benjamin Kendzia, Per Gustavsson, Karl-Heinz Jöckel, Georg Johnen, Hermann Pohlabein, Ann Olsson, et al. 2012. "Cigarette Smoking and Lung Cancer–Relative Risk

Estimates for the Major Histological Types from a Pooled Analysis of Case-Control Studies." *Int J Cancer* 131 (5): 1210–9. <https://doi.org/10.1002/ijc.27339>.

Schenker, Nathaniel, and Jane F. Gentleman. 2001. "On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals." *The American Statistician* 55 (3). [American Statistical Association, Taylor & Francis, Ltd.]: 182–86. <http://www.jstor.org/stable/2685796>.

Schoenfeld, Jonathan D, and John P A Ioannidis. 2013. "Is Everything We Eat Associated with Cancer? A Systematic Cookbook Review." *Am J Clin Nutr* 97 (1): 127–34. <https://doi.org/10.3945/ajcn.112.047142>.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychol Sci* 22 (11): 1359–66. <https://doi.org/10.1177/0956797611417632>.

Smaldino, Paul E, and Richard McElreath. 2016. "The Natural Selection of Bad Science." *R Soc Open Sci* 3 (9): 160384. <https://doi.org/10.1098/rsos.160384>.

Stigler, Stephen M. 2016. *The Seven Pillars of Statistical Wisdom*. Harvard University Press.

Sullivan, Gail M, and Richard Feinn. 2012. "Using Effect Size-or Why the P Value Is Not Enough." *J Grad Med Educ* 4 (3): 279–82. <https://doi.org/10.4300/JGME-D-12-00156.1>.

Teicholz, Nina. 2014. *The Big Fat Surprise*. Simon & Schuster.

Wakefield, A J. 1999. "MMR Vaccination and Autism." *Lancet* 354 (9182): 949–50. [https://doi.org/10.1016/S0140-6736\(05\)75696-8](https://doi.org/10.1016/S0140-6736(05)75696-8).

Wansink, Brian, David R Just, and Collin R Payne. 2012. "Can

Branding Improve School Lunches?" *Arch Pediatr Adolesc Med*
166 (10): 1–2. <https://doi.org/10.1001/archpediatrics.2012.999>.