11-13-2018

An Introduction to Psychological Statistics

Garett C. Foster
University of Missouri-St. Louis, fostergc@umsl.edu

David Lane
Rice University, lane@rice.edu

David Scott
Rice University

Mikki Hebl
Rice University

Rudy Guerra
Rice University

See next page for additional authors

Authors
Garett C. Foster, David Lane, David Scott, Mikki Hebl, Rudy Guerra, Dan Osherson, and Heidi Zimmer

An Introduction to Psychological Statistics

Department of Psychological Sciences
University of Missouri – St Louis

This work was created as part of the University of Missouri's Affordable and Open Access Educational Resources Initiative (https://www.umsystem.edu/ums/aa/oer).

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

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

Changes to the original works were made by Dr. Garett C. Foster in the Department of Psychological Sciences to tailor the text to fit the needs of the introductory statistics course for psychology majors at the University of Missouri – St. Louis. 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.

Prologue: A letter to my students

Dear Students:

I get it.

Please believe me when I say that I completely understand, from firsthand experience, that statistics is rough. I was forced to take an introductory statistics course as part of my education, and I went in to it with dread. To be honest, for that first semester, I hated statistics. I was fortunate enough to have a wonderful professor who was knowledgeable and passionate about the subject. Nevertheless, I didn't understand what was going on, why I was required to take the course, or why any of it mattered to my major or my life.

Now, almost ten years later, I am deeply in love with statistics. Once I understood the logic behind statistics (and I promise, it is there, even if you don't see it at first), everything became crystal clear. More importantly, it enabled me to use that same logic not on numerical data but in my everyday life.

We are constantly bombarded by information, and finding a way to filter that information in an objective way is crucial to surviving this onslaught with your sanity intact. This is what statistics, and logic we use in it, enables us to do. Through the lens of statistics, we learn to find the signal hidden in the noise when it is there and to know when an apparent trend or pattern is really just randomness.

I understand that this is a foreign language to most people, and it was for me as well. I also understand that it can quickly become esoteric, complicated, and overwhelming. I encourage you to persist. Eventually, a lightbulb will turn on, and your life will be illuminated in a way it never has before.

I say all this to communicate to you that I am on your side. I have been in your seat, and I have agonized over these same concepts. Everything in this text has been put together in a way to convey not just formulae for manipulating numbers but to make connections across different chapters, topics, and methods, to demonstrate that it is all useful and important.

So I say again: I get it. I am on your side, and together, we will learn to do some amazing things.

Garett C. Foster, Ph.D.

Table of Contents

Unit 1 – Fundamentals of Statistics

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 will conclude with a brief introduction to concepts in probability that underlie how scientists perform data analysis. The material in this unit will serve 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.

Chapter 1: Introduction

This chapter provides an overview of statistics as a field of study and presents terminology that will be used throughout the course.

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. For example, 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.
>
> 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 Chapter 6, 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, these examples show that statistics are *not only facts and figures*; they are something more than that. 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 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 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 (students who do not believe this are welcome to ask the professor what

matrix algebra is). 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 television 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, beyond preparing you for a career in psychology, is to help you learn statistical essentials. It 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 will likely be the course you use most in your day to day life, even if you do not ever run a formal analysis again.

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. In psychology, we are interested in people, so we might get a group of people together and measure their levels of stress (one variable), anxiety (a second variable), and their physical health (a third variable). Once we have data on these three variables, we can use statistics to understand if and how they are related. Before we do so, we need to understand the nature of our data: what they represent and where they came from.

Types of Variables

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.

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. 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.

1. What is the independent variable? (dietary supplement: none, blueberry, strawberry, and spinach)
2. 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.

1. What is the independent variable? (supplements: beta-carotene or placebo)
2. 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 a following car to realize that the car in front is stopping and to hit the brakes.

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

Levels of an Independent Variable

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. 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.

Qualitative and Quantitative Variables

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. The values of a qualitative variable do not imply a numerical ordering. Values of the variable "religion" differ qualitatively; no ordering of religions is implied. Qualitative variables are sometimes referred to as categorical variables. Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

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

Before we can conduct a statistical analysis, we need to measure our dependent variable. 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 stop watch. Stop watches 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. 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.

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

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).

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.

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 ordinally-measured variable is a meaningful statistic. However, there are extreme situations in which computing the mean of an ordinally-measured variable can be very misleading.

Collecting Data

We are usually interested in understanding a specific group of people. This group is known as the population of interest, or simply the population. The population is the collection of all people who have some characteristic in common; it can be as broad as "all people" if we have a very general research question about human psychology, or it can be extremely narrow, such as "all freshmen psychology majors at Midwestern public universities" if we have a specific group in mind.

Populations and samples

In statistics, we often rely on a sample --- that is, a small subset of a larger set of data --- to draw inferences about the larger set. The larger set is known as the population from which the sample is drawn.

> 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 competing 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

Not all sampling methods are perfect, and sometimes that's okay. For example, if we are beginning research into a completely unstudied area, we may sometimes take some shortcuts to quickly gather data and get a general idea of how things work before fully investing a lot of time and money into well-designed research projects with proper sampling. This is known as convenience sampling, named for its ease of use. In limited cases, such as the one just described, convenience sampling is okay because we intend to follow up with a representative sample. Unfortunately, sometimes convenience sampling is used due only to its convenience without the intent of improving on it in future work.

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: experimental, quasi-experimental, and non-experimental.

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. An experiment is defined by the use of random assignment to treatment conditions and manipulation of the independent variable. 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.

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.

Non-Experimental Designs

Finally, non-experimental research (sometimes called correlational research) involves observing things as they occur naturally and recording our observations as data. 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.

Here, it is not possible to manipulate conscientious, so the researcher must gather data from employees as they are in order to find a relation between her variables.

Although this technique 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.

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. 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 1. The table shows the average salaries for various occupations in the United States in 1999.

Salary	Occupation
$112,760	pediatricians
$106,130	dentists
$100,090	podiatrists
$76,140	physicists
$53,410	architects,
$49,720	school, clinical, and counseling psychologists
$47,910	flight attendants
$39,560	elementary school teachers
$38,710	police officers
$18,980	floral designers

Table 1. Average salaries for various occupations in 1999.

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 2. It shows the number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990. From this table we see that men outnumber women most in Jacksonville, NC, and women outnumber men most in Sarasota, FL. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner! (These data come from the Information Please Almanac.)

Cities with mostly men	Men per 100 Women	Cities with mostly women	Men per 100 Women
1. Jacksonville, NC	224	1. Sarasota, FL	66
2. Killeen-Temple, TX	123	2. Bradenton, FL	68
3. Fayetteville, NC	118	3. Altoona, PA	69
4. Brazoria, TX	117	4. Springfield, IL	70
5. Lawton, OK	116	5. Jacksonville, TN	70
6. State College, PA	113	6. Gadsden, AL	70
7. Clarksville-Hopkinsville, TN-KY	113	7. Wheeling, WV	70
8. Anchorage, Alaska	112	8. Charleston, WV	71
9. Salinas-Seaside-Monterey, CA	112	9. St. Joseph, MO	71
10. Bryan-College Station, TX	111	10. Lynchburg, VA	71

Table 2. Number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990. *NOTE: Unmarried includes never-married, widowed, and divorced persons, 15 years or older.*

These descriptive statistics may make us ponder why the numbers are so disparate in these cities. One potential explanation, for instance, as to why there are more women in Florida than men may involve the fact that elderly individuals tend to move down to the Sarasota region and that women tend to outlive men. Thus, more women might live in Sarasota than men. However, in the absence of proper data, this is only speculation.

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
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

Table 3. 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 3 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 is $23.87. 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? 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

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:

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

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:

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

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:

$$\sum XY = 28$$

Exercises – Ch. 1

1. In your own words, describe why we study statistics.
2. For each of the following, determine if the variable is continuous or discrete:
    a. Time taken to read a book chapter
    b. Favorite food
    c. Cognitive ability
    d. Temperature
    e. Letter grade received in a class
3. For each of the following, determine the level of measurement:
    a. T-shirt size
    b. Time taken to run 100 meter race
    c. First, second, and third place in 100 meter race
    d. Birthplace
    e. 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
2	8
3	8
7	4
5	1
9	4

   a. ΣX
   b. ΣY^2
   c. ΣXY
   d. $(\Sigma Y)^2$

10. What are the most common measures of central tendency and spread?

Answers to Odd-Numbered Exercises – Ch. 1

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:
   a. Ordinal
   b. Ratio
   c. Ordinal
   d. Nominal
   e. 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:
   a. 26
   b. 161
   c. 109
   d. 625

Chapter 2: Describing Data using Distributions and Graphs

Before we can understand our analyses, we must first understand our data. The first step in doing this is using tables, charts, graphs, plots, and other visual tools to see what our data look like.

Graphing Qualitative Variables

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. 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. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

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.

Pie Charts

The pie chart in Figure 1 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

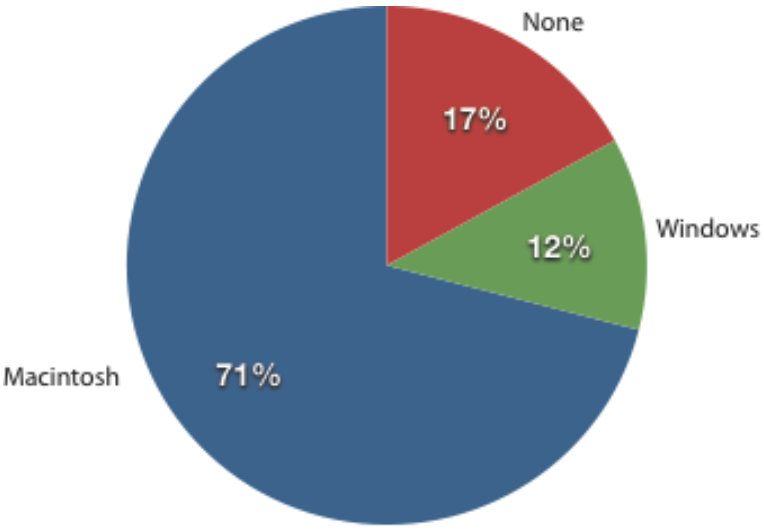


Figure 1. Pie chart of iMac purchases illustrating frequencies of previous computer ownership.

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, 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."
Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar charts

Bar charts can also be used to represent frequencies of different categories. 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.

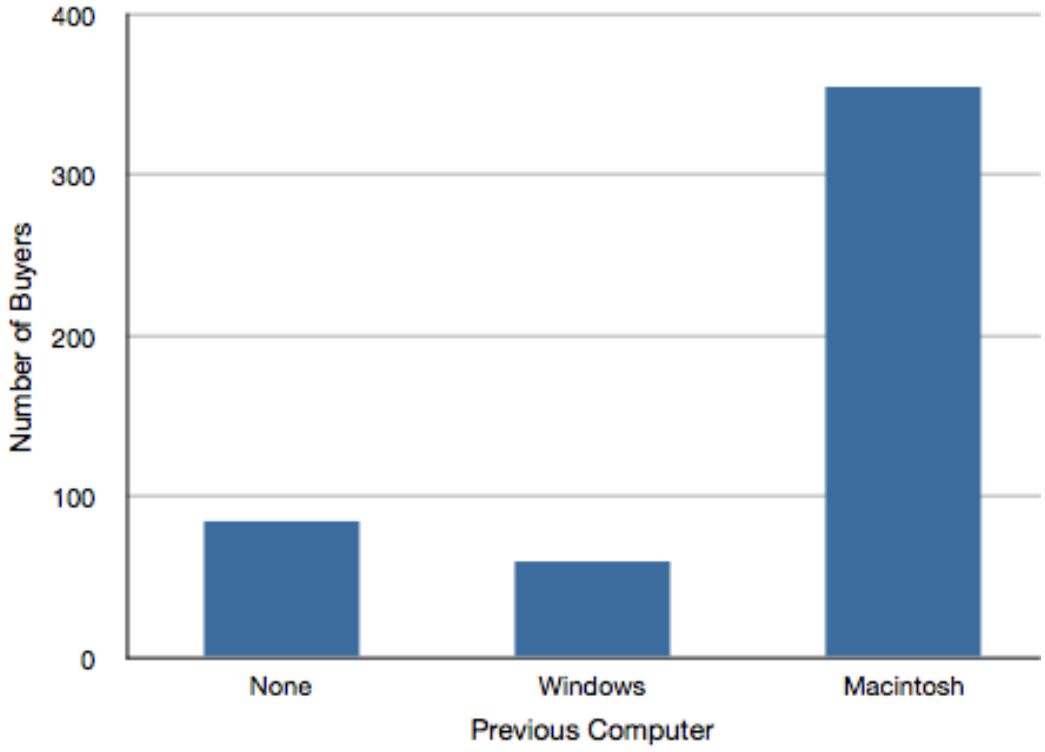


Figure 2. Bar chart of iMac purchases as a function of previous computer ownership.

Comparing Distributions

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 web site 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.

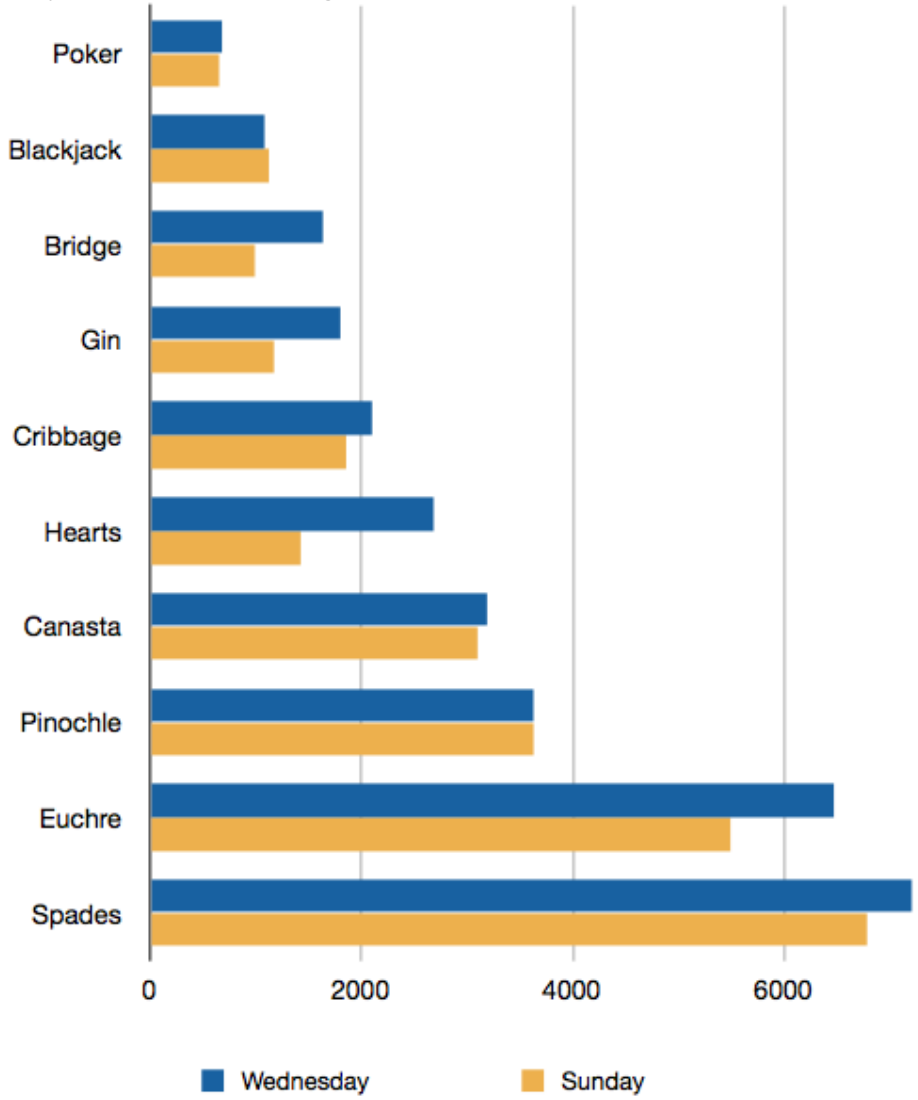


Figure 3. A bar chart of the number of people playing different card games on Sunday and Wednesday.

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

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts such as the one shown in Figure 4 are usually not as effective as their two-dimensional counterparts.

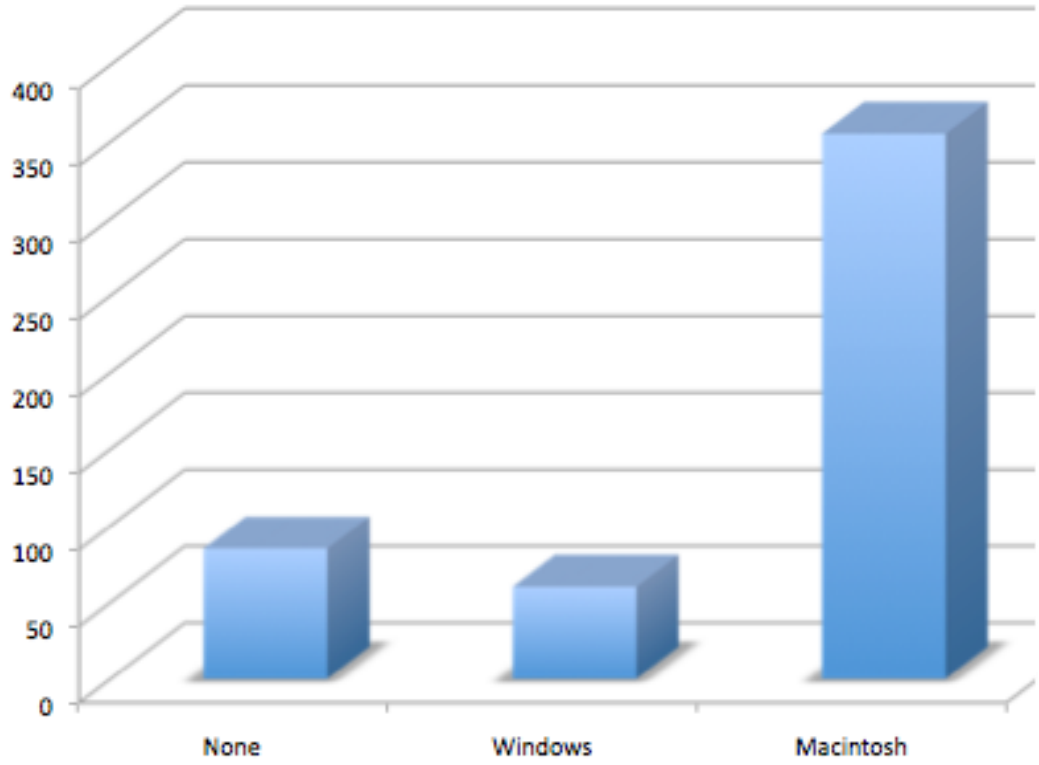


Figure 4. A three-dimensional version of Figure 2.

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 5 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 5 is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 5 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 5 instead of Figure 2! 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. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

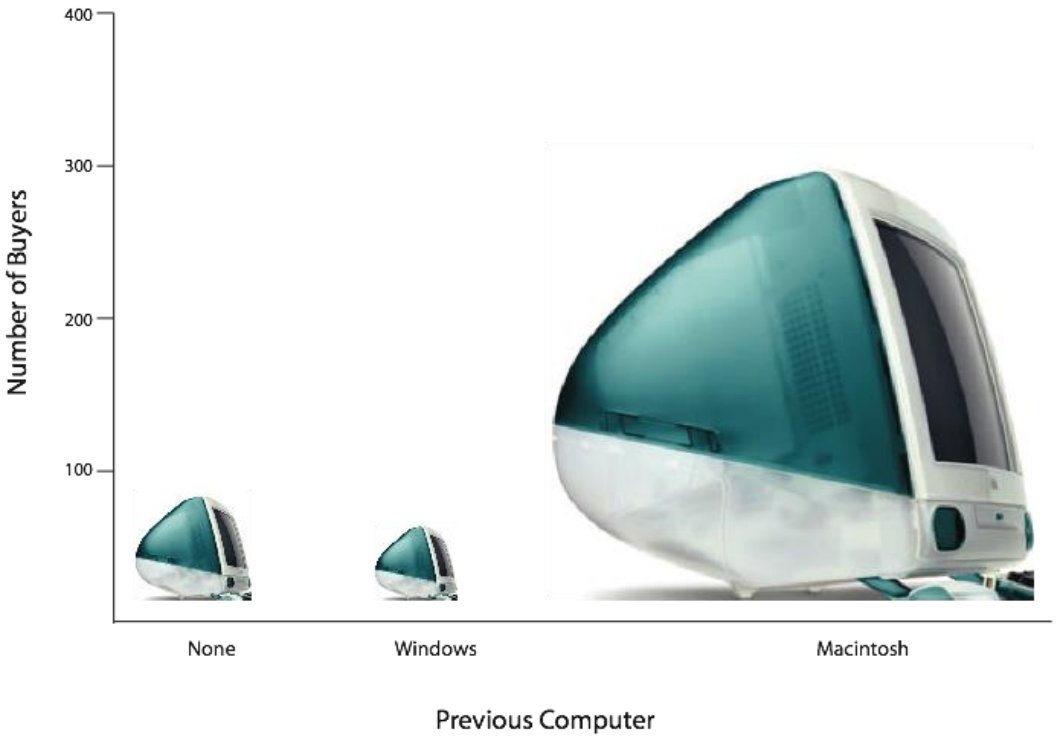


Figure 5. A redrawing of Figure 2 with a lie factor greater than 8.

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 6 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%.

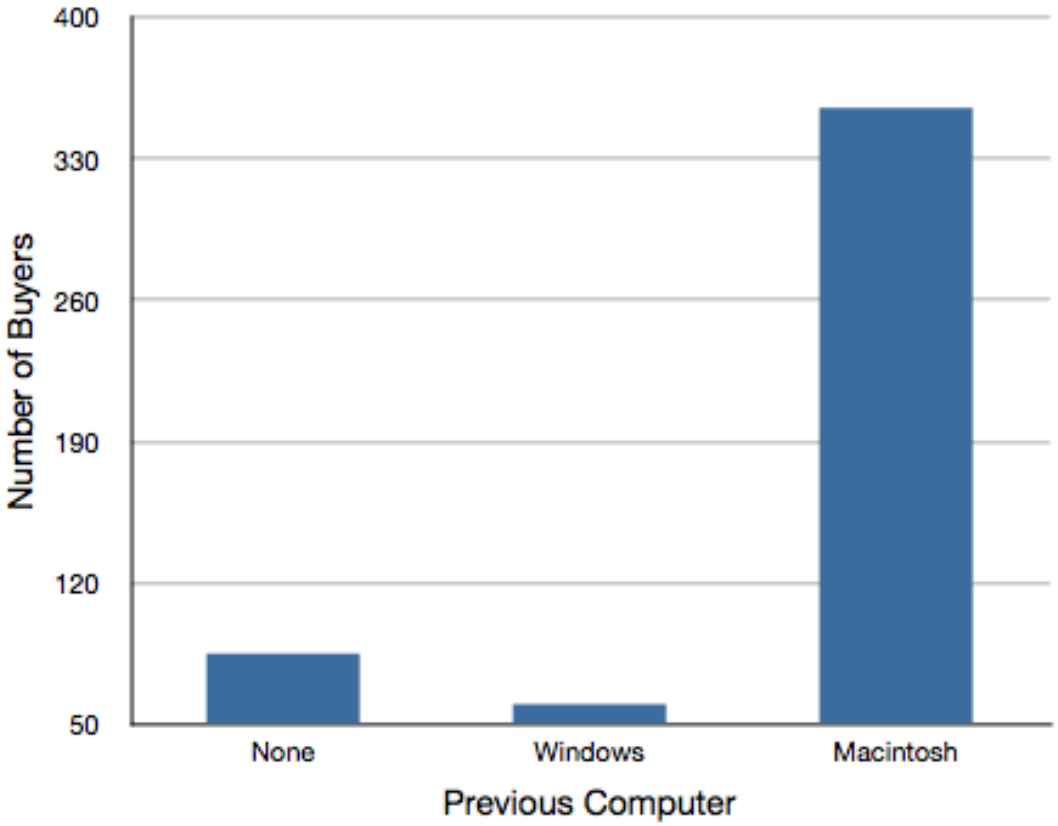


Figure 6. A redrawing of Figure 2 with a baseline of 50.

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative 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 7 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 7 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

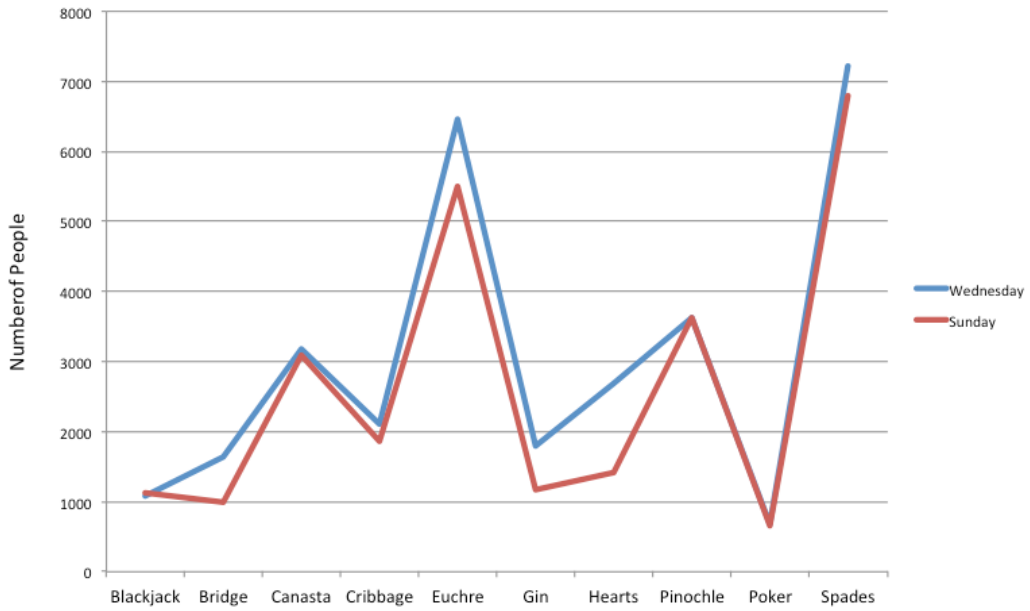


Figure 7. A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.

Summary

Pie charts and bar charts can both 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. The upcoming sections cover the following types of graphs: (1) stem and leaf displays, (2) histograms, (3) frequency polygons, (4) box plots, (5) bar charts, (6) line graphs, (7) dot plots, and (8) 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.

Stem and Leaf Displays

A stem and leaf display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, we will start with an example. Consider Table 2 that shows the number of touchdown passes (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 2. Number of touchdown passes.

A stem and leaf display of the data is shown in Figure 7. The left portion of Figure 1 contains the stems. They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10's digits. A stem of 3, for example, can be used to represent the 10's digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the 1's digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

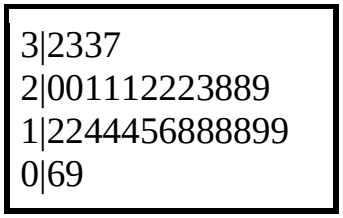


Figure 7. Stem and leaf display of the number of touchdown passes.

To make this clear, let us examine Figure 1 more closely. In the top row, the four leaves to the right of stem 3 are 2, 3, 3, and 7. Combined with the stem, these leaves represent the numbers 32, 33, 33, and 37, which are the numbers of TD

passes for the first four teams in Table 1. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Table 1, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure 1 than in Table 1. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TD's, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. Figure 2 shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table 2. The second row is reserved for the numbers from 30 to 34 and holds the 32, 33, and 33 TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

```
3|7
3|233
2|889
2|001112223
1|56888899
1|22444
0|69
```

Figure 8. Stem and leaf display with the stems split in two.

Figure 8 is more revealing than Figure 7 because the latter figure lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common

column of stems. The result is a “back-to-back stem and leaf display.” Figure 9 shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with 11, 12, and 13 TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69

Figure 9. Back-to-back stem and leaf display. The left side shows the 1998 TD data and the right side shows the 2000 TD data.

Figure 9 helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled. Let us look at another example.

Table 3 shows data from the case study Weapons and Aggression. Each value is the mean difference over a series of trials between the times it took an experimental subject to name aggressive words (like “punch”) under two conditions. In one condition, the words were preceded by a non-weapon word such as “bug.” In the second condition, the same words were preceded by a weapon word such as “gun” or “knife.” The issue addressed by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.

43.2, 42.9, 35.6, 25.6, 25.4, 23.6, 20.5, 19.9, 14.4, 12.7, 11.3, 10.2, 10.0, 9.1, 7.5, 5.4, 4.7, 3.8, 2.1, 1.2, -0.2, -6.3, -6.7, -8.8, -10.4, -10.5, -14.9, -14.9, -15.0, -18.5, -27.4

Table 3. The effects of priming (thousandths of a second).

You see that the numbers range from 43.2 to -27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value -27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in Figure 10. Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf) the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number –27. The second-to-last row represents the numbers -10, -10, -15, etc. Once again, we have rounded the original values from Table 3.

```
 4|33
 3|6
 2|00456
 1|00134
 0|1245589
-0|0679
-1|005559
-2|7
```

Figure 10. Stem and leaf display with negative numbers and rounding.

Observe that the figure contains a row headed by “0” and another headed by “-0.” The stem of 0 is for numbers between 0 and 9, whereas the stem of -0 is for numbers between 0 and -9. For example, the fifth row of the table holds the numbers 1, 2, 4, 5, 5, 8, 9 and the sixth row holds 0, -6, -7, and -9. Values that are exactly 0 before rounding should be split as evenly as possible between the “0” and “-0” rows. In Table 3, none of the values are 0 before rounding. The “0” that appears in the “-0” row comes from the original value of -0.2 in the table.

Although stem and leaf displays are unwieldy for large data sets, they are often useful for data sets with up to 200 observations. Figure 11 portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between 100,000 and 500,000 residents.

```
4|899
4|6
4|4455
4|333
4|01
3|99
3|677777
3|55
3|223
3|111
2|8899
2|666667
2|444455
2|22333
2|000000
1|8888888888899999999999
1|666666777777
1|444444444445555555555
1|22222222222222222222333333333
1|00000000000000011111111111111111111111111111
```

Figure 11. Stem and leaf display of populations of 185 US cities with populations between 100,000 and 500,000 in 1988.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest 10,000. For example the largest number (493,559) was rounded to 490,000 and then plotted with a stem of 4 and a leaf of 9. The fourth highest number (463,201) was rounded to 460,000 and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of 100,000 and the leaves represent units of 10,000. Notice that each stem value is split into five parts: 0-1, 2-3, 4-5, 6-7, and 8-9.

Whether your data can be suitably represented by a stem and leaf display depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in Figure 11 fit into the 10,000 and 100,000 places (for leaves and stems, respectively). Deciding what

kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

Histograms

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items each graded as "correct" or "incorrect." The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, 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.

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.

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.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure 12.

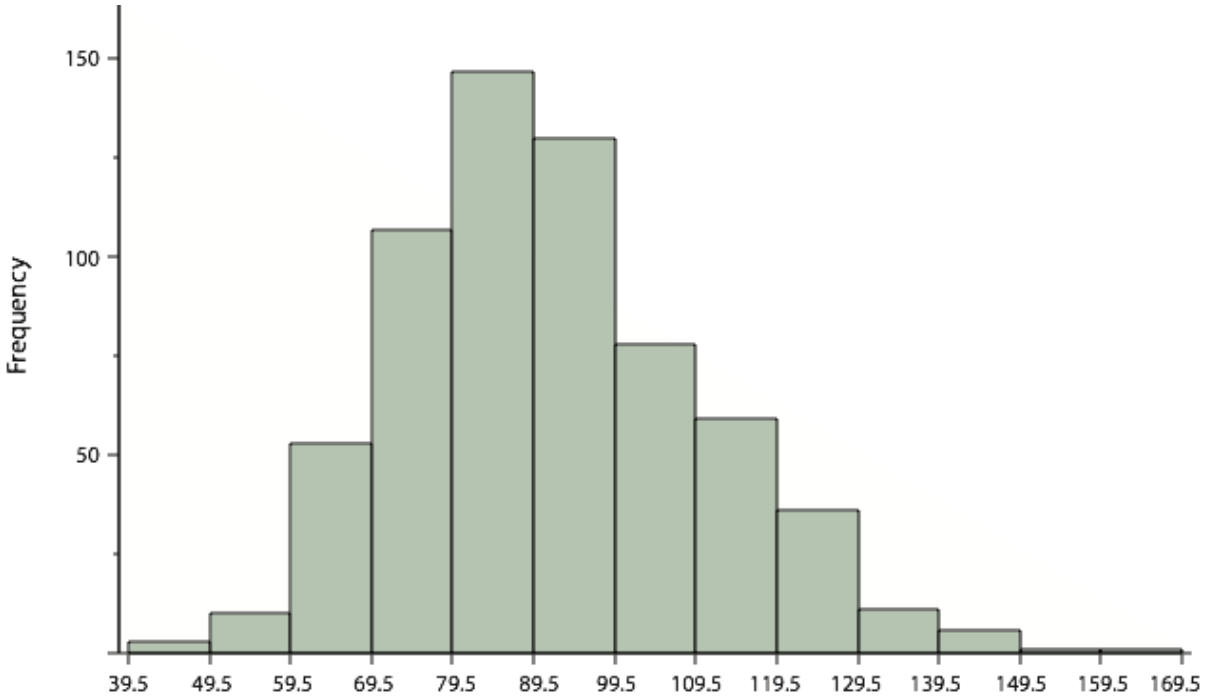


Figure 12. Histogram of scores on a psychology test.

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 in Chapter 3.)

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. Note also that some computer programs label the middle of each interval rather than the end points.

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 Count
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 13. 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). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is skewed.

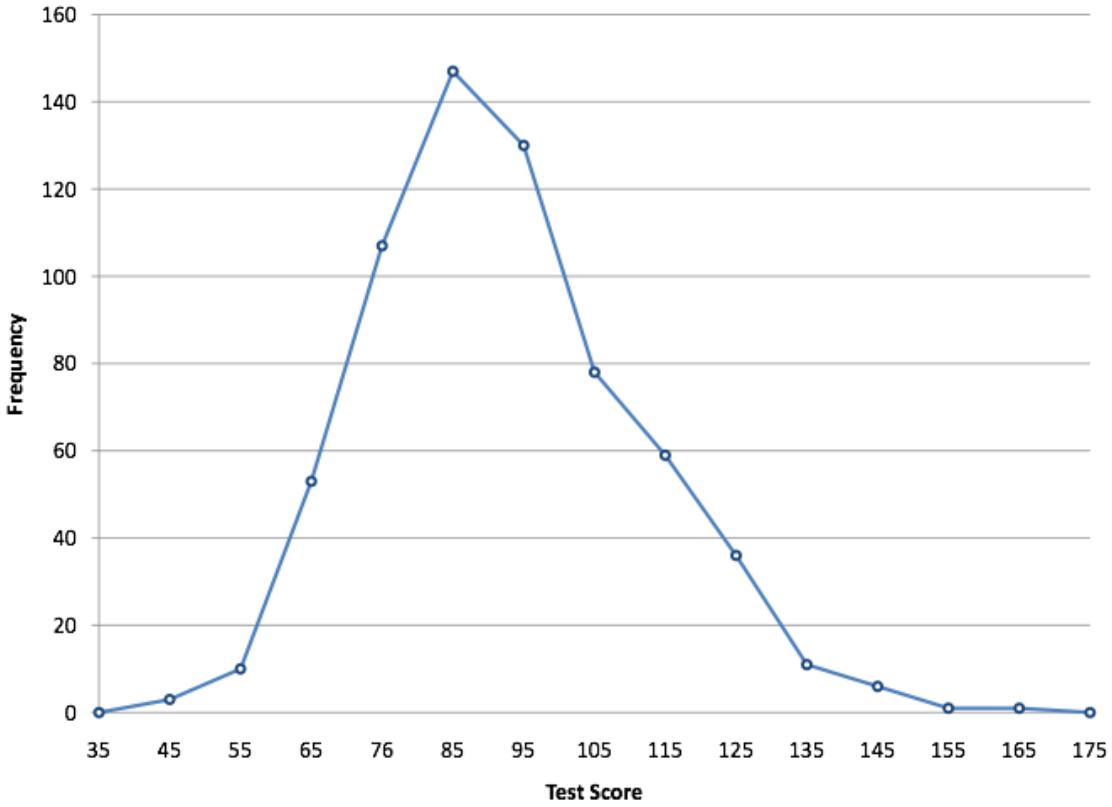


Figure 13. Frequency polygon for the psychology test scores.

A cumulative frequency polygon for the same test scores is shown in Figure 14. 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.

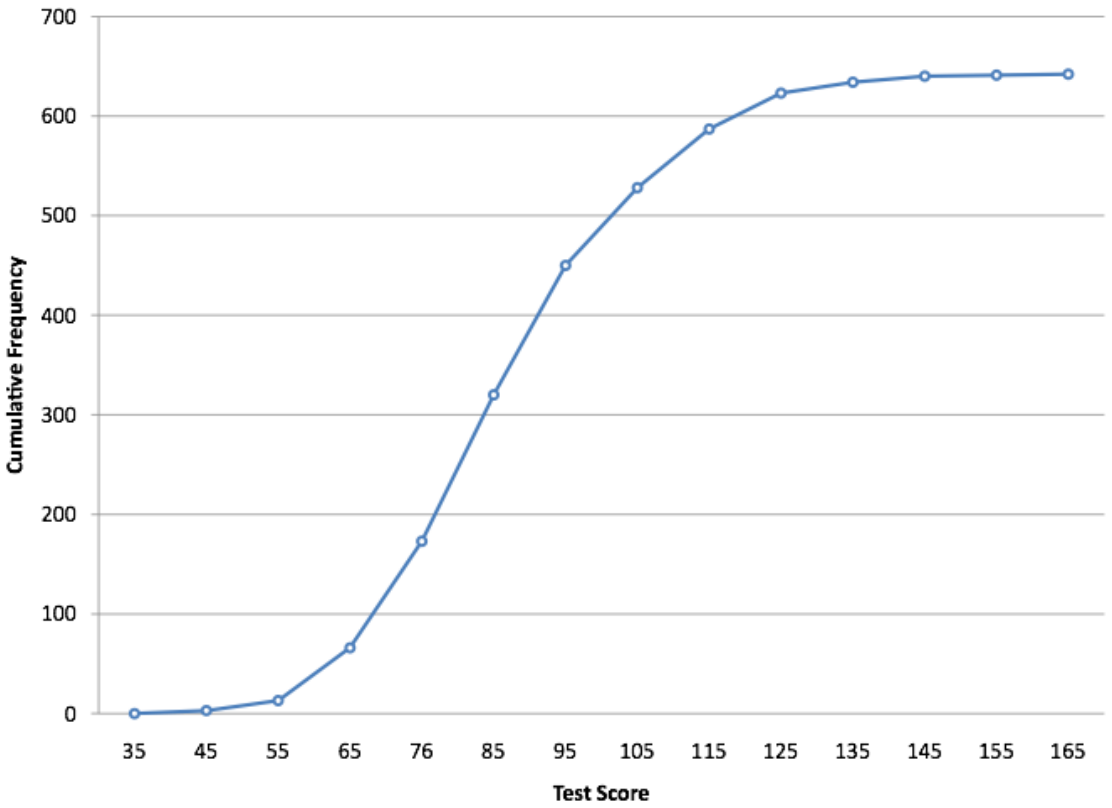


Figure 14. Cumulative frequency polygon for the psychology test scores.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 3 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.

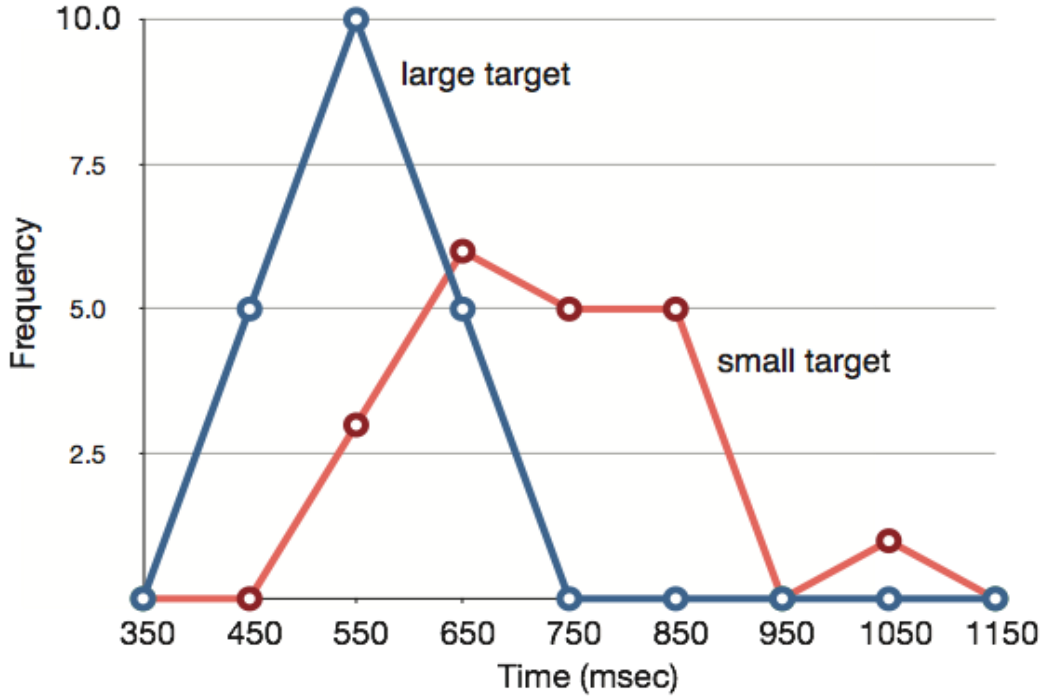


Figure 15. Overlaid frequency polygons.

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 16 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

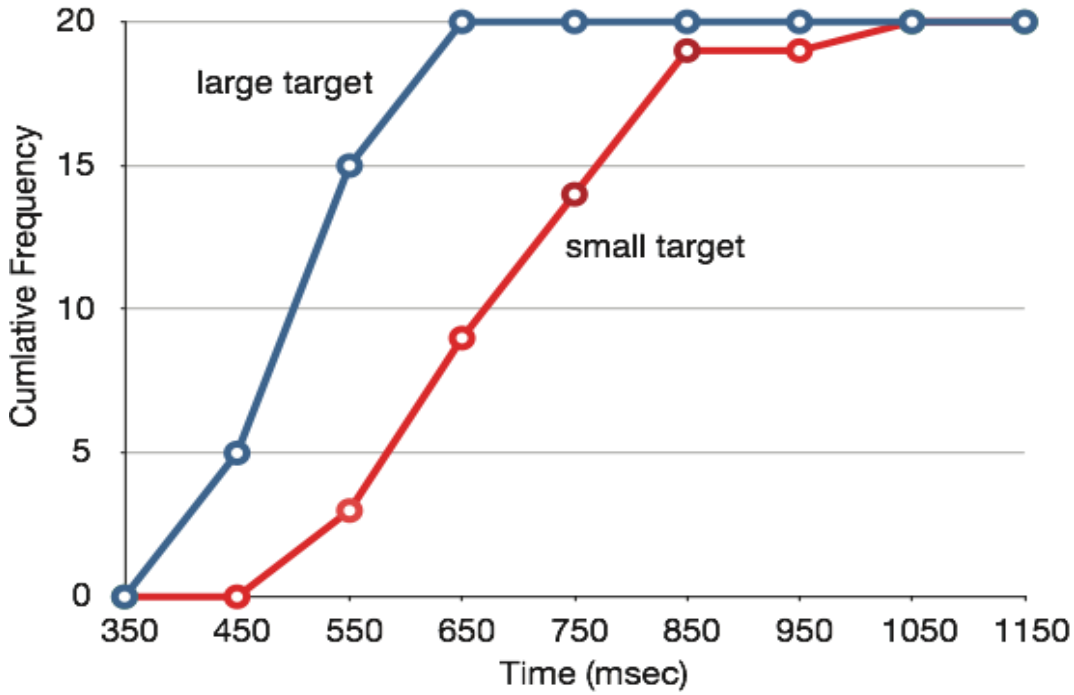


Figure 16. Overlaid cumulative frequency polygons.

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 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 17 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, 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.

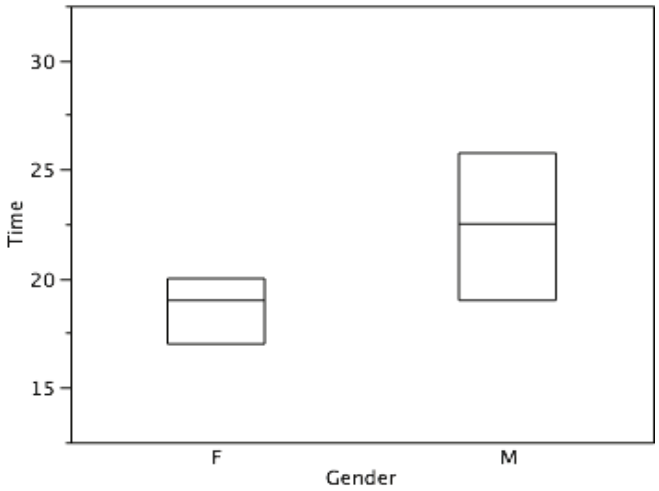


Figure 17. The first step in creating box plots.

Before proceeding, the terminology in Table 7 is helpful.

Name	Formula	Value
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 Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

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 18.

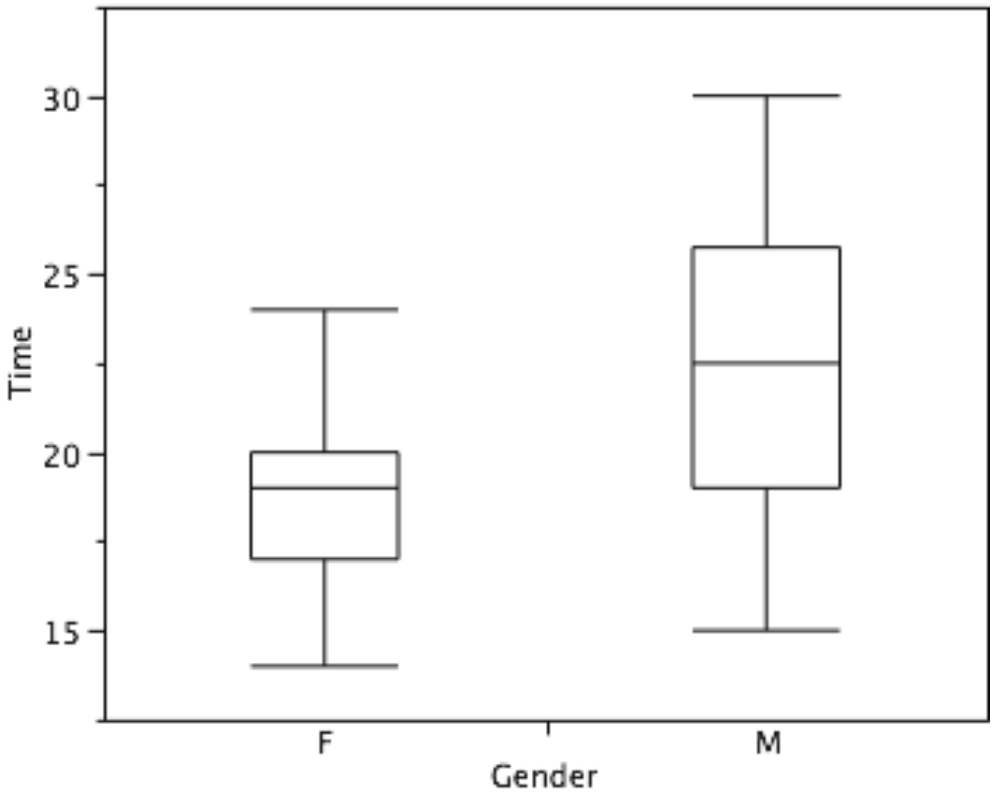


Figure 18. The box plots with the whiskers drawn.

Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small "o's" and far out 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 19.

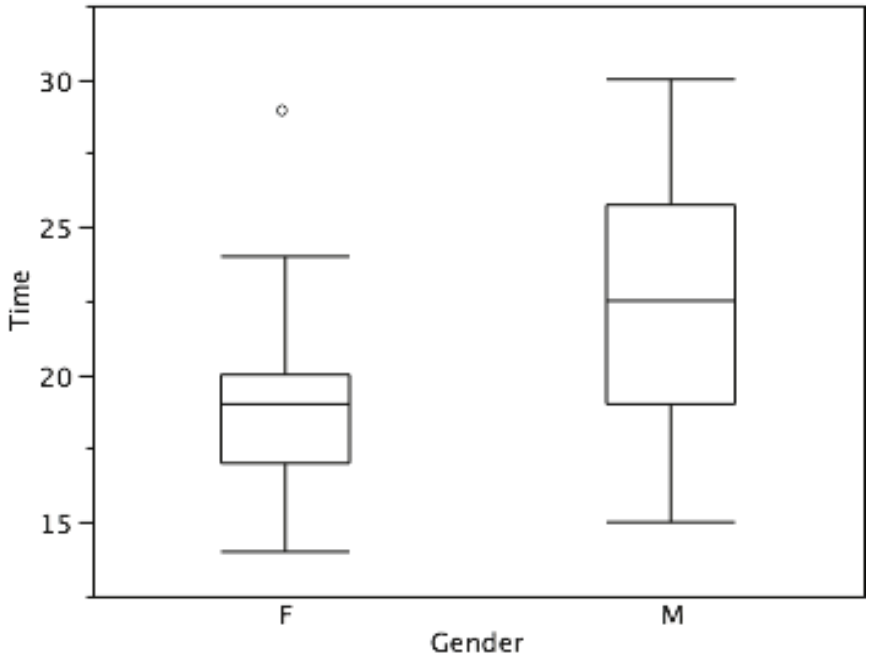


Figure 19. The box plots with the outside value shown.

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. Figure 20 shows the result of adding means to our box plots.

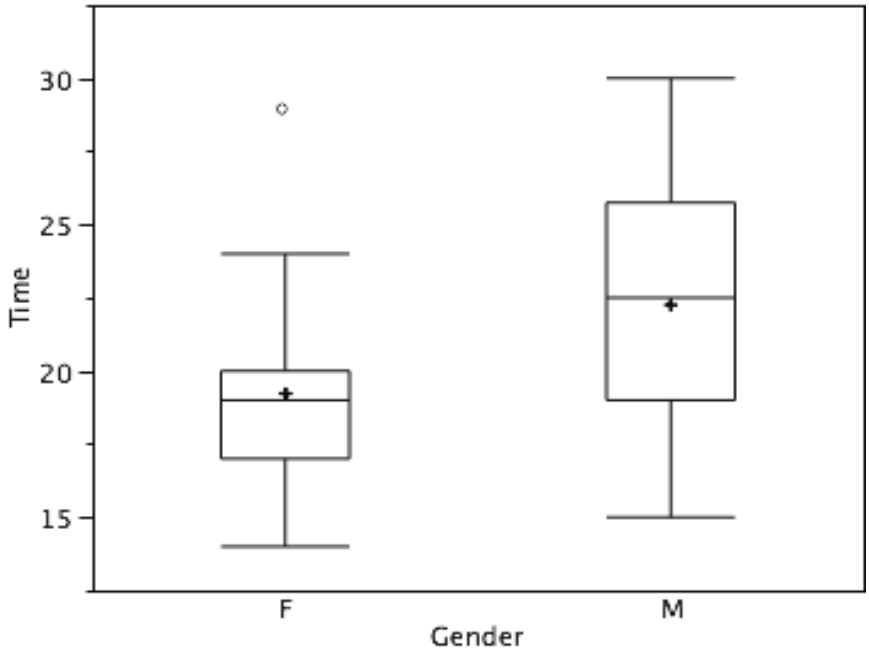


Figure 20. The completed box plots.

Figure 20 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. Figure 21 shows the box plot for the women's data with detailed labels.

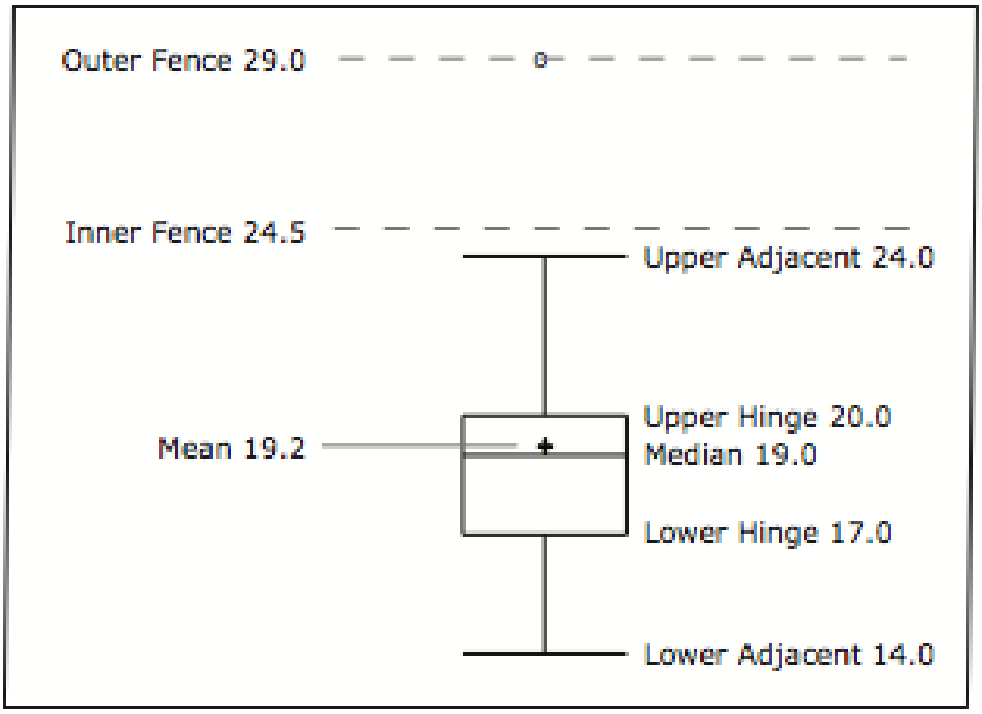


Figure 21. The box plots for the women's data with detailed labels.

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. 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 display.

Bar Charts

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, the bar chart shown in Figure 22 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

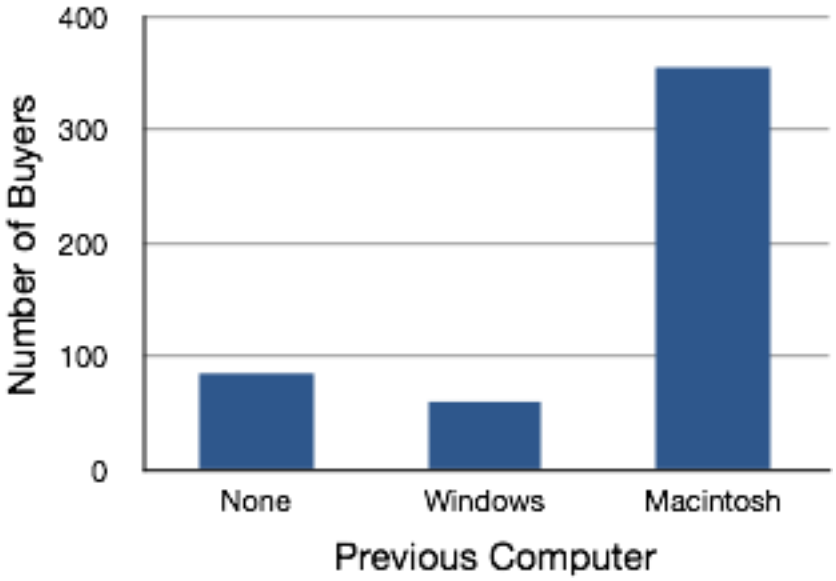


Figure 22. iMac buyers as a function of previous computer ownership.

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 23 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*.

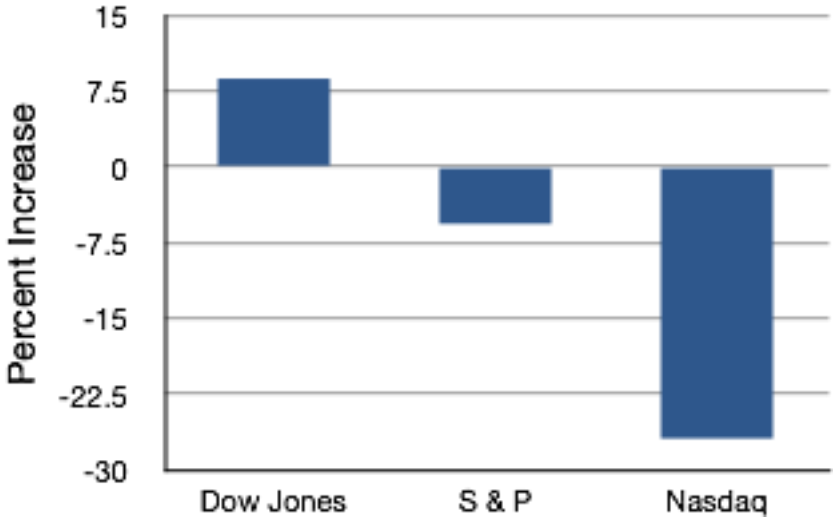


Figure 23. Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

Bar charts are particularly effective for showing change over time. Figure 24, 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.

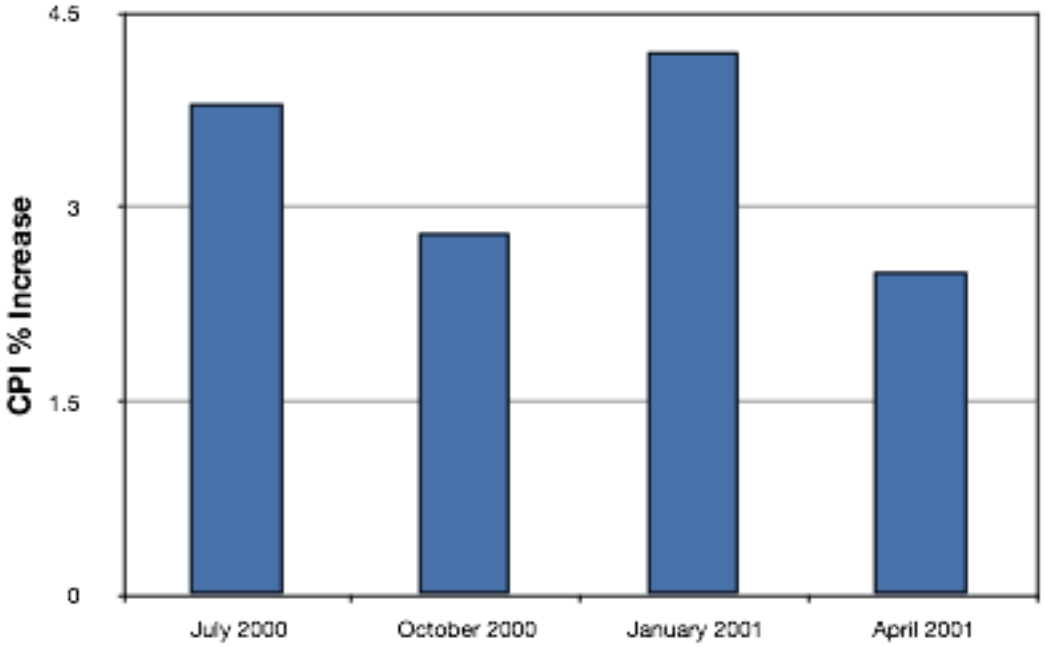


Figure 24. Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Bar charts are often used to compare the means of different experimental conditions. Figure 4 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.

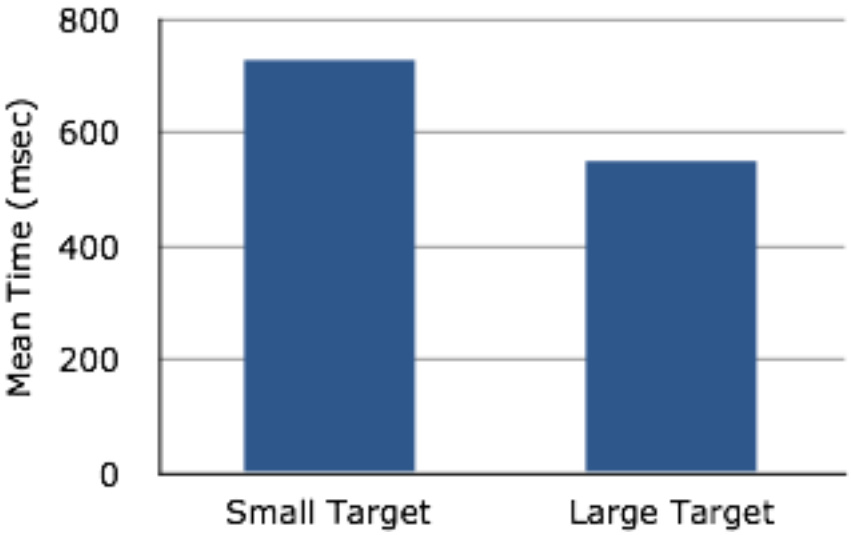


Figure 25. Bar chart showing the means for the two conditions.

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 26. You can see that Figure 26 reveals more about the distribution of movement times than does Figure 25.

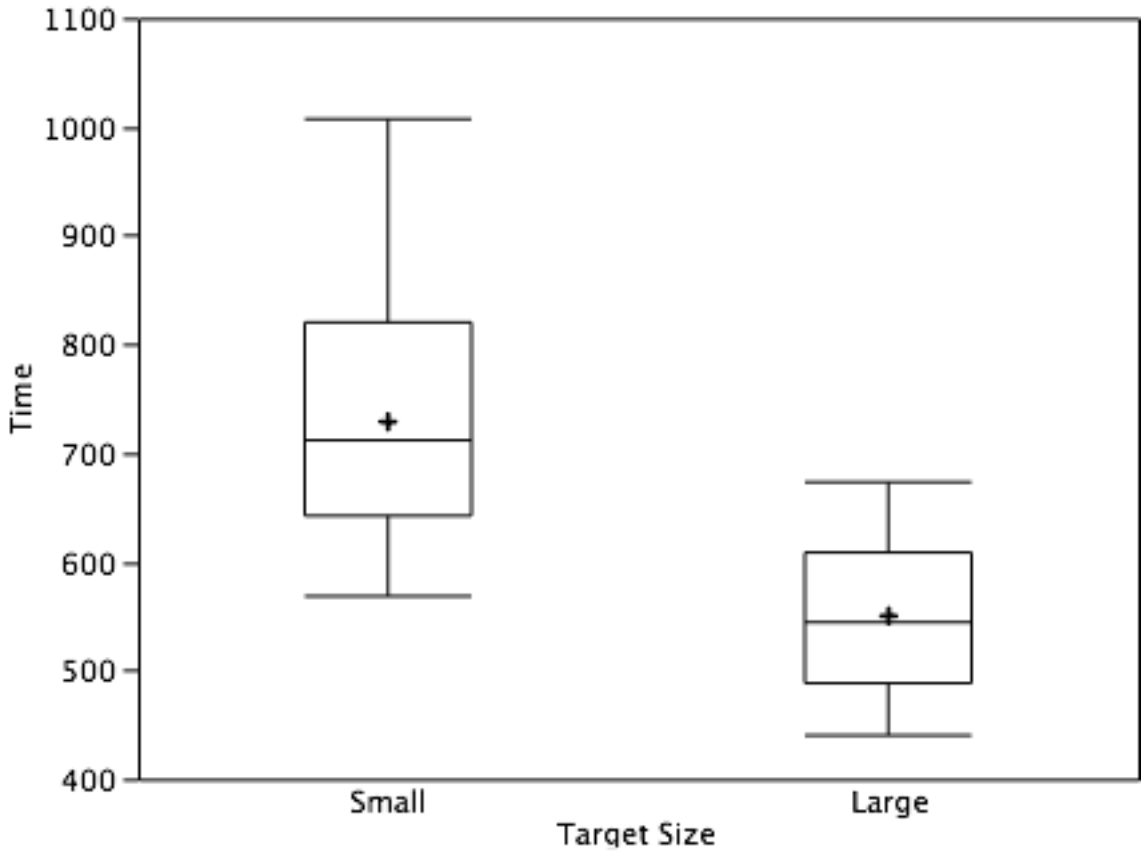


Figure 26. Box plots of times to move the cursor to the small and large targets.

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

Line Graphs

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 27 was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

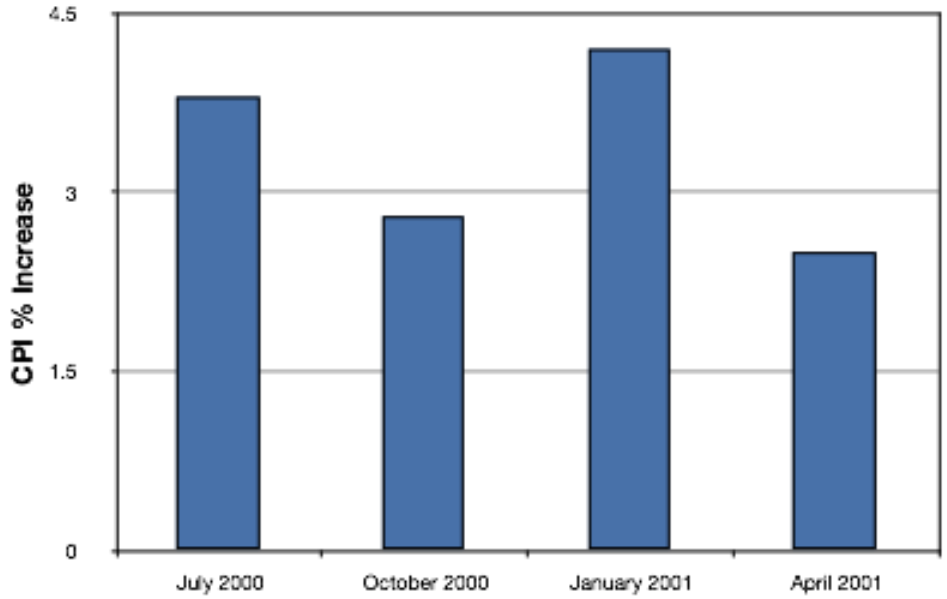


Figure 27. A bar chart of the percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

A line graph of these same data is shown in Figure 28. Although the figures are similar, the line graph emphasizes the change from period to period.

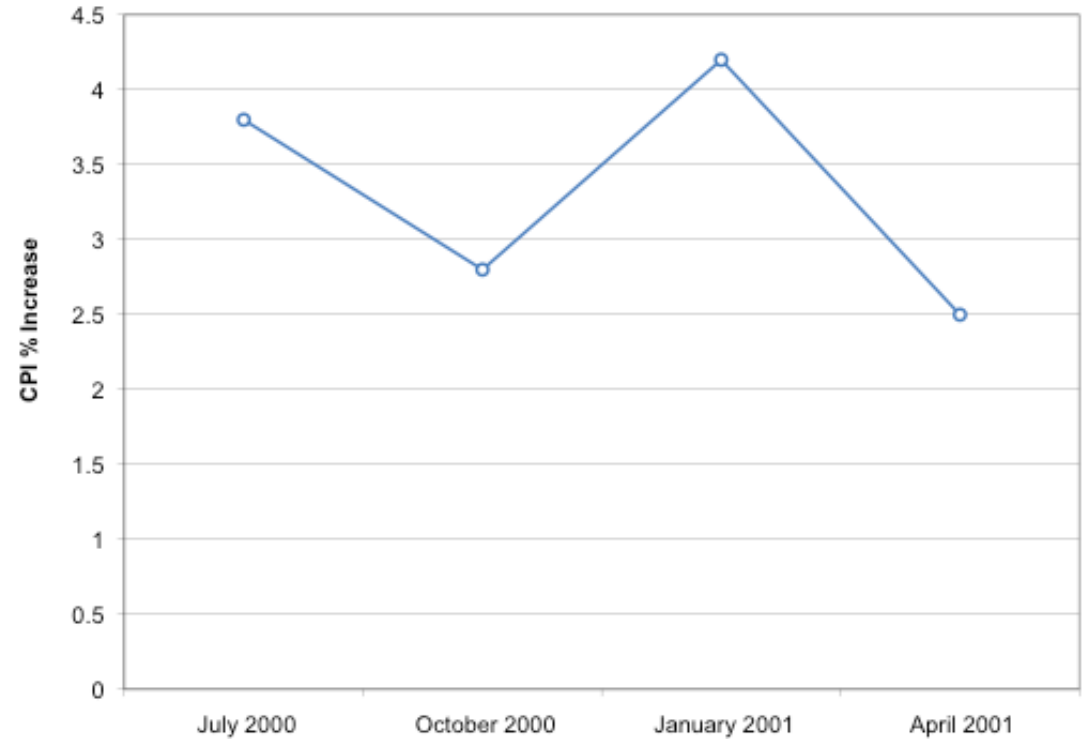


Figure 28. A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.

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 29, 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.

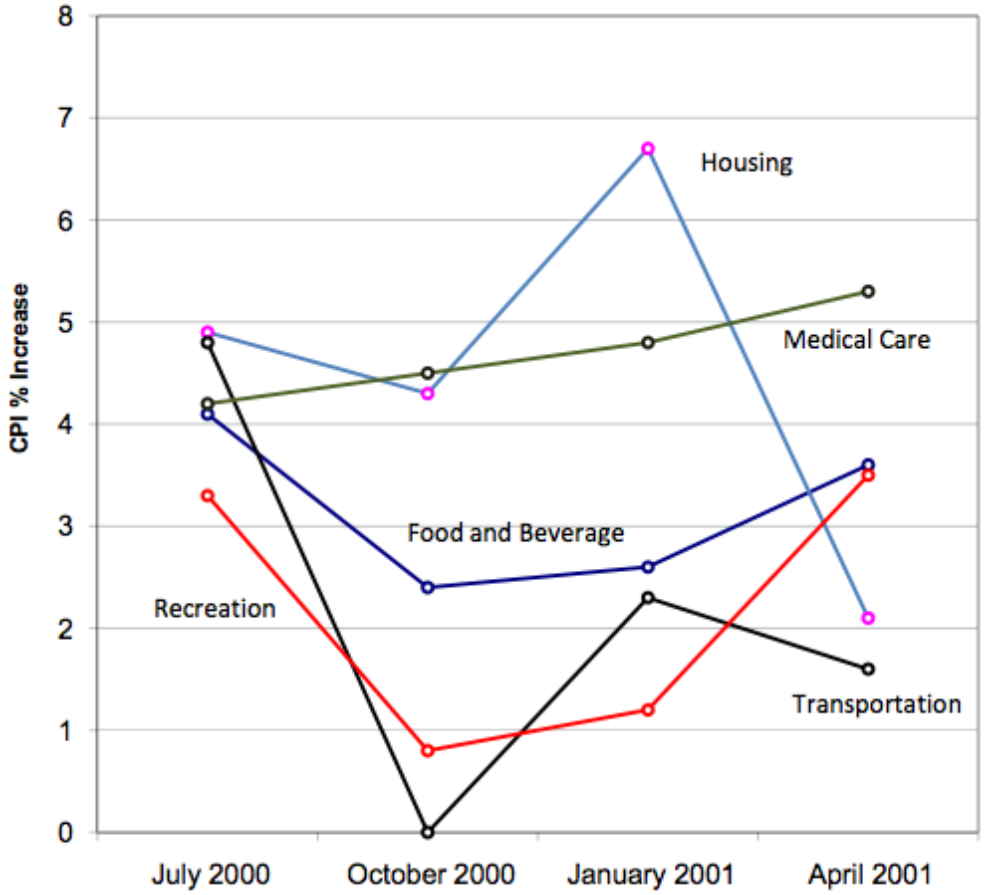


Figure 29. A line graph of the percent change in five components of the CPI over time.

Let us stress that it is misleading to use a line graph when the X-axis contains merely qualitative variables. Figure 30 inappropriately shows a line graph of the card game data from Yahoo, discussed in the section on qualitative variables. The defect in Figure 30 is that it gives the false impression that the games are naturally ordered in a numerical way.

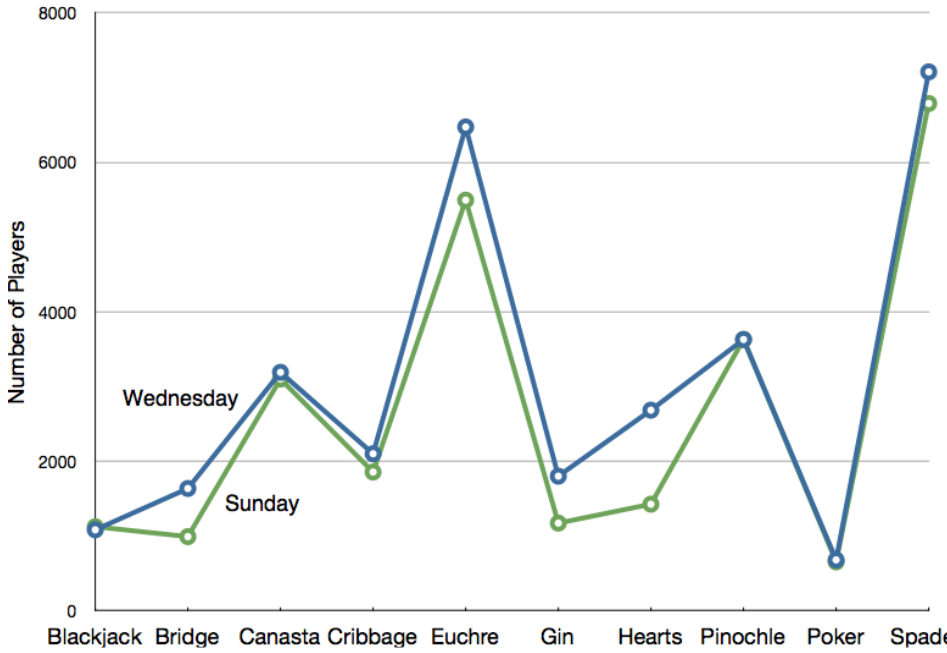


Figure 30. A line graph, inappropriately used, depicting the number of people playing different card games on Wednesday and Sunday.

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 31. 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.

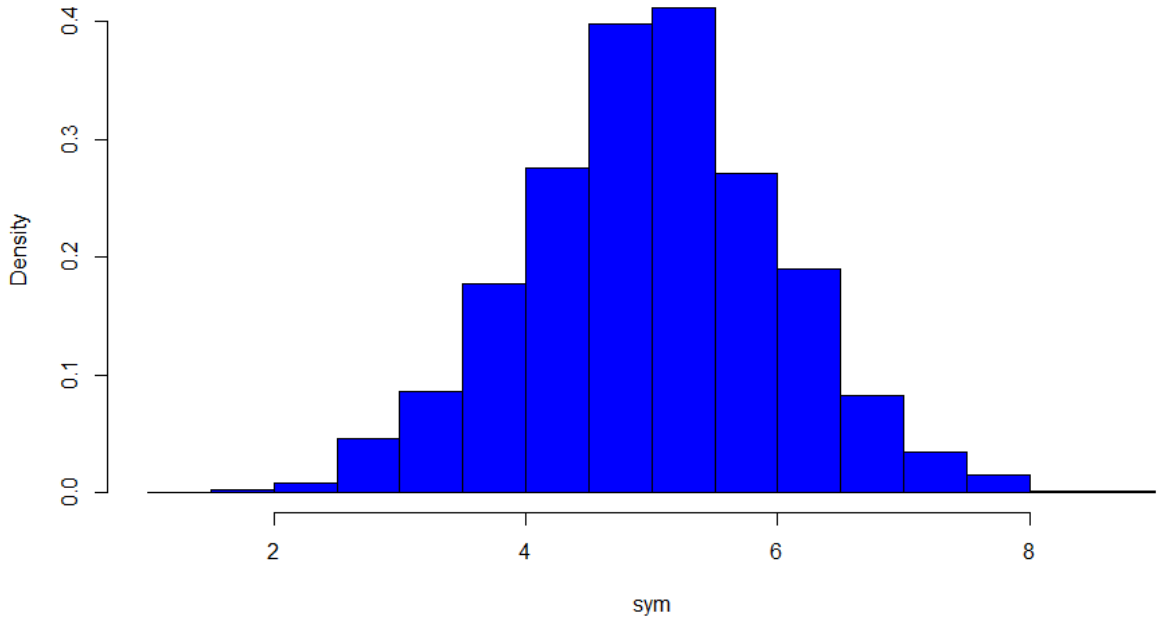


Figure 31. A symmetrical distribution

Symmetrical distributions can also have multiple peaks. Figure 32 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.

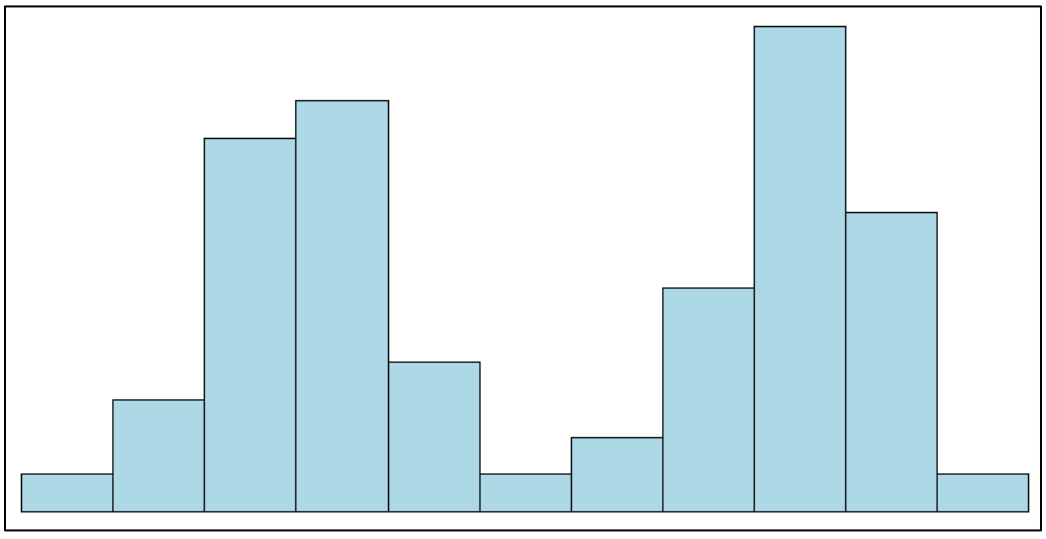


Figure 32. A bimodal distribution

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 33 and 34 show positive (right) and negative (left) skew, respectively.

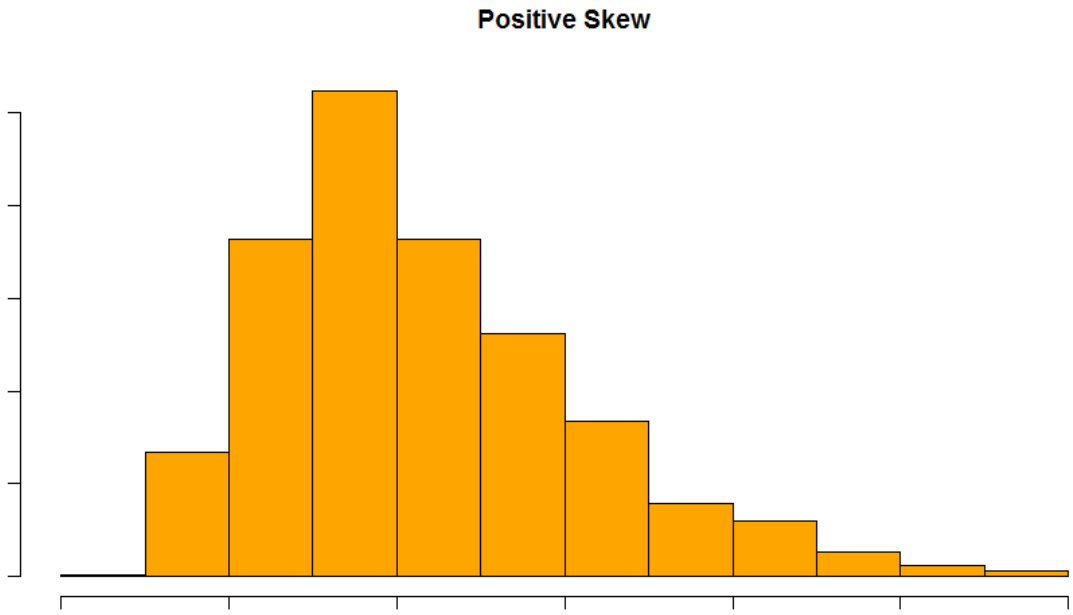


Figure 33. A positively skewed distribution

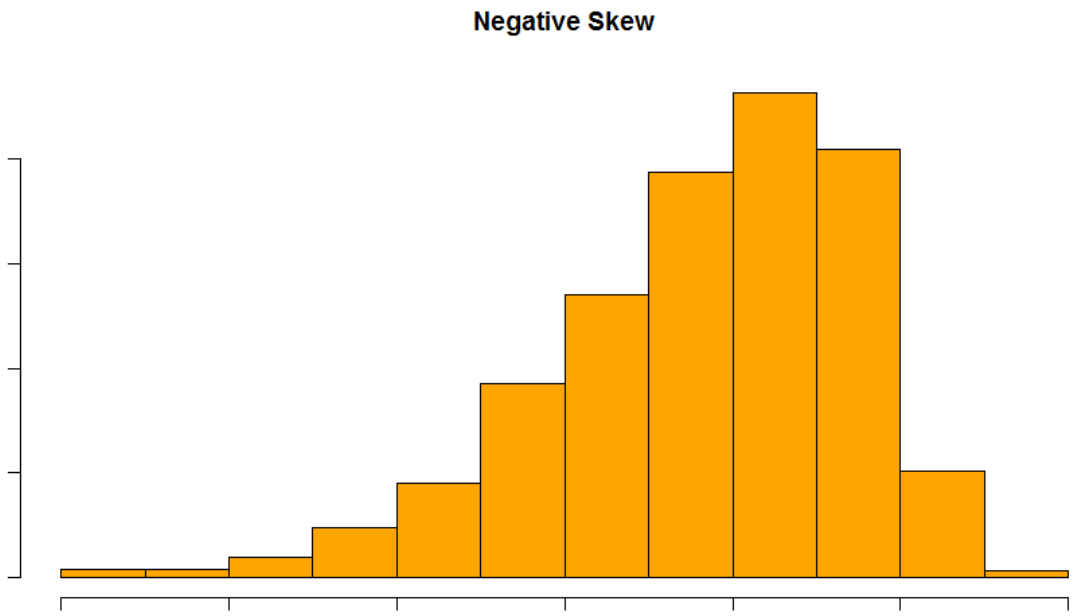


Figure 34. A negatively skewed distribution

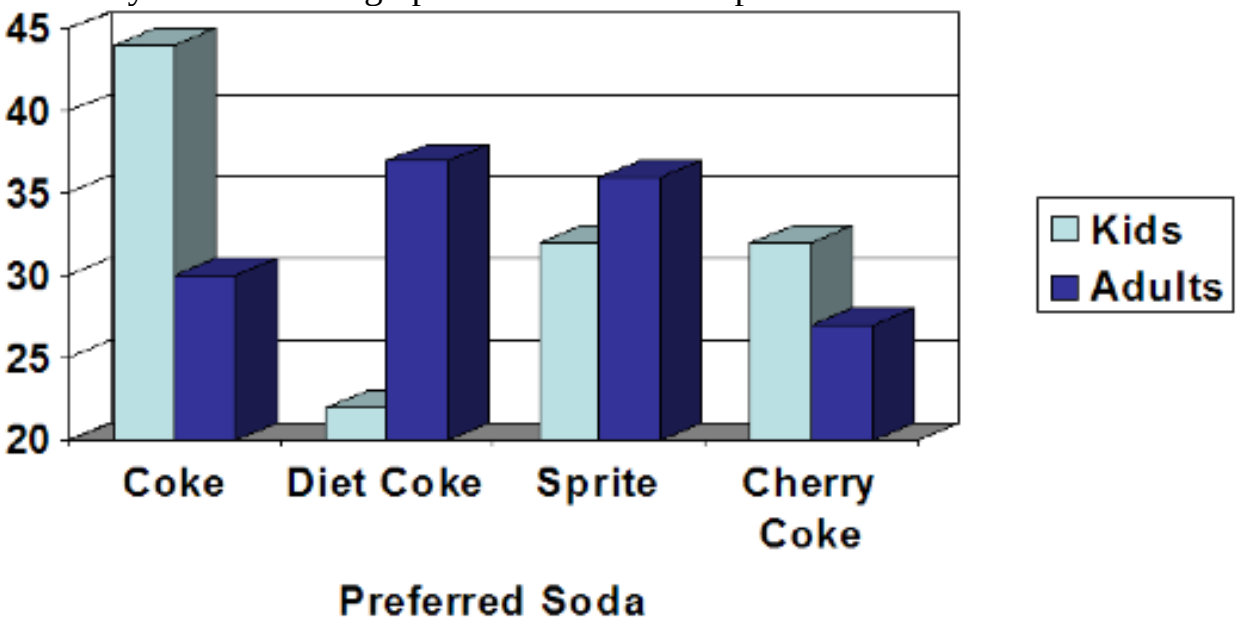
Exercises – Ch. 2

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?

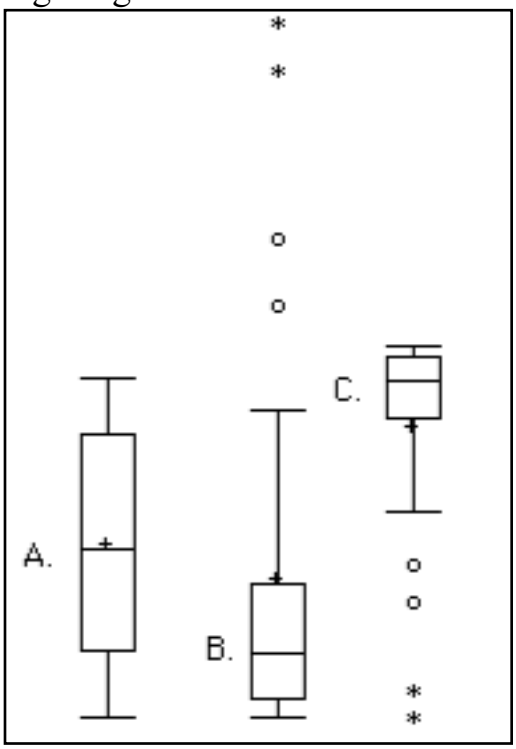
Favorite Movie Genre	Freq.
Comedy	14
Horror	9
Romance	8
Action	12

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.
    a. What is on the Y-axis? Explain.
    b. What is on the X-axis? Explain.
    c. 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.



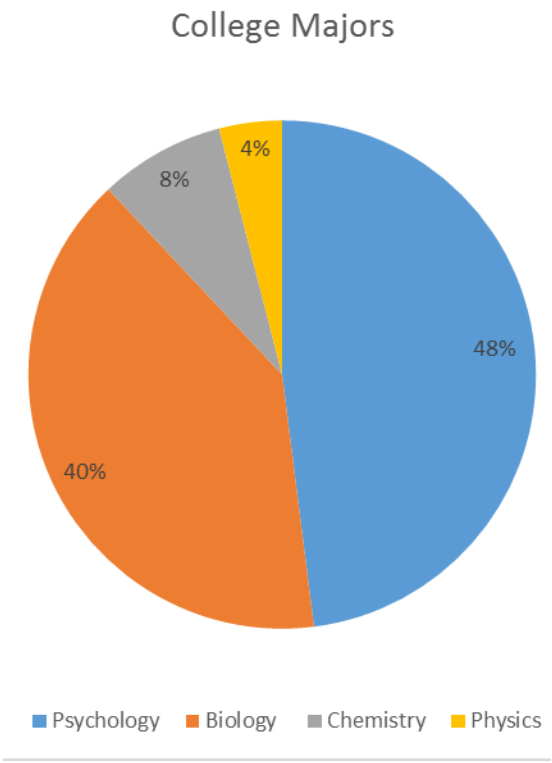
5. Which of the box plots on the graph has a large positive skew? Which has a large negative skew?



6. Create a histogram of the following data representing how many shows children said they watch each day:

Number of TV Shows	Frequency
0	2
1	18
2	36
3	7
4	3

7. Explain the differences between bar charts and histograms. When would each be used?
8. Draw a histogram of a distribution that is
    a. Negatively skewed
    b. Symmetrical
    c. 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.



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. 2

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]
    a. The Y-axis would have the frequency or proportion because this is always the case in histograms
    b. The X-axis has income, because this is out quantitative variable of interest
    c. 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

Chapter 3: Measures of Central Tendency and Spread

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 describe the distribution of our data objectively and mathematically – out first step into statistical analysis! The topics here will serve as the basis for everything we do in the rest of the course.

What is Central Tendency?

What is "central tendency," and why do we want to know the central tendency of a group of scores? Let us first try to answer these questions intuitively. Then we will proceed to a more formal discussion.

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 1. 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 1. 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. Figure 1 shows the results of an experiment on memory for chess positions. Subjects were shown a chess position and then asked to reconstruct it on an empty chess board. The number of pieces correctly placed was recorded. This was repeated for two more chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions. The maximum possible score was 89.

	8	05
	7	156
	6	233
	5	168
330	4	06
9420	3	
622	2	

Figure 1. 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.

Two groups are compared. On the left are people who don't play chess. 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.

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 attempt to communicate the idea behind each concept. In the succeeding sections we will give statistical measures for these concepts of central tendency.

Definitions of Center

Now we explain the three different ways of defining the center of a distribution. All three are called measures of central tendency.

Balance Scale

One definition of central tendency is the point at which the distribution is in balance. Figure 2 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.

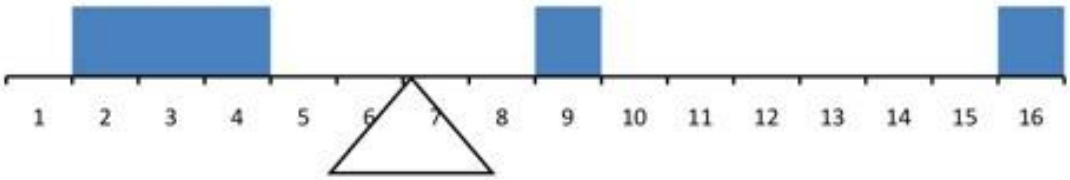


Figure 2. A balance scale.

For another example, consider the distribution shown in Figure 3. It is balanced by placing the fulcrum in the geometric middle.

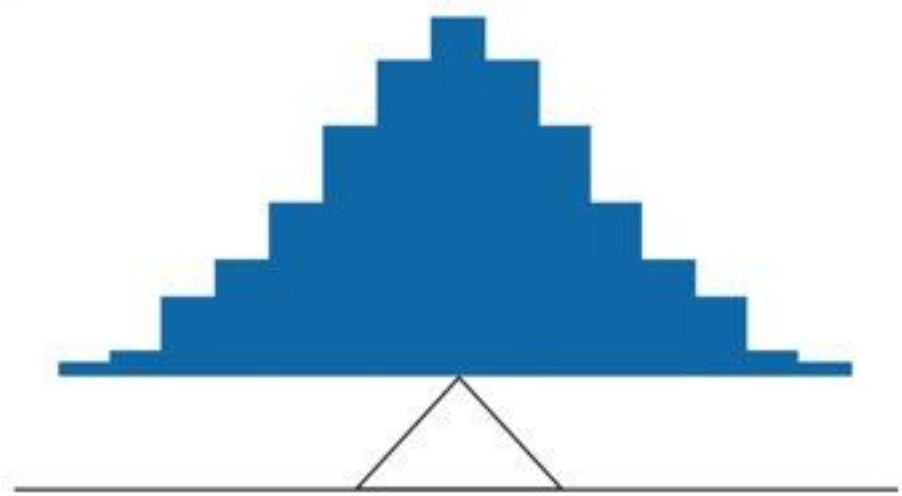


Figure 3. A distribution balanced on the tip of a triangle.

Figure 4 illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center.

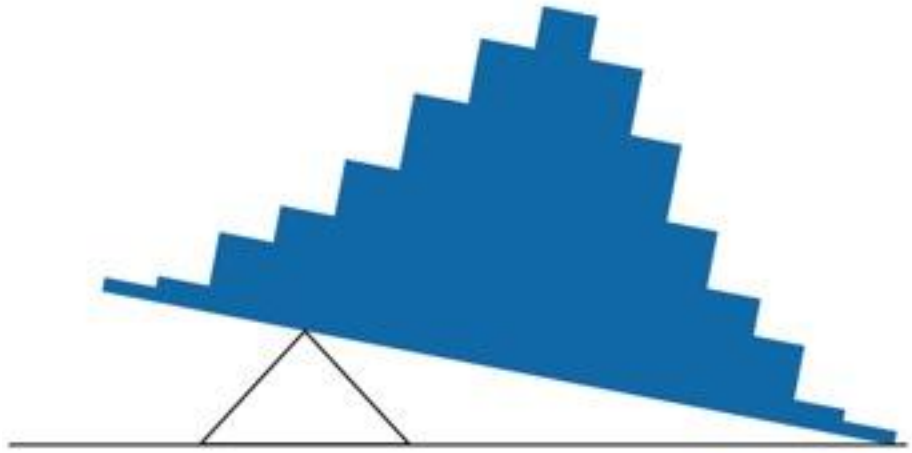


Figure 4. The distribution is not balanced.

Figure 5 shows an asymmetric distribution. To balance it, we cannot put the fulcrum halfway between the lowest and highest values (as we did in Figure 3). Placing the fulcrum at the "half way" point would cause it to tip towards the left.

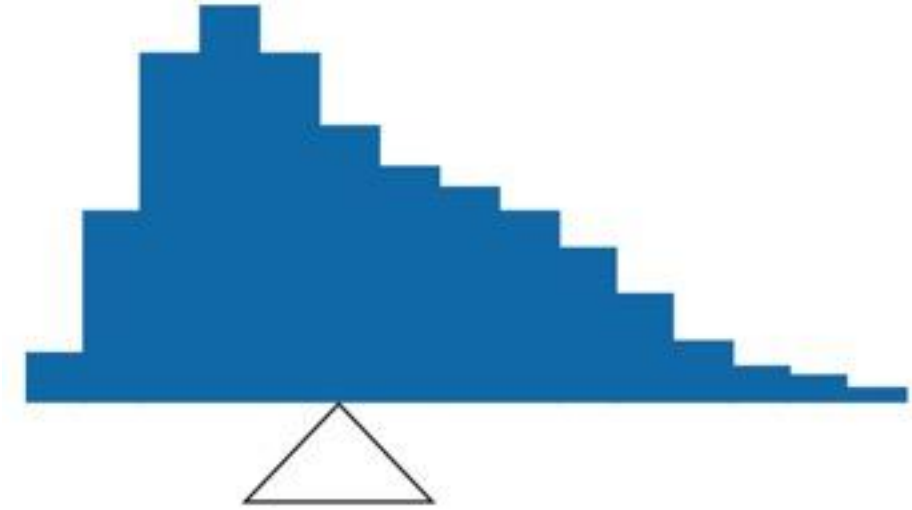


Figure 5. An asymmetric distribution balanced on the tip of a triangle.

Smallest Absolute Deviation

Another way to define the center of a distribution is based on the concept of the sum of the absolute deviations (differences). Consider the distribution made up of the five numbers 2, 3, 4, 9, 16. Let's see how far the distribution is from 10 (picking a number arbitrarily). Table 2 shows the sum of the absolute deviations of these numbers from the number 10.

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	28

Table 2. An example of the sum of absolute deviations

The first row of the table shows that the absolute value of the difference between 2 and 10 is 8; the second row shows that the absolute difference between 3 and 10 is 7, and similarly for the other rows. When we add up the five absolute deviations,

we get 28. So, the sum of the absolute deviations from 10 is 28. Likewise, the sum of the absolute deviations from 5 equals 3 + 2 + 1 + 4 + 11 = 21. So, the sum of the absolute deviations from 5 is smaller than the sum of the absolute deviations from 10. In this sense, 5 is closer, overall, to the other numbers than is 10.

We are now in a position to define a second measure of central tendency, this time in terms of absolute deviations. Specifically, according to our second definition, the center of a distribution is the number for which the sum of the absolute deviations is smallest. As we just saw, the sum of the absolute deviations from 10 is 28 and the sum of the absolute deviations from 5 is 21. Is there a value for which the sum of the absolute deviations is even smaller than 21? Yes. For these data, there is a value for which the sum of absolute deviations is only 20. See if you can find it.

Smallest Squared Deviation

We shall discuss one more way to define the center of a distribution. It is based on the concept of the sum of squared deviations (differences). Again, consider the distribution of the five numbers 2, 3, 4, 9, 16. Table 3 shows the sum of the squared deviations of these numbers from the number 10.

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
Sum	186

Table 3. An example of the sum of squared deviations.

The first row in the table shows that the squared value of the difference between 2 and 10 is 64; the second row shows that the squared difference between 3 and 10 is 49, and so forth. When we add up all these squared deviations, we get 186. Changing the target from 10 to 5, we calculate the sum of the squared deviations from 5 as 9 + 4 + 1 + 16 + 121 = 151. So, the sum of the squared deviations from 5 is smaller than the sum of the squared deviations from 10. Is there a value for which the sum of the squared deviations is even smaller than 151? Yes, it is

possible to reach 134.8. Can you find the target number for which the sum of squared deviations is 134.8?

The target that minimizes the sum of squared deviations provides another useful definition of central tendency (the last one to be discussed in this section). It can be challenging to find the value that minimizes this sum.

Measures of Central Tendency

In the previous section we saw that there are several ways to define central tendency. This section defines the three most common measures of central tendency: the mean, the median, and the mode. The relationships among these measures of central tendency and the definitions given in the previous section will probably not be obvious to you.

This section gives only the basic definitions of the mean, median and mode. A further discussion of the relative merits and proper applications of these statistics is presented in a later section.

Arithmetic Mean

The arithmetic mean is the most common measure of central tendency. 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 "$\overline{X}$" (pronounced "X-bar") is used for the mean of a sample. The formula for μ is shown below:

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

where ΣX is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for $\overline{X}$ is essentially identical:

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

where ΣX 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 $\bar{X}$).

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Table 4 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.45 as shown below.

$$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.45$$

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 4. Number of touchdown passes.

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, the geometric mean, or some other 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. For the data in Table 1, 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. 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 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:

$$\frac{4 + 7}{2} = 5.5$$

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. For the data in Table 1, 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

More on the Mean and Median

In the section "What is central tendency," we saw that the center of a distribution could be defined three ways: (1) the point on which a distribution would balance, (2) the value whose average absolute deviation from all the other values is minimized, and (3) the value whose squared difference from all the other values is minimized. 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.

Table 6 shows the absolute and squared deviations of the numbers 2, 3, 4, 9, and 16 from their median of 4 and their mean of 6.8. You can see that the sum of absolute deviations from the median (20) is smaller than the sum of absolute deviations from the mean (22.8). On the other hand, the sum of squared deviations from the median (174) is larger than the sum of squared deviations from the mean (134.8).

Value	Absolute Deviation from Median	Absolute Deviation from Mean	Squared Deviation from Median	Squared Deviation from Mean
2	2	4.8	4	23.04
3	1	3.8	1	14.44
4	0	2.8	0	7.84
9	5	2.2	25	4.84
16	12	9.2	144	84.64
Total	20	22.8	174	134.8

Table 6. Absolute & squared deviations from the median of 4 and the mean of 6.8.

Figure 6 shows that the distribution balances at the mean of 6.8 and not at the median of 4. The relative advantages and disadvantages of the mean and median are discussed in the section "Comparing Measures" later in this chapter.

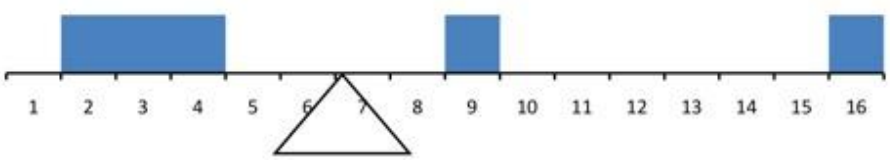


Figure 6. The distribution balances at the mean of 6.8 and not at the median of 4.0.

When a distribution is symmetric, then the mean and the median are the same. Consider the following distribution: 1, 3, 4, 5, 6, 7, 9. The mean and median are both 5. The mean, median, and mode are identical in the bell-shaped normal distribution.

Comparing Measures of Central Tendency

How do the various measures of central tendency compare with each other? For symmetric distributions, the mean and median, as is the mode except in bimodal distributions. Differences among the measures occur with skewed distributions. Figure 7 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

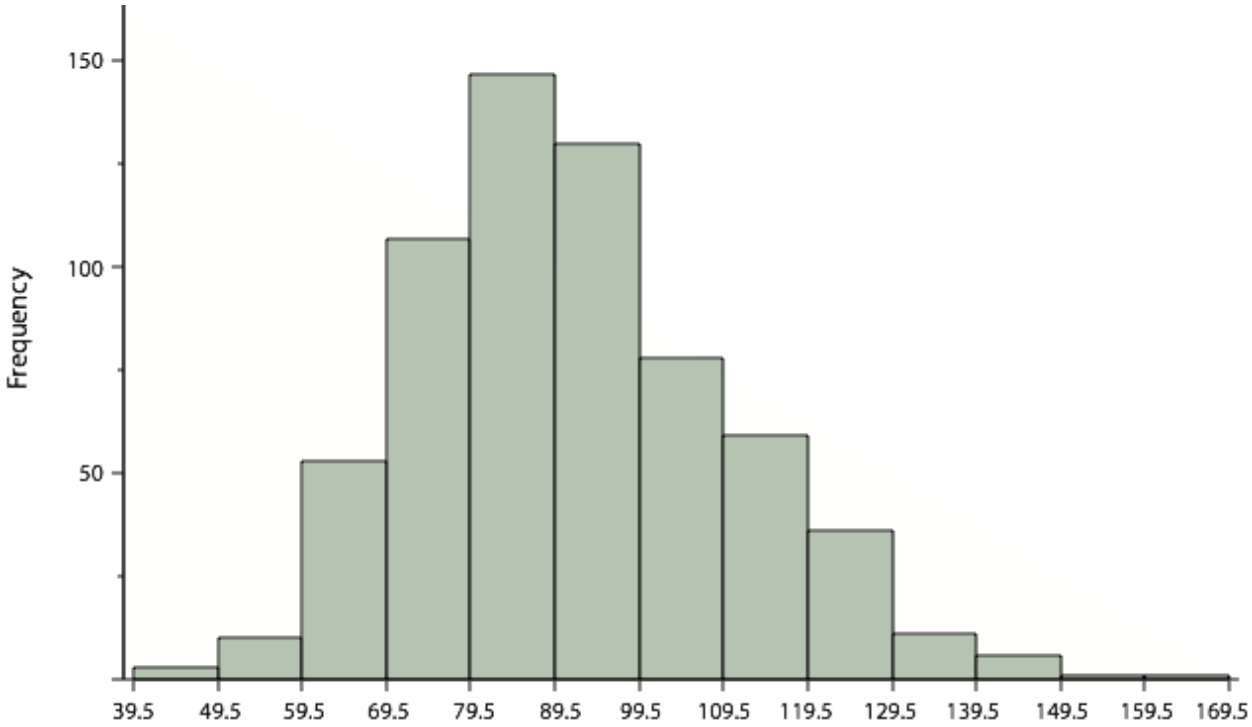


Figure 7. A distribution with a positive skew.

Measures of central tendency are shown in Table 7. 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	84.00
Median	90.00
Mean	91.58

Table 7. Measures of central tendency for the test scores.

The distribution of baseball salaries (in 1994) shown in Figure 8 has a much more pronounced skew than the distribution in Figure 7.

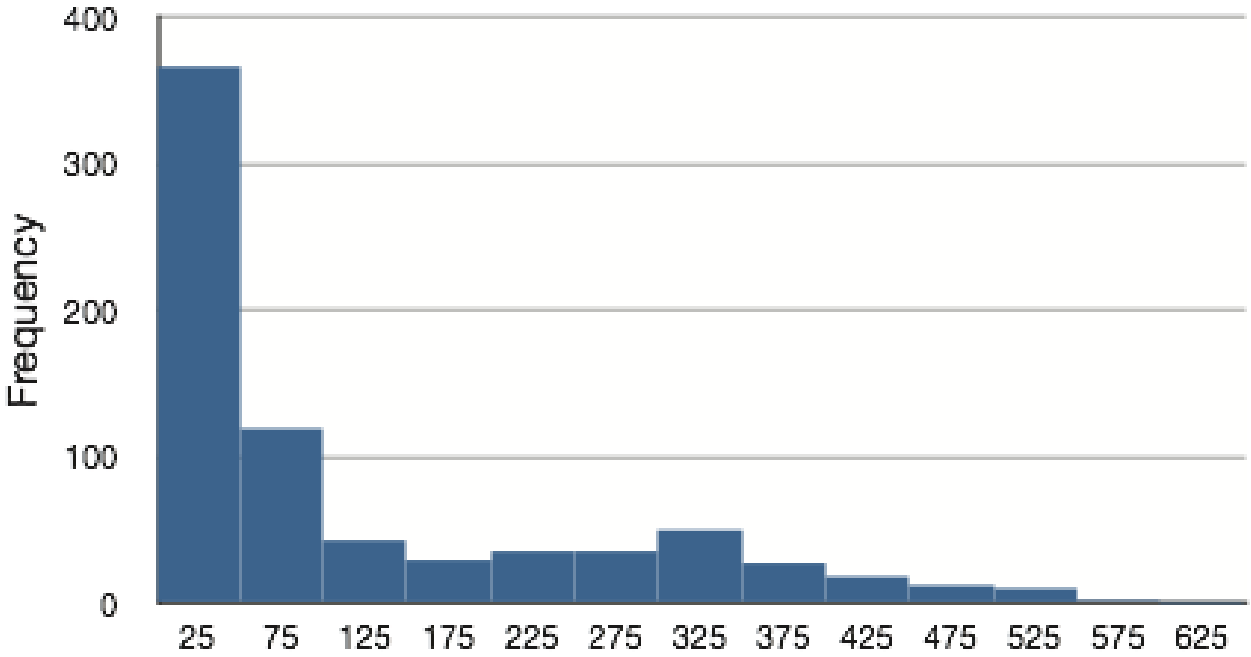


Figure 8. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Table 8 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
Mode	250
Median	500
Mean	1,183

Table 8. Measures of central tendency for baseball salaries (in thousands of dollars).

Spread and Variability

Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider graphs in Figure 9. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

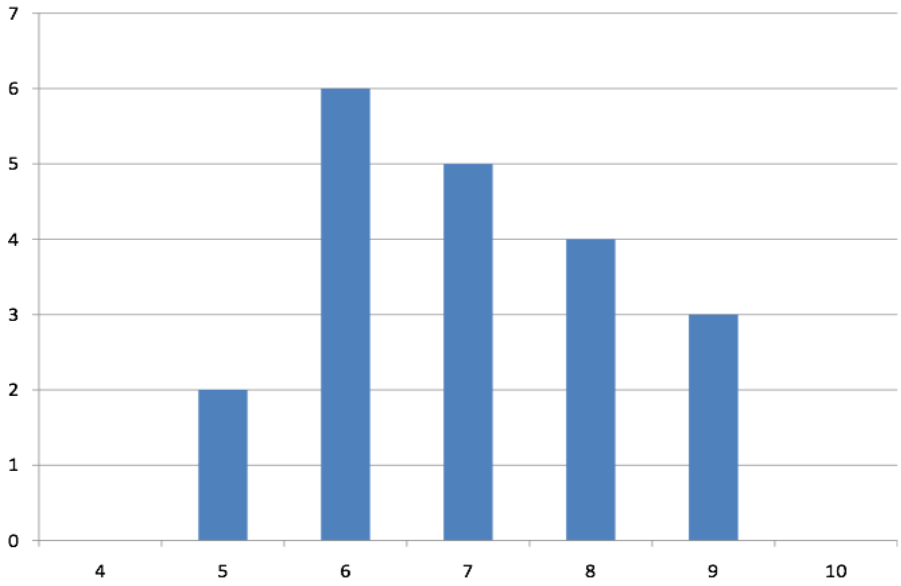


Figure 9.1. Bar chart of quiz one.

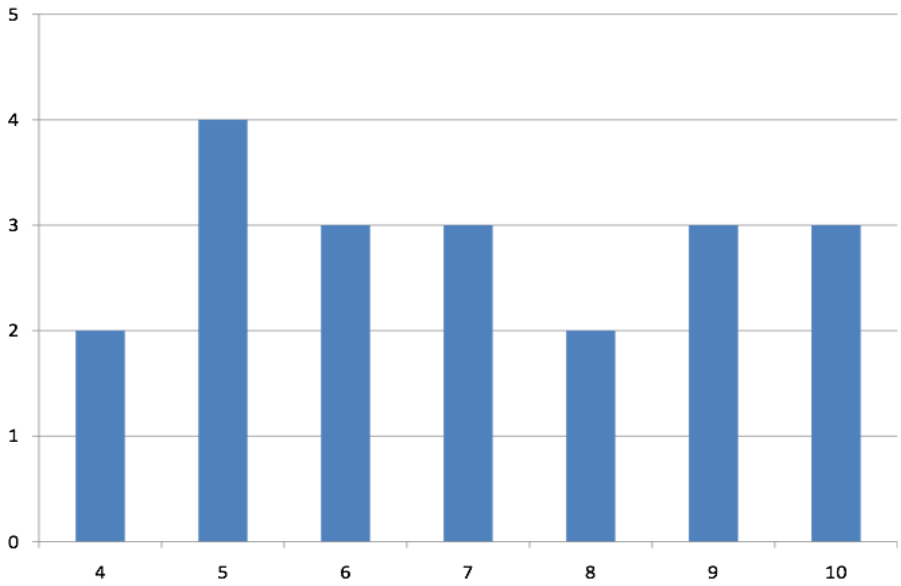


Figure 9.2. Bar chart of quiz two.

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. There are three frequently used measures of variability: range, variance, and standard deviation. In the next few paragraphs, we will look at each of these measures of variability in more detail.

Range

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. 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. Now consider the two quizzes shown in Figure 1. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

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

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 = 75th percentile - 25th percentile
```

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. Recall that in the discussion of box plots, 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.

Sum of Squares

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 middle of the distribution, we can see how far, on average, each data point is from the center. The data from Quiz 1 are shown in Table 9. The mean score is 7.0 ($\Sigma X/N = 140/20 = 7$). Therefore, the column "$X - \overline{X}$" contains deviations (how far each score deviates from the mean), here calculated as the score minus 7. The column "$(X - \overline{X})^2$" has the "Squared Deviations" and is simply the previous column squared.

There are a few things to note about how Table 9 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). 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 $\sum(X - \overline{X})^2$). 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.

X	$X - \bar{X}$	$(X - \bar{X})^2$
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	-2	4
5	-2	4
Σ = 140	Σ = 0	Σ = 30

Table 9. Calculation of Variance for Quiz 1 scores.

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. The variance is defined as the average squared difference of the scores from the mean. 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.

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

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

Notice that the numerator that formula is identical to the formula for Sum of Squares presented above with $\bar{X}$ 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. If we assume that the values in Table 9 represent the full population, then we can take our value of Sum of Squares and divide it by N to get our population variance:

$$\sigma^2 = \frac{30}{20} = 1.5$$

So, on average, scores in this population are 1.5 squared units away from the mean. This 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 sample statistic used to estimate the variance is s^2 ("s-squared"):

$$s^2 = \frac{\sum(X - \bar{X})^2}{N - 1}$$

This formula is very similar to the formula for the population variance with one change: 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 ($\bar{X}$) 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.

Going back to the values in Table 9 and treating those scores as a sample, we can estimate the sample variance as:

$$s^2 = \frac{30}{20 - 1} = 1.58$$

Notice that this value is 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.

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:

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

Back to our earlier example from Table 9: $\sigma = \sqrt{\frac{30}{20}} = \sqrt{1.5} = 1.22$

The sample statistic follows the same conventions and is given as s:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{N - 1}} = \sqrt{\frac{SS}{df}}$$

The sample standard deviation from Table 9 is: $s = \sqrt{\frac{30}{20-1}} = \sqrt{1.58} = 1.26$

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 example, 68% of the distribution is within one standard deviation (above and below) of the mean and approximately 95% of the distribution is within two standard deviations of the mean. Therefore, 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 x 10 = 30 and 50 + 2 x 10 = 70.

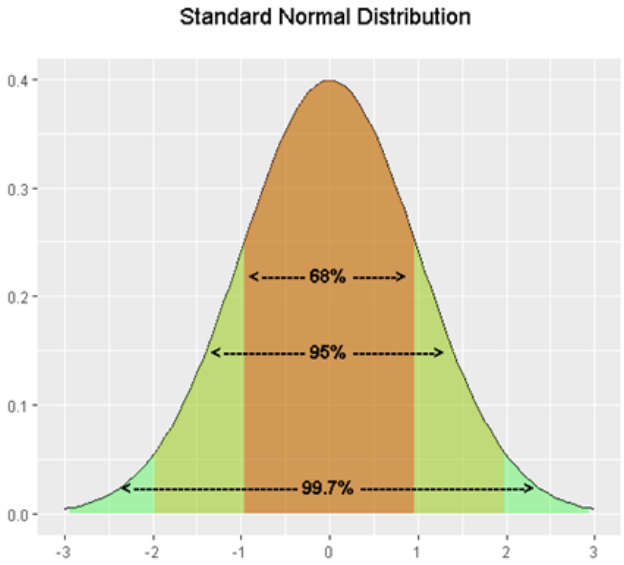


Figure 10: Percentages of the normal distribution

Figure 11 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 12 presents several more examples of this effect.

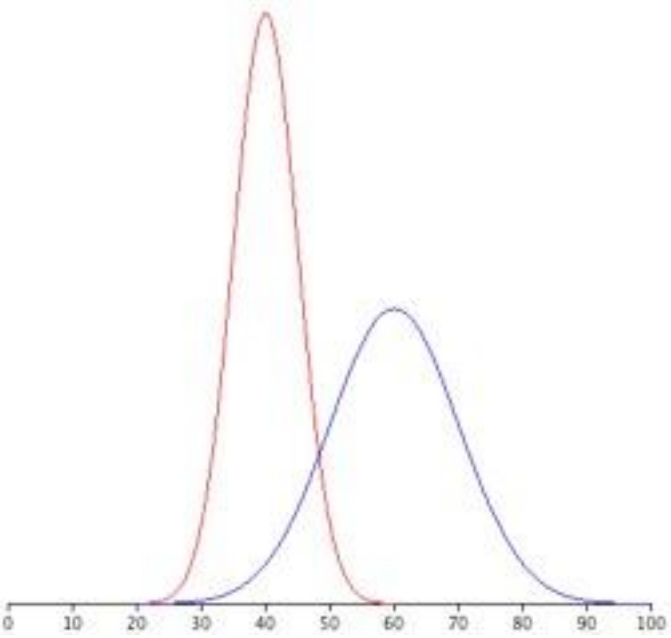


Figure 11. Normal distributions with standard deviations of 5 and 10.

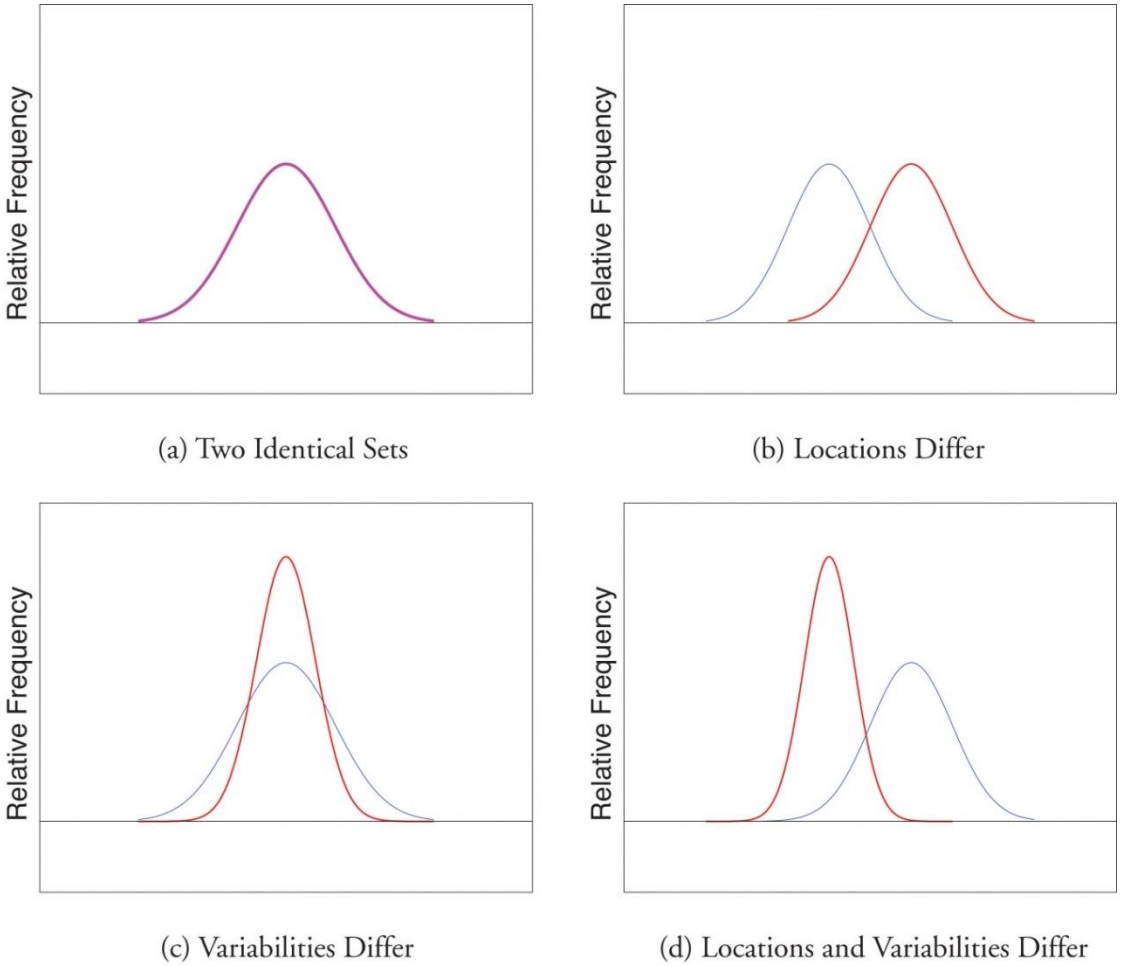


Figure 12. Differences between two datasets.

Exercises – Ch. 3

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:
    a. the same mean but different standard deviations.
    b. the same mean but different medians.
    c. the same median but different means.
5. Compute the population mean and population standard deviation for the following scores (remember to use the Sum of Squares table):
   5, 7, 8, 3, 4, 4, 2, 7, 1, 6
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
    a. Create a histogram of these data. What is the shape of this histogram?
    b. How do you think the three measures of central tendency will compare to each other in this dataset?
    c. Compute the sample mean, the median, and the mode
    d. Draw and label lines on your histogram for each of the above values. Do your results match your predictions?
7. 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
8. Using the same values from problem 7, 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?
9. 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?
10. Compute the sample mean and sample standard deviation for the following scores: -8, -4, -7, -6, -8, -5, -7, -9, -2, 0

Answers to Odd-Numbered Exercises – Ch. 3

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.
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. $\mu = 4.80$, $\sigma^2 = 2.36$
7. range = 16, $s^2 = 18.40$, $s = 4.29$
9. 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

Chapter 4: 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.

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 Karl Friedrich Gauss.

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).

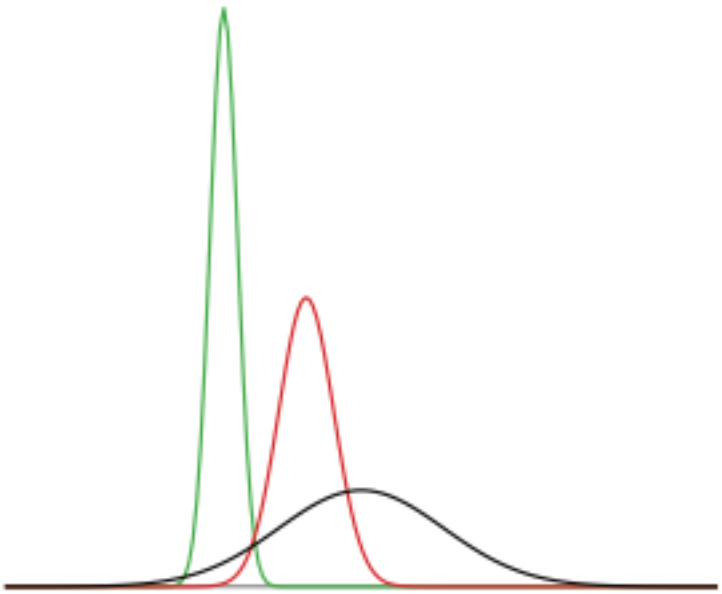


Figure 1. Normal distributions differing in mean and standard deviation.

Seven features of normal distributions are listed below.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.
7. 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

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. The formula for converting a raw score into a z-score is:

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

for values from a population and

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

for values from a sample.

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.

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.

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. 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. In Unit 2, we will learn to formalize the distinction between what we consider "close" to the center or "far" from the center. For now, we will use a rough cut-off of 1.5 standard deviations in either direction as the difference between close scores (those within 1.5 standard deviations or between z = -1.5 and z = 1.5) and extreme scores (those farther than 1.5 standard deviations – below z = -1.5 or above z = 1.5).

We can also 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. 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 of 54 with a standard deviation of 8. To find out our relative location within this distribution, we simply convert our test score into a z-score.

$$z = \frac{X - \mu}{\sigma} = \frac{68 - 54}{8} = 1.75$$

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!

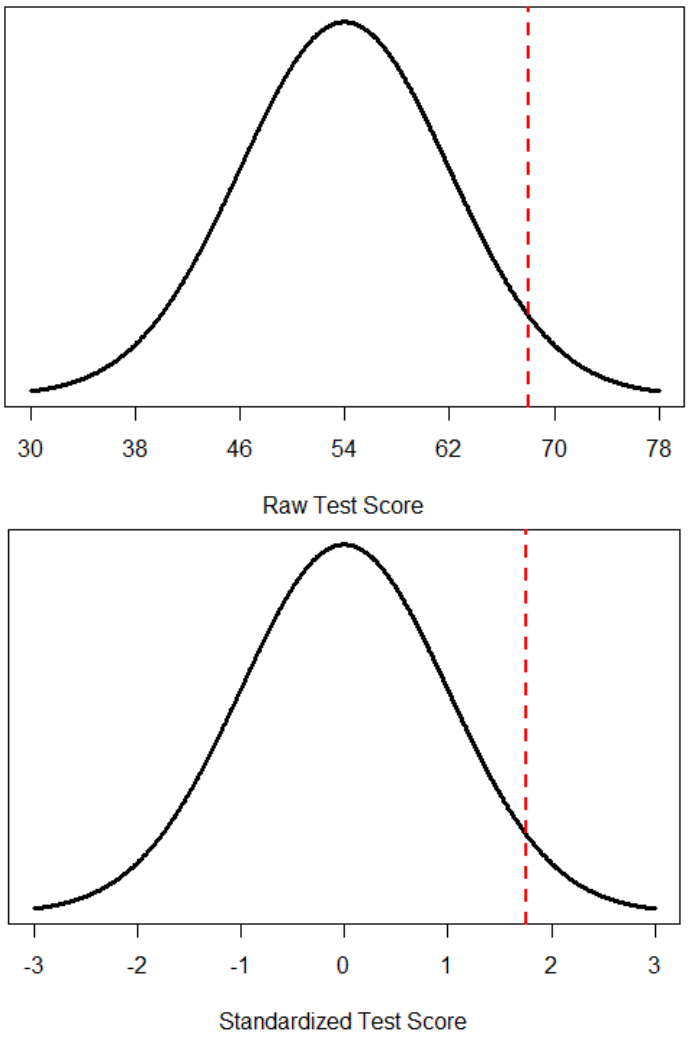


Figure 2. Raw and standardized versions of a single score

Figure 2 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.

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

$$z_{math} = \frac{501 - 511}{120} = -0.08$$

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

$$z_{CR} = \frac{501 - 495}{116} = 0.05$$

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 1:

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 1. 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 2.

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 2. 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). This can be very useful if we don't want to work with negative numbers or if we have a specific range we would like to present. The formulas for transforming z to x are:

$$x = z\sigma + \mu$$

for a population and

$$x = zs + \bar{X}$$

for a sample. Notice that these are just simple rearrangements of the original formulas for calculating z from raw scores.

Let's say we create a new measure of intelligence, and initial calibration finds that our scores have a mean of 40 and standard deviation of 7. Three people who have scores of 52, 43, and 34 want to know how well they did on the measure. We can convert their raw scores into z-scores:

$$z = \frac{52 - 40}{7} = 1.71$$

$$z = \frac{43 - 40}{7} = 0.43$$

$$z = \frac{34 - 40}{7} = -0.80$$

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), or 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:

$$\text{IQ} = 1.71 * 16 + 100 = 127.36$$

$$\text{IQ} = 0.43 * 16 + 100 = 106.88$$

$$\text{IQ} = -0.80 * 16 + 100 = 87.20$$

We would also likely round these values to 127, 107, and 87, respectively, for convenience.

Z-scores and the Area under the Curve

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 saw in the previous chapter that standard deviations can be used to divide the normal distribution: 68% of the distribution falls within 1 standard deviation of the mean, 95% within (roughly) 2 standard deviations, and 99.7% within 3 standard deviations. 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. 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 3 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.

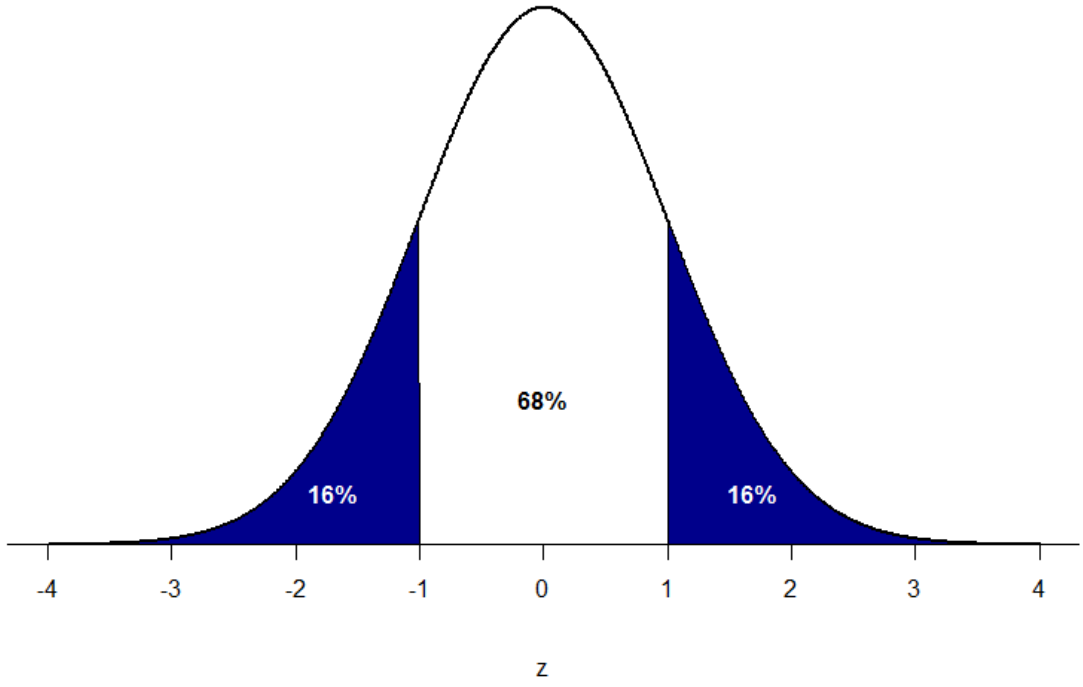


Figure 3. Shaded areas represent the area under the curve in the tails

We will have much more to say about this concept 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.

Exercises – Ch. 4

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:
    a. All normal distributions are symmetrical
    b. All normal distributions have a mean of 1.0
    c. All normal distributions have a standard deviation of 1.0

   d. 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:
   a. -2.00
   b. 1.25
   c. 3.50
   d. -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:
   a. One-half of a standard deviation below the mean
   b. 5 points above the mean
   c. Three standard deviations above the mean
   d. 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:
   a. 4.0
   b. 2.2
   c. -1.3
   d. 0.46

Answers to Odd-Numbered Exercises – Ch. 4

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. $\overline{X}$ = 4.2, s = 1.64; z = -1.34, -0.73, 0.49, 0.49, 1.10
5.
   a. 2 standard deviations below the mean, far
   b. 1.25 standard deviations above the mean, near
   c. 3.5 standard deviations above the mean, far
   d. 0.34 standard deviations below the mean, near
7. z = 0.75, 0.56, -2.75, -0.75, 2.19, -0.06

9. If
    a. -0.50
    b. 0.25
    c. 3.00
    d. 1.10

Chapter 5: Probability

Probability can seem like a daunting topic for many students. In a mathematical statistics course this might be true, as the meaning and purpose of probability gets obscured and overwhelmed by equations and theory. 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?

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. For instance, what is the probability that it will rain? That is, how likely do we think it is that it will rain today under the circumstances or conditions today? To define or understand the conditions that might affect how likely it is to rain, we might look out the window and say, "it's sunny outside, so it's not very likely that it will rain today." Stated using probability language: given that it is sunny outside, the probability of rain is low. "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."

In these examples, we spoke about whether or not it is going to rain. Raining is an example of an event, which is the catch-all term we use to talk about any specific thing happening; it is a generic term that we specified to mean "rain" in exactly the same way that "conditions" is a generic term that we specified to mean "sunny" or "cloudy and windy."

It should also be noted that the terms "low" and "high" are relative and vague, and they will likely be interpreted different by different people (in other words: given how vague the terminology was, the probability of different interpretations is high). 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.

Let's look at a slightly deeper example. Say we have a regular, six-sided die (note that "die" is singular and "dice" is plural, a distinction that Dr. Foster has yet to get

correct on his 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:

$$Probability\ of\ an\ event = \frac{number\ of\ outcomes\ that\ satisfy\ our\ criteria}{total\ number\ of\ possible\ outcomes}$$

We can also using P() as shorthand for probability and A as shorthand for an event:

$$P(A) = \frac{number\ of\ outcomes\ that\ count\ a\ A}{total\ number\ of\ possible\ outcomes}$$

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

$$P(Even\ Number) = \frac{2, 4, or\ 6}{1, 2, 3, 4, 5, or\ 6} = \frac{3}{6} = \frac{1}{2}$$

So we have a 50% chance of rolling an even number of this die. 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. Now, let's turn back to more familiar topics.

Probability in Graphs and Distributions

We will see shortly that the normal distribution is the key to how probability works for our purposes. To understand exactly how, let's first look at a simple, intuitive example using pie charts.

Probability in Pie Charts

Recall that a pie chart represents how frequently a category was observed and that all slices of the pie chart add up to 100%, or 1. This means that if we randomly select an observation from the data used to create the pie chart, the probability of it taking on a specific value is exactly equal to the size of that category's slice in the pie chart.

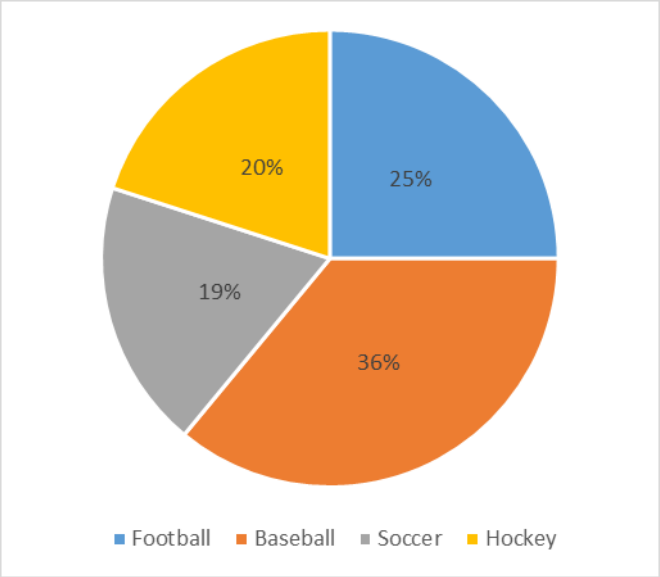


Figure 1. Favorite sports

Take, for example, the pie chart in Figure 1 representing the favorite sports of 100 people. If you put this pie chart on a dart board and aimed blindly (assuming you are guaranteed to hit the board), the likelihood of hitting the slice for any given sport would be equal to the size of that slice. So, the probability of hitting the baseball slice is the highest at 36%. The probability is equal to the proportion of the chart taken up by that section.

We can also add slices together. For instance, maybe we want to know the probability to finding someone whose favorite sport is usually played on grass. The outcomes that satisfy this criteria are baseball, football, and soccer. To get the probability, we simply add their slices together to see what proportion of the area

of the pie chart is in that region: 36% + 25% + 20% = 81%. We can also add sections together even if they do not touch. If we want to know the likelihood that someone's favorite sport is not called football somewhere in the world (i.e. baseball and hockey), we can add those slices even though they aren't adjacent or continuous in the chart itself: 36% + 20% = 56%. We are able to do all of this because 1) the size of the slice corresponds to the area of the chart taken up by that slice, 2) the percentage for a specific category can be represented as a decimal (this step was skipped for ease of explanation above), and 3) the total area of the chart is equal to 100% or 1.0, which makes the size of the slices interpretable.

Probability in Normal Distributions

If the language at the end of the last section sounded familiar, that's because its 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.

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.

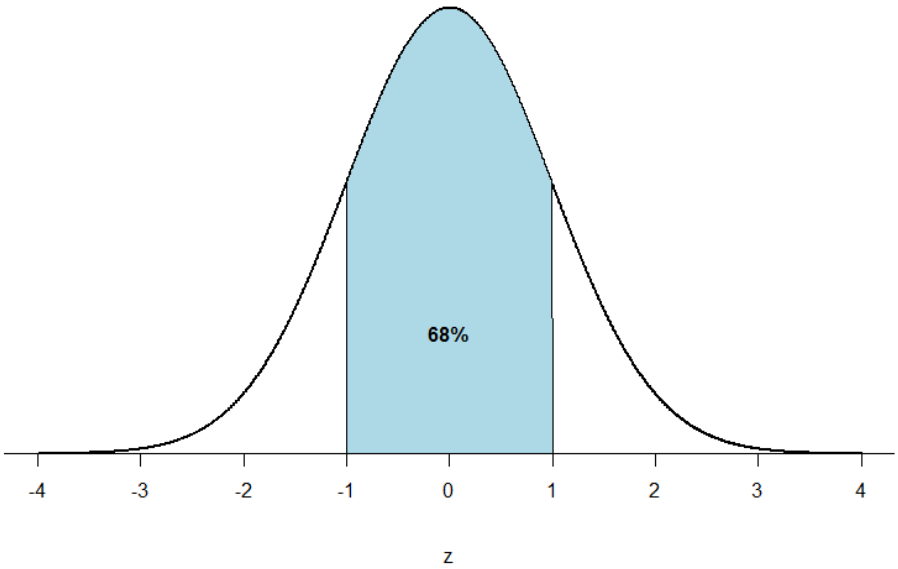


Figure 2: There is a 68% chance of selection a z-score from the blue-shaded region

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.

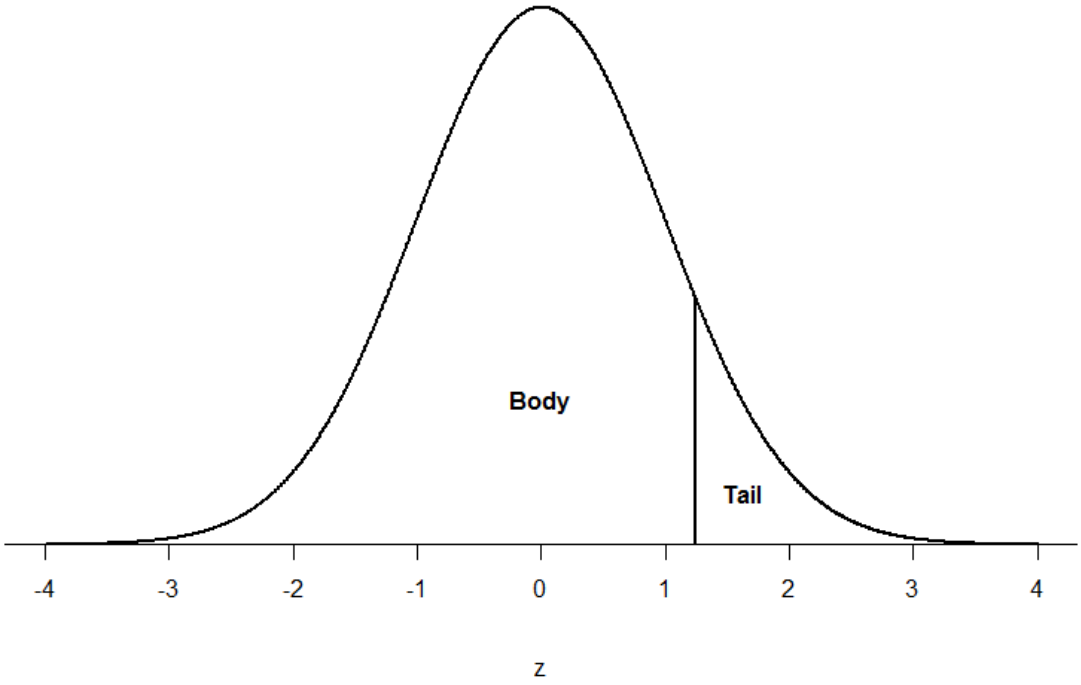


Figure 3. Body and tail of the normal distribution

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 you to you in the Standard Normal Distribution Table, also known at 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 0.02 (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	0.9463	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.

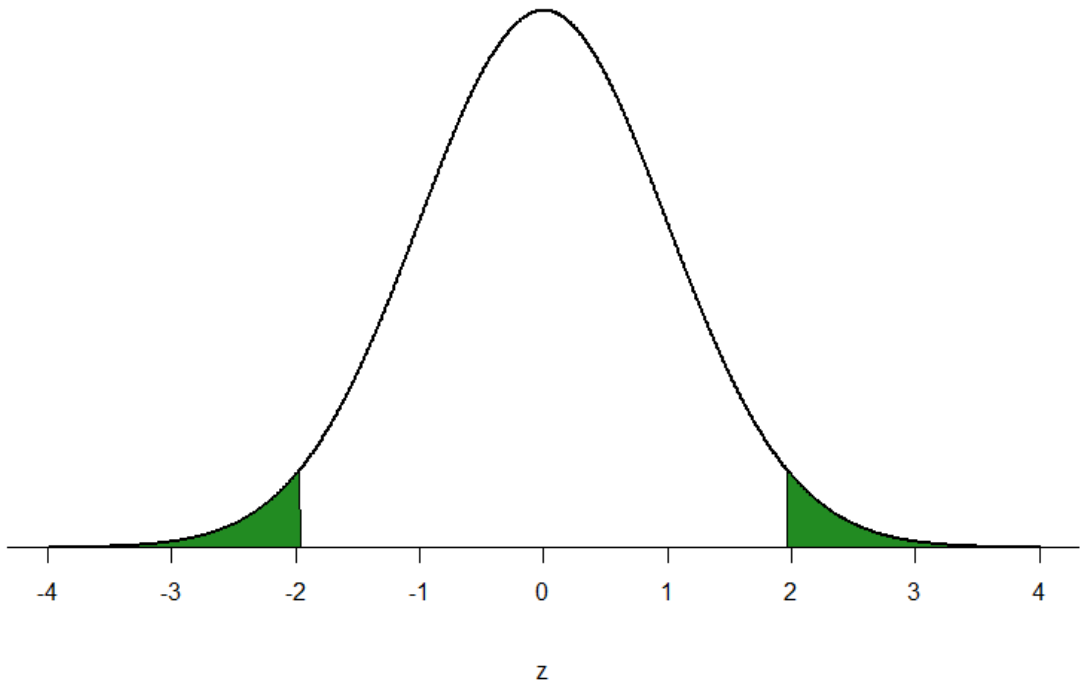


Figure 5. Area in the tails beyond z = -1.96 and z = 1.96

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.

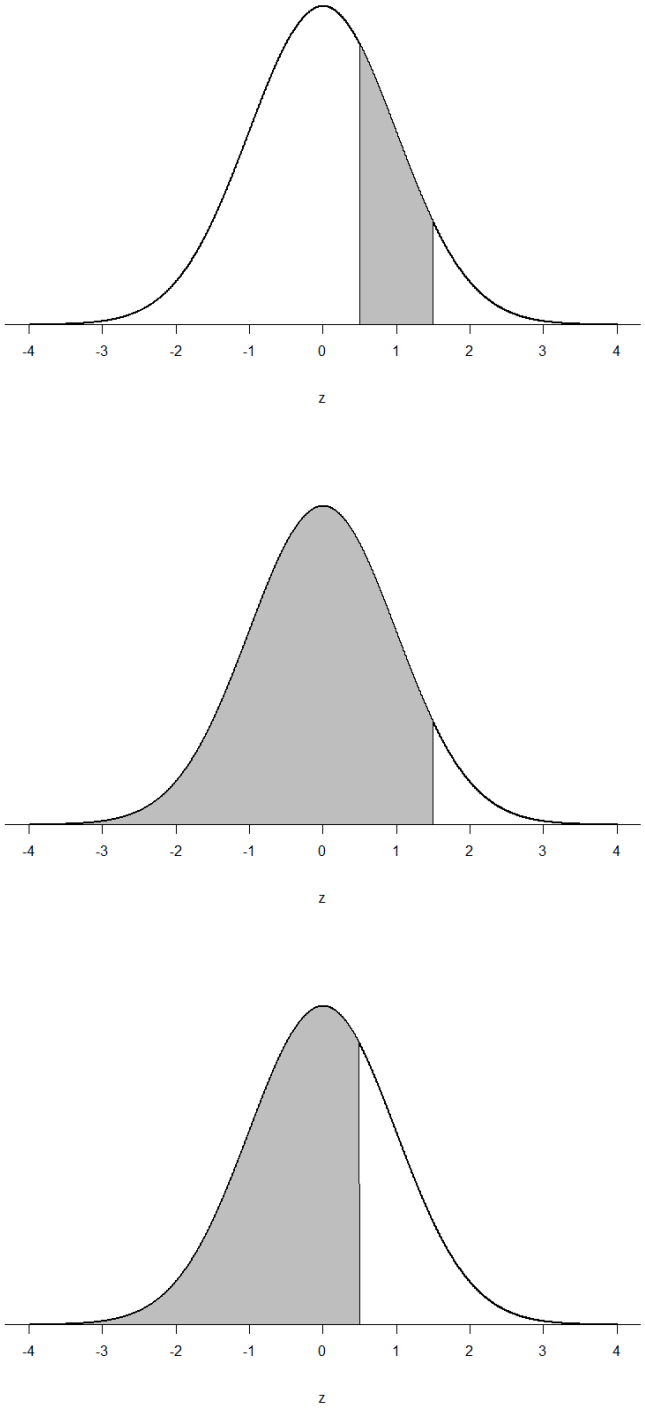


Figure 6. Area between z = 0.50 and 1.50, along with the corresponding areas in the body

Probability: The Bigger Picture

The concepts and ideas presented in this chapter are likely not intuitive at first. 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.

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). This extends to graphs like a pie chart, where the biggest slices take up more of the area and are therefore more likely to be chosen at random. 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.

Exercises – Ch. 5

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)
   Getting a z-score greater than z = 2.75
   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. 5

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)

Chapter 6: Sampling Distributions

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, especially chapters 4 and 5.

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.

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. 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. This will sometimes be written as $\mu_{\bar{X}}$ to denote it as the mean of the sample means. The spread of the sampling distribution is called the standard error, the quantification of sampling error, denoted $\sigma_{\bar{X}}$. The formula for standard error is:

$$\sigma_{\bar{X}} = {}^{\sigma}/_{\sqrt{n}}$$

Notice that the sample size is in this 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 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. Figure 1 displays the principles stated here in graphical form.

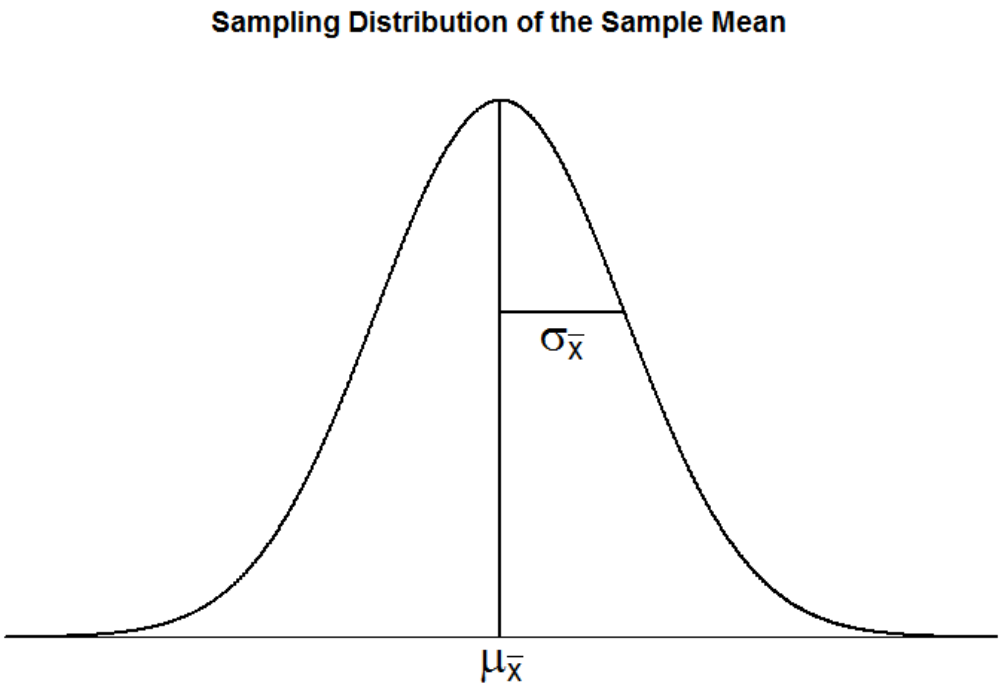


Figure 1. The sampling distribution of sample means

Two Important Axioms

We just learned that the sampling distribution is theoretical: we never actually see it. If that is true, then how can we know it works? How can we use something that we don't see? The answer lies in 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 states:

> 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 $\sigma_{\bar{X}}^2 = \sigma^2/n$. This distribution will approach normality as *n* increases.

From this, we are able to find the standard deviation of our sampling distribution, the standard error. As you can see, just like any other standard deviation, the standard error is simply the square root of the variance of the distribution.

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.

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 2 and 3.

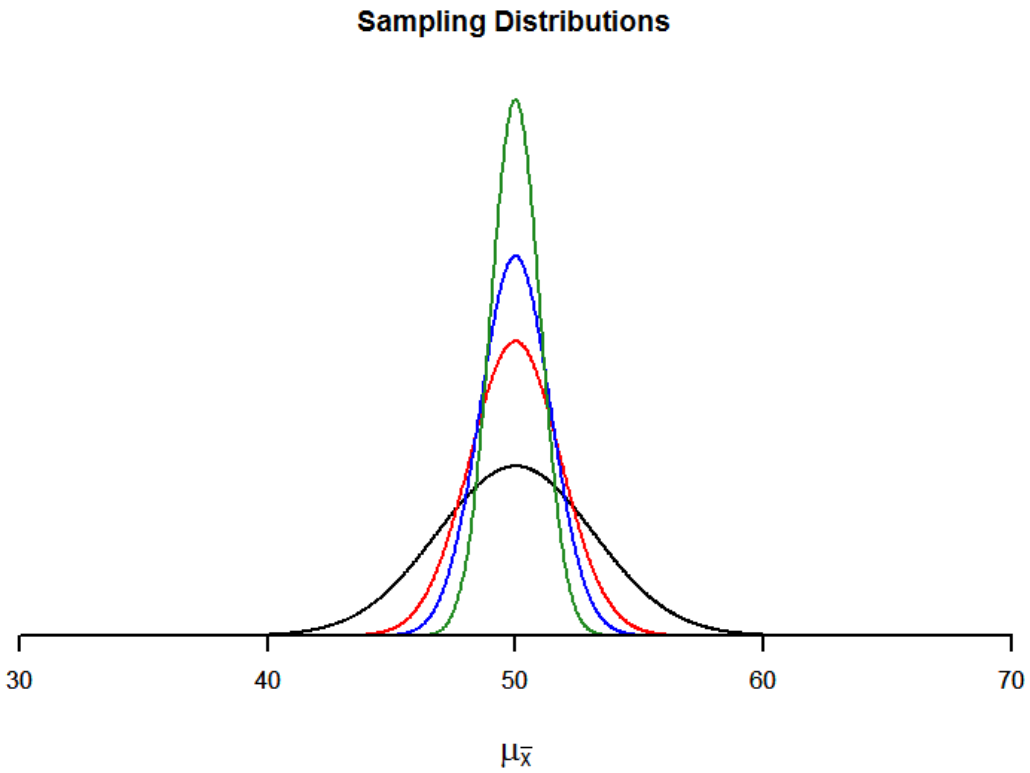


Figure 2. Sampling distributions from the same population with μ = 50 and σ = 10 but different sample sizes (N = 10, N = 30, N = 50, N = 100)

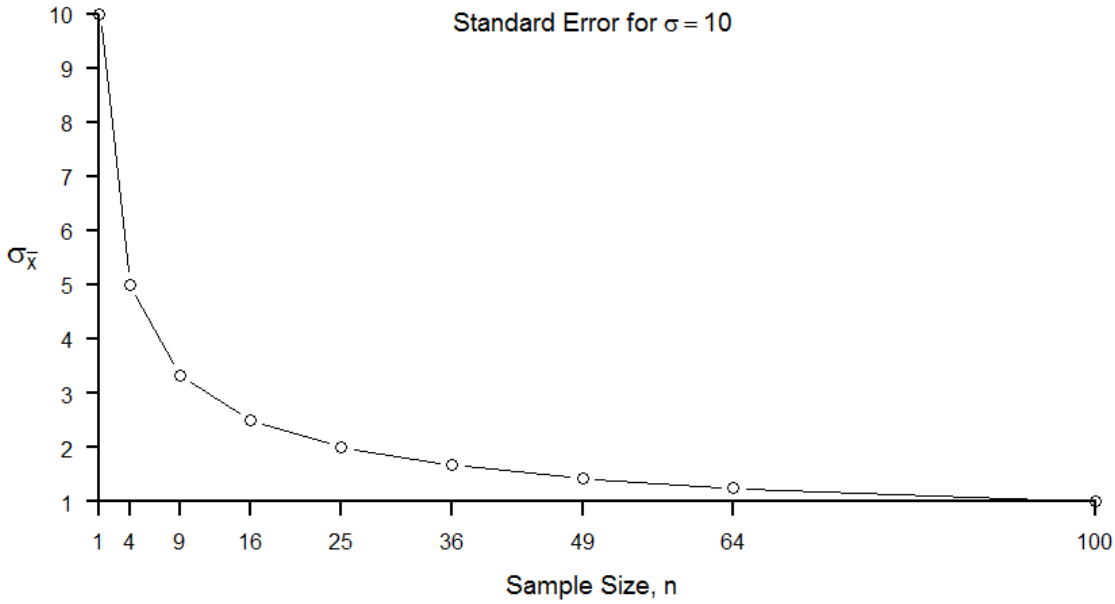


Figure 3. Relation between sample size and standard error for a constant σ = 10

Using Standard Error for Probability

We saw in chapter 6 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.

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

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 2). 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:

$$z = \frac{55 - 50}{10 / \sqrt{10}} = \frac{5}{3.16} = 1.58$$

Now we need to shade the area under the normal curve corresponding to scores greater than z = 1.58 as in Figure 4:

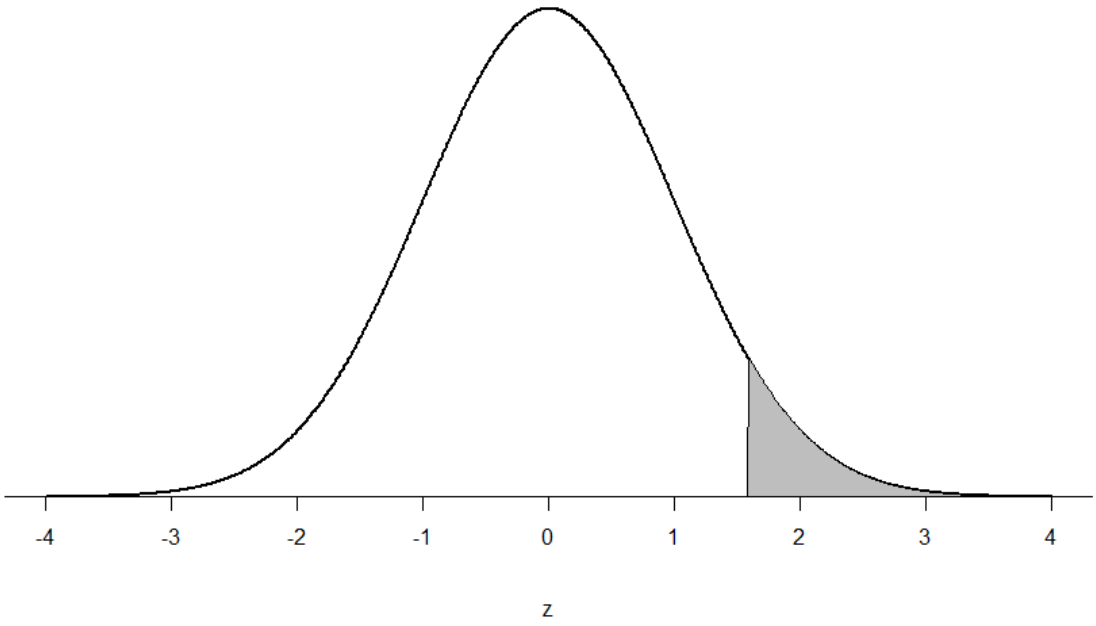


Figure 4: Area under the curve greater than z = 1.58

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:

$$z = \frac{55 - 50}{10/\sqrt{50}} = \frac{5}{1.41} = 3.55$$

Then we shade the appropriate region of the normal distribution:

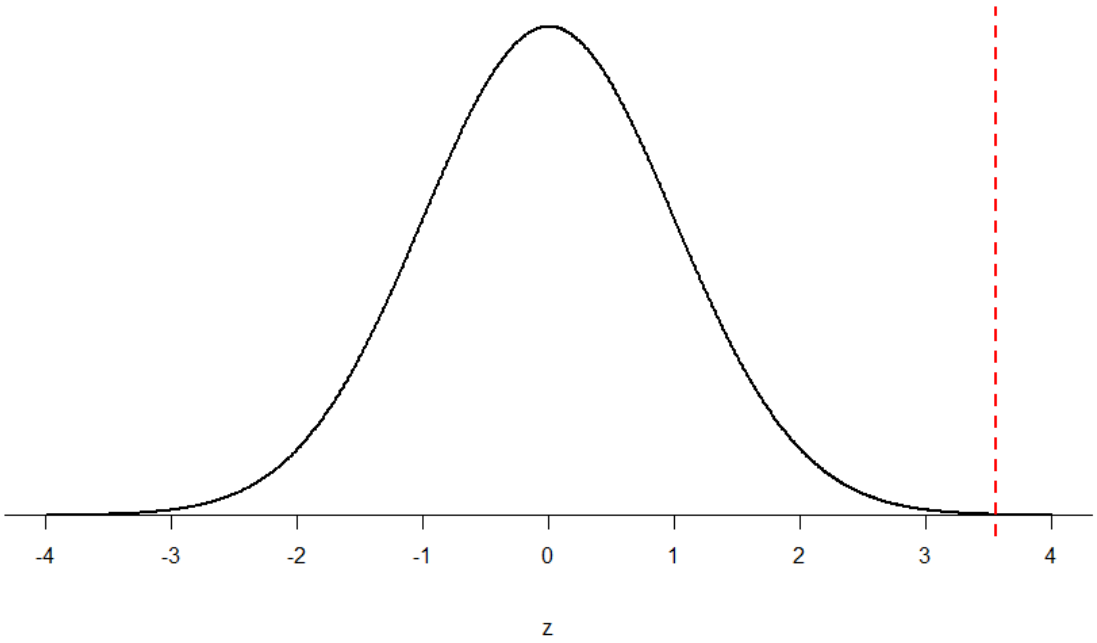


Figure 5: Area under the curve greater than z = 3.55

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. All of these pieces fit together, and the relations will always be the same: $n\uparrow\ \sigma_{\bar{X}}\downarrow\ z\uparrow\ p\downarrow$

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 $\bar{X}$ 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 6. 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.

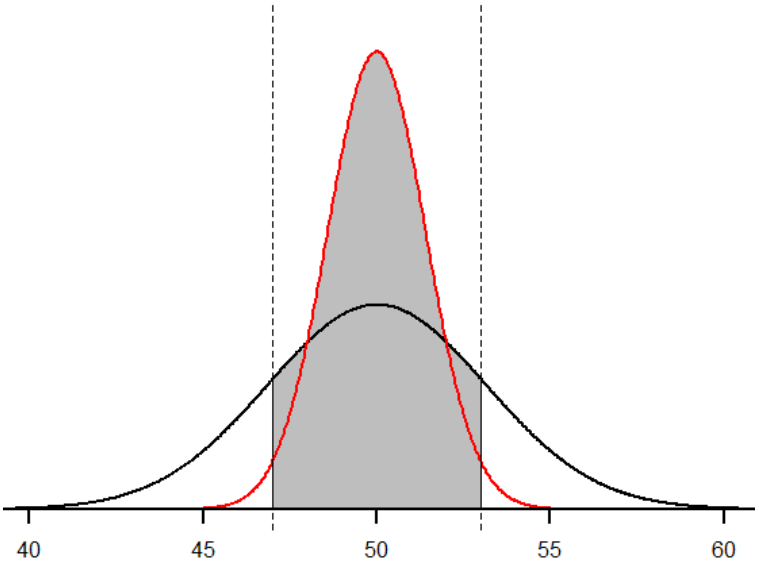


Figure 6. Areas between 47 and 53 for sampling distributions of n = 10 and n = 50

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 $(\bar{X} - \mu)$ 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.

Exercises – Ch. 6

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:
    a. $\overline{X} = 8$, n = 12
    b. $\overline{X} = 8$, n = 30
    c. $\overline{X} = 20$, n = 4
    d. $\overline{X} = 20$, 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. 6

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

Unit 2 – Hypothesis Testing

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.

Chapter 7: Introduction to Hypothesis Testing

This chapter lays out the basic logic and process of hypothesis testing. We will perform z-tests, which use the z-score formula from chapter 6 and data from a sample mean to make an inference about a population.

Logic and Purpose of Hypothesis Testing

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who 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. 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. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. 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.

Let's consider another example. The case study Physicians' Reactions sought to determine whether physicians spend less time with obese patients. 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.

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. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. Random assignment of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

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. Using methods presented in later chapters, this probability can be 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 and is not due to chance.

The Probability Value

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). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that

the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. We can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the probability value is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). 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.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference.

The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the null hypothesis, written H_0 ("H-naught"). 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:

$$H_0: \mu_{obese} - \mu_{average} = 0.$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$H_0: \rho = 0$$

where ρ is the population correlation, which we will cover in chapter 12.

Although the null hypothesis is usually that the value of a parameter is 0, there are occasions 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 birthweight 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:

$$H_0: \mu = 7.47$$

The number on the right hand side is our null hypothesis value that is informed by our research question. Notice that we are testing the value for μ, the population parameter, NOT the sample statistic $\bar{X}$. This is for two reasons: 1) once we collect data, we know what the value of $\bar{X}$ is – it's not a mystery or a question, it is observed and used for the second reason, which is 2) we are interested in understanding the population, not just our sample.

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 is put forward with the hope that it can be discredited and therefore rejected. 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.

In general, the null hypothesis 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. This is always our baseline starting assumption, and it is what we seek to reject. If we are trying to treat depression, we want to find a difference in average symptoms between our treatment and control groups. If we are trying to predict job performance, we want to find a relation between conscientiousness and evaluation scores. However, until we have evidence against it, we must use the null hypothesis as our starting point.

The 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, and there are three options, depending on where we expect the difference to lie. Thus, our alternative hypothesis is the mathematical way of stating our research question. If we expect our obtained sample mean to be above or below the null hypothesis value, which we call a directional hypothesis, then our alternative hypothesis takes the form:

$$H_A: \mu > 7.47 \qquad \text{or} \qquad H_A: \mu < 7.47$$

based on the research question itself. We should only use a directional hypothesis if we have 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:

$$H_A: \mu \neq 7.47$$

We will set different criteria for rejecting the null hypothesis based on the directionality (greater than, less than, or not equal to) of the alternative. To understand why, we need to see where our criteria come from and how they relate to z-scores and distributions.

Critical values, p-values, and significance level

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. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. 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 $\alpha = 0.05$.

The significance level is a threshold we set 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. 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).

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 α = .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 1.

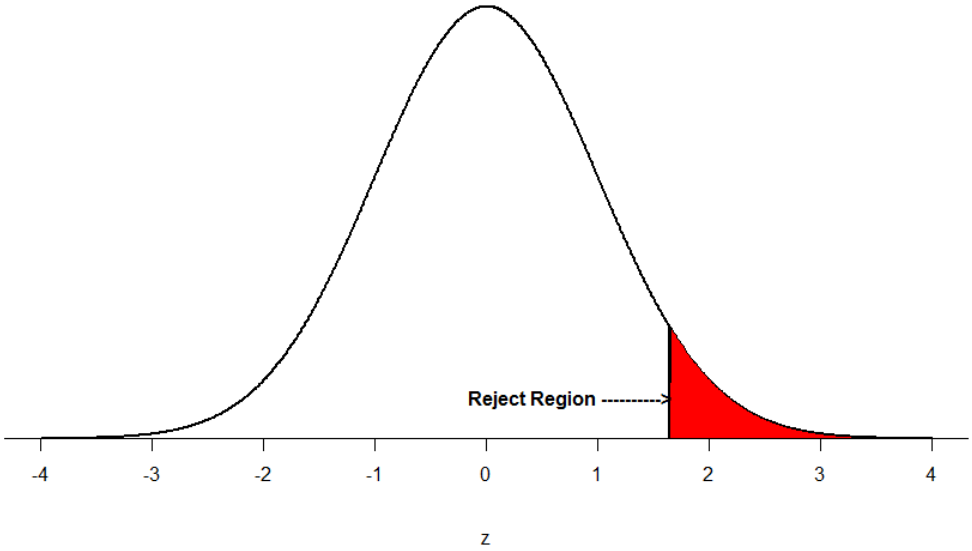


Figure 1: The rejection region for a one-tailed test
The shaded rejection region takes us 5% of the area under the curve. Any result which falls in that region is sufficient evidence to reject the null hypothesis.

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 2.

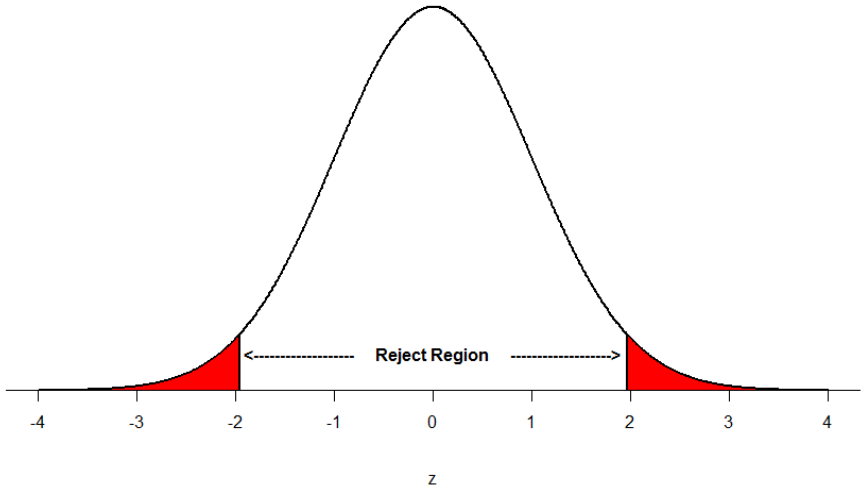


Figure 2: Two-tailed rejection region

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. The formula for our z-statistic has not changed:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

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 3.

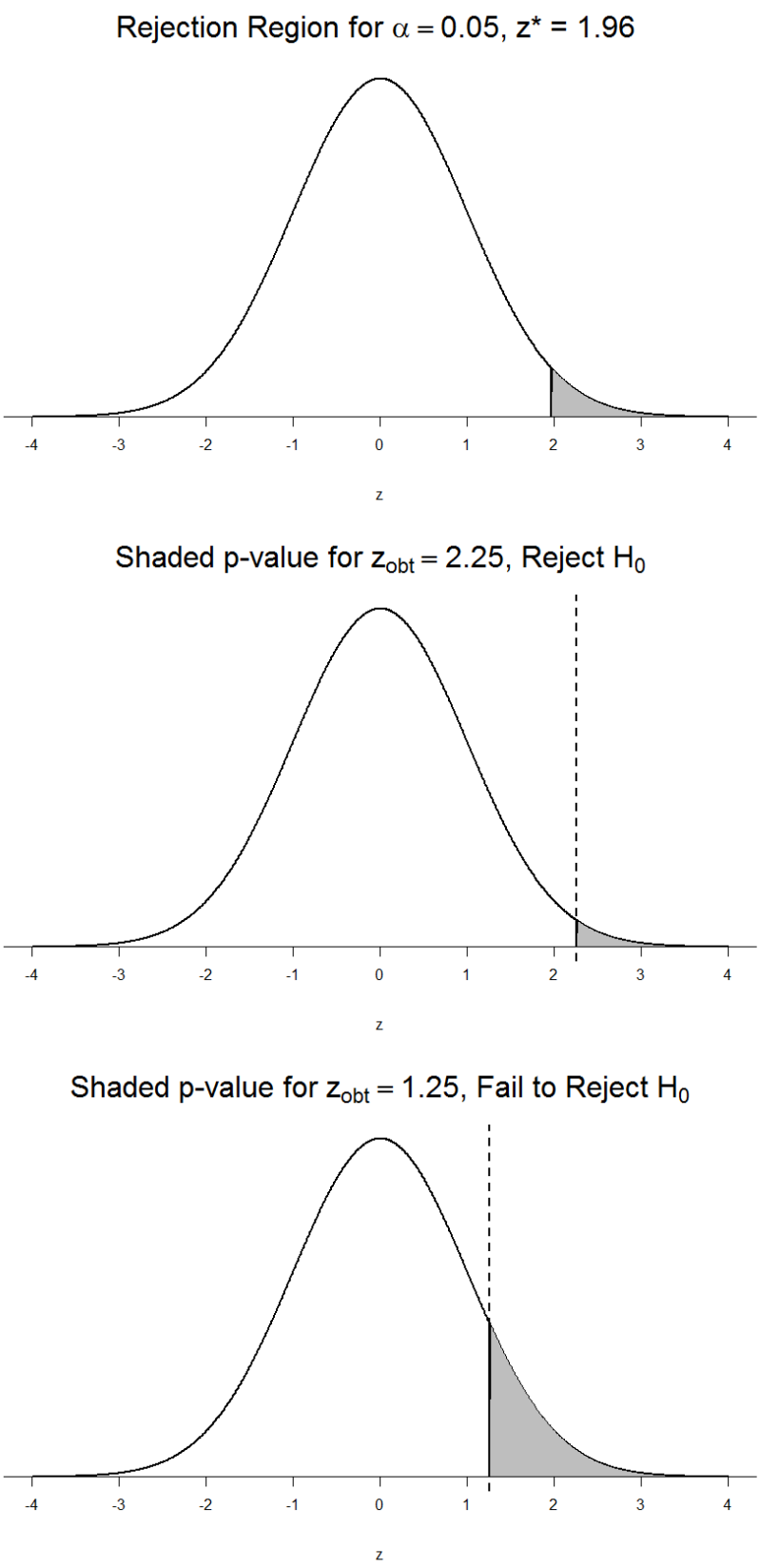


Figure 3. Relation between α, z_{obt}, and p

When the null hypothesis is rejected, the effect is said to be *statistically significant*. For example, in the Physicians Reactions case 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.

> 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.

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 remained 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.

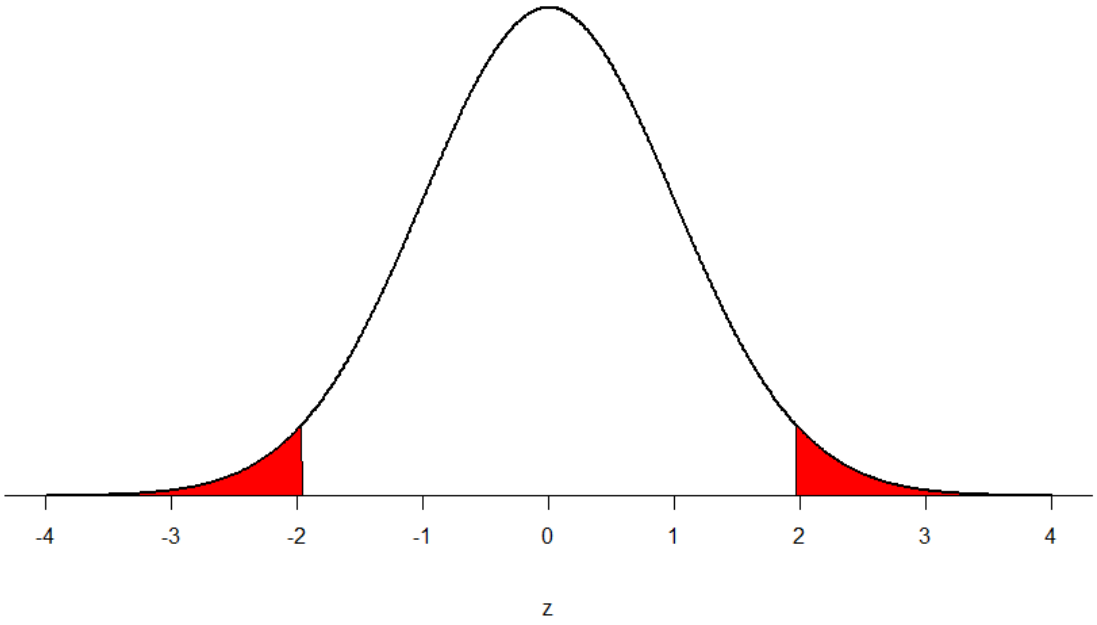


Figure 4: Rejection region for $z^* = \pm 1.96$

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 $\overline{X} = 7.75$ 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:

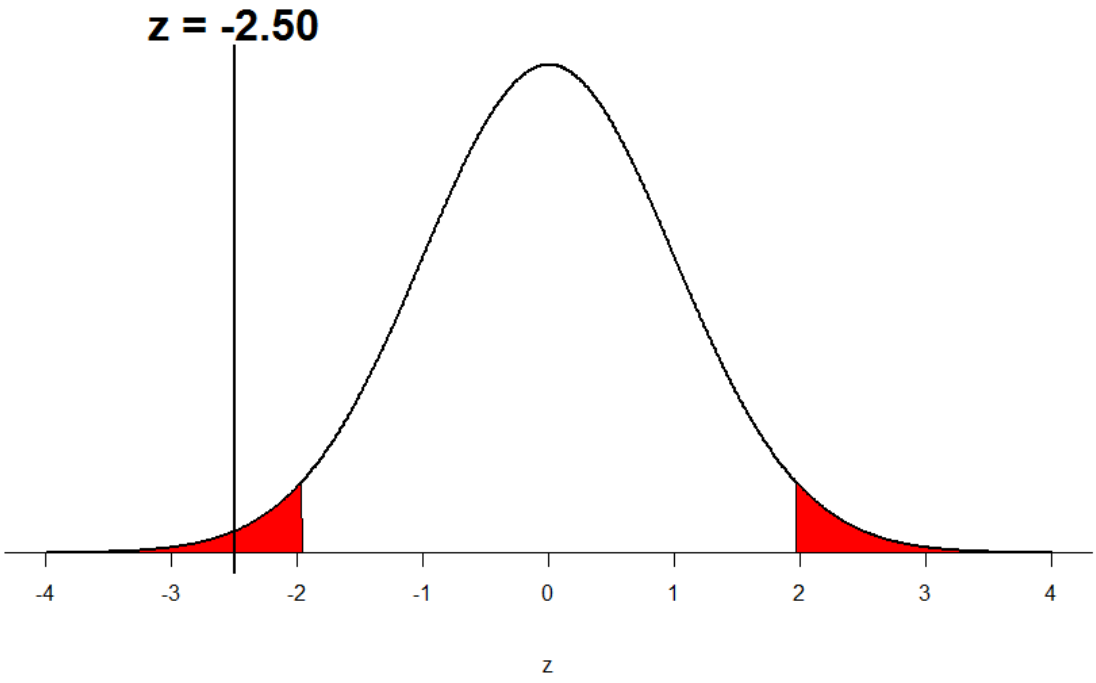


Figure 5: Test statistic location

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*:

$$d = \frac{\bar{X} - \mu}{\sigma}$$

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:

$$d = \frac{7.75 - 8.00}{0.50} = \frac{-0.25}{0.50} = 0.50$$

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.

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 $z^* = 1.645$. To keep track of the directionality of the test and rejection region, you draw out your distribution:

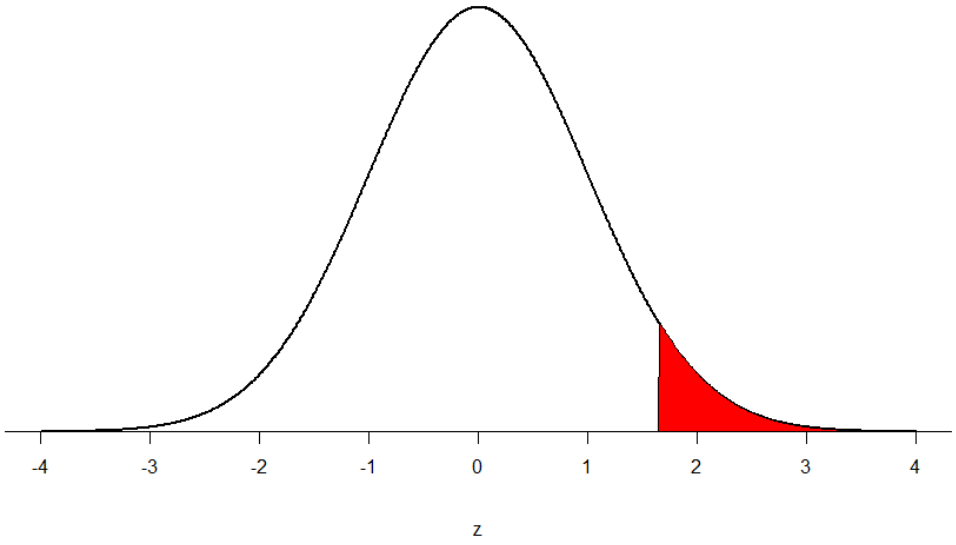


Figure 6: Rejection region

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 μ = 74 (the supposed average temperature), σ = 1.00 (how much the temperature should vary), and n = 5 (how many data points you collected):

$$z = \frac{76.60 - 74.00}{1.00/\sqrt{5}} = \frac{2.60}{0.45} = 5.78$$

This value falls so far into the tail that it cannot even be plotted on the distribution!

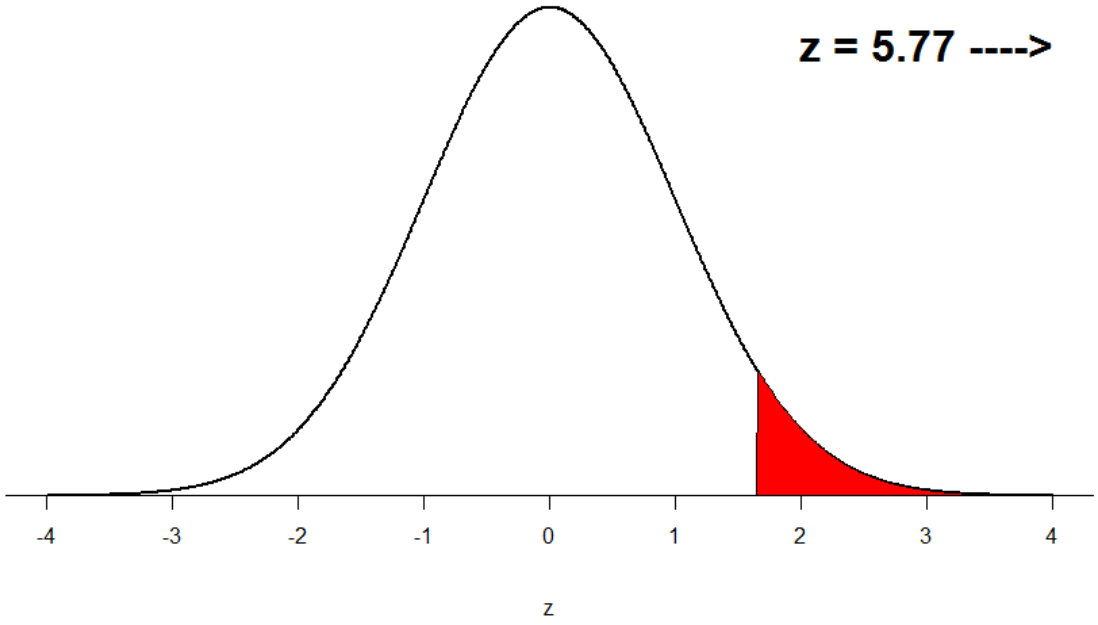


Figure 7: Obtained z-statistic

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 ($\bar{X}$ = 76.6 degrees) is statistically significantly higher than it is supposed to be, z = 5.77, p < .05.

Because the result is significant, you also calculate an effect size:

$$d = \frac{76.60 - 74.00}{1.00} = \frac{2.60}{1.00} = 2.60$$

The effect size you calculate is definitely large, meaning someone has some explaining to do!

Example: Different Significance Level

Finally, 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, α = 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 α = 0.05 levels of significance several times. To find the values for α = 0.01, we will go to the standard normal table and find the z-score cutting of 0.005 (0.01 divided by 2 for a two-tailed test) of the area in the tail, which is z* = ±2.575. Notice that this cutoff is much higher than it was for α = 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. We will use σ = 10 as our known population standard deviation and the following data to calculate our sample mean:

61	62
65	61
58	59
54	61
60	63

The average of these scores is $\bar{X}$ = 60.40. From this 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 ($\bar{X}$ = 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.

Finally, because we did not detect a statistically significant effect, we do not need to calculate an effect size.

Other Considerations in Hypothesis Testing

There are several other considerations we need to keep in mind when performing hypothesis testing.

Errors in Hypothesis Testing

In the Physicians' Reactions case study, the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a Type I error. More generally, a Type I error occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05 then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α level or simply α. It is also called the significance level. As discussed in the introduction to hypothesis testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a reject or do-not-reject decision. Therefore, 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- β and is called power. 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).

Misconceptions in Hypothesis Testing

Misconceptions about significance testing are common. This section lists three important ones.

1. Misconception: The probability value is the probability that the null hypothesis is false.
Proper interpretation: 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. Misconception: A low probability value indicates a large effect.
Proper interpretation: A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true. A low probability value can occur with small effect sizes, particularly if the sample size is large.

3. Misconception: A non-significant outcome means that the null hypothesis is probably true.
Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

Exercises – Ch. 7

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 "H0: μ = 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?

8. Determine whether you would reject or fail to reject the null hypothesis in the following situations:
    a. $z = 1.99$, two-tailed test at $\alpha = 0.05$
    b. $z = 0.34$, $z^* = 1.645$
    c. $p = 0.03$, $\alpha = 0.05$
    d. $p = 0.015$, $\alpha = 0.01$
9. 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.
10. 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. 7

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. $H_A: \mu \neq 40$, $H_A: \mu > 40$, $H_A: \mu < 40$
7. We calculate an effect size when we find a statistically significant result to see if our result is practically meaningful or important
9. 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: $\bar{X} = 88.75$, $\sigma_{\bar{X}} = 4.24$, $z = 2.54$. Step 4: $z > z^*$, Reject H_0. Based on 8 weeks of games, we can conclude that our average score ($\bar{X} = 88.75$) is higher now that the new person is on the team, $z = 2.54$, $p < .05$. Since the result is significant, we need an effect size: Cohen's $d = 0.90$, which is a large effect.

Chapter 8: Introduction to *t*-tests

Last chapter 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.

The *t*-statistic

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 populations standard deviation, σ, which is rarely going to be the case. Instead, we will estimate that parameter σ using the sample statistic s in the same way that we estimate μ using $\overline{X}$ (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:

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

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 $s_{\overline{X}}$ to indicate that we use the sample statistic for standard deviation, *s*, instead of the population parameter σ. The process of using and interpreting the standard error and the full test statistic remain exactly the same.

In chapter 3 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*. 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. Two things should stand out: First, for lower degrees of freedom (e.g. 2), the tails of the distribution are much fatter, meaning the 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 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.

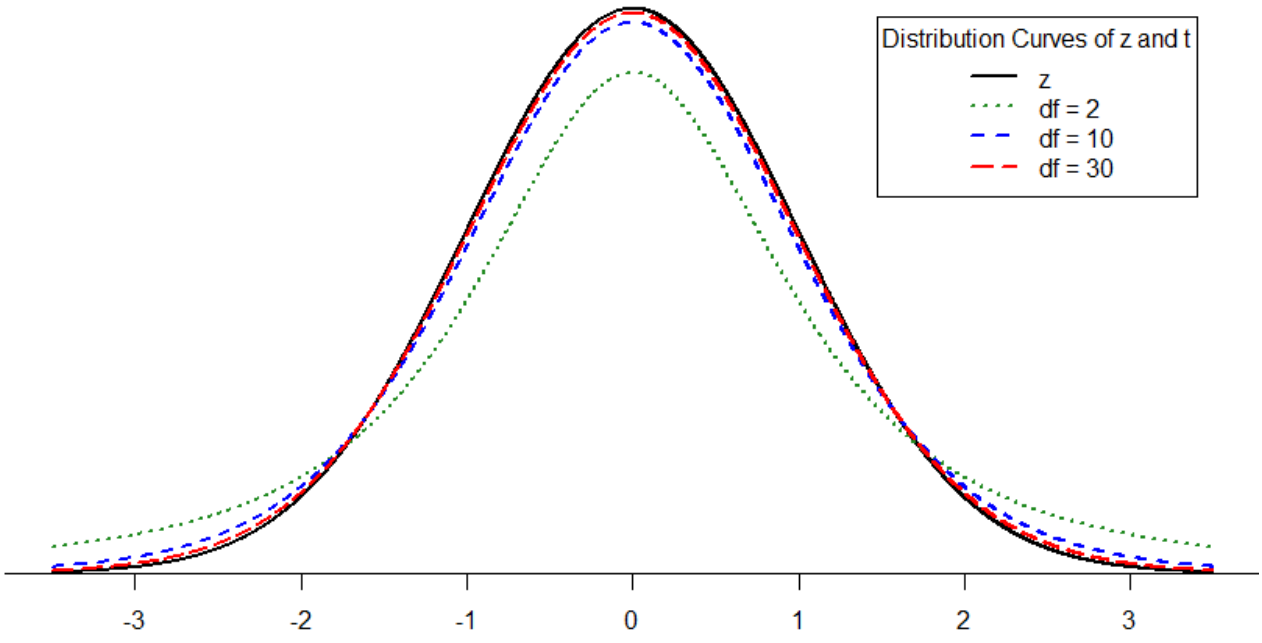


Figure 1. Distributions comparing effects of degrees of freedom

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 7) 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. We will work though an example: let's say that you move to a new city and find a 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. For our example:

H_0: There is no difference in the average time to change a car's oil
$H_0: \mu = 30$

H_A: This shop takes longer to change oil than your old mechanic
$H_A: \mu > 30$

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 rejection region

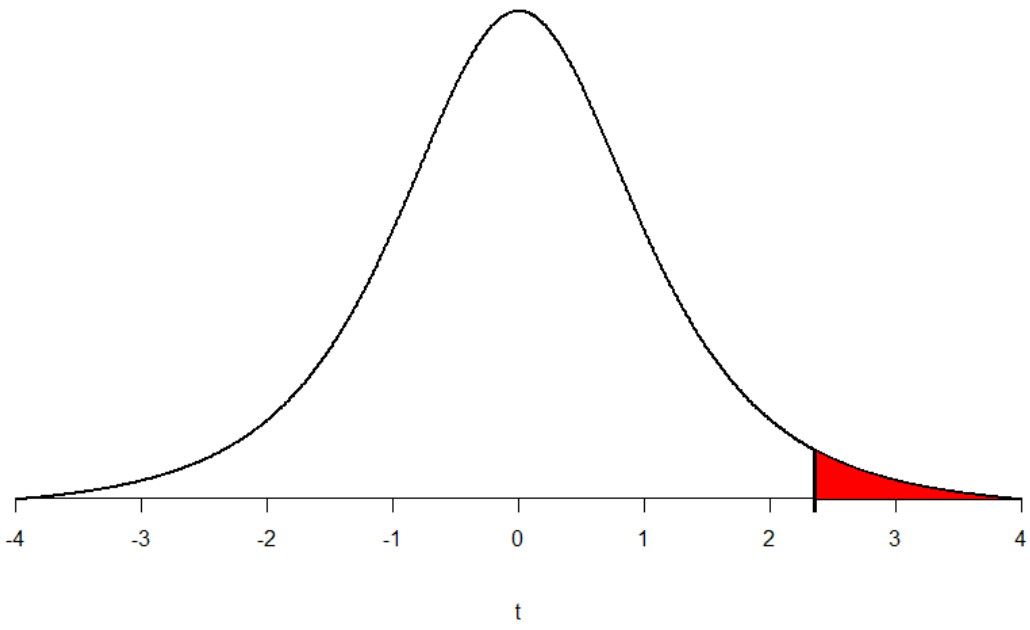


Figure 3. Rejection Region

Step 3: Compute the Test Statistic

The four wait times you experienced for your oil changes are 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:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{N - 1}} = \sqrt{\frac{SS}{df}} = \sqrt{\frac{564.74}{3}} = 13.72$$

Next, we take this value and plug it in to the formula for standard error:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{13.72}{2} = 6.86$$

And, finally, we put the standard error, sample mean, and null hypothesis value into the formula for our test statistic *t*:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{53.75 - 30}{6.86} = \frac{23.75}{6.68} = 3.46$$

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 everything in correctly.

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*. And because we found a significant result, we need to calculate an effect size, which is still Cohen's *d*, but now we use *s* in place of σ:

$$d = \frac{\bar{X} - \mu}{s} = \frac{53.75 - 30.00}{13.72} = 1.73$$

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:

$$Margin\ of\ Error = t^* \left({}^{s}/_{\sqrt{n}} \right)$$

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:

$$Upper\ Bound = \bar{X} + Margin\ of\ Error$$
$$Lower\ Bound = \bar{X} - Margin\ of\ Error$$

Or simply:

$$Confidence\ Interval = \bar{X} \pm t^* \left({}^{s}/_{\sqrt{n}}\right)$$

To write out a confidence interval, we always use soft brackets and put the lower bound, a comma, and the upper bound:

$$Confidence\ Interval = (Lower\ Bound, Upper\ Bound)$$

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 α = 0.05 is the same as a two-tailed α = 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 α = 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:

$$95\%\ CI = 53.75 \pm 3.182(6.86)$$

$$Upper\ Bound = 53.75 + 3.182(6.86)$$
$$UB = 53.75 + 21.83$$
$$UB = 75.58$$

$$Lower\ Bound = 53.75 - 3.182(6.86)$$
$$LB = 53.75 - 21.83$$
$$LB = 31.92$$

$$95\%\ CI = (31.92, 75.58)$$

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, $\overline{X} = 53.75$, 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.

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, 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. 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: $\mu \neq 38$

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:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{5.61}{5.48} = 1.02$$

Now we can put that value, our point estimate for the sample mean, and our critical value from step 2 into the formula for a confidence interval:

$$95\%\ CI = \ 39.85 \pm 2.045(1.02)$$

$$Upper\ Bound = \ 39.85 + 2.045(1.02)$$
$$UB = \ 39.85 + 2.09$$
$$UB = \ 41.94$$

$$Lower\ Bound = 39.85 - 2.045(1.02)$$
$$LB = 39.85 - 2.09$$
$$LB = 37.76$$

$$95\%\ CI = (37.76, 41.94)$$

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.

Confidence Intervals using 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 $s_{\bar{X}}$ with $\sigma_{\bar{X}}$.

Exercises – Ch. 8

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 $\overline{X} = 25$ for the following conditions:
    a. N = 25, s = 15, 95% confidence level
    b. N = 25, s = 15, 90% confidence level
    c. $s_{\overline{X}} = 4.5$, $\alpha = 0.05$, df = 20
    d. 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.

Year	Temp	Year	Temp	Year	Temp	Year	Temp
1981	9.301	1991	9.336	2001	9.542	2011	9.65
1982	8.788	1992	8.974	2002	9.695	2012	9.635
1983	9.173	1993	9.008	2003	9.649	2013	9.753
1984	8.824	1994	9.175	2004	9.451	2014	9.714
1985	8.799	1995	9.484	2005	9.829	2015	9.962
1986	8.985	1996	9.168	2006	9.662	2016	10.16
1987	9.141	1997	9.326	2007	9.876	2017	10.049
1988	9.345	1998	9.66	2008	9.581		
1989	9.076	1999	9.406	2009	9.657		
1990	9.378	2000	9.332	2010	9.828		

8. Determine whether you would reject or fail to reject the null hypothesis in the following situations:
    a. t = 2.58, N = 21, two-tailed test at $\alpha = 0.05$
    b. t = 1.99, N = 49, one-tailed test at $\alpha = 0.01$
    c. $\mu = 47.82$, 99% CI = (48.71, 49.28)
    d. $\mu = 0$, 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 $\overline{X}$ = 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 $\overline{X}$ = 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. 8

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. $\overline{X}$ = 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: $\overline{X}$ = 1.10, $s_{\overline{X}}$ = 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 ($\overline{X}$ = 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.

Chapter 9: Repeated Measures

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 time points. 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. The absolute value of our measurements does not matter – all that matters is the change. 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 1, we calculate our difference scores by taking the time 2 score and subtracting the time 1 score. That is:

$$X_d = X_{T2} - X_{T1}$$

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, will be the data we use to test for improvement or 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. 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.

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.

Hypotheses of Change and Differences

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:

$$H_0: \text{There is no change or difference}$$
$$H_0: \mu_D = 0$$

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
$H_A: \mu_D < 0$

As before, you 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.

Critical Values and Decision Criteria

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 $t_{obt} > t^*$.

Test Statistic

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.

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.

$$\overline{X_D} = \frac{\sum X_D}{n}$$

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:

$$s_D = \sqrt{\frac{\sum(X_D - \overline{X_D})^2}{n-1}} = \sqrt{\frac{SS}{df}}$$

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:

$$s_{\overline{X_D}} = {s_D}/{\sqrt{n}}$$

Finally, our test statistic *t* has the same structure as well:

$$t = \frac{\overline{X_D} - \mu_D}{s_{\overline{X_D}}}$$

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 40 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

First, we state our null and alternative hypotheses:

H_0: There is no change in average job satisfaction
H_0: $\mu_D = 0$

H_A: There is an increase in average job satisfaction
H_A: $\mu_D > 0$

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. Thus, we will use a one-directional alternative hypothesis.

Step 2: Find the Critical Values

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 one-tailed to the right, 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 40 people, so our degrees of freedom are 39. Going to our *t*-table, we find that the critical value is $t^* = 1.685$ as shown in Figure 1.

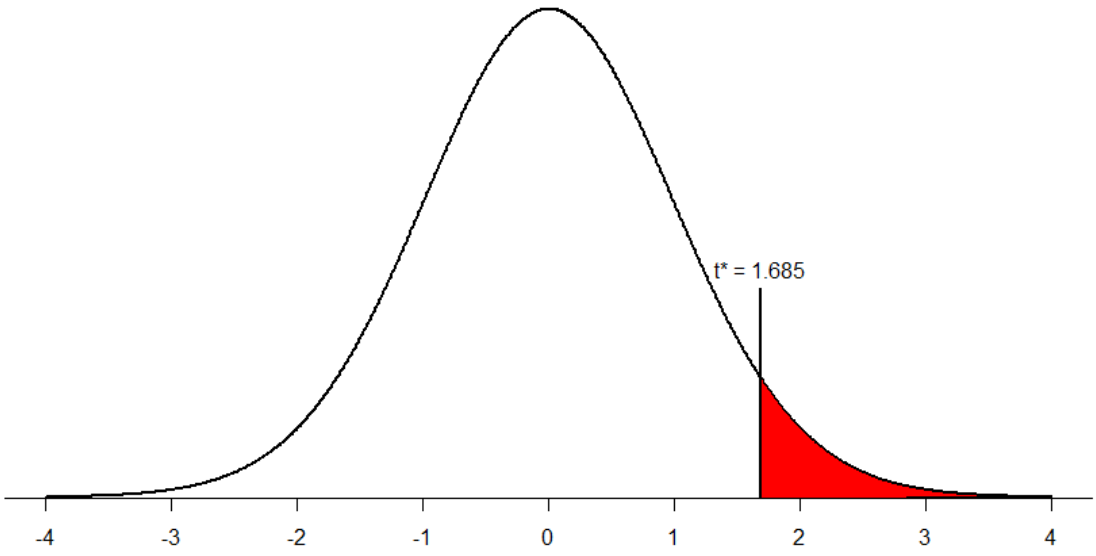


Figure 1. Critical region for one-tailed *t*-test at $\alpha = 0.05$

Step 3: Calculate the 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 $\overline{X_D} = 2.96$ and a standard deviation of $s_D = 2.85$. Using this information, plus the size of the sample (N = 40), we first calculate the standard error:

$$s_{\overline{X_D}} = {}^{s_D}/_{\sqrt{n}} = {}^{2.85}/_{\sqrt{40}} = {}^{2.85}/_{6.32} = 0.46$$

Now, we can put that value, along with our sample mean and null hypothesis value, into the formula for *t* and calculate the test statistic:

$$t = \frac{\overline{X_D} - \mu_D}{s_{\overline{X_D}}} = \frac{2.96 - 0}{0.46} = 6.43$$

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 the Decision

We have obtained a test statistic of $t = 6.43$ that we can compare to our previously established critical value of $t^* = 1.685$. 6.43 is larger than 1.685, so $t > t^*$ and we reject the null hypothesis:

> Reject H_0. Based on the sample data from 40 workers, we can say that the intervention statistically significantly improved job satisfaction ($\overline{X_D} = 2.96$) among the workers, $t(39) = 6.43$, $p < 0.05$.

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:

$$t = \frac{\overline{X_D} - \mu_D}{s_D} = \frac{2.96}{2.85} = 1.04$$

This is a large effect size. Notice again that we can omit the null hypothesis value here because it is always equal to 0.

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, which was just a small change in how *z*-tests were performed in chapter 7. 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.

Example: 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 $\alpha = 0.10$. 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:

H_0: There is no change in how people view the bank
$H_0: \mu_D = 0$

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 $t^* = 1.943$.

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 are calculated as Time 2 – Time 1.

Time 1	Time 2	X_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:

$$\overline{X_D} = \frac{\sum X_D}{n} = \frac{4}{7} = 0.57$$

The standard deviation will be solved by first using the Sum of Squares Table:

X_D	$X_D - \overline{X_D}$	$(X_D - \overline{X_D})^2$
-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
$\Sigma = 4$	$\Sigma = 0$	$\Sigma = 65.68$

$$s_D = \sqrt{\frac{SS}{df}} = \sqrt{\frac{65.68}{6}} = \sqrt{10.95} = 3.31$$

Finally, we find the standard error:

$$s_{\overline{X_D}} = {}^{s_D}/_{\sqrt{n}} = {}^{3.31}/_{\sqrt{7}} = 1.25$$

We now have all the pieces needed to compute our confidence interval:

$$95\%\ CI = \ \overline{X_D} \pm t^*\left(s_{\overline{X_D}}\right)$$
$$95\%\ CI = \ \ 0.57 \pm 1.943(1.25)$$

$$Upper\ Bound = \ 0.57 + 1.943(1.25)$$
$$UB = \ 0.57 + 2.43$$
$$UB = \ 3.00$$

$$Lower\ Bound = \ 0.57 - 1.943(1.25)$$
$$LB = \ 0.57 - 2.43$$
$$LB = \ -1.86$$

$$95\%\ CI = (-1.86, 3.00)$$

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 ($\overline{X_D}$ = 0.57) was any better or worse after viewing the commercial, CI: (-1.86, 3.00).

As with before, we only report the confidence interval to indicate how we performed the test.

Exercises – Ch. 9

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 $\overline{X_D}$ = 12.10 with a standard deviation of s_D = 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:
    a. $\overline{X_D}$ = 3.50, s_D = 1.10, n = 12, $\alpha = 0.05$, two-tailed test
    b. 95% $CI = (0.20, 1.85)$
    c. t = 2.98, t^* = -2.36, one-tailed test to the left
    d. 90% CI = (-1.12, 4.36)
7. Calculate difference scores for the following data:

Time 1	Time 2	X_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 $\overline{X_D}$ = -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 of $\overline{X_D}$ = 1.25, standard error of $s_{\overline{X_D}}$ = 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	89	
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. 9

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: $\overline{X_D} = 12.10$, $s_{\overline{X_D}} = 2.94$, t = 4.12. Step 4: t > t*, Reject H_0. Based on opinions from 35 people, we can conclude that people trust algorithms more ($\overline{X_D} = 12.10$) 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.

Time 1	Time 2	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.

Chapter 10: 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. Now, we will learn how to compare two separate means from 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 repeated measures/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

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 therefore 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 that 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).

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.

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:

$$H_0: \mu_1 = \mu_2$$

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:

$$H_0: \textit{There is no difference between the means of the two groups}$$

Our alternative hypotheses are also unchanged: we simply replace the equal sign (=) with one of the three inequalities (>, <, ≠):

$$H_A: \mu_1 > \mu_2$$
$$H_A: \mu_1 < \mu_2$$
$$H_A: \mu_1 \neq \mu_2$$

Or

$$H_A: \mu_1 - \mu_2 > 0$$
$$H_A: \mu_1 - \mu_2 < 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

Whichever formulation you chose for the null hypothesis should be the one you use for the alternative hypothesis (be consistent), and the interpretation of them is always the same:

$$H_A: \textit{There is a difference between the means of the two groups}$$

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 and control groups

$$H_0: \mu_{treatment} = \mu_{control}$$

H_A: There is a difference between the means of the treatment and control groups

$$H_A: \mu_{treatment} \neq \mu_{control}$$

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, which for an independent samples *t*-test are:

$$df = n_1 + n_2 - 2$$

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. What is of greater concern is keeping track of which is which using the subscripts.

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:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{s_{\overline{X_1} - \overline{X_2}}}$$

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:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{s_{\overline{X_1} - \overline{X_2}}}$$

Our standard error in the denomination is still standard deviation (s) with a subscript denoting what it is the standard error of. 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:

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

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:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

Using this 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. Thus, when we use this formula, the pooled variance is not nearly as intimidating as it might have originally seemed.

Once we have our pooled variance calculated, we can drop it into the equation for our standard error:

$$s_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

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. An alternative but mathematically equivalent way of writing our old standard error is:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

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

$H_0: \mu_1 - \mu_2 = 0$

or

$H_0: \mu_1 = \mu_2$

H_A: The comedy film will give a better average mood than the horror film

$H_A: \mu_1 - \mu_2 > 0$

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 out *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 that because 62 does not appear on the 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 values for the Comedy group, and Table 2 shows the values for the Horror group. Values for both have already been placed in the Sum of Squares tables since we will need to use them for our further calculations. As always, the column on the left is our raw data.

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
$\Sigma = 840$	$\Sigma = 0$	Σ=5061.60

Table 1. Raw scores and Sum of Squares for Group 1

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
$\Sigma = 478.6$	$\Sigma = 0.10$	Σ=3896.45

Table 2. Raw scores and Sum of Squares for Group 1.

Using the sum of the first column for each table, we can calculate the mean for each group:

$$\overline{X_1} = \frac{840}{35} = 24.00$$

And

$$\overline{X_2} = \frac{478.60}{29} = 16.50$$

These values were used to calculate the middle rows of each table, which sum to zero as they should (the middle column for group 2 sums to a very small value instead of zero due to rounding error – the exact mean is 16.50344827586207, but that's far more than we need for our purposes). Squaring each of the deviation scores in the middle columns gives us the values in the third columns, which sum to our next important value: 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 – no new computations yet. 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:

$$s_1 = \sqrt{\frac{5061.60}{34}} = 12.20$$

And

$$s_2 = \sqrt{\frac{3896.45}{28}} = 11.80$$

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:

$$s_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{144.48}{35} + \frac{144.48}{29}} = \sqrt{4.13 + 4.98} = \sqrt{9.11} = 3.02$$

Finally, we can use our standard error and the means we calculated earlier to compute our test statistic. Because the null hypothesis value of $\mu_1 - \mu_2$ is 0.00, we will leave that portion out of the equation for simplicity:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{s_{\overline{X_1}-\overline{X_2}}} = \frac{24.00 - 16.50}{3.02} = 2.48$$

The process of calculating our obtained test statistic $t = 2.48$ followed the same sequence of steps as before: use raw data to compute the mean and sum of squares (this time for two groups instead of one), use the sum of squares and degrees of freedom to calculate standard error (this time using pooled variance instead of standard deviation), and use that standard error and the observed means to get *t*. 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 ($\overline{X_1} = 24.00$) is better than the average mood after a horror movie ($\overline{X_2} = 16.50$), $t(62) = 2.48$, $p < .05$.

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*. This gives us:

$$d = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{s_p^2}}$$

For our example above, we can calculate the effect size to be:

$$d = \frac{24.00 - 16.50}{\sqrt{144.48}} = \frac{7.50}{12.02} = 0.62$$

We interpret this using the same guidelines as before, so we would consider this a moderate or moderately large effect.

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 xbar1 minus xbar2. 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. That is:

$$Confidence\ Interval = (\overline{X_1} - \overline{X_2}) \pm t^*\left(s_{\overline{X_1} - \overline{X_2}}\right)$$

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). 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 $\overline{X_1}$=28.65 s_1 = 12.40 n_1 = 40 and $\overline{X_2}$ = 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, and the key lies in rearranging the formulas to work backwards through them. We need the Sum of Squares and degrees of freedom to calculate our pooled variance. Degrees of freedom is very simple: we just take the sample size minus 1.00 for each group. Getting the Sum of Squares is also easy: remember that variance is standard deviation squared and is the Sum of Squares divided by the degrees of freedom. That is:

$$s^2 = (s)^2 = \frac{SS}{df}$$

To get the Sum of Squares, we just multiply both sides of the above equation to get:

$$s^2 * df = SS$$

Which is the squared standard deviation multiplied by the degrees of freedom (n – 1) equals the Sum of Squares.

Using our example data:

$$(s_1)^2 * df_1 = SS_1$$
$$(12.40)^2 * (40 - 1) = 5996.64$$

$$(s_2)^2 * df_2 = SS_2$$
$$(15.68)^2 * (42 - 1) = 10080.36$$

And thus our pooled variance equals:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{5996.64 + 10080.36}{39 + 41} = \frac{16077}{80} = 200.96$$

And our standard error equals:

$$s_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{200.96}{40} + \frac{200.96}{42}} = \sqrt{5.02 + 4.78} = \sqrt{9.89} = 3.14$$

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 out standard error, it's time to build our confidence interval.

$$95\%\ CI = \ 3.25 \pm 1.990(3.14)$$

$$Upper\ Bound = \ 3.25 + 1.990(3.14)$$
$$UB = \ 3.25 + 6.25$$
$$UB = \ 9.50$$

$$Lower\ Bound = \ 3.25 - 1.990(3.14)$$
$$LB = \ 3.25 - 6.25$$
$$LB = \ -3.00$$

$$95\%\ CI = (-3.00, 9.50)$$

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.

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.

Exercises – Ch. 10

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
    a. $s_1 = 24$, $s_2 = 21$, $n_1 = 36$, $n_2 = 49$
    b. $s_1 = 15.40$, $s_2 = 14.80$, $n_1 = 20$, $n_2 = 23$
    c. $s_1 = 12$, $s_2 = 10$, $n_1 = 25$, $n_2 = 25$
5. Determine whether to reject or fail to reject the null hypothesis in the following situations:
    a. $t(40) = 2.49$, $\alpha = 0.01$, one-tailed test to the right
    b. $\overline{X_1} = 64$, $\overline{X_2} = 54$, $n_1 = 14$, $n_2 = 12$, $s_{\overline{X_1}-\overline{X_2}} = 9.75$, $\alpha = 0.05$, two-tailed test
    c. 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.

   $\overline{X_1} = 16.25$, $\overline{X_2} = 20.95$, $s_1 = 4.00$, $s_2 = 5.10$, $n_1 = 29$, $n_2 = 25$

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?

   $\overline{X_1} = 3.78$, $\overline{X_2} = 2.23$, $s_1 = 2.60$, $s_2 = 1.15$, $n_1 = 45$, $n_2 = 40$

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.

    $\overline{X_H} = 47.50$, $\overline{X_A} = 45.70$, $s_H = 14.65$, $s_A = 13.20$, $n_H = 139$, $n_A = 150$

Answers to Odd- Numbered Exercises – Ch. 10

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. $SS_1 = 106.86$, $SS_2 = 78.86$, $s_p^2 = 15.48$
5. A) Reject B) Fail to Reject C) 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: $\overline{X_1} = 26.67$, $\overline{X_2} = 19.50$, $s_p^2 = 27.73$, $s_{\overline{X_1}-\overline{X_2}} =$, 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 ($\overline{X_1} = 26.67$) than night owls ($\overline{X_2} = 19.50$), 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. $\overline{X_1} - \overline{X_2} = 1.55$, t* = 1.990, $s_{\overline{X_1}-\overline{X_2}} = 0.45$, 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.

Unit 3 – 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. In this final unit, we will continue to use this same hypothesis testing logic and procedure on new types of data. We will start with group mean differences on more than two groups, then see how we can test hypotheses using only continuous data. We will wrap up this unit with a look at a different kind of test statistic: a non-parametric statistic for only categorical data.

Chapter 11: Analysis of Variance

Analysis of variance, often abbreviated to ANOVA for short, serves the same purpose as the *t*-tests we learned 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 (dependent 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.

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).

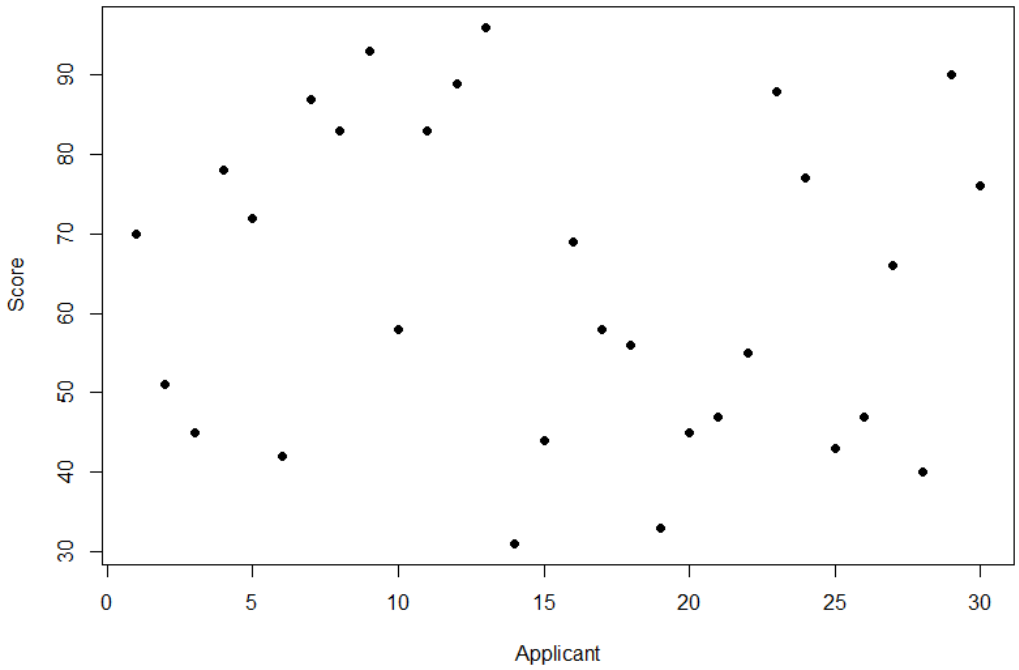


Figure 1. Scores on a job test

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.

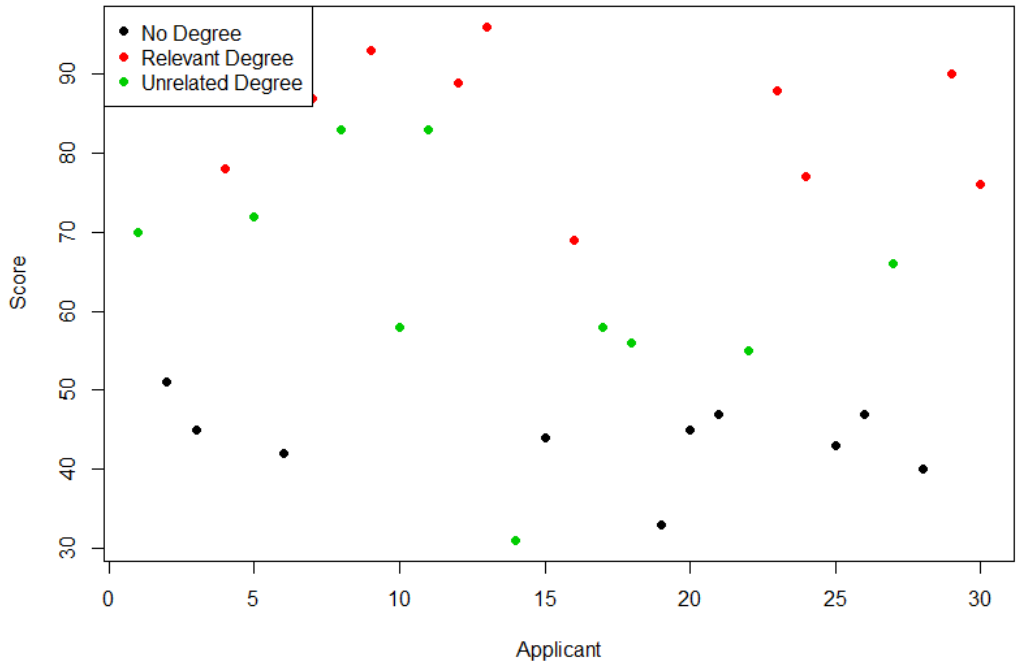


Figure 2. Applicant scores coded by degree earned

This pattern is even easier to see when the applicants are sorted and organized into their respective groups, as shown in Figure 3.

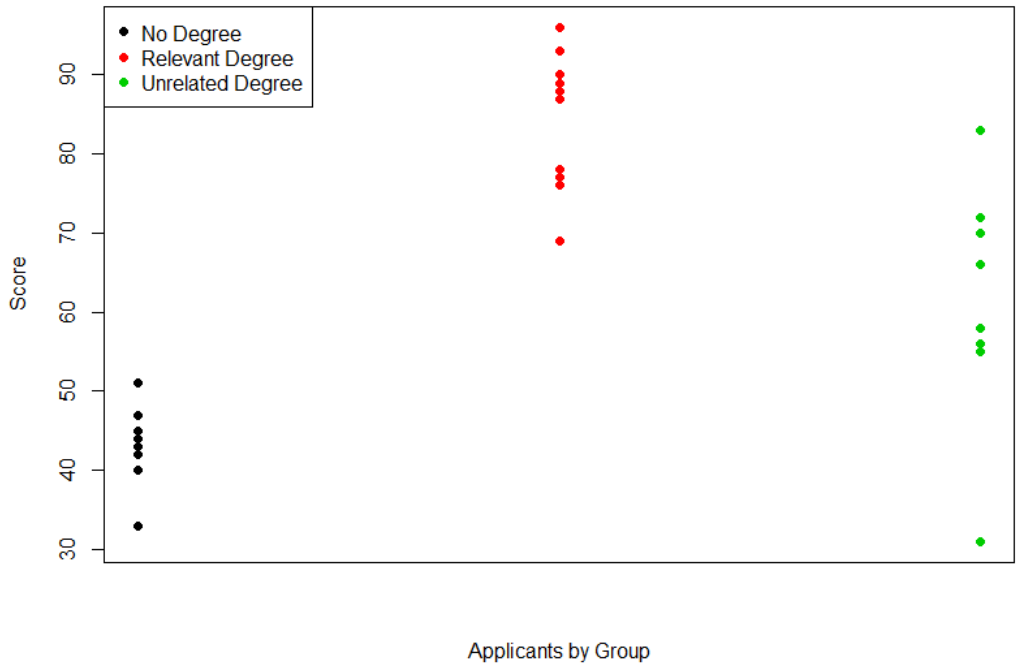


Figure 3. Applicant scores by group

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.

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 $\overline{X_1}$, $\overline{X_2}$, and $\overline{X_3}$. 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 $\overline{X_G}$. 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 is just the group sample sizes added together.

Between Groups Sum of Squares

One source of variability we can identified 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 multiple them by the number of people within that group. Incorporating this, we find our equation for Between Groups Sum of Squares to be:

$$SS_B = \sum n_j \left(\bar{X}_J - \overline{X_G}\right)^2$$

The subscript *j* refers to the "j^{th}" group where $j = 1 \ldots k$ 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:

$$SS_W = \sum \left(X_{ij} - \bar{X}_J\right)^2$$

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:

$$SS_T = \sum (X_i - \overline{X_G})^2$$

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	F
Between	SS_B	$k-1$	SS_B/df_B	MS_B/MS_W
Within	SS_W	$N-k$	SS_W/df_W	
Total	SS_T	$N-1$		

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, 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, 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 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.

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, 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:

$$H_0: There\ is\ no\ difference\ in\ the\ group\ means$$
$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

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:

$$H_A: At\ least\ one\ mean\ is\ different$$

The first difference in 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 in Unit 2. Due to this, our alternative hypothesis is always exactly the same: at least one mean is different.

In Unit 2, 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. 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. We will 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:

$$H_0: \textit{There is no difference between the means of the education groups}$$
$$H_0: \mu_1 = \mu_2 = \mu_3$$

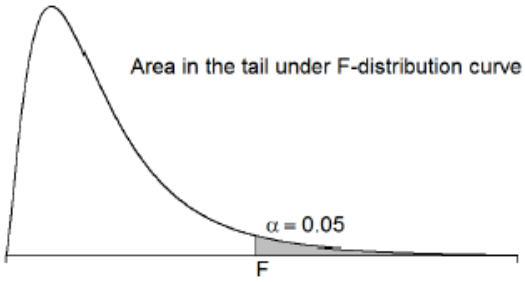
$$H_A: \textit{At least one mean is different}$$

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 Values

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:

F-Distribution Table



df	Degrees of Freedom: Numerator									
denom.	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	2.28
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.26
25	4.24	3.39	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20

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. We will assume that we are given the SS values as shown below:

Source	SS	df	MS	F
Between	8246			
Within	3020			
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.

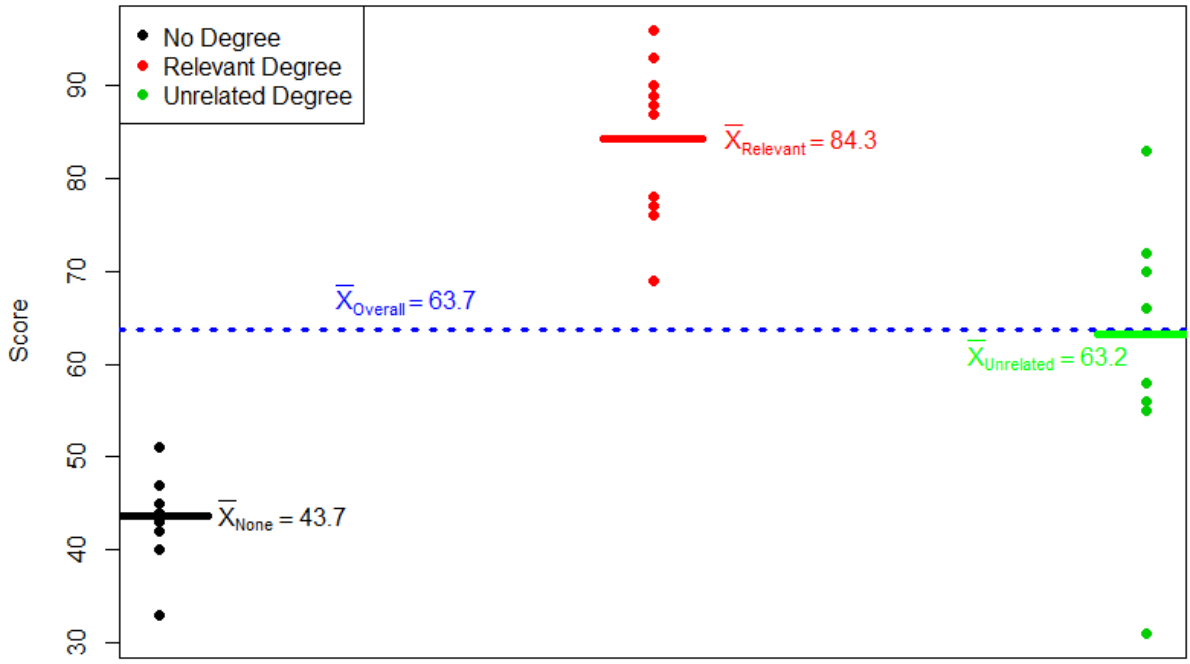


Figure 5. Means

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	df	MS	F
Between	8246			
Within	3020			
Total	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	MS	F
Between	8246	2		
Within	3020	27		
Total	11266	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	F
Between	8246	2	4123	
Within	3020	27	111.85	
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	8246	2	4123	36.86
Within	3020	27	111.85	
Total	11266	29		

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 the 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 H0. 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. Specifically:

$$\eta^2 = \frac{SS_B}{SS_T}$$

The effect size η^2 is called "eta-squared" and represents variance explained. For our example, our values give an effect size 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. Our guidelines for the size of our effects are:

η^2	Size
0.01	Small
0.09	Medium
0.25	Large

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. We will only focus here on the most commonly used ones. We will also only discuss the concepts behind each and will not worry about calculations.

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.

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:

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 in chapter 10: 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!

Exercises – Ch. 11

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:
   a. K = 4

Source	SS	df	MS	F
Between	87.40			
Within				
Total	199.22	33		

   b. N = 14

Source	SS	df	MS	F
Between		2	14.10	
Within				
Total	64.65			

   c.

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 xbar = $27.82, Store 2 xbar = $38.96, and Store 3 xbar = $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. 11

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:
   a. K = 4

Source	SS	df	MS	F
Between	87.40	3	29.13	7.81
Within	111.82	30	3.73	
Total	199.22	33		

   b. N = 14

Source	SS	df	MS	F
Between	28.20	2	14.10	4.26
Within	36.45	11	3.31	
Total	64.65	13		

   c.

Source	SS	df	MS	F
Between	210.10	2	105.05	42.36

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.

Chapter 12: Correlations

All of our analyses thus far have focused on comparing the value of a continuous variable across different groups via mean differences. We will now turn away from means and look instead at 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, but we will now have the ability to assess an entirely new data structure.

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 variance on a single variable:

$$s^2 = \frac{\sum(X - \bar{X})^2}{N - 1}$$

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 as:

$$s^2 = \frac{\sum\big((X - \bar{X})(X - \bar{X})\big)}{N - 1}$$

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:

$$cov_{XY} = \frac{\sum\left((X - \bar{X})(Y - \bar{Y})\right)}{N - 1}$$

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.

$$SP = \sum\left((X - \bar{X})(Y - \bar{Y})\right)$$

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	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$

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.

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 2 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 1 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.

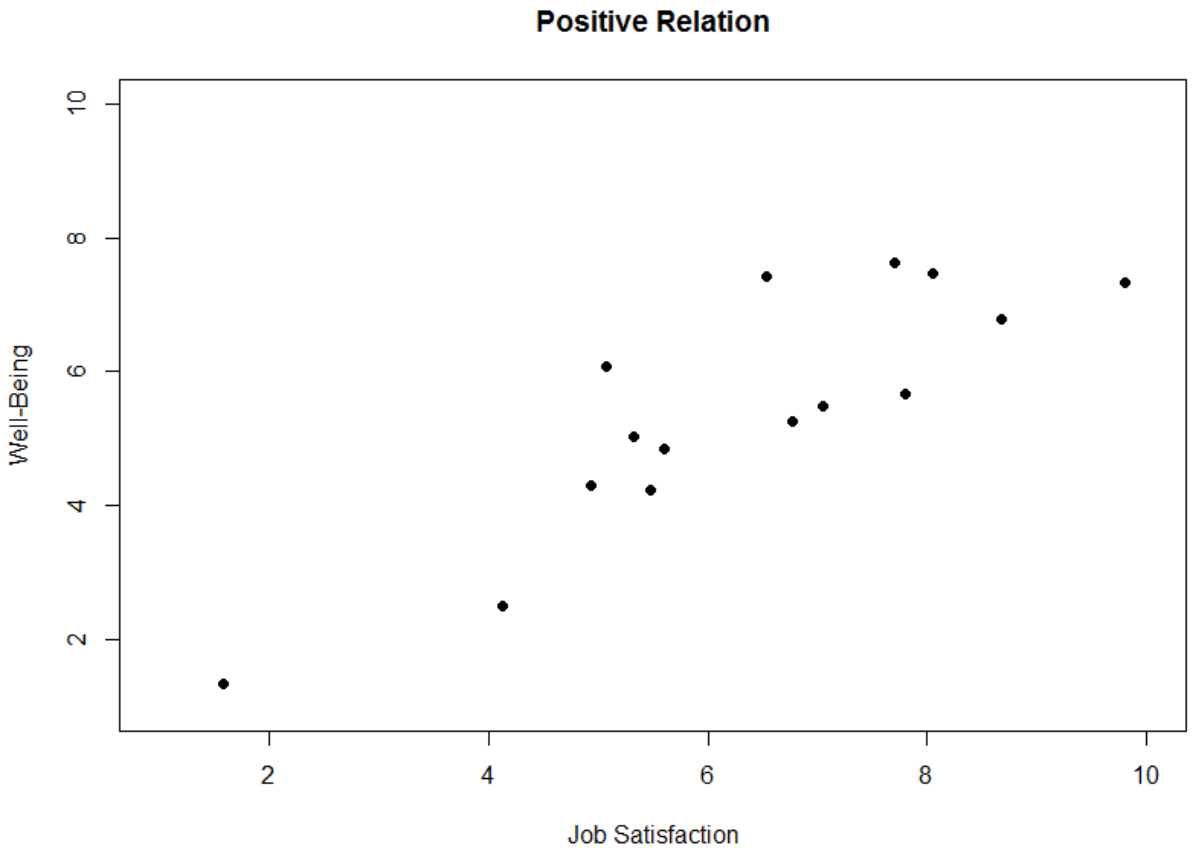


Figure 1. Plotting satisfaction and well-being scores.

Figure 1 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 2 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 1, 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.

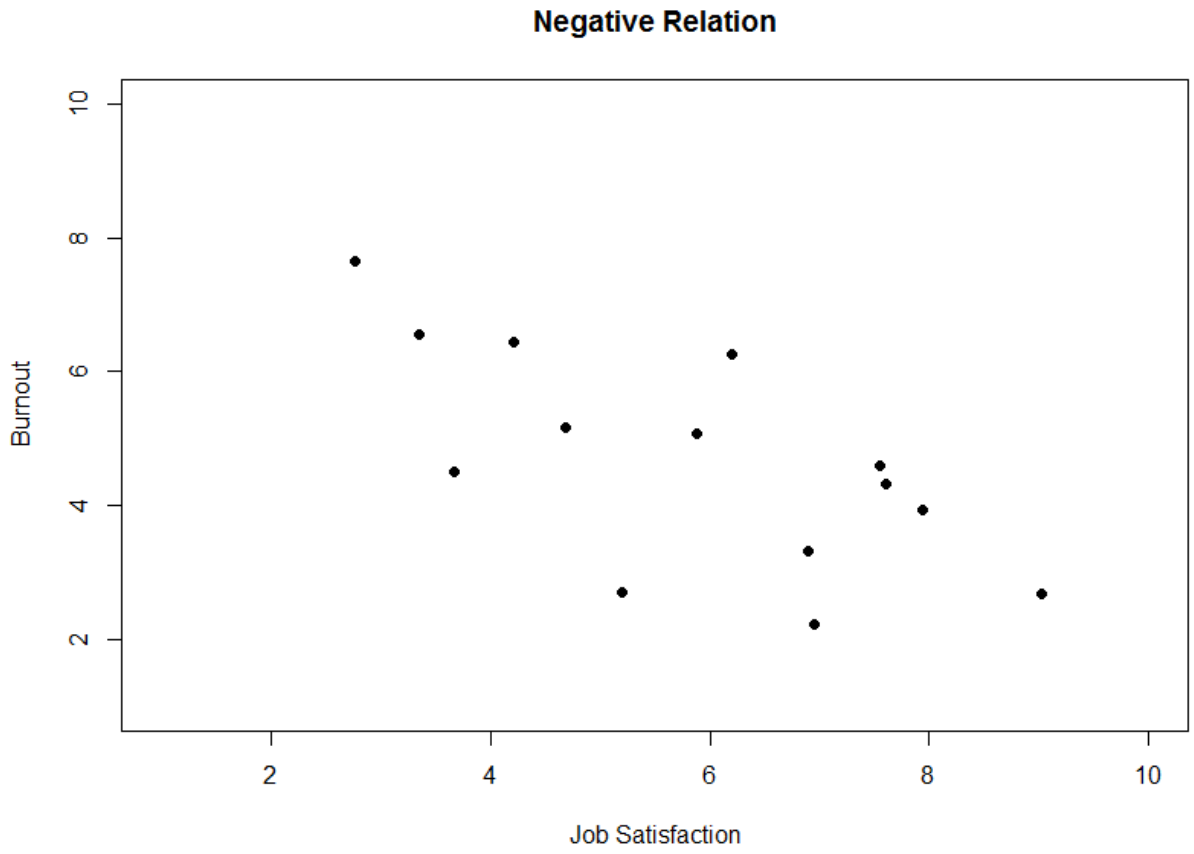


Figure 2. Plotting satisfaction and burnout scores.

Scatterplots can also indicate that there is no relation between the two variables. In these scatterplots (an example is shown below in figure 3 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.

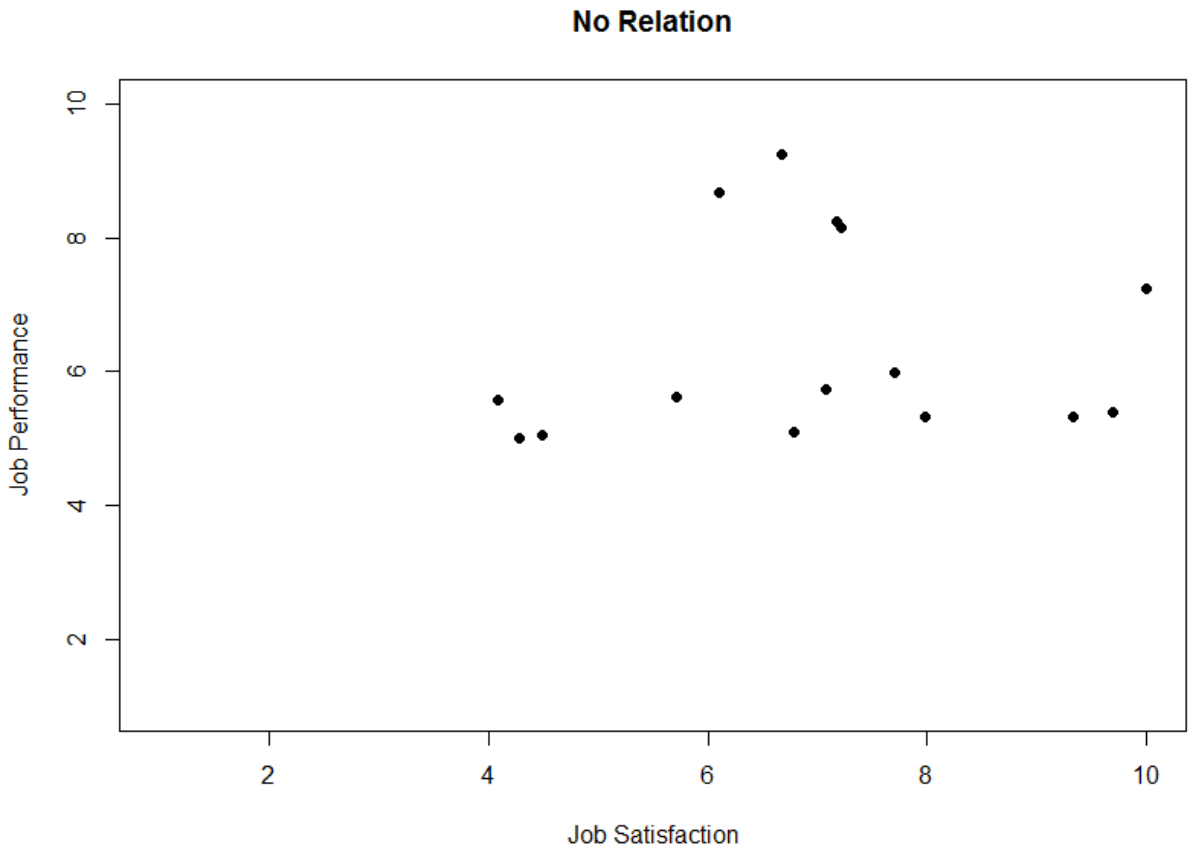


Figure 3. Plotting no relation between satisfaction and job performance.

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 4 and 5.

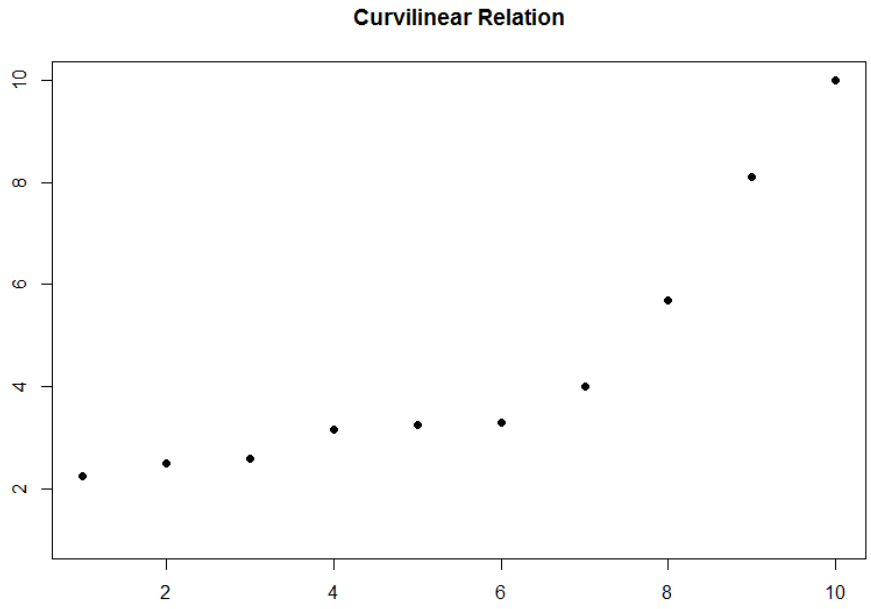


Figure 4. Exponentially increasing curvilinear relation

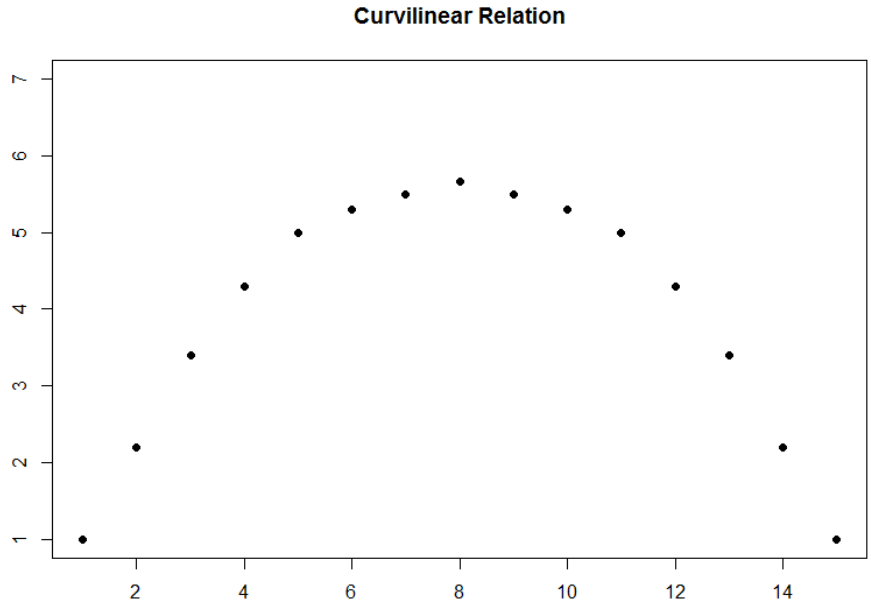


Figure 5. Inverted-U curvilinear relation.

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 6.

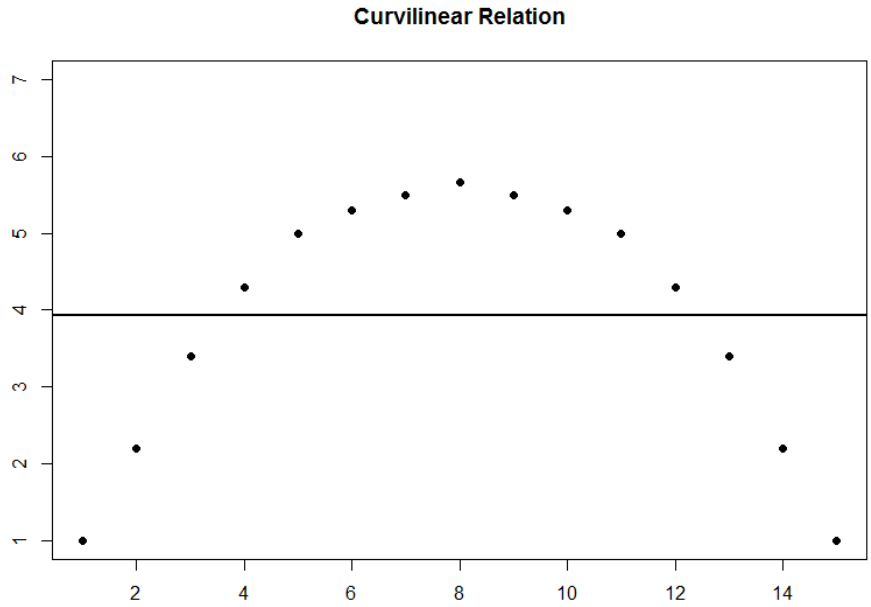


Figure 6. Overlaying a straight line on a curvilinear relation.

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 7. 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.

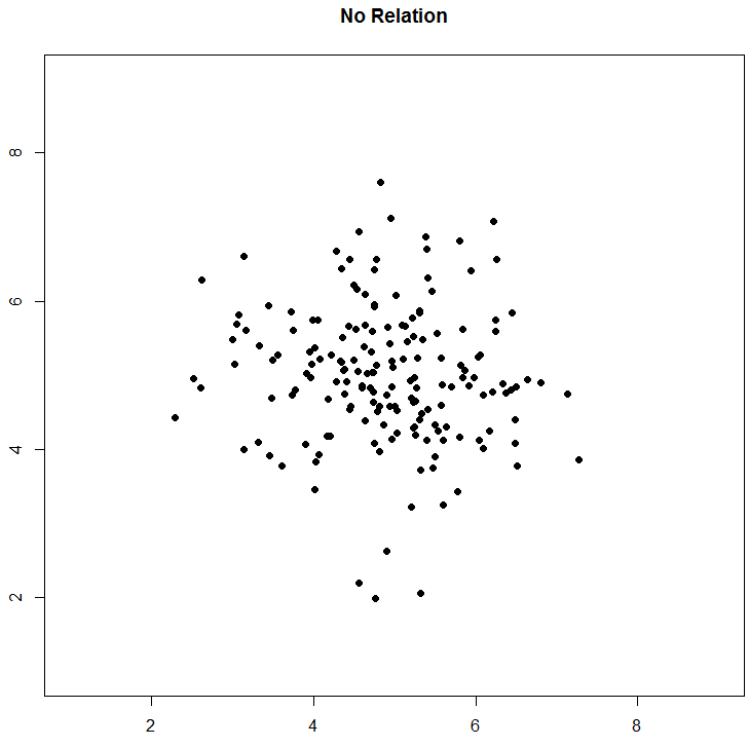


Figure 7. No relation

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 it 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 8 and 9, 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.

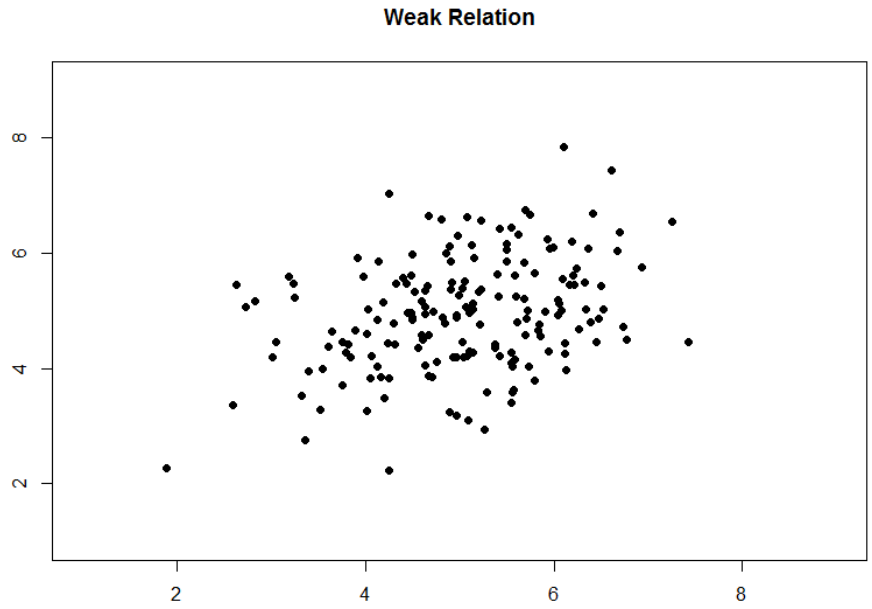


Figure 8. Weak positive correlation.

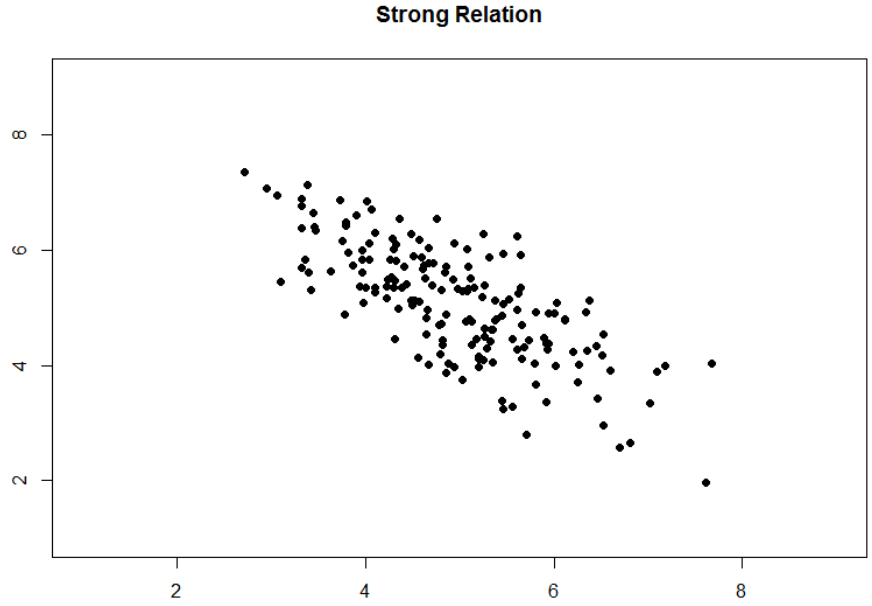


Figure 9. Strong negative correlation.

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 $\overline{X}$) 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 formula for *r* is very simple: it is just the covariance (defined above) divided by the standard deviations of X and Y:

$$r = \frac{cov_{XY}}{s_X s_Y} = \frac{SP}{\sqrt{SSX * SSY}}$$

The first formula gives a direct sense of what a correlation is: a covariance standardized onto the scale of X and Y; the second formula is computationally simpler and faster. Both of these equations will give the same value, and as we saw at the beginning of the chapter, all of these values are easily computed by using the sum of products table. 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.

It was stated earlier that *r* is a descriptive statistic like $\overline{X}$, and just like $\overline{X}$, 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 $\overline{X}$ 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.

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:

$$H_0: There\ is\ no\ relation\ between\ anxiety\ and\ depression$$
$$H_0: \rho = 0$$
$$H_A: There\ is\ a\ positive\ relation\ between\ anxiety\ and\ depression$$
$$H_0: \rho > 0$$

Remember that ρ ("rho") is our population parameter for the correlation that we estimate with *r*, just like $\overline{X}$ 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 10). 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). For our example, we have 10 people, so our degrees of freedom = 10 – 2 = 8.

Critical Values for Pearson's *r*

df	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.01	1-tailed α 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 10. 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 11 below shows our data plotted out, and it looks like they are, in fact, linearly related, so Pearson's *r* is appropriate.

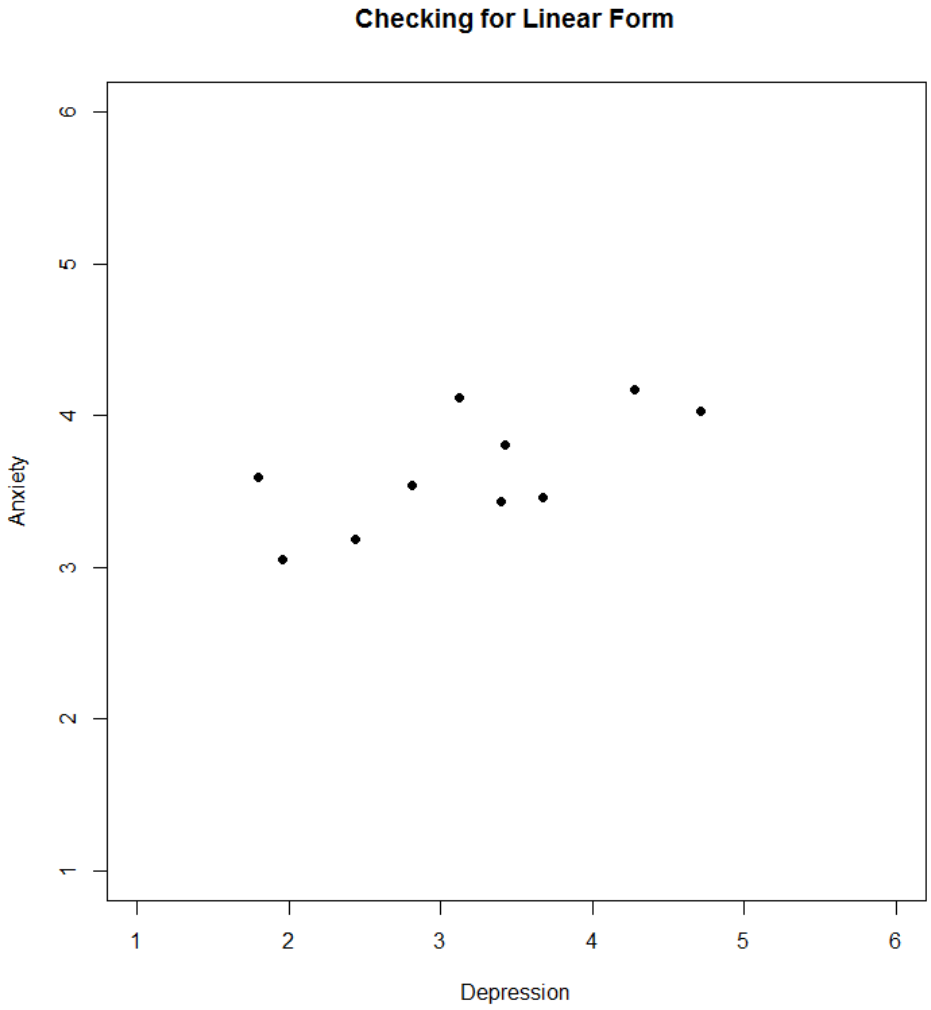


Figure 11. Scatterplot of anxiety and depression

The data we gather from our participants 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

We will need to put these values into our Sum of Products table to calculate the standard deviation and covariance of our variables. 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
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.

There are now three pieces of information we need to calculate before we compute our correlation coefficient: the covariance of X and Y and the standard deviation of each.

The covariance of two variable, remember, is the sum of products divided by N – 1. For our data:

$$cov_{XY} = \frac{SP}{N-1} = \frac{2.22}{9} = 0.25$$

The formula for standard deviation are the same as before. Using subscripts X and Y to denote depression and anxiety:

$$s_X = \sqrt{\frac{\sum (\mathrm{X} - \overline{\mathrm{X}})^2}{N-1}} = \sqrt{\frac{7.97}{9}} = 0.94$$

$$s_Y = \sqrt{\frac{\sum (\mathrm{Y} - \overline{\mathrm{Y}})^2}{N-1}} = \sqrt{\frac{1.33}{9}} = 0.38$$

Now we have all of the information we need to calculate *r*:

$$r = \frac{cov_{XY}}{s_X s_Y} = \frac{0.25}{0.94 * 0.38} = 0.70$$

We can verify this using our other formula, which is computationally shorter:

$$r = \frac{SP}{\sqrt{SSX * SSY}} = \frac{2.22}{\sqrt{7.97 * 1.33}} = .70$$

So our observed correlation between anxiety and depression is $r = 0.70$, 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 $r = 0.70$. 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 and 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.

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 12 and 13 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.

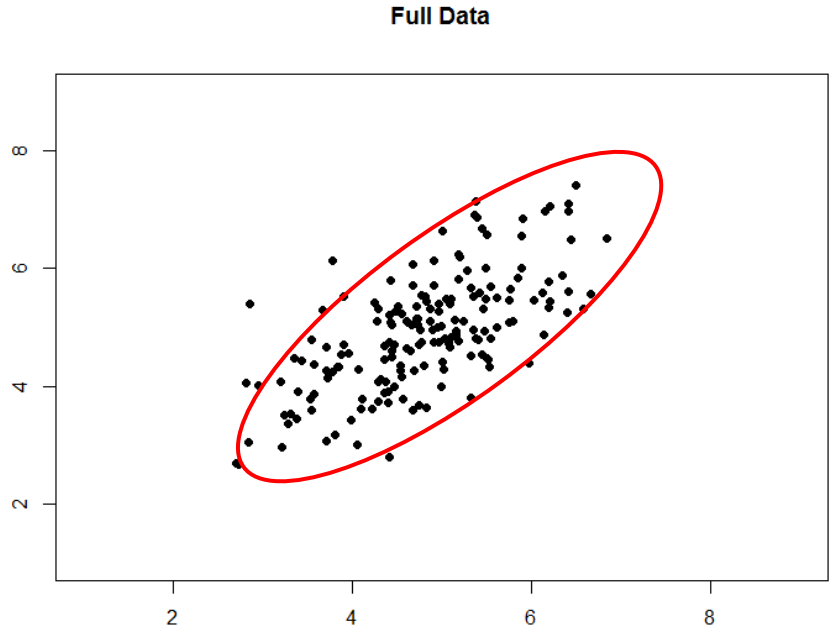


Figure 12. Strong, positive correlation.

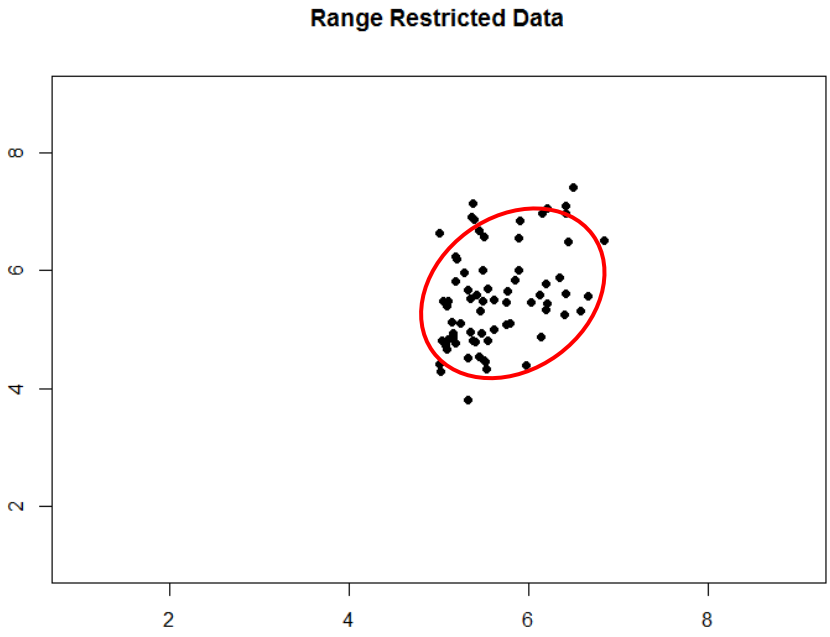


Figure 13. Effect of range restriction.

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).

The plots below in figure 14 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).

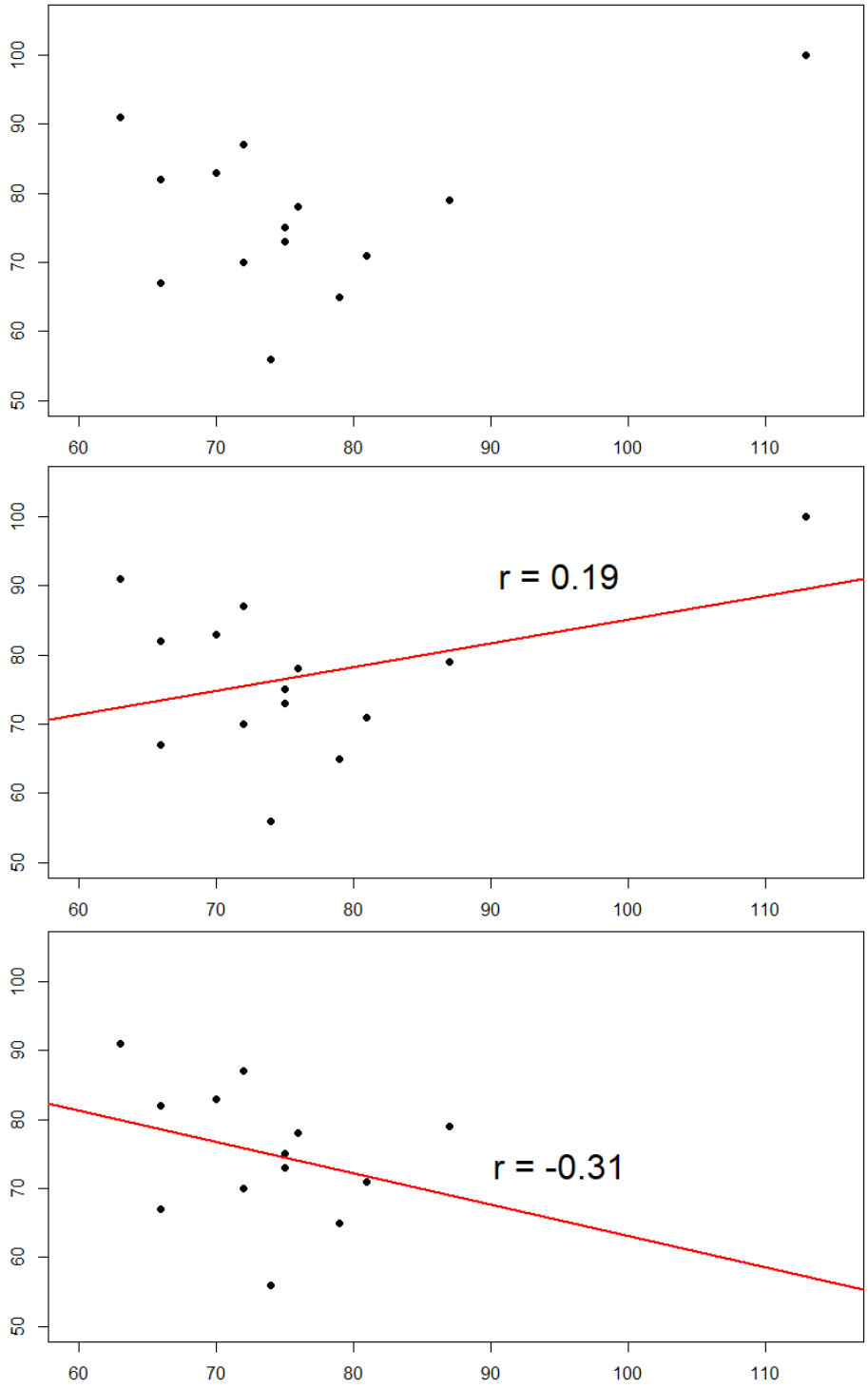


Figure 14. Three plots showing correlations with and without outliers.

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 our 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

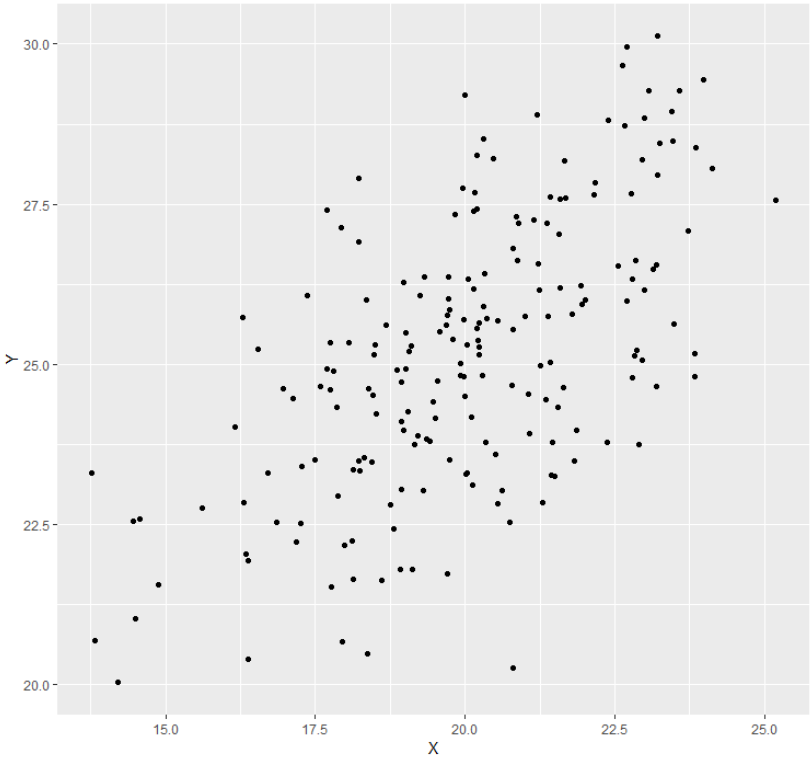
Notice that there are values of 1.00 where each row and column of the same variable intersect. This is because a variable correlates perfectly with itself, so the value is always exactly 1.00. Also notice that the upper cells are left blank and only the cells below the diagonal of 1s are filled in. This 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.

Exercises – Ch. 12

1. What does a correlation assess?
2. What are the three characteristics of a correlation coefficient?
3. What is the difference between covariance and correlation?
4. Why is it important to visualize correlational data in a scatterplot before performing analyses?
5. What sort of relation is displayed in the scatterplot below?



6. What is the direction and magnitude of the following correlation coefficients
    a. -0.81
    b. 0.40
    c. 0.15
    d. -0.08
    e. 0.29
7. 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

8. In the following correlation matrix, what is the relation (number, direction, and magnitude) between…
    a. Pay and Satisfaction
    b. Stress and Health

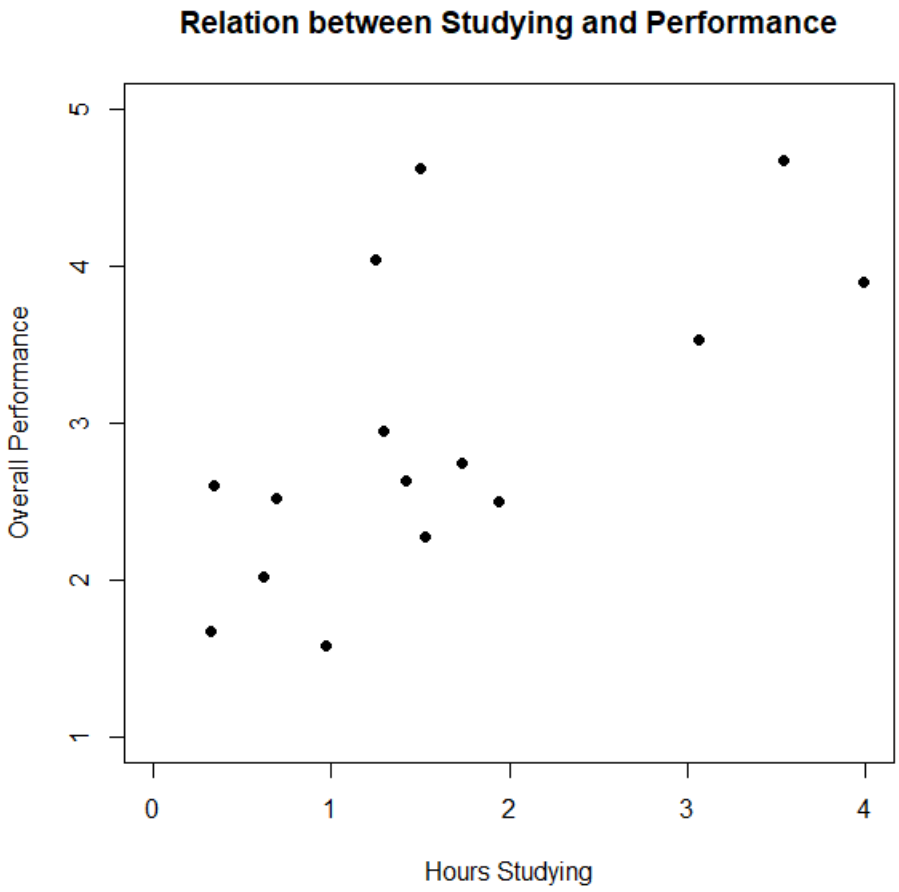
Workplace	Pay	Satisfaction	Stress	Health
Pay	1.00			
Satisfaction	.68	1.00		
Stress	0.02	-0.23	1.00	
Health	0.05	0.15	-0.48	1.00

9. Using the data from problem 7, test for a statistically significant relation between the variables.

10. 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: $\overline{X} = 4.02$, $s_x = 1.15$, $\overline{Y} = 15.92$, $s_y = 5.01$, $SS_X = 130.93$, $SS_Y = 2484.91$, $SP = 159.39$. Test for a significant relation using the four step hypothesis testing procedure.

Answers to Odd- Numbered Exercises – Ch. 12

1. Correlations assess the linear relation between two continuous variables
3. Covariance is an unstandardized measure of how related two continuous variables are. Correlations are standardized versions of covariance that fall between negative 1 and positive 1.
5. Strong, positive, linear relation
7. Your scatterplot should look similar to this:



9. 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: $\overline{X} = 1.61$, $SS_X = 17.44$, $\overline{Y} = 2.95$, $SS_Y = 13.60$, 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, $r(13) = 0.65$, $p < .05$.

Chapter 13: Linear Regression

In chapter 11, we learned about ANOVA, which involves a new way a looking at how our data are structured and the inferences we can draw from that. In chapter 12, 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.

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 2 and 12, 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:

$$\hat{Y} = a + bX$$

The terms in the equation are defined as:

$\hat{Y}$: the predicted value of Y for an individual person
a: the intercept of the line
b: the slope of the line
X: the observed value of X for an individual person

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). The formulae for these parameter estimates use very familiar values:

$$a = \overline{Y} - b\overline{X}$$

$$b = \frac{cov_{XY}}{s_X^2} = \frac{SP}{SSX} = r\left(\frac{s_y}{s_x}\right)$$

We have seen each of these before. $\overline{Y}$ and $\overline{X}$ are the means of Y and X, respectively; cov_{XY} is the covariance of X and Y we learned about with correlations; and s_X^2 is the variance of X. The formula for the slope is very similar to the formula for a Pearson correlation coefficient; the only difference is that we are dividing by the variance of X instead of the product of the standard deviations of X and Y. Because of this, our slope is scaled to the same scale as our X variable and is no longer constrained to be between 0 and 1 in absolute value. This formula provides a clear definition of the slope of the line of best fit, and just like with correlation, this definitional formula can be simplified into a short computational formula for easier calculations. In this case, we are simply taking the sum of products and dividing by the sum of squares for X.

Notice that there is a third formula for the slope of the line that involves the correlation between X and Y. This is because regression and correlation look for the same thing: a straight line through the middle of the data. The only difference between a regression coefficient in simple linear regression and a Pearson correlation coefficient is the scale. So, if you lack raw data but have summary information on the correlation and standard deviations for variables, you can still compute a slope, and therefore intercept, for a line of best fit.

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 ($\widehat{Y}$) 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 ($\widehat{Y}$). 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 ($Y - \widehat{Y}$) 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.

Let's look at this from an applied example. 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.

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.

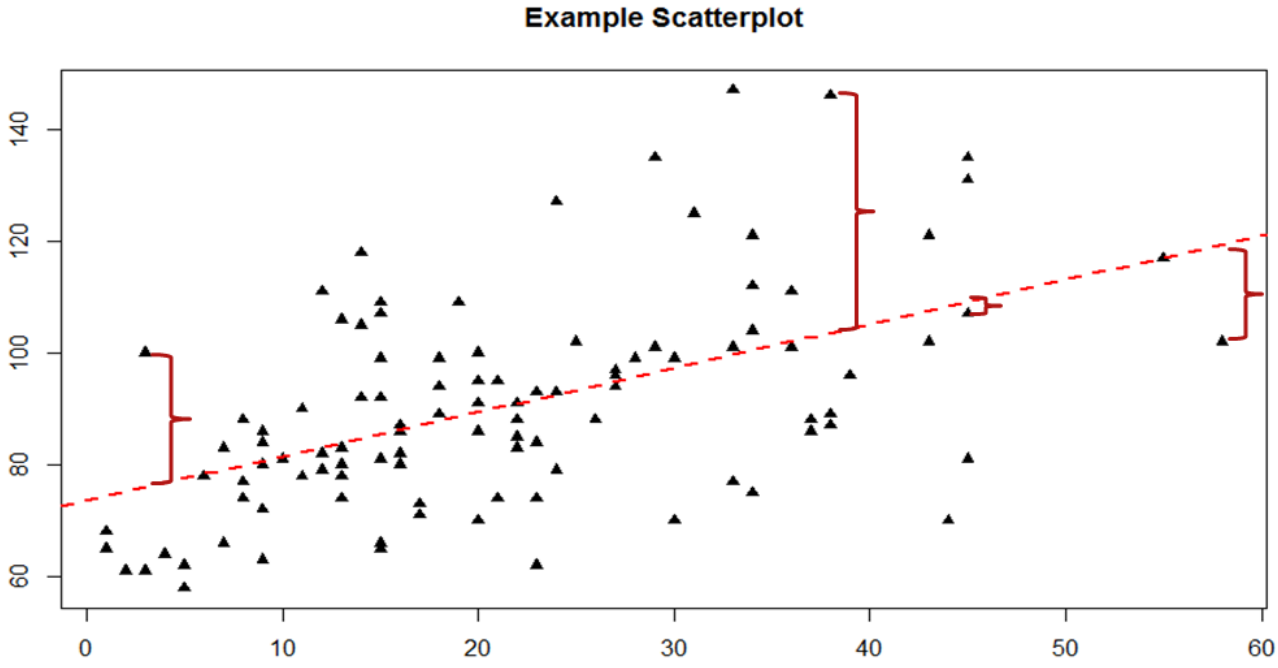


Figure 1. Scatterplot with residuals

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.

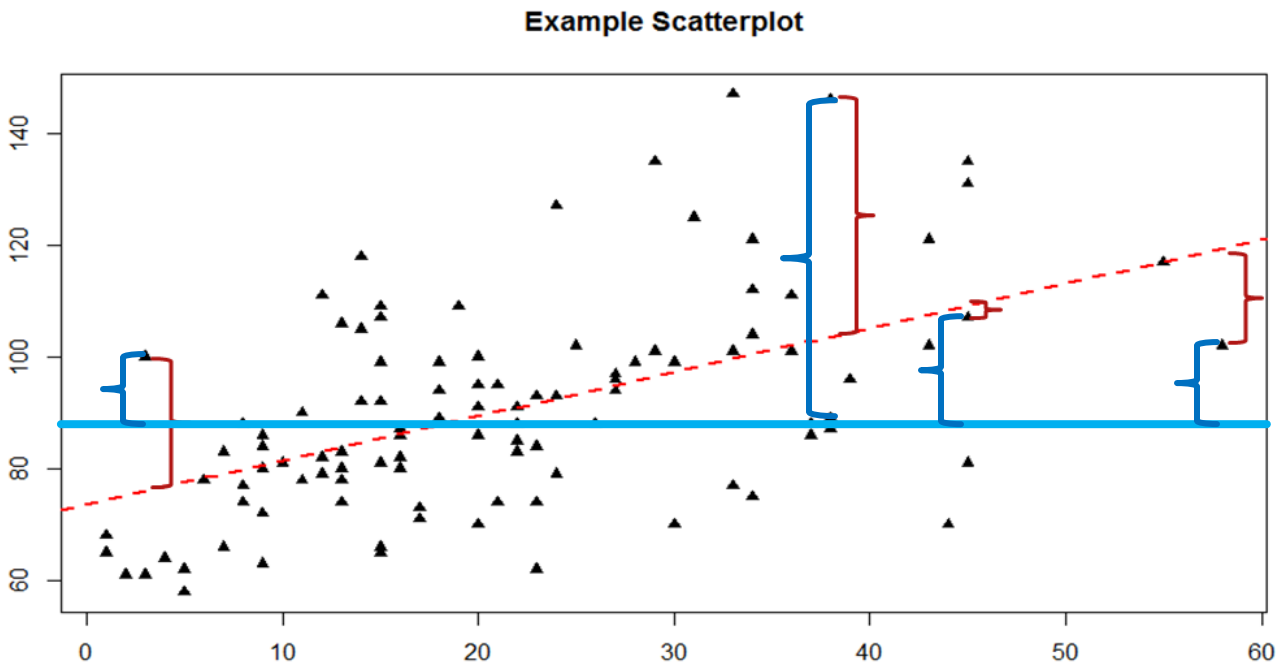


Figure 2. Scatterplot with residuals and deviation scores.

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	$\sum(\hat{Y} - \bar{Y})^2$	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:

$$H_0: \textit{There is no explanatory relation between our variables}$$
$$H_0: \beta = 0$$

$$H_A: \textit{There is an explanatory relation between our variables}$$
$$H_A: \beta > 0$$
$$H_A: \beta < 0$$
$$H_A: \beta \neq 0$$

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.

$$H_0: \textit{There is no explanatory relation between health and happiness}$$
$$H_0: \beta = 0$$

$$H_A: \textit{There is an explanatory relation between health and happiness}$$
$$H_A: \beta \neq 0$$

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	Degrees of Freedom: Numerator						
denom.	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
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 $cov_{XY} = 3.43$, 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:

$$b = \frac{SP}{SSX} = \frac{58.29}{76.14} = 0.77$$

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:

$$a = \overline{Y} - b\overline{X}$$
$$a = 17.45 - 0.77 * 19.78$$
$$a = 17.45 - 15.03 = 2.42$$

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.

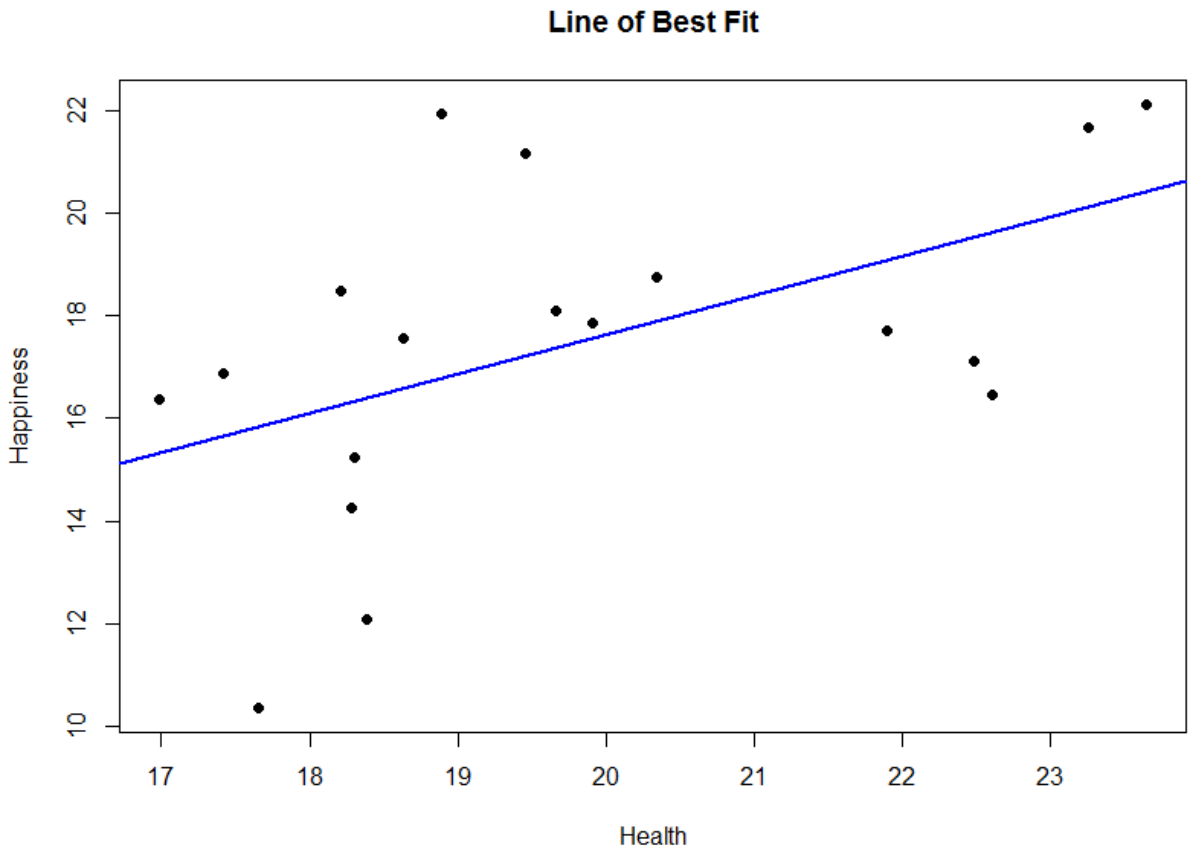


Figure 4. Health and happiness data and line.

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 us R^2, as we saw in correlation (yet more evidence that all of these are the same analysis). Variance explained is still the ratio of SS_M to SS_T:

$$R^2 = \frac{SS_M}{SS_T} = \frac{44.62}{173.99} = 0.26$$

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 ($s_{(Y-\hat{Y})}$), which is given by the formula

$$s_{(Y-\hat{Y})} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{N - 2}}$$

This formula is almost identical to our standard deviation formula, and it follows the same logic. We square our residuals, add them up, then divide by the degrees of freedom. Although this sounds like a long process, we already have the sum of the squared residuals in our ANOVA table! In fact, the value under the square root sign is just the SS_E divided by the df_E, which we know is called the mean squared error, or MS_E:

$$s_{(Y-\hat{Y})} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N-2}} = \sqrt{MSE}$$

For our example:

$$s_{(Y-\hat{Y})} = \sqrt{\frac{129.37}{16}} = \sqrt{8.09} = 2.84$$

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.

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.

Exercises – Ch. 13

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			
Error				
Total	66.12	54		

b.

Source	SS	df	MS	F
Model			6.03	
Error		16		
Total	19.98			

7. In chapter 12, 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: $\overline{X} = 1.61$, $s_X = 1.12$, $\overline{Y} = 2.95$, $s_Y = 0.99$, $r = 0.65$.
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: $\overline{X} = 12.58$, $s_X = 4.65$, $\overline{Y} = 7.44$, $s_Y = 2.12$, $r = 0.34$, N = 67, SSM = 19.79, SSE = 215.77.

Answers to Odd- Numbered Exercises – Ch. 13

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.

6. Least Squares Error Solution; the line that minimizes the total amount of residual error in the dataset.
7. $b = r^*(s_y/s_x) = 0.65*(0.99/1.12) = 0.72$; $a = \overline{Y} - b\overline{X} = 2.95 - (0.72*1.61) = 1.79$; $\widehat{Y} = 1.79 + 0.72X$
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 : $\overline{X} = 8.50$, $\overline{Y} = 8.50$, $SS_X = 339.86$, $SP = 29.99$, giving us a line of best fit of: $b = 29.99/339.86 = 0.09$; $a = 8.50 - 0.09*8.50 = 7.74$; $\widehat{Y} = 7.74 + 0.09X$. 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, $F(1,14) = 0.11$, $p > .05$

Chapter 14. Chi-square

We come at last to our final topic: 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:

$$df = C - 1$$

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.

	1-tailed α			
df	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. This is shown in the formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

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 pineapple on pizza*

H_A: *A significant majority of people will agree one way or the other*

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.

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			
		Primary	Somewhat	No	Total
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. Specifically:

$$E_{ij} = \frac{R_i C_j}{N}$$

The subscripts *i* and *j* indicate which row and column, respectively, correspond to the cell we are calculating the expected frequency for, and the R_i and C_j are the row and column marginal values, respectively. *N* is still the total sample size. Using the data from table 3, we can calculate the expected frequency for cell (1,1), the college sport watchers who used sports at their primary criteria, to be:

$$E_{1,1} = \frac{87 * 68}{168} = 35.21$$

We can follow the same math to find all the expected values for this table:

Expected Values		Affected Decision			
		Primary	Somewhat	No	Total
Watched	Yes	35.21	25.38	26.41	*87*
	No	32.79	23.62	24.59	*81*
	Total	*68*	*49*	*51*	

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. 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.

Example: College Sports

We will follow the same 4 step procedure as we have since chapter 7.

Step 1: State the 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:

$$H_0: \textit{College choice criteria is independent of college sports viewership as a child}$$

$$H_A: \textit{College choice criteria is related to college sports viewership as a child}$$

Step 2: Find the Critical Value

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 use information on both:

$$df = (R - 1)(C - 1)$$

In our example:

$$df = (2 - 1)(3 - 1) = 1 * 2 = 2$$

Based on our 2 degrees of freedom, our critical value from our table is 5.991.

Step 3: Calculate the Test Statistic

The same formula for χ^2 is used once again:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$\chi^2 = \frac{(47 - 35.21)^2}{35.21} + \frac{(26 - 25.38)^2}{25.38} + \frac{(14 - 26.41)^2}{26.41} + \frac{(21 - 32.79)^2}{32.79} + \frac{(23 - 23.62)^2}{23.62} + \frac{(37 - 24.59)^2}{24.59}$$

$$\chi^2 = 3.94 + 0.02 + 5.83 + 4.24 + 0.02 + 6.26 = 20.31$$

Step 4: Make the 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. Like any other correlation coefficient (e.g. Pearson's *r*), the cutoffs for small, medium, and large effect sizes of Cramer's *V* are 0.10, 0.30, and 0.50, respectively. The calculation of Cramer's *V* is very simple:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

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 we calculate during step 3 of the hypothesis testing procedure. 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.

Exercises – Ch. 14

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:
   a. N = 19, R = 3, C = 2, χ^2 (2) = 7.89, α = .05
   b. N = 12, R = 2, C = 2, χ^2 (1) = 3.12, α = .05
   c. N = 74, R = 3, C = 3, χ^2 (4) = 28.41, α = .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. 13

1. 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.
3. 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	

χ^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 = $\sqrt{\frac{7.13}{48(3-1)}}$ = 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, χ^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

χ^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

Epilogue: A Brave New World

This textbook has covered quite a bit of ground, and we are light-years away from where we began. When we started in Unit 1, numbers were foreign, data was an abstract idea, and statistics was dark magic used to trick us into thinking a certain way. However, we soon learned that statistics is neither magic nor evil. Instead, it is a way of thinking objectively about the information that surrounds us in our everyday lives and enables us to critically examine whether or not effects are real.

We have only scratched the very surface of statistics in this book. Each of the topics covered in the fourteen chapters can be dived into to a depth far beyond what we could imagine now. There is nuance to each of our analyses, alternative ways of looking at our data, and countless extensions of everything we have done.

It is a very exciting time to be in statistics and data analysis. Techniques from statistics, math, engineering, computer science, physical science, and behavioral science have all come together into data science, opening up new worlds of possibilities for data collection, visualization, and interpretation. This budding field is growing quickly and promises to be among the most important players of the 21st century. The skills you have learned in this book are the foundation upon which data science is built. Understanding the logic and process behind where our statistics come from and the different forms data can take is the first necessary step into this broader world of data science. Even if you don't pursue a career in this area, it will undoubtedly influence your life in many ways.

It is the idealistic (that is, foolishly optimistic) hope of every educator that their students leave class with a deep understanding and newfound passion for the subject that permeates the rest of the students' lives and careers. For some of you, this will be the case, and I welcome you to the wonderful world of data analysis. For many of you, though, your journey understandably ends here. I do hope that you nonetheless take with you a general understanding of the principles underlying statistics, data analysis, and scientific inquiry: numbers and people differ naturally every day, observed differences may be true and important or they may just be caused by random chance, and – most importantly – CORRELATION DOES NOT PROVE CAUSATION. ☺

To all my students, I wish you the best. May you continue on to accomplish wondrous things I could only every dream of.

Garett C. Foster, Ph.D.