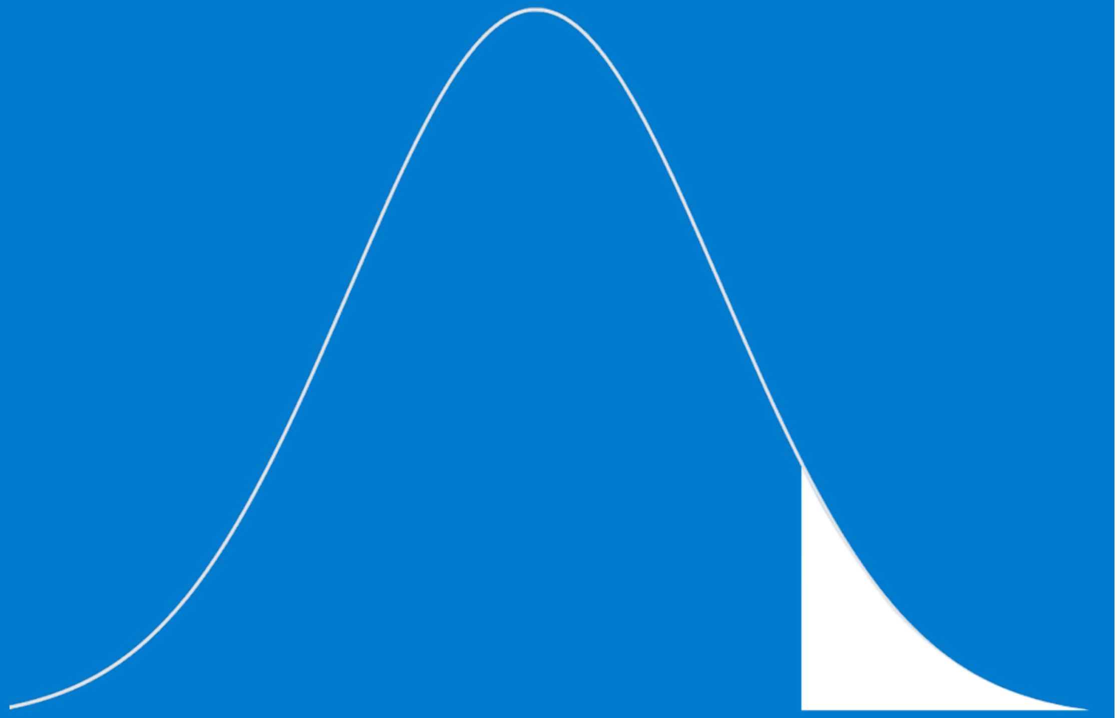
Fundamentals of Statistics



Reagan Brown

About the Author

Reagan Brown is a professor at Western Kentucky University. He earned his B.A. in Psychology from the University of Texas in 1993 and his Ph.D. in Industrial/Organizational Psychology from Virginia Tech in 1997. You can contact Dr. Brown at reagan.brown@wku.edu with comments or questions about this book. Dr Brown is also the author of *Fundamentals of Psychological Measurement* and *Fundamentals of Correlation and Regression*.

A Note on the Writing Style

I wrote this book in a conversational style. Apparent grammatical errors are merely symptomatic of this writing style. I might happen to lapse into a formal writing style on occasion. Do not be alarmed if such an event should occur.

Introduction

1

The Wonderful World of Statistics.

Not all of It.

Introduction

Consider a world without statistics. No way to quantify the magnitude of a problem ("It's a big, big problem. Really big. Gigantic even.") or the effectiveness of a treatment for the problem ("The experimental group had better results than the control group. Not a lot better but not nothing either. There was an improvement. Some."). In this statistics-free world bad science abounds with only the simplest or crudest of ways to identify and correct problems. We need statistics.

We also need sound measurement (see a textbook on measurement theory) and sound research designs to go with our statistics if the field is going to function as a science. The need for the former should be clear: lousy measurement means that nothing of value can be salvaged from the study (one of my professors used to say that there is no statistical fix for bad measurement). The issue of experimental design is the heart of a research methods class. We'll leave those topics to those classes – after a very brief discussion of research methods. Sorry. Has to be done.

Predictive Versus Explanatory Research

Statistics are tools we use to analyze data, nothing more. *Why* we analyze these data is another story. Psychological research can serve two general purposes: prediction and explanation. Predictive research is conducted to see how well a variable (or set of variables) predicts another variable. If we determine that the relationship between these variables is strong enough for applied purposes, then predictive research is also concerned with establishing the means for making these predictions in the future.

Explanatory research is concerned with causal issues. "Explanation is probably the ultimate goal of scientific inquiry, not only because it satisfies the need to understand phenomena, but also because it is the key for creating the requisite conditions for the achievement of specific objectives" (Pedhazur, 1997, p. 241). Stated differently, understanding causality is important because if we understand *how* something occurs, we have the means to change *what* occurs. That's powerful stuff.

Thus, with explanatory research we seek to understand why something is occurring. Why do children succeed or fail in school? Why do people feel satisfied or dissatisfied with their job? Why do some people continually speak in the form of questions? It should be obvious that explanation is more difficult than mere prediction. With prediction we don't care why something is happening. All we want to do is predict it. Understanding why something is occurring may help to predict it, but it's not necessary. Explanation requires more than simply finding variables related to the dependent variable – it requires the identification of the variables that actually cause the phenomenon. Many variables, although not actually causing the phenomenon, will predict simply because they are related to causal variables. Many variables predict, but only a subset of these variables are the actual causes.

The Role of Theory

You might ask, how then is predictive research different from explanatory research, aside from their end goals? The answer is they can involve different analytic tools, but there are some other important differences. Foremost among these is the role of theory. Theory need not play any role at all in predictive research. It's possible to go completely theory-free and have successful predictive research. Just try a bunch of variables and see what works. Because it doesn't matter *why* something predicts, we don't have to possess a good reason for trying and using a variable if it predicts. That said, predictive research based on a sound theory is more likely to succeed than theory-free predictive research.

The situation is completely different for explanatory research. A sound theoretical basis is essential for explanatory research. Because explanatory research is all about why different outcomes occur, we must include all of the relevant variables in our analysis. No throwing a bunch of variables in the experiment just to see what works. A set of variables, chosen with little regard to any previous work, will not likely include the actual cause. (Also, including too many irrelevant variables can corrupt our analysis in other ways.) Furthermore, there is no way to fix explanatory research that was incorrectly conceived. “Sound thinking within a theoretical frame of reference and a clear understanding of the analytic methods used are probably the best safeguards against drawing unwarranted, illogical, or nonsensical conclusions” (Pedhazur, 1997, p. 242). I don’t know about you, but I don’t want to draw unwarranted, illogical, or nonsensical conclusions.

A hypothetical study may help illustrate the differences between predictive and explanatory research. In this study, researchers measured the number of classical music CDs, books, computers, and desks in the houses of parents of newborns. Ten years later they measured the mathematical intelligence of these children. An analysis revealed that the combined number of classical music CDs and desks strongly correlated with mathematical intelligence.

The first issue to address is: Is this sound predictive research? Yes, the number of classical CDs and desks are strongly related to mathematical intelligence and can be used to predict math IQ scores with excellent accuracy. (Just a reminder, this study is not real. I had a lot of fun making it up.)

A second question is: Is this sound explanatory research? No, and it's not even close. These variables were chosen simply because they correlated with the dependent variable, not because there was a logical reason for them to affect math ability. To think that the possession of these items is the cause of mathematical intelligence for these children is to make the classic mistake of equating a strong relationship with a causal relationship. If you're still not convinced, ask yourself this: Would supplying classical music and furniture to households of newborns that didn't have those items raise the math scores of children living in those households? The cause of a given variable is also the means for changing the status of people on that variable.

Research Designs

No chapter that even mentions causal, or explanatory, research would be complete without a short discussion of research design. Statistics are fun and all, but it is the research design (and associated methodology) that allows us to draw, or prevents us from drawing, clear conclusions about causality.

The three basic research designs are: the true experiment, the quasi experiment, and the non experiment (also called a correlational study, but that's a terrible name). These three designs differ in two aspects: how subjects are assigned to conditions (through a random or non random process) and whether the independent variables are manipulated by the experimenter. Some variables can be manipulated, like type of reinforcement schedule, and some can't, like height or SAT score.

Put these two factors together, and we get our three basic types of experimental designs (Chart 1). The true experiment has random assignment to groups and a manipulated independent variable. Due to the random assignment, the groups likely begin the study equal on all relevant vari-

CHART 1 Experimental Design Characteristics

	Random Assignment	Manipulation
True Experiment	✓	✓
Quasi Experiment	✗	✓
Non Experiment	✗	✗

ables, meaning that after the manipulation has occurred the differences observed between the groups on the dependent variable are the result of the experimenter's manipulations (i.e., the independent variable). The great advantage of this design is that, if done correctly, causal claims are clear and easy to substantiate. There are some disadvantages to this design, but let's not concern ourselves with those.

In the quasi experiment, people are not randomly assigned to groups, but there is a manipulated independent variable. Aside from the lack of random assignment, the quasi experiment is like the true experiment. However, that one difference makes all of the difference. The non random assignment to groups is a fundamental weakness. Only random assignment offers any assurance that the groups start out equal. And if the groups start out unequal, there is no way to know if the observed differences on the dependent variable are due to the manipulated variable or to pre-existing differences. To summarize, there are an infinite number of possible causes for the differences observed on the dependent variable, of which the independent variable is but one. At least, however, the manipulated variable is a good candidate for the cause. So there's that. You may be asking, "If there are so many problems that result from not randomly assigning people to groups, why would anyone ever fail to randomly assign?" The answer is sometimes we are simply unable to randomly assign people to groups. The groups are pre-existing (i.e., they were formed before the study) and unalterable. An example would be the effect of two different teaching techniques on classes of introductory psychology students. The students picked the class (including instructor, dates, times and locations); it is not possible for the researcher to assign them, randomly or otherwise, to one class or the other. That's the real world, and sometimes it constrains our research.

In the third design, the non experiment, people are not randomly assigned to groups; there is also no manipulation. In fact, there are often not even groups. A classic example of this type of design is a study designed to determine what causes success in school. The dependent variable is scholastic achievement, and the independent variable is any number of things (IQ, SES, various personality traits). You will note that all of these various independent variable are continuous variables – there are no groups. And of course, nothing is manipulated; the people in the study bring their own IQ status (or SES or what have you) with them. As with the quasi experimental design, there are an infinite number of possible causes for the differences observed on the dependent variable. However, because nothing was manipulated in the non experiment, there isn't even a good candidate for causality. Every possible cause must be evaluated in light of theory and previous research. It's an enormous chore. So why would anyone use this design? Well, some variables can't be manipulated for ethical reasons (e.g., the effects of smoking on human health) or practical reasons (e.g., height). Conducting research on topics where the independent variable can't be eliminated requires researchers to make the best of a bad hand (to use a poker metaphor).

Terminology: Variables

I've used the term independent variable in the previous section without defining it. Independent variable has a twofold definition. An independent variable is a variable that is manipulated by the researcher; it is also a presumed cause of the dependent variable. You can now guess what a dependent variable is: it's the variable that is the outcome of the study, presumed to be the effect of the independent variable.

There is another way to think of variables, this time along a fixed-random axis. A fixed variable is one whose values are determined by the researcher. This can only apply to an independent variable in a true or quasi experiment. In the case of a fixed (independent) variable, the researcher decides what values the variable will have in the study. For example, if the study involves time spent studying a new language, the researcher might assign one group of subjects to study for zero minutes, another group to study for 10 minutes, and a third group to study for 20 minutes. Why these values and not some other values? Ask the experimenters – they're the ones who chose them.

A random variable is a variable whose values were not the result of a choice made by the experimenter. Aside from cases involving complete scientific fraud (which does happen), all dependent variables are random variables. This whole fixed-random variable thing will come up again later in an unexpected way. Something to look forward to, right? As a final note on random variable in the context of experimental design or probability analysis, we shouldn't confuse random variable with the general concept of randomness. When we discuss a random variable, we are not necessarily saying that the variable is comprised of random data – we are simply saying that the values of this random variable were not chosen by the experimenter.

Terminology: Samples vs Populations

Population. Everyone relevant to a study. If your study is about people in general, then your population consists of every person on the planet.

If your study is about students in an art history class being taught a certain way at a certain place, then your population is everyone in that class. Aside from studies with narrowly defined populations, we never measure the entire population. Sometimes researchers like to pretend that they have measured a population just because their sample is big, but they're just pretending.

Sample: A subset of the population. If there are ten million in the population, and you measure all but one, you've measured a sample. Samples can be small (*N* = 23) or large (*N* = 10,823). (I was once involved in a study in which the population was 27 people. That's it. Only 27 people on the planet were relevant to the study. The study concerned attitudes about a proposed change in departmental policy, something not relevant to anyone outside of this department. So the population for that study was 27. We received data from 26 people – one person did not respond to the questionnaire. Did we measure a sample or a population?)

If you want to know the characteristics of everyone in the population, then you have to measure everyone in the population. That sounds like work. For many types of studies, it's an impossible task. So we measure samples because we are either lazy or are dealing with a population so big that it's a practical impossibility to measure it all.

Sampling Error. Now for the bad news. Because we measure samples instead of the population, there will be error in our results, a type of error called sampling error. That is, because we measured a part of the group and not the whole group, the results we get for the part will not equal the results for the whole. This is true for every type of statistic in the known universe.

Sampling error is unavoidable when you measure samples. The vast majority of what we cover in statistics (in particular, most of the topics covered in this book) involves methods for addressing sampling error. Understanding sampling error is one of the keys to understanding statistics. Dealing with the error that comes from measuring samples instead of populations is the single biggest problem facing social scientists.

One Last Thing Before We Proceed

I didn't invent any of this stuff. In this book I am merely explaining concepts and principles that long ago entered into the body of foundational statistics knowledge.

Probability

2

Just a few basic principles and concepts. Nothing too big to deal with.

Overview

So much of what we will do in statistics involves probability analyses. I should rephrase that and say probability *analysis* (singular, not plural) because it's the same type of analysis repeated until the end of time. That's actually good news as it simplifies our task considerably. Before we get to that part, let's take some time to discuss the general nature of probability. Fun fact: probability analyses began as a way to understand the likelihood of various gambling outcomes.

We need a definition of probability. How about this: probability describes the likelihood that a certain outcome (or range of outcomes) will occur. If it is impossible for a certain outcome to occur (e.g., rolling a 2.3 on a six-sided die), then the probability of that outcome is zero. If it is certain that a certain outcome will occur (e.g., rolling a number greater than zero and less than seven on a six-sided die), then the probability is one. All other probabilities fall between zero and one.

Discrete vs. Continuous Variables

We mentioned random variables in the previous chapter, where random is defined as a variables whose value is not chosen or determined by the experimenter. There are actually two types of random variables: discrete and continuous. And, you guessed it, they have different probability models. So we need to define them both.

Discrete Random Variable. A discrete random variable is a random variable with discrete values. Sorry about that definition. I'll try to improve it. It's a random variable whose values are not continuous. Neither of these definitions are as good as an example: the roll of a six-sided die. There are only six possible values for this variable. There is not an infinite number of values; no chance of rolling a 2.3 or a 5.4. There are just six discrete possible scores. That's a discrete random variable. (By the way, the real definition is: a variable with countably infinite, but usually finite, number of values. Helpful, right?)

Continuous Random Variable. A continuous random variable is a random variable with an infinite number of possible values. How many possible heights are there for humans? How many possible heights are there between 6 feet even and 6 feet, one inch? An infinite number.

Now, we don't measure variables like height or time in their full, continuous glory, but that doesn't mean they are not continuous variables. Even if we measured time in milliseconds, there are still an infinite number of possible time values in between 9.580 and 9.581 seconds.

Probability: Discrete Case

Probability is definitely easier to understand for the discrete case. We will refer to this probability as $pr(X)$, the probability of a certain value of X, a variable with discreet (i.e., non-continuous) values. If we use our six-sided die example, the probability of rolling a five, $pr(5)$, is $^1/_6$.

Let's discuss a few probability concepts for discrete variables. First, as mentioned, the probability of a certain value of X, a discrete random variable, ranges from 0 to 1. No negative probabilities and no probabilities greater than 1 are allowed.

$$0 \leq pr(X) \leq 1$$

Second, the total probability across of possible values of $X = 1$. Once again, think of the roll of a six-sided (fair) die. The probability of rolling a 1 or a 2 or 3 or a 4 or a 5 or a 6 are all $^1/_6$. The sum of all of those probabilities equals 1.

$$\sum_{all\ X} pr(X) = 1$$

Finally, if the outcome is known (because the event has occurred or the opportunity for the event to occur has passed), then $pr(X) = 0$ or 1. So before I flip a (fair) coin, the probability of a heads, $pr(H)$, is .5. After the coin has been flipped, with a the result being a tail, the probability of having tossed a heads is zero. It may seem rather pointless to discuss probabilities for events that have already occurred, but, believe me, there's a place for this (e.g., maximum likelihood score estimation in Item Response Theory).

Now it's time to introduce a statistics term with a special meaning: expectation, which has the symbol *E* followed by some parentheses indicating the statistic for which you want the expected value. If you want the expected value of variable *X*, then it's $E(X)$. There are a number of ways to define expectation – we'll limit ourselves to just two of them. There's the statistical interpretation in which expectation refers to long run average. As in really long run. As in an infinite number of times. And no fair limiting the data in question to a subset of possible values and computing the average of that subset – that's not expectation (I actually had to refute this attempted end-run around the definition once). Second, and far less important to us (but worth mentioning anyway), is the probability definition of expectation in which expectation means all possible outcomes weighted by their probabilities. For the expected value of *X*, this probability definition looks like this:

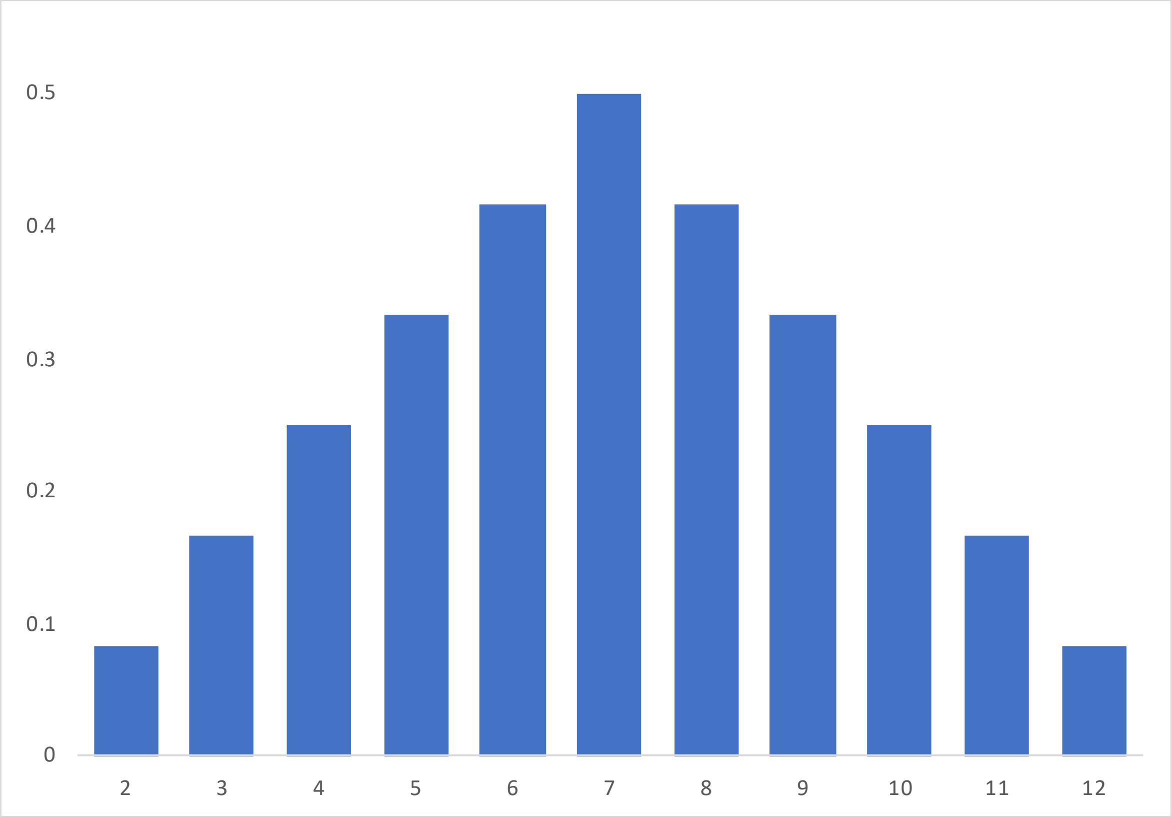
$$E(X) = \Sigma(Xpr(X))$$

That said, let's just remember expectation as long run average value – add up all of the scores and divide by *N*. Two final notes on the expected value of *X*. First, the expected value may not be an observable value. For example, the expected value for the roll of a six-sided die is 3.5 (i.e., if all scores are equally likely and the value of X can only be 1, 2, 3, 4, 5, or 6, then the long run average is 3.5). So 3.5 is the expected value but try rolling a 3.5 – it's not easy. The second point on the expected value of X is that it equals the population mean, μ.

$$E(X) = \mu$$

And finally, how about a graphical representation of scores for a discrete random variable? This graph is the Probability Mass Function (PMF). The PMF shows the probability for various outcomes of a discrete random variable. Let's consider a PMF for the roll of two six-sided dice (think Monopoly dice rolls).

FIGURE 1 Probability Mass Function for Dice Roll



Probability is on the *y*-axis; Total score on the two dice is on the x-axis. Total from two six-sided dice is a discrete variable – the values from 2 to 12 are the only possible values.

Notice a few things in this PMF displayed in Figure 1. The height of the bars indicates probability, higher bar = greater probability. No surprise that it's far more probable that a person will roll a seven than a two. Second, we see that in this case, the PMF is symmetrical – one half is a mirror image of the other half. Third, the total equals 1.0 (add 'em up!).

Probability: Continuous Case

The rules for probability for a continuous random variable are familiar, yet different enough to be annoying. Probably the strangest thing to understand is that for a continuous variable the probability that *X* will equal any specific value is zero. For example, height is a continuous variable. What is the probability that a person's height will equal six feet exactly? The answer is zero. Why? Because with measurement of sufficient precision and enough decimal places, no one is exactly six feet tall. Remember the nature of a continuous variable: there are an infinite number of possible values between any two values (e.g., an infinite number of possible heights between five feet eleven inches and six feet one inch). So that's one reason. There is also the calculus reason. We'll come to that in due time.

So if the probability that *X* equals any specific value is zero, what probability analysis can we do with a continuous variable? The answer is rather than define probability of a single value (i.e., $pr(X)$), we must define a probability function that covers a range of *X* values. This function, called a Probability Density Function (PDF), has the following characteristics. First the function will be non-negative for all values of *X*.

$$f(x) \geq 0 \; for \; all \; X$$

Second, the total probability across the range of values of *X* equals one.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

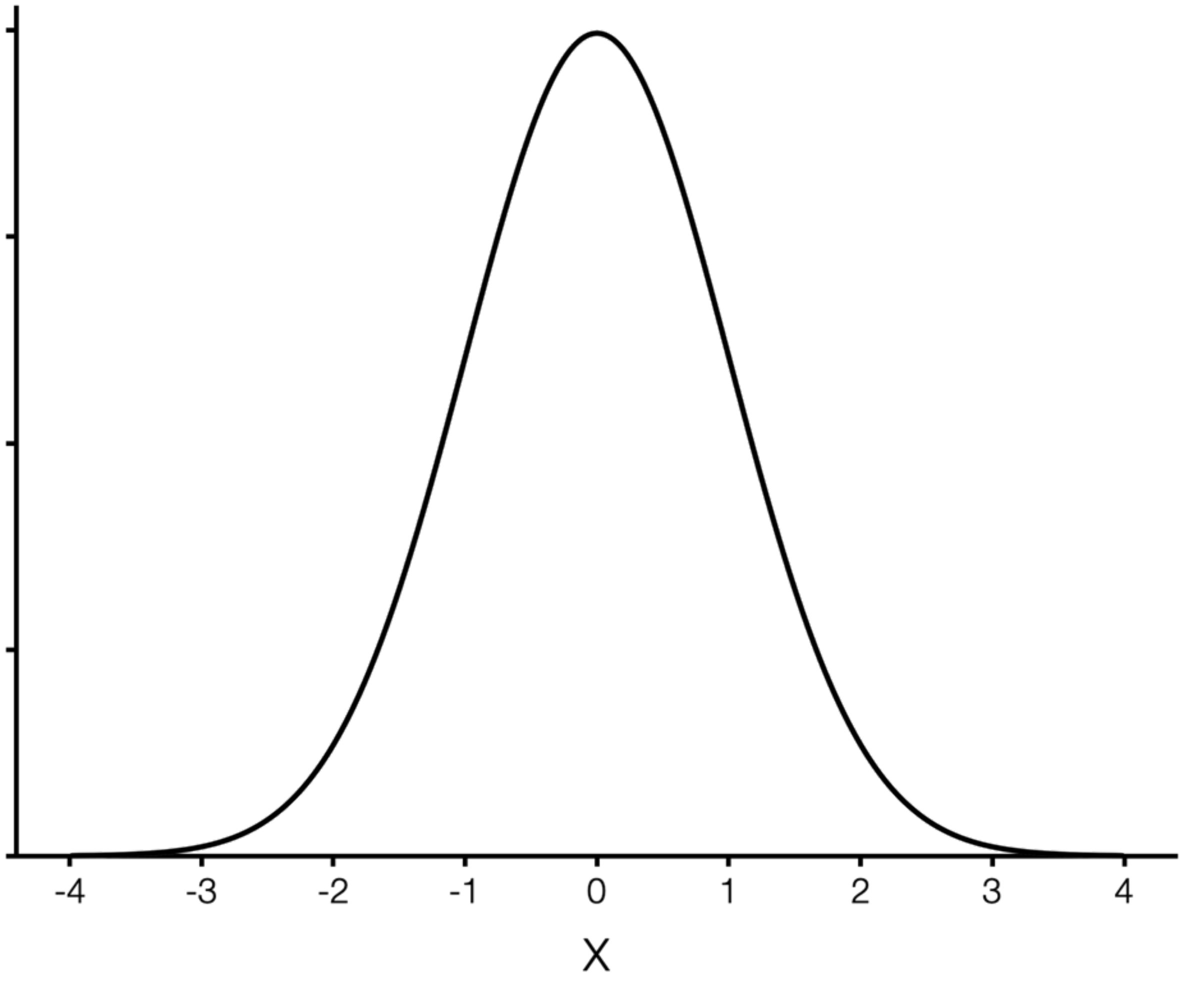
This is where calculus enters. Notice that to compute probability of a continuous function, we must integrate over a range of values. In the above case we are integrating across all possible values of *X*, resulting is a probability of 1.0. This calls back to the discrete variable case where we said the total probability equals one.

To compute the probability for a certain range of values of a continuous variable we perform the same operation with the beginning and end points being those particular values instead of the total range of values.

$$pr(a < X < b) = \int_a^b f(x)dx$$

As I'm sure you remember from your days as a calculus student, we integrate to find the area under the curve defined by $f(x)$ for a range of scores (in this case ranging from a to b) Thus, whereas the discrete probability case was determined by the height of the density function for a given score on X, the continuous case is defined by the area under the curve defined by $f(x)$ for a range of scores on X. Thus, the graphical representation of the PDF is some curve defined by $f(x)$. Figure 2 displays a rather common PDF.

FIGURE 2 Probability Density Function



(And now the calculus reason for why the probability for a specific value of X is zero. Integrating in calculus is a way to find the area under the function for a range of values of X. Well, what happens if the range of values is from a to a? The area is zero.Think of computing the area of a rectangle – it's length times width. And what is that area if the width is zero? That's what you're doing when you say integrate at this one point, having a range of zero.)

The expected value of *X* for a continuous variable takes the same general form as before: *X* times its probability across the entire range of values for *X*. Of course, this being a continuous variable, we have to calculus it up a bit.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Thus, expectation is a weighted average of scores on *X*, where scores on *X* are weighted by the likelihood (i.e., probability as defined by $f(x)$) of occurrence.

Finally, as with the discrete case, the expected value of *X* for a continuous variable equals the population mean.

$$E(X) = \mu$$

And as long we're discussing expectation, we should introduce the concept of variance (symbolized as σ^2 for the population value), which we'll define as the expected value of the squared difference between *X* and the population mean, μ.

$$\sigma^2 = E(X - \mu)^2$$

And since μ is the expected value of *X*, we could rewrite this as:

$$\sigma^2 = E(X - E(X))^2$$

There is so much more to say about variance (and we will in the next chapter). For now just think of variance as an index of differences in scores (calculated as the average squared difference between scores on *X* and the mean of *X*). Also note that variance can't be negative ($\sigma^2 \geq 0$) because the scores are either different from each other or they're not; there can't be negative differences.

One last statistic for now is standard deviation. Standard deviation is the square root of variance and has the symbol σ for the population value.

$$\sigma = \sqrt{\sigma^2}$$

It's nice that the symbols make sense. That seldom happens in statistics.

The Normal Distribution

The normal distribution is a theoretical concept not found in real data (the dataset would have to be fully continuous – literally no two scores the same – with $N = \infty$). So no one has ever had a normally distributed set of data. But real data often approximate the normal distribution, hence its usefulness.

The PDF for a normal distribution having a mean of μ and a variance of σ^2 is defined by the following function.

$$f(x) = \frac{1}{e^{(X-\mu)^2/2\sigma^2}\sqrt{2\pi\sigma^2}}$$

As before, area under curve equates to probability.

The normal distribution is symmetrical and unimodal (one peak) in which μ equals the center of distribution and σ^2 (or σ) equals the spread of distribution. The normal distribution is actually a family of distributions having the same shape but varying on mean and variability. That said, we are going to make our lives easier by designating one type of normal distribution as the official standard. This standard version is called the standard normal distribution, and it has a mean of zero and a standard deviation of one. As nice as that is, the even better news is that when $\mu = 0$ and $\sigma = 1$, then the above equation simplifies to the following slightly less intimidating equation.

$$f(x) = \frac{1}{e^{.5(X^2)}\sqrt{2\pi}}$$

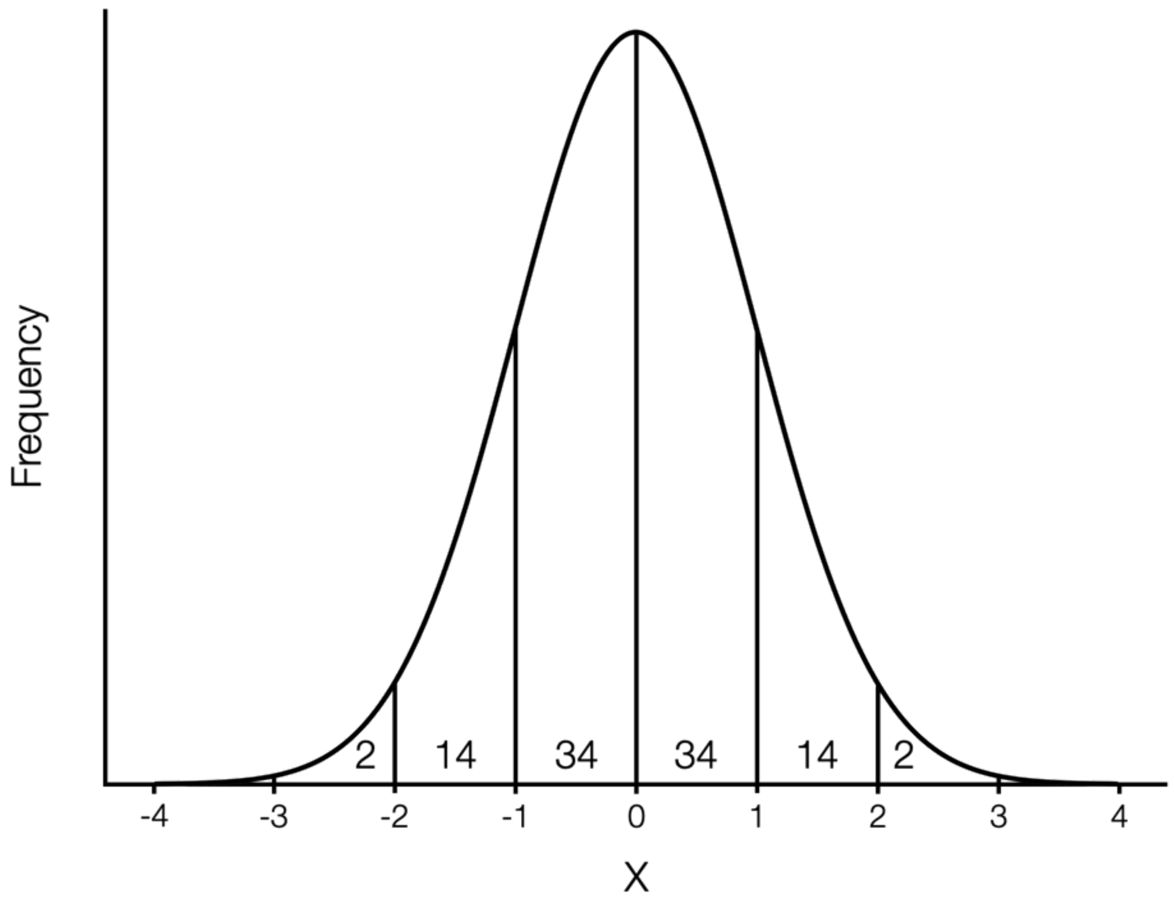
We'll be referring to the normal distribution quite often in the coming chapters. So there's something to which you can look forward. For now, let's just do a bit of basic navigation of the normal distribution.

Because normally distributed data always have the same shape, the normal distribution has many desirable properties. First, because the normal distribution is symmetrical, we can describe a score's relative position to the mean. That is, is a score above the mean or below the mean? How far above or below? Both questions are important and both have quantifiable answers that have the same meaning for all normally distributed data. (What if the data are not normally distributed? Well then, life's not so simple. We have the good fortune that most variables are approximately normally distributed.)

Here's what you'll find if you examine a set of normally distributed data. A certain percentage of scores will always be the same distance from the mean. For example, let's say you have a set of normally distributed data from a sample of 100 people. Let's also say the mean of this data is 0 and the standard deviation is 1. If you count how many people have scores between the mean and one standard deviation above the mean, you will find 34. If you count the number of people with scores between one and two standard deviations above the mean, you'll find 14. Finally, if you count the number of people with scores between two and three standard deviations above the mean, you find just two people. And because the distribution is symmetrical, things are the same for scores below the mean.

A few words about the preceding numbers. First, because I used a sample of 100 people, the numbers are rounded. Second, the previous paragraph also makes it appear that the normal distribution goes no further than three standard deviations above or below the mean. The truth is that

FIGURE 3 Areas of the Normal Distribution



the normal distribution is without bounds; in theory, you could find someone with a score so high that they are seven standard deviations above the mean (or nine below the mean or whatever). These are scores so rare that we will not concern ourselves with them; we'll just focus on the world that is three standard deviations above and below the mean. It is this area that contains 99.7% of all scores. One last note, if a person's score is at the mean, their score is at the 50th percentile, meaning that it is higher than 50% of the scores. Figure 3 is a diagram of the normal distribution, divided into sections by standard deviation, showing the percentages in each section.

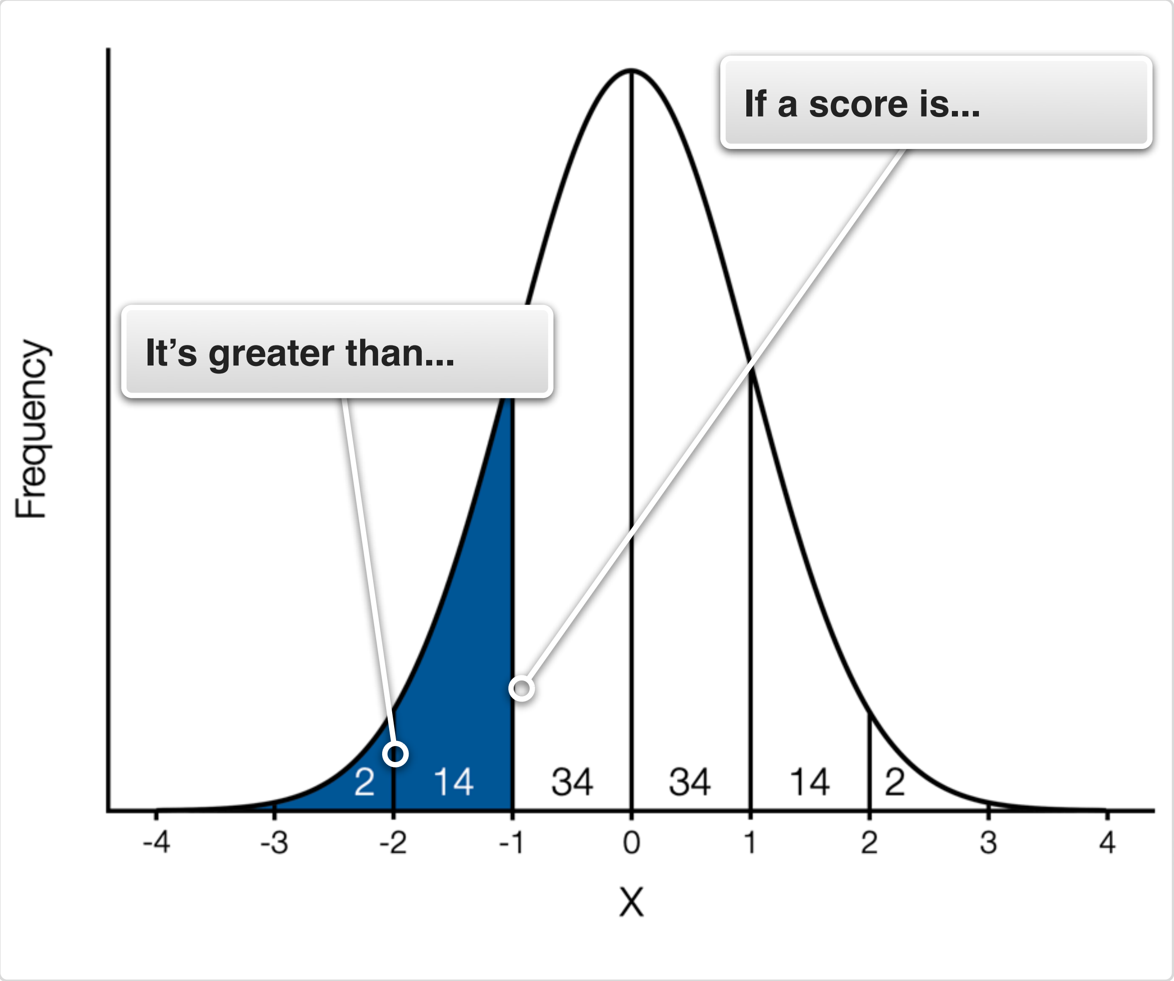
Now, what does all of this buy us? It allows us to quickly and easily attach meaning to a score. All you have to do is remember three numbers: 34, 14, and 2. If I told you that my score on a test was one standard deviation below the mean, what do we know about it? Obviously, it's below average. Using the 34/14/2 rule, we can estimate my percentile rank (the percent of people at or below a given score – we'll discuss percentile ranks in greater detail later). Now how do we figure this out? The only people with scores worse than mine are those with scores even lower than one standard deviation below the mean. A quick calculation shows that my score is greater than 2% +

14% of the scores. Thus, my percentile rank is 16%. This example is shown in Figure 4. So knowing the properties of the normal distribution helps us interpret test scores without much work. And it's all because normally distributed data always has the same shape.

I should stress one point that with this 34/14/2 rule for normal distributions: these percentile ranks are only crude estimates. If any precision is needed, consult a *z* table. Also, if the number of standard deviations above or below the mean for the score in question isn't a nice round number (e.g., 1.7 standard deviations above the mean), we'll need to consult a *z* table. And if the dataset isn't normally distributed, then forget 34/14/2 rule. And forget the *z* table. The z table is descriptive of the normal distribution only.

The only lingering question is this one: How do we know the number of standard deviations above or below the mean a score lies? As an example, if someone's score on a test is a 23, how do we know the number of standard deviations above or below the mean? We'll have to compare that score to the mean score and use the standard deviation of the test to compute something.

FIGURE 4 Using the 34/14/2 Rule to Estimate Percentile Ranks for a Normally Distributed Set of Data



Standard Scores: Linear z Scores

There are many types of standard scores, but the most popular is the linear z score (often referred to as just *z score*). Whether you are computing z scores for population data

$$z_X = \frac{(X - \mu_X)}{\sigma_X}$$

or for sample data

$$z_X = \frac{(X - \bar{X})}{S_X}$$

the equations are fundamentally the same: a given score minus the mean of the scores divided by the standard deviation of the scores.

How about an example? Let's say that I took the SAT, and my verbal score (SAT-V) is a 400. The mean of the SAT-V section is 500, and the standard deviation is 100. Now we're ready to go. Plugging in these values into the *z* score equation, we find that my 400 on the SAT-Verbal becomes a *z* score of -1.0.

Let's take a closer look at my *z* score of -1.0. My *z* score is negative. The negative sign tells you something – my score is below average. If my score were above the mean, my *z* score would have been positive. If my score had been exactly the same as the mean, my *z* score would have been 0.0. The difference between my score of 400 and the mean is 100 points. The standard deviation is 100 points. Thus, my score of 400 is exactly one standard deviation below the mean. The *z* score is -1.0. Do you see where this is going? I'm not this redundant on accident. Here it comes: A z score is literally the number of standard deviations a score deviates from the mean. In case that's not clear, I'll restate the definition in the form of a question: How far (in terms of number of standard deviations) from the mean (above or below) is this score? If the *z* score is -2.0, then the person's score is two standard deviations below the mean. If the *z* score is 1.5, then the person's score is one and a half standard deviations above the mean. If the *z* score is 0.0, then the person's score is zero standard deviations above the mean – it is right on the mean. So when we discuss the number of standard deviations a score is from the mean, we are using *z* scores. Very convenient. The last thing to mention is that if our data are normally distributed, then we can quickly and easily attach meaning to the score with the 34/14/2 rule we learned earlier. Take my score of 400. In *z* score terms it is -1.0. If the data are normally distributed, that means my score is better than only 16% of the test takers (see Figure 4 again). More precise estimates require a *z* table.

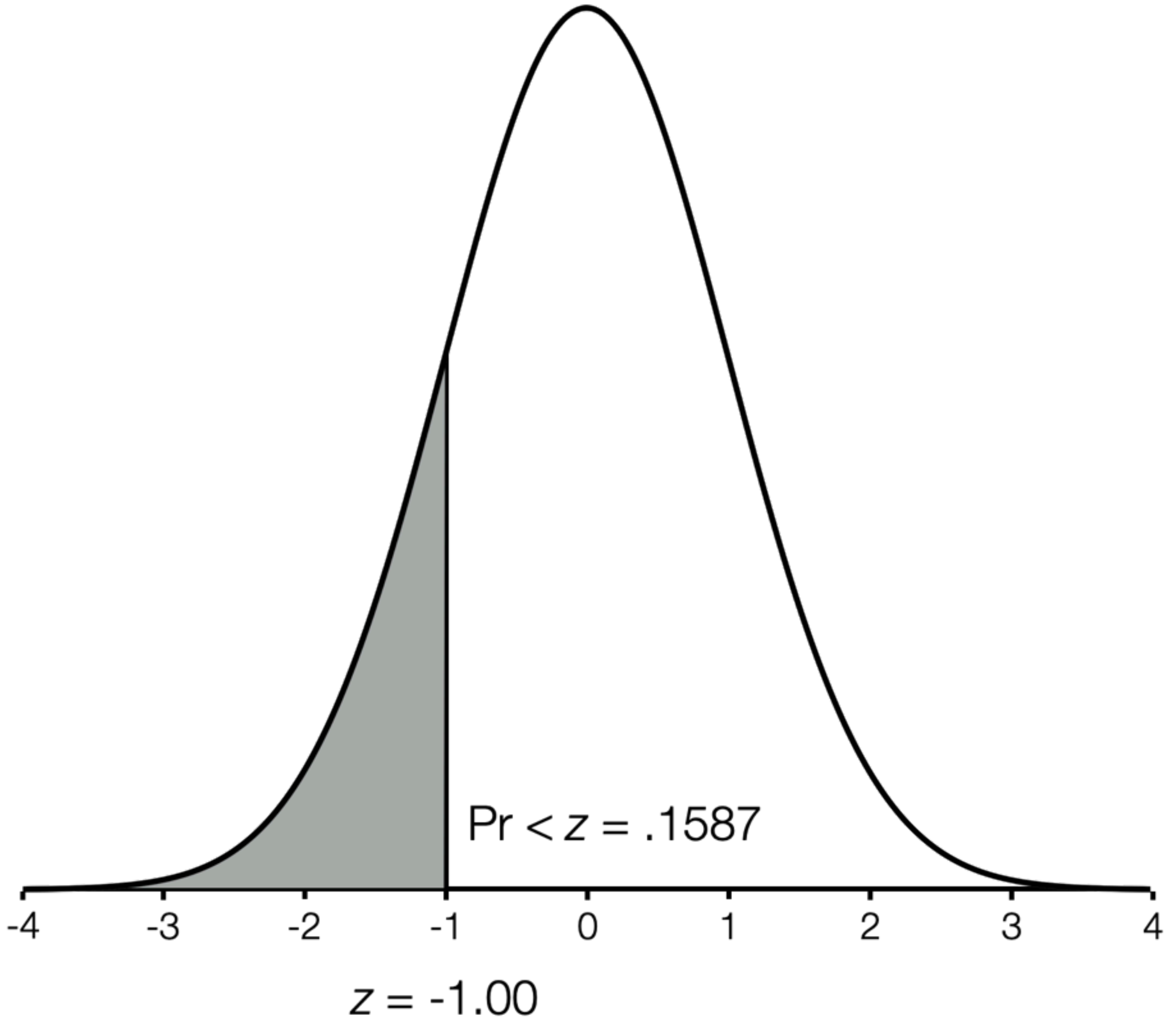
How to Read a z Table

One problem with z tables is that there is a remarkable lack of uniformity to their structure. It's quite annoying. If you understand their design, you can always read the correct value, but some of them seem (much like this sentence) to have a design that can be fairly described as intentionally obfuscatory.

The abbreviated z table given in Table 1 has the structure shown in Figure 5. That is, each entry in the table indicates the proportion of scores (visually: area under the curve) that are less than the given z score. In the case of Figure 5, a z score of -1.00 is greater than .1587 (or 15.87%) of the scores in a normal distribution. Due to space limitations, only selected values from a z table are shown in Table 1.

Let's read a few values from Table 1. If you want the z score that is greater than 95% of the scores, then find the one with a probability (the *pr* < z column) that is closest to .95. In this case, that's 1.645. Only 5% of the scores in a normal distribution are greater than 1.645. We could say that the z score that separates the top five percent of scores from the bottom 95% is 1.645. This num-

FIGURE 5 Reference Guide for z Table Structure



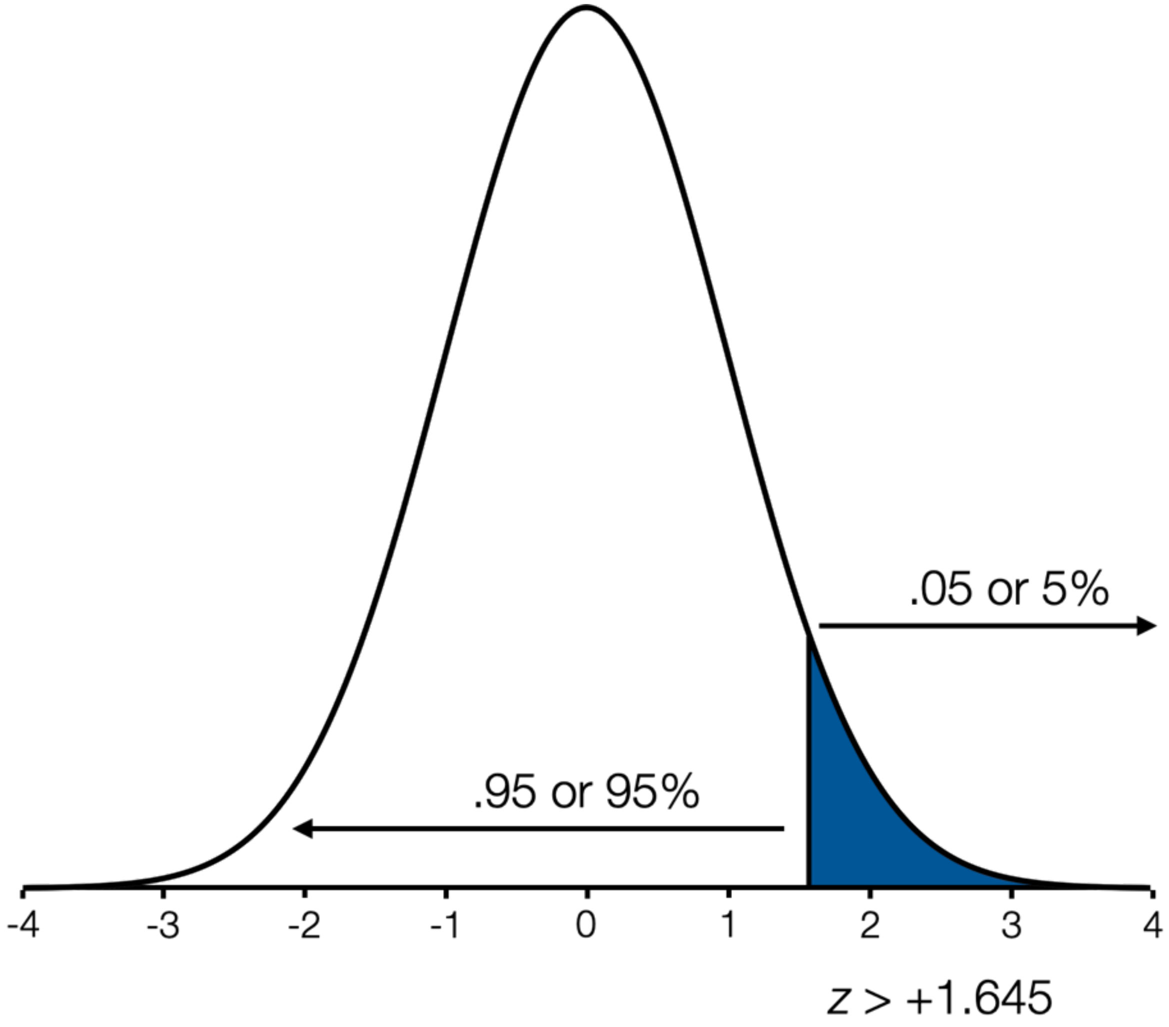
Values in z table indicate proportion of scores below z value. In the diagram, a *z* of -1.00 is a score that is greater than .1587 (i.e., 15.87%) of the scores in a normal distribution.

TABLE 1 Selected *z* Table Values

z	***pr* < *z***	***z***	***pr* < *z***
-3.00	0.0013	0.50	0.6915
-2.00	0.0228	1.00	0.8413
-1.96	0.025	1.50	0.9332
-1.645	0.0500	1.645	0.9500
-1.50	0.0668	1.96	0.975
-1.00	0.1587	2.00	0.9772
-0.50	0.3085	3.00	0.9987
0.0	0.5		

ber will turn out to be moderately important later. It also works in the reverse: -1.645 is greater than only 5% of the scores; so it separates the bottom five percent from the top 95%. In both the positive and negative cases, we are identifying this extreme five percent of the distribution – it's in one tail of the distribution or the other (but not both). This is referred to as a one-tailed situation (except we don't say situation; we say one-tailed test when

FIGURE 6 Upper 5% of Normal Distribution (i.e., One-Tailed Region)



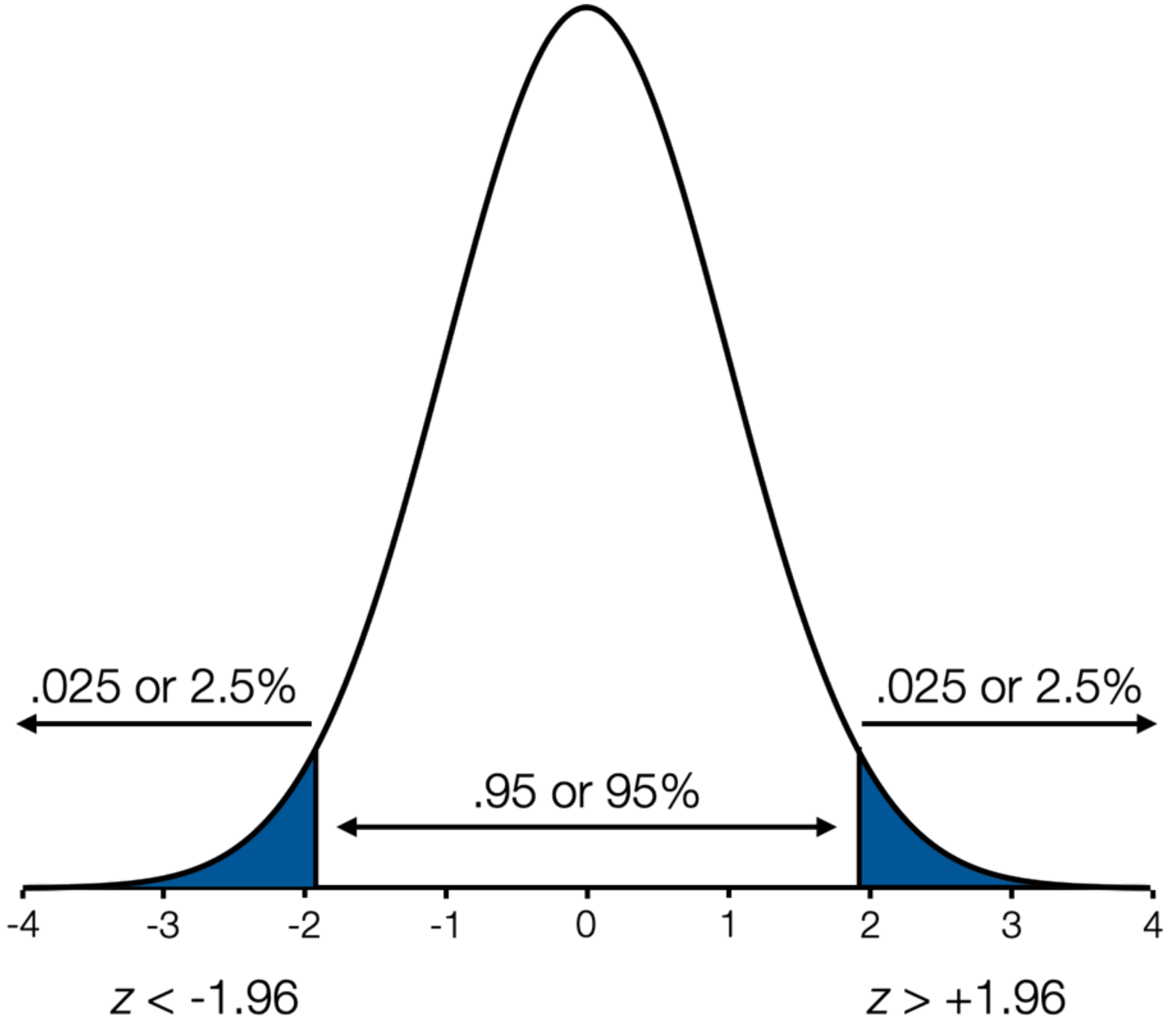
95% of scores distributed normally are between -1.96 and +1.96 *z*. This two tailed region of a symmetrical distribution is commonly used in hypothesis testing.

it's done as part of hypothesis testing). Figure 6 displays this one-tailed region of the normal for which +1.645 z separates the top 5% of scores from the bottom 95% of scores.

One other z score worth mentioning is 1.96. A z score of 1.96 is greater than 97.5% of the scores. That doesn't sound important yet. Once again, because of the symmetrical nature of the normal distribution, -1.96 is greater than 2.5% of the scores. Put these two together and you have this: 95% of the scores in a normal distribution are between -1.96 and +1.96 (because 2.5% are less than -1.96 and 2.5% are greater than +1.96). This region is shown in Figure 7. Because we are splitting the five percent into both tails, you can probably guess that this will be called a two-tailed something or other (two-tailed test for hypothesis testing purposes).

The point of this is that 5% will be a big deal in much of what we do in statistics (i.e., if a result would occur less often than 5% of the time if our hypothesis were incorrect, then we will conclude

FIGURE 7 Upper 2.5% and Lower 2.5% of Normal Distribution (i.e., Two-Tailed Region)



95% of scores distributed normally are between -1.96 and +1.96 *z*. This two tailed region of a symmetrical distribution is commonly used in hypothesis testing.

that our hypothesis is not incorrect – this will be explained at length in Chapter 4). Because we so often use this 5% standard, there are really only two values from a z table that really matter: 1.645 and 1.96 (in both the positive and negative forms).

Estimation

3

If only we knew...

Then we wouldn’t have to estimate.

The Need for Estimation

Why are we discussing something called estimation? Estimation is what you do when you don't know something. We do experiments. We gather data. We have the information. We compute the statistics. We don't need to estimate anything. We know the thing.

Right?

Well, no. The reason is that the results from our experiments are intended to generalize beyond the specific group of people in our study. For example, let's say you do a medical study involving 200 people from a coastal city in America and you find that 58% possess antibodies for a certain virus. You know the percentage for that group of 200 people. You don't have to estimate it – you have the data, and you know it. But the purpose of the study is to use the information from that group and generalize to the rest of the city. What's true of those 200 people is only moderately interesting in itself – what's true of the people of the city as a whole is far more important. The 200 are just a sample from the population of the city, and we use sample statistics as estimates of population statistics.

As we mentioned in the first chapter, none of this would be a problem if we just measured the entire population. We would know the population values and we wouldn't have to estimate anything. But that is work. Way more than we want to do. And most of the time we don't have the resources to measure an entire population even if we wanted to. Which we don't.

To summarize, we will be using statistics computed from samples of data to estimate values for unmeasured populations.

This, of course, leads to no end of problems.

Point Estimation: Means

You know how to compute the mean of a set of data; add up the scores and divide by the number of scores. As mentioned in the previous chapter, the population mean has the symbol μ. So what about the mean of a sample? Once again, add ‘em up and divide by N.

But how well does this sample mean equation work as an estimator of the population mean? It would be bad news if it were biased in some way. The good news is that the sample mean is an unbiased estimator of the population mean. Now, unbiased doesn’t mean accurate. There will be sampling error. It just means that the error won’t consistently favor one direction – the positive errors (where the population mean is greater than the sample mean) will equal the negative errors (where the population mean is less than the sample mean).

The sample estimator of population mean has two symbols. The formal symbol, which we will never mention again, is $\hat{\mu}$, which indicates that it is an estimate of μ. (One imagines that the formal symbol is reserved for state dinners and other fancy affairs.) The common symbol is $\bar{X}$. This sample (estimator of the population) mean is computed just like every other mean.

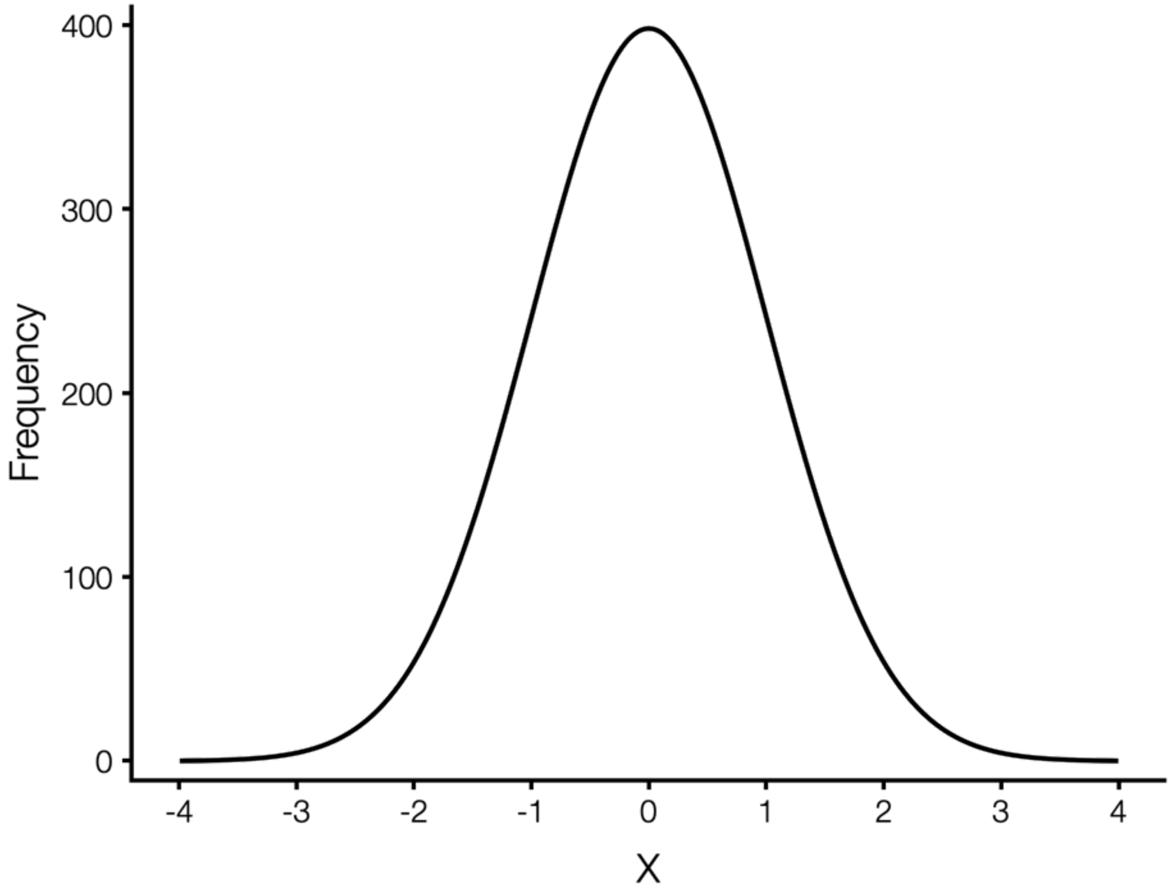
$$\bar{X} = \frac{\Sigma X}{N}$$

In summary, μ is the symbol for population mean, $\bar{X}$ is the symbol for sample mean, and they are both calculated the same way (like every mean you’ve ever calculated in your life). Finally, the sample mean is an unbiased estimator of the population mean, which is good news. But there will always be sampling error (the inevitable bad news that comes with every sample). Sampling error results in sample means that are greater than or less than the population value by some unknown amount.

A Brief Discussion of Variance

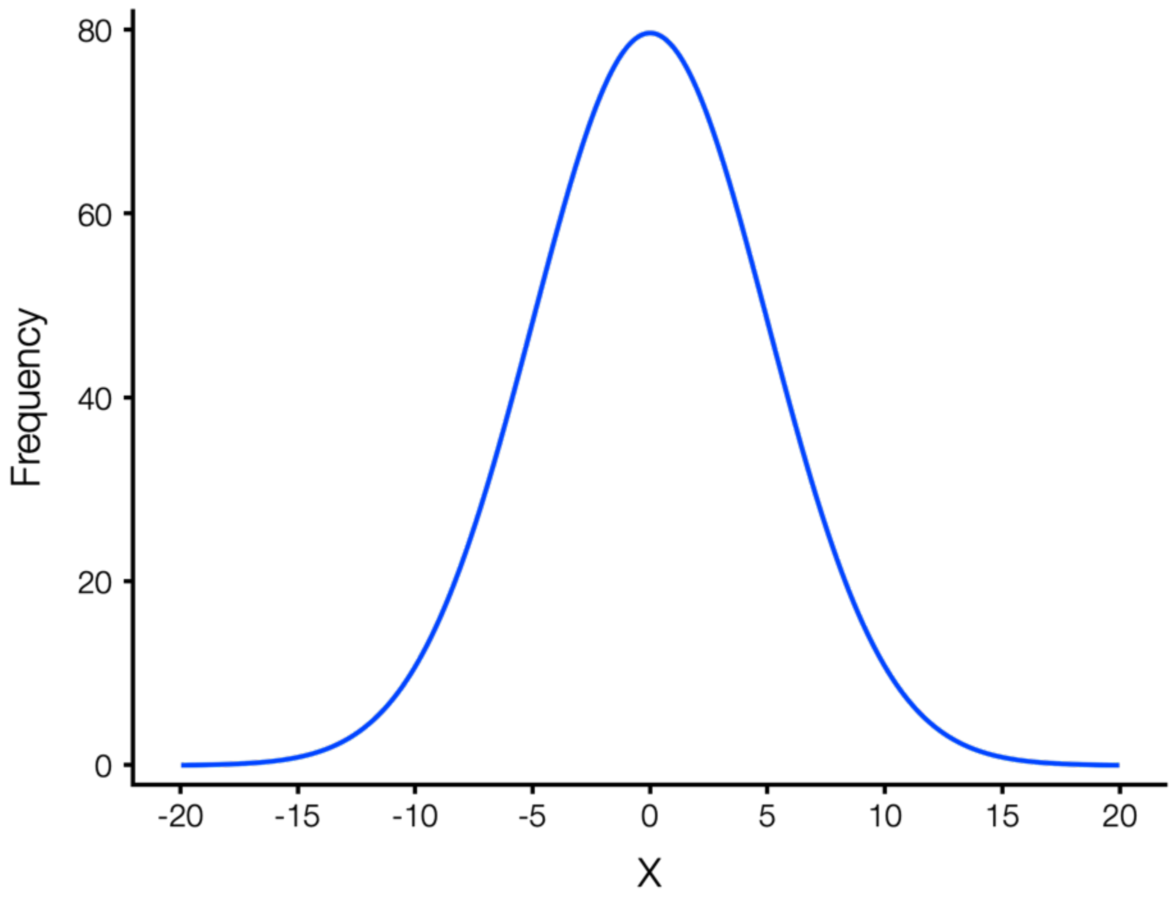
Distributions for two different datasets are displayed in Figure 1 and Figure 2. What's the difference between the two distributions? When they are shown on separate graphs they appear to be the same. They have the same mean score. Notice how the midpoint of each is zero. They have the same sample size (trust me on this). If you've read the title of this section, then you've guessed that the difference is variance. In the Figure 1 distribution (in black), most (approximately two-thirds) of the scores are within one point of the mean (the mean plus or minus one point), whereas in the Figure 2 distribution (in blue), very few of the scores are within one point of the mean. You have to move out to five points away from the mean (the mean plus or minus five points) in order capture most of the scores. If we place both datasets on the same scale (Figure 3), it's clear that the scores are not spread out in the same way (if Figure 3 seems like a massive cheat, pay careful attention to the scale on the *x*- and *y*- axes on the three graphs).

FIGURE 1 Variability Comparison: Low Variance



Variance is greater for the blue distribution than for the black distribution. Variance is all about the differences between the scores

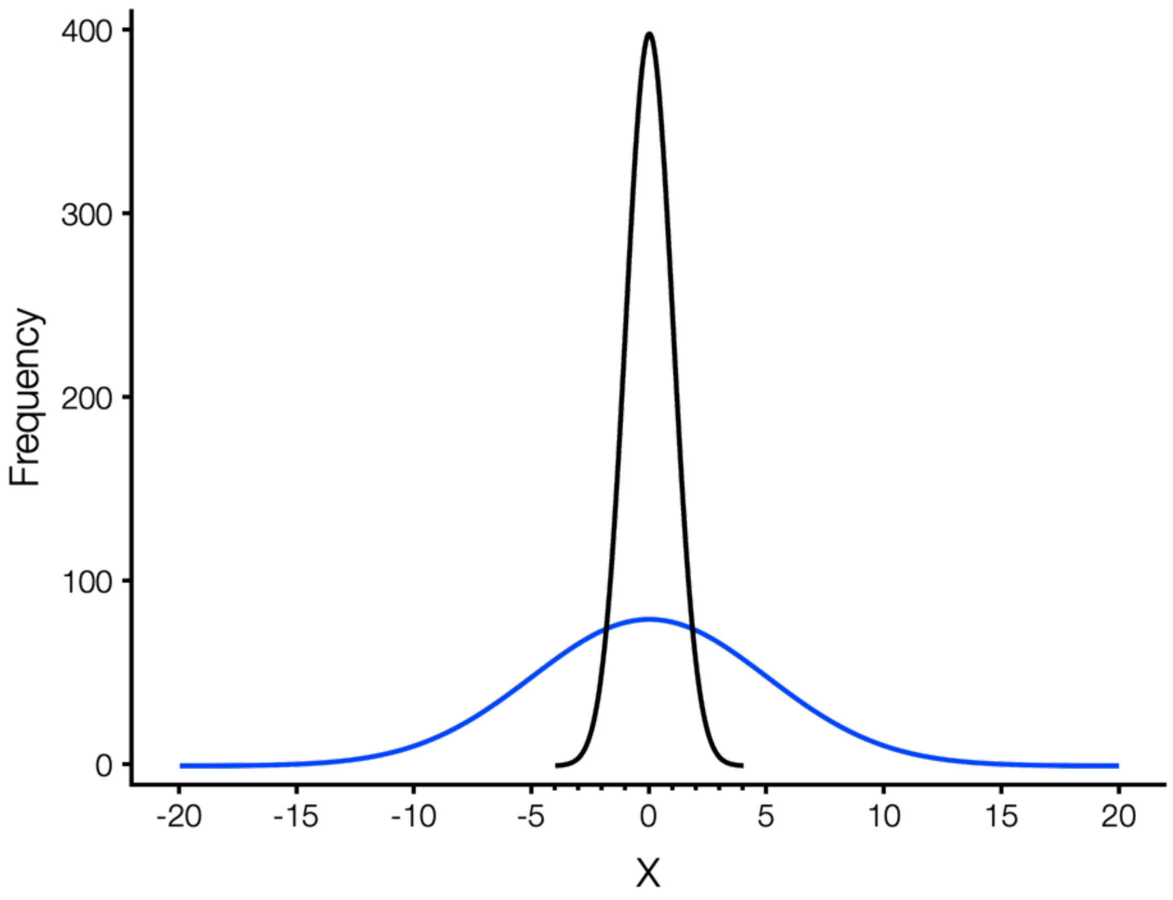
FIGURE 2 Variability Comparison: High Variance



Variance (along with its twin cousin, standard deviation) is our preferred index of variability (symbolized for populations as σ^2). Listed below is the equation to compute variance for a population of data.

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

FIGURE 3 Variance Comparison: Both Distributions



This equation isn't that bad. In fact, it is really similar to the equation for a mean. To see that, take all the parenthetical stuff and call it Q (just to give it a name). The equation is now $\frac{\sum Q}{N}$. Essentially, it is the mean of this Q variable. So variance is the mean of something. Now let's look at the parenthetical component. It's $(X - \mu)^2$. Forget the squared part, focus on $(X - \mu)$. This is called a mean-deviation score and it is the difference between a score on X and the mean score. If X equals the mean score, then the mean-deviation score is zero. If X is greater than the mean score, then the mean-deviation score is positive. You get the idea. We'll be computing mean-deviation scores for all people in our dataset. An example is presented below. The mean of X is 6.

Person	X	(X - Mean)
Bennett	3	-3
Tommy	9	3
Todd	4	-2
Matt	8	2

Now to deal with the squared part, we'll simply square those mean-deviation scores.

Person	X	(X - Mean)	(X - Mean)^2
Bennett	3	-3	9
Tommy	9	3	9
Todd	4	-2	4
Matt	8	2	4

Remember that Q thing we made up? That's the last column, the squared mean-deviation scores. As we said, variance is just the mean of this thing.

So variance is the mean of the squared mean-deviation scores. In this case, it's (9+9+4+4)/4 = 6.5. Another way to describe it: variance represents the average squared difference between each score and the mean. Here's another example.

Person	X	(X - Mean)	(X - Mean)^2
Julianna	9	0	0
Paul	9	0	0
Jennifer	9	0	0
Anthony	9	0	0
Brenden	9	0	0

Variance is – you guessed it – zero. Why? Every score is the same. Thus, the average distance between each score and the mean is zero. Just for fun, diagram the frequency distribution of this dataset.

Point Estimation: Variance

A quick review of the sample estimator for means. To estimate the population mean from sample data, we took the same formula that we used to compute population means and applied that formula to sample data. We can't exactly do that for variance because the population variance equation requires the population mean $(\Sigma(X-\mu)^2/N)$. And because we're stuck in sample-land we don't know the population mean – the sample mean is all we have. Not to worry, we'll just use the sample mean (which we already said is an unbiased estimator of the population mean) in its place. So to estimate the population variance from sample data you would think a simple modification to population variance equation to incorporate sample mean (i.e., turning this: $\Sigma(X-\mu)^2/N$ into this: $\Sigma(X-\bar{X})^2/N$) would get the job done. This approach seems logical and all, but such an approach would underestimate the population variance (if we only knew the population mean, this would not be a problem).

To obtain an unbiased estimate of the population variance (which has the formal symbol $\hat{\sigma}^2$ and will never be mentioned again) we must make an adjustment to the equation for sample data. That adjustment is a very simple change to the denominator of the equation so that N becomes $N-1$.

Thus, the sample (estimator of the population) variance takes the following form.

$$S_X^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

The above equation, popularly called the sample variance equation (with the symbol S_X^2), is what we use to obtain an unbiased estimate of the population variance from sample data.

Even though these are samples, standard deviation is still the square root of variance. Thus, the sample estimator of population standard deviation (with this never to be repeated symbol $\hat{\sigma}$) is the square root of the sample estimator of variance.

$$S_X = \sqrt{S_X^2}$$

And of course nothing changes just because we are dealing with measures of variability (variance, standard deviation) instead of central tendency (mean); the probability that these (unbiased) sample-based estimators will equal the population parameters is still zero. Sampling error affects all sample-based statistics.

Interval Estimation: Overview

Yes, $\bar{X}$ is an unbiased estimator of μ, but $\bar{X}$ will never equal μ. (As mentioned, the probability that $\bar{X} = \mu$ is zero; see discussion of continuous data for why.) Thus, our estimate of the population mean will always be incorrect. Always incorrect. Think about that.

(By the way, one counter argument to the above paragraph that I've heard is this: The sample statistic, or point estimate, is the *best estimate* of the population parameter. Best estimate? Best compared to what? Random guesses? Throwing darts? Is anyone arguing for those? As far as I can tell, the sample statistic is our *only* estimate, so it's best by default. If I'm the only one who shows up at the Olympics to run the 100 meter sprint, I am officially the fastest runner. I finished first in a field of one. Big win for me. So this *best estimate* claim is a hollow one and is a bit of a distraction from the real issue: Just how much error is likely inherent to this estimate?)

To have a probability greater than zero of locating a population parameter, we must create an interval around the sample statistic. This interval will have an upper and lower boundary that indicates the likely location of the population parameter. We don't know what the population parameter is, and we can't be 100% certain that it's in this interval. All we can say is that it is likely (how likely is up to us, but 95% is common) between these two boundaries. So instead of saying, "Here's my sample mean which has a zero percent chance of equaling the population mean," we say, "Here's my sample statistic and lower and upper values around it which give the likely location of the population mean."

We call these intervals confidence intervals. To form these intervals, we have to take a brief side trip to the standard error of the mean.

The Standard Error of the Mean

As you may have guessed by now, assuming the sample was collected in the right fashion (this is such a big issue, we'll wait to discuss it later), there is a relationship between sample size and the likely extent to which sampling error affects a mean. Greater sample sizes should lead to less sampling error (again, assuming that the sample was collected in the correct manner).

The statistic that describes how much error is likely associated with a given sample mean is called the standard error of the mean (and yes, if you're doing this for variance, there's a standard error for variance as well). The equation for the standard error of the mean (when the population standard deviation, σ, is known) is as follows.

$$\text{Standard Error of Mean} = \frac{\sigma_X}{\sqrt{N}}$$

That's it, just the standard deviation over the square root of the sample size. Definitely one of the simplest equations you'll see.

Interval Estimation: Confidence Intervals

So that answers the question of how much error we can expect with our sample estimate of the population mean. Now let's take that information and turn it into the answer to a better question: Given our sample mean (and the standard error of the mean), what is the likely population mean? To answer this question, we'll create an interval around the sample mean called a confidence interval (and now you can see why this section is called interval estimation).

How does the standard error of the mean help us form this confidence interval? Imagine that you drew a sample from a population and computed the mean on some variable. That's pretty easy to imagine since it describes a pile of studies. Now repeat this study over and over with the same sample size. Sampling error will affect every one of these sample means. Sometimes they will be less than the population mean; other times they will be greater. Do it enough times and these means will form a distribution of their own, called a sampling distribution. These sample means will be normally distributed with a certain mean and standard deviation. Here's the kicker: the mean of this distribution will equal the population mean and the standard deviation of these sample means will equal the standard error of the mean (not really, but go with this concept).

Now that we know that these sample means are normally distributed with a mean equal to the population mean and a standard deviation equal to the standard error of the mean, we can use some basic normal distribution principles and state that 95% of the sample means are less than 1.96 standard errors of the mean (i.e., $\pm 1.96 \times SEM$) away from the population mean (see Chapter 2 for a discussion of the normal distribution). Somehow (and I have never worked this out) this operation is reversible and we can say that for a given sample mean a bi-directional interval with a width that is 1.96 times the standard error of the mean (on each side) will contain the population mean 95% of the time.

The confidence interval for a mean when population standard deviation (σ) is known is computed as follows for the bi-directional or 2-tailed interval.

$$CI = \bar{X} \pm \left(z_{1-\frac{\alpha}{2}}\right) \frac{\sigma_X}{\sqrt{N}}$$

The last part of the equation ($\sigma_X/\sqrt{N}$) is just the standard error of the mean. No problem there. That other thing next to it ($z_{1-\frac{\alpha}{2}}$) is likely confusing. Not to worry, that's just statistician language for "find the value on a z table (i.e., normal distribution table of values) that demarcates the upper $\alpha/2$ percentile from the top." So if α is .05 (our normal value for α; more on α in the next chapter), you would look for the z value that separates the top .025 (or 2.5%) of the distribution; that is, the value at the 1-.025 percentile, .975. The z score at the .975 percentile happens to be 1.96. (Note: If this interval were one-tailed, then just do the top α. That is, don't divide α by 2. The z score you want in that case is 1.65). The truth is that as long as $\alpha = .05$, the only two values you need to know from a z table are 1.96 and 1.65.

Because we are almost always computing a bi-directional 95% confidence interval, we can simplify the previous equation to this:

$$CI_{95} = \bar{X} \pm 1.96 \times SEM$$

Let's compute an example. We did a study where we collected data from 100 people randomly sampled from the relevant population. The mean score in our sample was 109. Finally, the population standard deviation is known to be 15. To compute the 95% confidence interval (bidirectional), simply plug in the values into the above equation.

$$CI_{95} = 109 \pm 1.96 \times \frac{15}{\sqrt{100}}$$

Which works out to...

$$CI_{95} = 109 \pm 2.94$$

So that's 106.6 on the low end and 111.94 on the high end.

Now let's make sure that we interpret this confidence interval properly. It would be a mistake to say that we are 95% confident that the sample mean is between 106.6 and 111.94. It's a mistake because we know the sample mean – we can be 100% confident that it's 109. There was never any doubt about the sample mean. We're trying to determine the population mean. Thus, the point of the interval is this: we are 95% confident that the *population mean* is between 106.6 and 111.94. (Although you never hear it put this way, the textbook accurate version of the statement is that the population mean will be between 106.6 and 111.94 95% of the time.)

So that's interval estimation. We make confidence intervals around sample statistics so that we can have a non-zero probability of locating the relevant population parameter. This is your regular reminder that we wouldn't have this problem if we just measured populations instead of samples. We don't do that of course, because it's far more convenient to measure samples. But that convenience comes at a price, sampling error. Finally, although we limited our discussion to the confidence interval for the mean, confidence intervals constructed for the purpose of identifying the likely value of the population parameter can be calculated for pretty much every statistic (e.g., variance, correlations, regression coefficients).

Independent and Identically Distributed

There is a concept in statistics that is rather difficult to explain. It refers to data being independent and identically distributed (shorthand *iid* – yes, lowercase, italicized, and without punctuation).

This *iid* thing a statistical term of art with a complicated meaning. For our purposes it is sufficient to say this when data are randomly sampled from a single population, the scores are *iid*. The *independent* part deserves special attention. Scores are said to be independent when one observation is not related to another. The simplest way to understand this is that each case of data comes from independent subjects. This means that if there are 30 scores averaged to compute a mean or whatever, there are 30 separate subjects providing one score each. There are not 15 subjects who are providing two scores each – those scores would not be fully independent (but you could average each subjects two scores into one score and have 15 independent cases...).

This *iid* concept is important because it's an assumption for just about every statistical test that we have. As mentioned, the good news is that when a very basic condition (data are randomly sampled from a single population) is met, we can rest assured that our data is *iid*.

A few other *iid* points of note. For a sample of *iid* data, the mean is $\bar{X}$, the expected value of $\bar{X}$ is μ, the variance of $\bar{X}$ is $\frac{\sigma_X^2}{N}$, and the standard deviation of $\bar{X}$ is $\frac{\sigma_X}{\sqrt{N}}$. Nothing in that previous sentence is new to us. That's what we get with *iid* data.

Central Limit Theorem

The central limit theorem states that for *iid* data the distribution of sample means approaches normality as *N* increases regardless of original distribution of *X*. This is quite big, so let's break it down. The central limit theorem is addressing the sampling distribution of sample means (i.e., $\bar{X}$), and it is saying that the distribution of these sample means will approach normality even if the distribution from which these samples are drawn is not normally distributed.

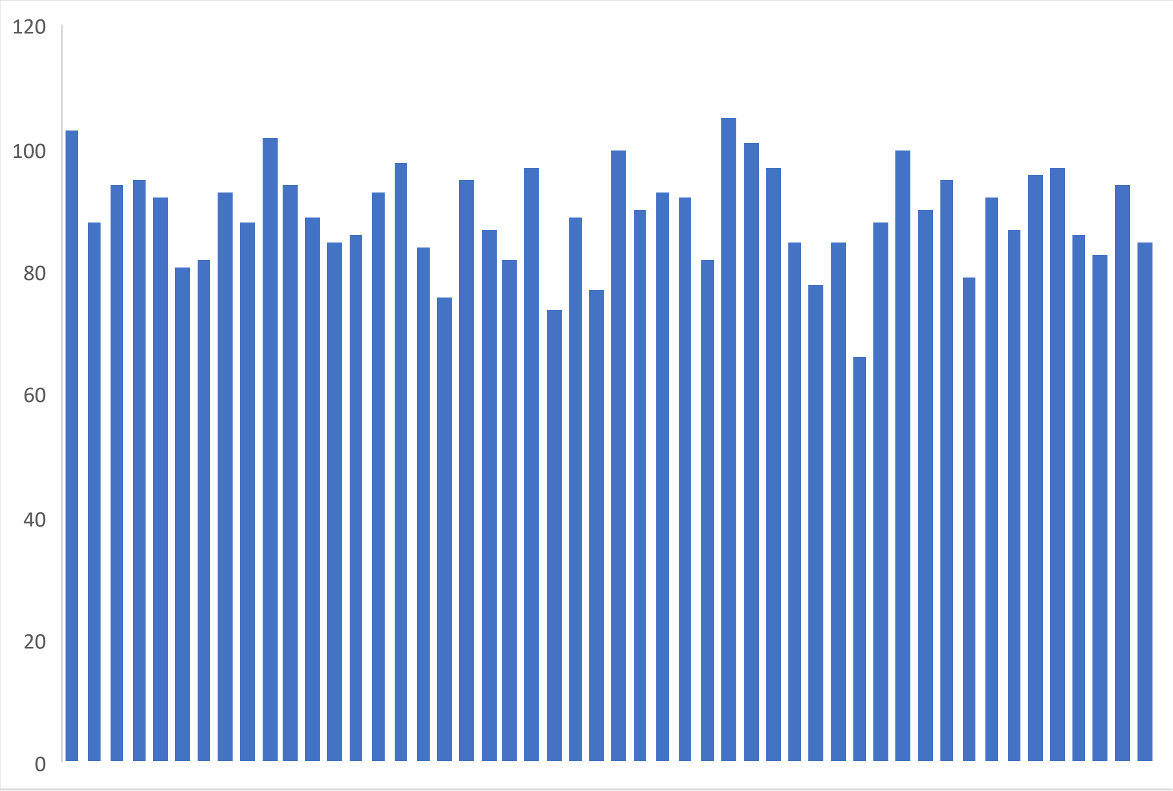
The reason why this is such a big deal is that you would think that if scores on *X* in the population are distributed in a serious non-normal fashion, then any sample means computed from those scores would also be non-normal. But that's not the case. The sample means will be normally distributed even for means based on small sample sizes. Studies have demonstrated that $N \geq 30$ is sufficient to make distribution of $\bar{X}$ approximately normal for any distribution of *X*. The central limit theorem is quite important as many statistical procedures assume normality. It's also quite powerful, as we will see in the next section.

The Central Limit Theory in Action

Here's a quick demonstration of the power of the central limit theory. Lottery numbers have an equal chance of being selected. (At least they had better have an equal chance or someone has some serious explaining to do.) Thus, the resulting distribution of the numbers chosen should have a rectangular shape, which is about as far from normally distributed as you can get with real data.

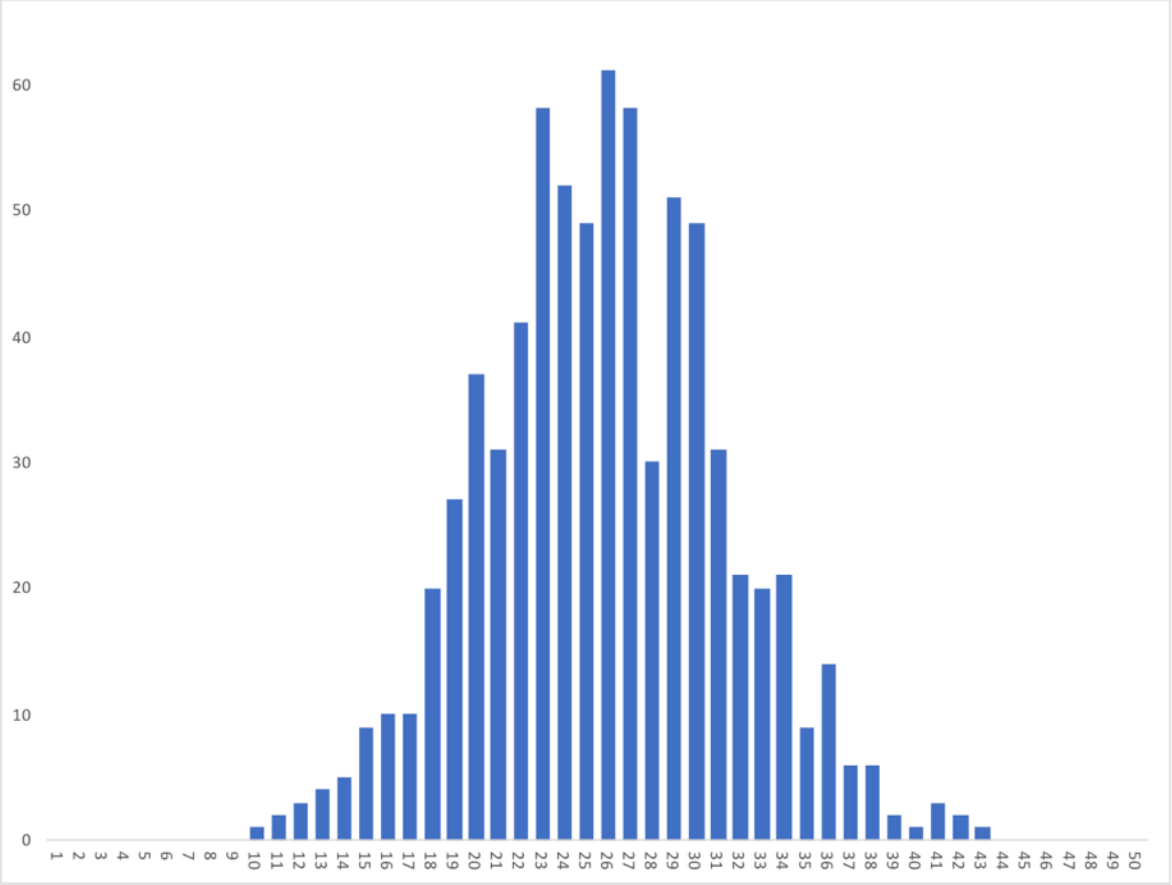
I pulled a record of the selected lottery numbers from a lottery database for an unnamed state lottery. These dataset includes all of the numbers drawn over a seven year stretch when the lottery had a simple 50-6 format (randomly pick six balls from 50). All told, 745 draws of 6 balls each resulted in 4,470 lottery balls picked. Figure 4 is the distribution of the numbers chosen. That's pretty close to a rectangular distribution. Nothing normal about it.

Figure 4 Frequency Distribution of Individual Lottery Balls (Numbered 1-50) from 745 Draws



Now watch what happens when I compute the mean value of the six numbers chosen in each drawing (Figure 5). We now have 745 sample means, each based on an *N* of 6 (a very small sample size). (This is analogous to a sampling distribution in which a study with *N* = 6 has been repeated 745 times.) The distribution shown in Figure 5 is not normal, but it's not bad. It's a world closer to normal than was the distribution of individual scores on *X*. And remember, these means were based on an *N* of only 6.

Figure 5 Frequency Distribution of the Mean Lottery Number (*N* = 6 for Each) Across 745 Draws



Sampling Distributions and You

The concept of sampling distributions deserves more attention. Here's what we said earlier (about two pages earlier).

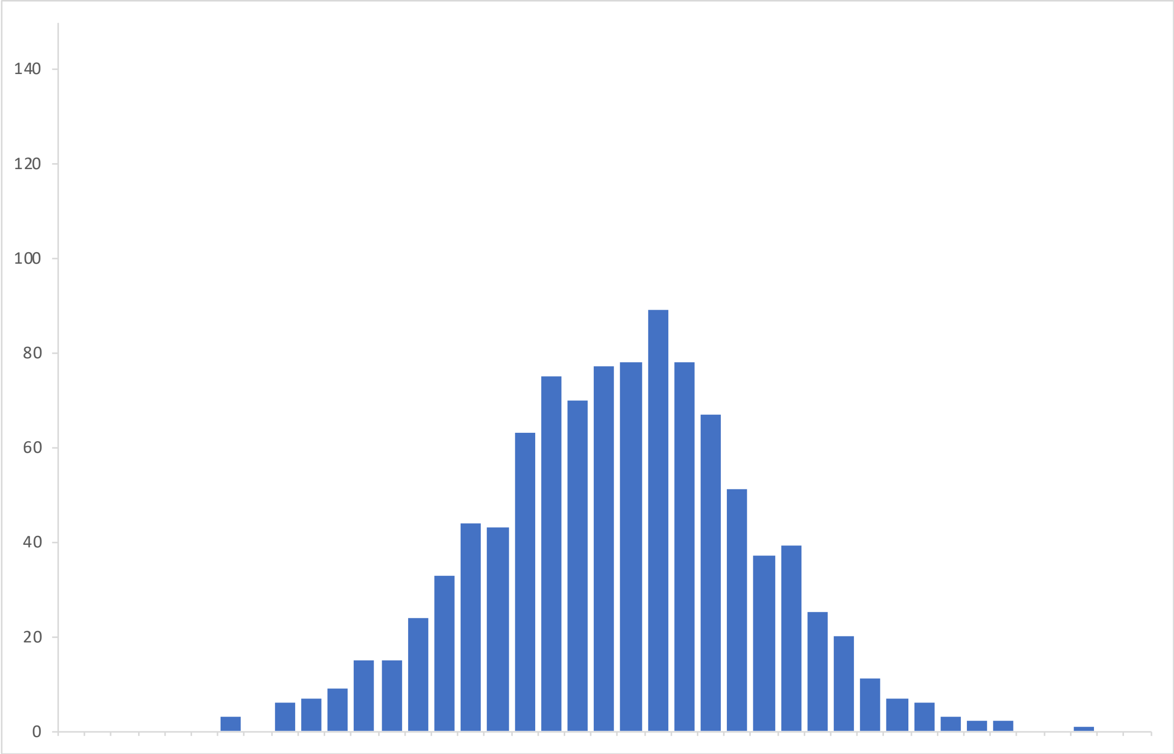
> Imagine that you drew a sample from a population and computed the mean on some variable... Now repeat this study over and over with the same sample size. Sampling error will affect every one of these sample means. Sometimes they will be less than the population mean; other times they will be greater. Do it enough times and these means will form a distribution of their own, called a sampling distribution. These sample means will be normally distributed with a certain mean and standard deviation. Here's the kicker: the mean of this distribution will equal the population mean and the standard deviation of these sample means will equal the standard error of the mean ($\sigma_X/\sqrt{N}$).

The summary of that paragraph is a sampling distribution is a distribution of sample statistics computed on repeated samples of size N. You can make one yourself right now via a coin-flipping study. Here's how: a fair coin will result in a heads 50% (or .5) of the time. This implies that we expect 50 heads in 100 tosses. So think of our sample size as 100 and our sample mean as #heads/100 (it's also worth noting that $\sigma = .5$ for a fair coin). So if you do that study (flip a coin 100 times), the expected mean is .5 (i.e., 50 heads out of 100 tosses). But do you think you'll get exactly 50 heads if you do this experiment one time? It's possible, but it's far more likely that you'll some other value close 50 (e.g., 48, 53, 51, etc.) because of sampling error. This is probably familiar to you already – you know about sampling error. (By the way, what is the size of the population for our coin-flipping study?)

So flip a coin 100 times and compute the mean (or proportion) of heads; there's your study done once. Now repeat that study, I don't know, 1,000 times. Plot those means. That's a sampling distribution. The mean of this sampling distribution (a mean of sample means) is μ, the population and the standard deviation of the sampling distribution is the population standard deviation divided by the sample size (i.e., $\sigma_X/\sqrt{N}$), something we know as the standard error of the mean.

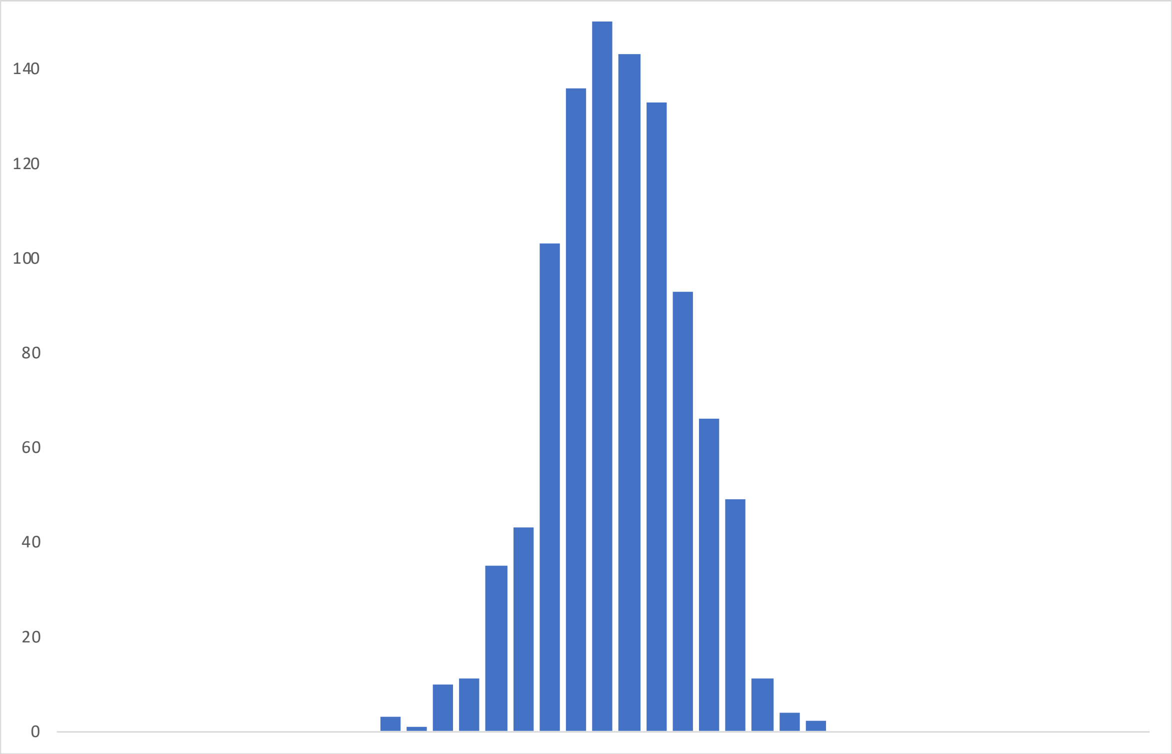
That last part has some major implications. In the version of the study we just described, σ was .5 and N was 100. Thus, the standard deviation of our sampling distribution is $.5/\sqrt{100} = .05$. If this sampling distribution is normally distributed, then we expect to see 95% of the sample means to be no more than $1.96 \times .05$ units away from the population mean in either direction. All well and good. Now do the study where the sample size for each experiment is 400 flips. Larger sample sizes mean less sampling error. So now the sampling distribution has a standard deviation of $.5/\sqrt{400} = .025$, effectively cutting the width of the sampling distribution in half. See Figure 6 and Figure 7 where the sampling distributions for this experiment are plotted with the same scale on the x-axis.

FIGURE 6 Sampling Distribution for Hypothetical Coin-Flipping Study with N = 100



The range of values for the x-axis is .3 to .7. The y-axis indicates frequency and ranges from 0 to 150.

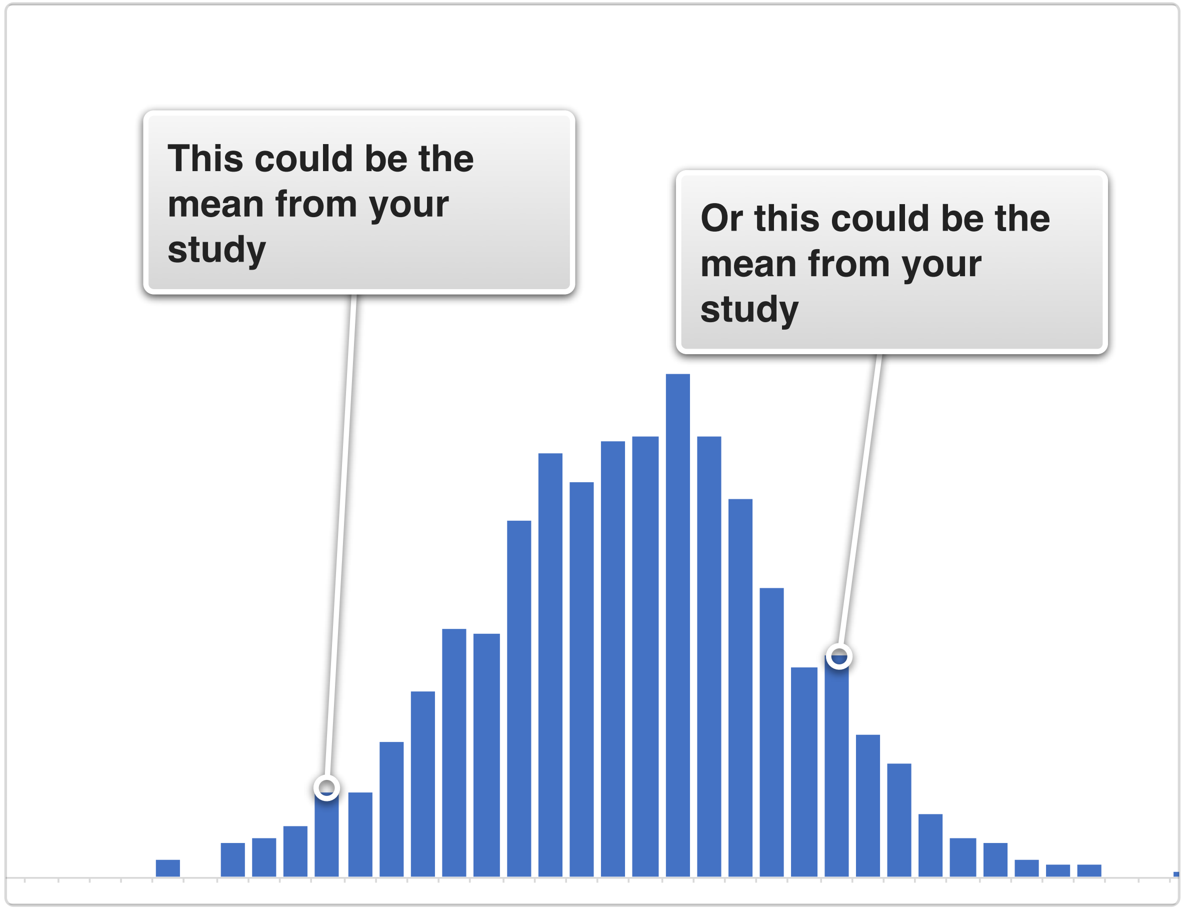
FIGURE 7 Sampling Distribution for Hypothetical Coin-Flipping Study with N = 400



The range of values for the *x*-axis is .3 to .7. The *y*-axis indicates frequency and ranges from 0 to 150.

Now here's the kicker; when you do a study one time, that's just one result in a sampling distribution. Your mean is probably in the thick part of the distribution, making it moderately representative of the population mean. Probably. But it could be anywhere (see Figure 8).

FIGURE 8 One Last Sampling Distribution Point



The population mean may be .5, but sampling error causes results for individual sample-based studies to be higher or lower than the population mean (with a zero probability that any one result will exactly equal the population mean).

This worst part of all of this sampling error/sampling distribution stuff is that short of repeating your study many, many times, you'll never know how well a sample statistic represents the population parameter. Depressing, right? There is at least one bit of good news we can take away from this: as *N* increases, the spread of this distribution decreases, which increases the probability that a sample mean is close (in an absolute sense) to the population value.

Null Hypothesis Significance Testing

4

Not having it would be worse than having it

Why Is This Necessary in Life?

You do a study. You get a result. The treatment mean is greater than the control mean. Case closed, right? The treatment works.

Not so fast, you say. There are about a thousand threats to internal validity (see Cook & Campbell, 1979, for an exploration of these many threats). Any one of these could cause the treatment mean to be greater than (or less than or simply different from) the control group mean. But even if the study were magically free from all of these confounds, the fact remains that we didn't measure the entire population and the difference between the two means could be simply due to sampling error. That is, in reality the treatment has no effect at all, but the treatment mean is greater than the control mean due to purely random effects of sampling error. (Remember this common example of sampling error: Flip a fair coin ten times. Ten is the sample size; infinity is the population size. More often that not, you will observe some result other than the expected value of five heads in ten tosses.)

OK, fine. We'll need a way to deal with sampling error. We need a way to determine just how big our observed differences in mean scores needs to be in order to conclude, yeah, these results are probably not due to sampling error. Something like: if there were no effect for our treatment, a result the size of the one we observed would occur very rarely simply due to sampling error.

Hypothesis Testing Logic

Null Hypothesis Significance Testing (NHST), the logic of which is known as the Fisher Inferential Model (inference: a conclusion drawn based on some idea, evidence, reason,… something), is our preferred method for dealing with sampling error. It's far from perfect, but NHST (or something like it) needs to be conducted to help us determine when sampling error is versus is not a likely cause of our results. Even the proposed alternatives to NHST (e.g., confidence intervals) end up looking like the NHST model (and yield the same conclusions).

Before getting into the model, we need to define two terms. First, there is the null hypothesis (symbol: H_0). The null hypothesis is effectively the opposite of our study hypothesis. We should always hypothesize that there will be an effect (e.g., treatment mean is greater than control mean). The null hypothesis states that there is no effect (e.g., the treatment mean is not greater than the control mean – it may be equal to it or lower than it, but it's not greater). The second important term is the null distribution. The null distribution is a sampling distribution that has the characteristics of the null hypothesis; it's the sampling distribution that would be observed if the null were true.

The Model:

A. Data (D) from experiment is observed. This is easy enough to follow. Do an experiment. Get data.

B. Assume that there is no effect and determine the sampling distribution reflecting this lack of an effect. That is, assume that the null hypothesis is true and define the null distribution. Answer question: How unusual would this result (the one that I found) be in a null distribution?

C. If the answer is, "The result would be very rare" (i.e., results like this occur less than 5% of the time in a null distribution), then infer that the observed data did not arise from a null distribution and reject the null hypothesis. Accept the alternative hypothesis (H_1), which is (I know this sounds messed up) our actual hypothesis.

You might wonder if there is no effect, then wouldn't all of the results be the same? How could a result be considered rare in a null distribution? The answer is sampling error. Even if the null is true, some results are common and some are unusual due to sampling error.

One final note before we proceed concerns the *how rare in a null distribution is rare enough?* issue. This standard for *unusual enough in a null distribution that we can reject the null* is called α (i.e., alpha) and is chosen by the researcher. Sort of. Much like Henry Ford's line about paint colors on the Model T ("You can have any color you want as long as it's black."), you can choose any value for α that you want as along as it's not greater than .05. An α of .05 became industry standard on a long time ago and despite the occasional volley of journal articles arguing for less stringent alphas (it's easier to reject the null hypothesis for an α of .10 than .05), the consensus hasn't changed. You can choose to have a more stringent standard than .05 (.01 is somewhat common), but you can't relax the standard beyond a 5% chance of obtaining a given result in a null distribution.

NHST Example: Fisher's "Tea" Test

Perhaps one of the early examples of this NHST logic will help. The format is a bit strange compared to our experiments, but the logic is the same. It's something called Fisher's "Tea" test.

> Research question: Can a person identify whether milk was added to a cup before or after the tea? (This was the contentious issue of the day.)
>
> Experiment: Eight cups of tea, four with milk added first, four with milk added last. Subject tastes all eight cups and identifies the order of the milk/tea in each.
>
> Hypothesis Testing Logic: If a person is unable to tell the order of milk/tea (i.e., the null hy-

pothesis), then our subject's choices are indistinguishable from random choices. In this experiment there is only a 1/70 chance of identifying all four cups correctly if picked at random. Thus, if we test a person, and this person does correctly identify all eight cups, and we compare this result to a null distribution of data, this result would be very rare. In this event, we would infer that our subject was not picking cups at random. The alternative that we would accept is that our subject is capable of tasting the difference.

Stated in terms of the NHST structure:

A. Observed data: All cups identified correctly.

B. If cups picked at random, 100% correct is an unusual result.

C. Inference: Cups were not picked at random.

NHST Example: One Group Experiment

The study discussed here is a simple one group experiment in which there probably isn't a manipulation. It might be conducted simply to find the mean value of some variable for a defined population. For example, a researcher might want to know if the mean intelligence for students at a certain type of school is greater than the mean for all students of that age. Or it could be about whether students at a certain university have higher placement test scores than the national average. Or a thousand other questions.

The study (we'll do the university study) is conducted as follows. A single sample of placement test scores are collected (with a probability sampling technique) for students at Enormous State University and are compared to some known population mean (many standardized tests have their mean scores set to a certain value). The study hypothesis can be directional (i.e., one-tailed) or non-directional (i.e., two-tailed). We'll go with one-tailed for this example, making our hypothesis: Placement test scores for students at ESU are greater than the national average ($\mu_{ESU} > \mu_{national}$). We could have hypothesized it the other way, but we think the ESU kids are above the national value.

Null Hypothesis: Placement test scores for students at ESU are not greater than the national average ($\mu_{ESU} \leq \mu_{national}$). In other words, the ESU mean is equal to or less than the national average. Either way, they aren't greater.

Note that the null and alternative hypotheses are stated in terms of population parameters (the population means in this case). We use what is known (sample data) to draw inferences about what is unknown (population parameters). It would be silly to formulate our hypotheses around sample data (i.e., the sample mean for this group is greater than some other value); there's no need to test that with an inferential test – we know what the sample mean is (there's not really much of an inference there).

In NHST structure:

A. A difference between the means (the university sample vs the national data) is observed.

B. In a null distribution, differences of this magnitude are very rare (or not, but in this example we're saying that they are).

C. Inference: These data did not come from a null distribution in which there are no differences between the groups. Reject null hypothesis (university mean is not greater than national average) and accept alternative hypothesis (the university mean is greater than the national average).

NHST Example: True Experiment

The classic setup is as follows: two groups, treatment and control, with random assignment to groups. The study hypothesis can be directional (i.e., one-tailed) or non-directional (i.e., two-tailed). We'll go with two-tailed for this example, making our hypothesis: the treatment group's mean score on the dependent variable does not equal the control group's mean score ($\mu_1 \neq \mu_2$). We don't know if the treatment mean will be greater than or less than the control mean; we are just hypothesizing that the two means won't be the same.

Null Hypothesis: Treatment mean equals control mean ($\mu_1 = \mu_2$); in other words, there is no difference. (Still yet another way of stating this: $\mu_1 - \mu_2 = 0$.)

In NHST structure:

A. A difference between the means is observed.

B. In a null distribution, differences of this magnitude are very rare (or not, but in this example we're saying that they are).

C. Inference: These data did not come from a null distribution in which there are no differences between the groups. Reject null hypothesis (no differences between groups in population) and accept alternative hypothesis (yes, there are differences in the population).

Once again, the null and alternative hypotheses are stated in terms of population parameters (the population mean in this case).

What It Means When We Reject the Null

When we reject the null hypothesis (because the results of our study are unusual in a null distribution), we are doing a probability analysis. This “occurs less than 5% of the time” doesn’t tell us the probability that the null is true (or false) – rather, it tells us that a result like the one that we have would be improbable (less than 5%) if the null were true.

NHST-Speak

There are many different ways to say the same thing in NHST-land. Significance test results are often reported in the form of a *p*-value, where *p* is the probability that a result would be found in a null distribution. Assuming that $\alpha = .05$, then obtaining a *p*-value of less than .05 from a significance test means we reject the null hypothesis. So, reject the null = significant result (yet another way to say this) = $p < .05$. To the converse, failing to reject the null = non significant result = $p >$ .05.

To complicate matters, statistics programs like SAS and SPSS don’t report significance tests results in any of the preceding forms. Rather, they report an exact *p*-value. For example, in addition to the test statistic, stats programs will say $p = .03$. It’s up to you, the user of the program, to translate that into a conclusion: Because .03 is less than .05, I reject the null. Use caution though; blindly following *p*-values can lead to incorrect conclusions. Stat programs default to performing two-tailed (i.e., non-directional) tests. Furthermore, when there is a one-tailed option, the program doesn’t know which direction you hypothesized and assumes that it’s whatever direction the results show. *Caveat lector* indeed.

Sampling Theory, Techniques, and Issues

Before we discuss sampling techniques, first consider the type of research that we conduct. Quite a bit of our research can be placed into one of just two categories: random sampling experiments and random assignment experiments.

These are not formal names, just descriptions of how these experiments should operate.

Random sampling experiments. The main issue in a random sampling experiment concerns measuring the characteristics of the population as it exists. In this type of study there isn't a treatment being tested; we are simply trying to learn about characteristics of the population via samples. Here are two examples, a medical example and a social science example. What percent of people in a geographic region possess antibodies for a certain virus? What are college student attitudes toward university administration? In this type of study, everything depends on how the sample was collected. (Fun note: the single biggest research fraud of the 20th century concerned a failure to collect samples the right way for what should have been a simple random sampling study.)

Random assignment experiments. The main issue in a random assignment experiment concerns whether a given manipulation has an effect. Two more examples follow. Does this vaccine reduce the rate of spread of a certain virus? The social science version is very familiar: Are dependent variable scores higher for the treatment group than for the control group? (e.g., Does a training program improve employee performance?) For this type of research a failure to randomly sample from the relevant population isn't a threat to the internal validity of random assignment experiments (or to NHST), but it does limit their generalization. The key issue with this type of study is, as the name states, random assignment to groups.

As to how we collect our samples (i.e. sampling techniques), there are many types, but we can categorize them as being a probability sampling technique or a non-probability sampling technique.

Non-probability sampling. A non-probability sample is one that was not collected via a probability sampling technique. These are samples of convenience as well as volunteer samples. Such samples are wholly inappropriate to a random sampling experiment. Everything in those experiments depends on the sample representing the population, and there is no amount of statistical jujitsu that can fix this. Pedhazur and Schmelkin (1991, p. 321) said it best: “The incontrovertible fact is that, in non-probability sampling, it is not possible to estimate sampling errors. Therefore, validity of inferences to a population cannot be ascertained.” In addition, the problems that arise from using non-probability samples cannot be fixed via sample size. Larger samples are no more likely to be representative of the population than are smaller samples when the sample is a volunteer sample. What is true of the sample of volunteers cannot be validly inferred to be true of the population.

Probability sampling. There are a number of ways to collect a probability sample. The easiest to understand is simple random sampling in which each member of population has an equal likelihood of being sampled. Pedhazur and Schmelkin again (1991, p. 321): “With probability sampling, the various sample outcomes occur with known probabilities.” Thus, because each person in the population has the chance of being a member of the sample, we can estimate the probability that the sample characteristics represent the population characteristics.

Problems with NHST

1. NHST assumes that data were collected via a probability sampling technique. This means that for certain types of studies (studies where the entire point of the study is to understand population characteristics), significance testing is just a charade, and NHST is an empty process performed out of a belief that the process confers some protective power. A meaningless ritual. I know I’m sugarcoating it here, so I’ll be direct: If a re-

searcher is doing the sort of research that can be fairly categorized as a random sampling study and this researcher fails to collect data via a probability sampling technique, then there is no way to know how well the sample represents the population. Significance testing is just a sham, and the results are no different than anecdotal data.

2. The probability that a result (D) is found in a null distribution does not equal the probability that the null is false/true: $\Pr(D|H_0) \neq \Pr(H_0|D)$; see Cohen (1994). In other words, the unlikeliness of finding a given result in a null distribution does not mean that the null distribution is unlikely. Believe me, if we could estimate it the other way, we would. Let me state this a few ways to clarify. I'll make pairs of statements; the first one is correct, and the second is the all-too-common incorrect version.

What the *p*-value in NHST actually means ($\Pr(D|H_0)$): The probability that a given result is observed given the that the null is true (i.e., probability it will be observed in a null distribution).

What we hear/think it means/want it to mean ($\Pr(D|H_0)$): The probability that the null is true given our result.

When $p < .05$, NHST actually indicates: There is less than a five percent chance that a result like mine would be found in a null distribution (i.e., found if the null is actually true).

When $p < .05$, we hear/think/wish NHST indicates: There is less than a five percent chance that the null is true given my result.

Like I said, the second (and incorrect) version is far better, and if we could do it that way, we would. We can't because estimating that probability requires information which we do not have.

3. The point-null hypothesis is always false. There are two types of null hypotheses:

> Point-null: there is no difference between the groups ($\mu_1 = \mu_2$), the population correlation is zero ($\rho = 0$), etc. Tested with two-tailed significance test.
>
> Directional null: the experimental group will not have a higher score than the control group ($\mu_1 \leq \mu_2$), the population correlation is not greater than zero ($\rho \leq 0$), etc. Tested with one-tailed significance test.

So why is the point-null always false? Because of what we know about continuous variables (the probability that a continuous variable equals a specific value is zero), the probability that two continuous variables will equal each other (i.e., $\mu_1 \leq \mu_2$) is also zero. In other words, there will always be some (probably small) real difference between the groups, but that doesn't mean that the difference is due to our hypothesized cause.

As if that's not bad enough, the directional null fares only slightly better. If the direction is chosen at random, given a large enough sample size, the directional null will be rejected 50% of the time.

There is good news here though – there are slight adjustments to the process that we can make to eliminate this problem. These adjustments are a bit beyond the scope of this book, but they do exist (although they are hardly ever used form some reason).

NHST Silliness

Remember, this is the Fisher Inferential Model – we are using the data and the null distribution to draw an inference about the null hypothesis (we either reject it or we do not reject). When we have a result so unusual in a null distribution that we infer the result did not come from a null distribution (i.e., we have a significant result) we reject the null hypothesis. There can be no degree of rejection – it's a dichotomous decision: reject null or fail to reject null. Therefore, it's the height of idiocy to talk about degrees of significance: if we reject when $p < .05$, then we can't reject it even more ("I strenuously reject") when $p < .01$ (an even rarer result).

Final Thought on NHST

Significance tests exist to address sampling error – and nothing else. Rejecting the null hypothesis means that we are saying that the results we obtained are probably not due to sampling error.

That's all they do. Ruling out sampling error as a probable cause of our results is a big deal, but there are many other types of errors that could have caused our results. So don't invest too much significance (sorry) in the favorable result of a significance test.

One Sample Tests

5

Actual significance tests

One Sample Tests on a Mean

One sample tests are for experiments in which we collect data from a single sample; we then compare the sample mean to some reference population mean. These are not true experiments, in which there are two samples (commonly, a treatment and control). There is just one sample of data, and we make inferences from characteristics of this sample to a relevant population.

Formal Testing Procedure

In this section, we outline the official set of steps we will take when conducting our significance tests. These steps will be repeated in upcoming chapters as we learn new tests. At some point, we'll stop listing them because they are well known to us.

i . State H_0 and H_1

ii . Determine rejection region and critical value

iii . Compute test statistic

iv . State conclusion

The details of each of these steps will be given in an example in the next section.

One Sample, σ Known

This significance test is often referred to as a z test because the test statistic that it produces is a z value and it employs a z table (see Chapter 2 for more on the z tables) to determine significance.

Characteristics: *z* test; can be one- or two-tailed; designed to use information from a sample mean to draw inference about population mean; population standard deviation must be known. Test statistic:

$$z_{obs} = \frac{\bar{X} - \mu_c}{\sigma / \sqrt{N}}$$

(Does that denominator look familiar? I'm sure I've seen it somewhere before.)

The transformation of the sample mean into this test statistic (z) allows us to enjoy the benefits of a common, simple null distribution of scores as this particular test statistic is distributed as $z(0,1)$, the standard normal distribution. This is a big deal as it saves us a great deal of trouble.

Example: Test the hypothesis that scores for a given school are greater than the national average ($\mu_c = 500, \sigma = 100$). Data: a sample of 25 students at the school had a sample mean ($\bar{X}$) of 505. (Notation issue: Most presentations of this test refer to all of the populations means with the same symbol, μ. But there is often a difference between the population of interest to our study and the population to which we compare ours. We'll denote the former as μ_s and the latter as μ_c.)

i. List null and alternative hypotheses

H_0 = The school mean is not greater than the national average ($\mu_s \leq 500$).

H_1 = The school mean is greater than the national average ($\mu_s > 500$).

Note: I find it easier to list the alternative hypothesis first as it is the same as our actual study hypothesis. Once you have H_1 listed, then turn it into the "not" version to obtain H_0.

ii. Determine rejection region and critical value

I think the toughest part of this process is determining the rejection region. So let's spend some time on that. The rejection region refers to the part of the null distribution that is the "unusual result if the null is true" area.

To determine the rejection region, ask yourself: If the null hypothesis is true, where would the unusual result be found in a null distribution? Because our null hypothesis states that the mean will be less than or equal to 500, then an unusual result if this null hypothesis were true would be the opposite of that, a very high mean. Thus, our rejection region will be the top 5% (for the α =.05 case) of the null distribution. We will explore other rejection regions in future examples (not to spoil things, but there are really only three possibilities: the upper tail of the distribution, the lower tail, or both tails).

To demonstrate where the rejection region actually starts, called the critical value, I created a population dataset ($N = 1{,}000{,}000$) having a mean of 500 and a standard deviation of 100. I then generated a sampling distribution by randomly sampling 25 cases from this population, computing the mean of these 25 cases, repeating until I had 10,000 sample means. (This is what I do in my spare time.) The distribution of these means (i.e., the sampling distribution) is shown in Figure 1.

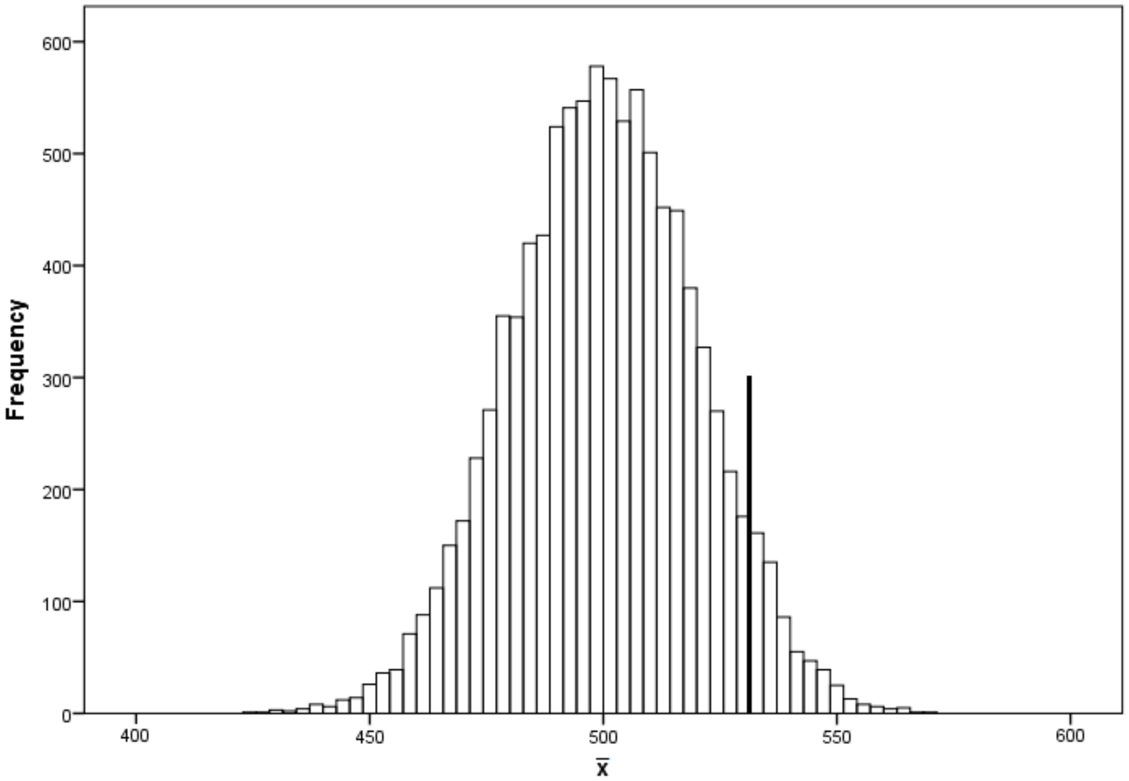


FIGURE 1 Sampling Distribution for 10,000 Sample Means (N = 25) from Population with μ = 500, σ = 100

Inspection of Figure 1 indicates that some of the sample means are as low as 450 and other are as high as 550. Bear in mind that all of these samples come from a population where the population mean is 500. Samples come with sampling error. It's unavoidable. Back the distribution; the value at the 95th percentile (bold line) is 532.97. Thus, based on this sampling distribution a sample mean must be greater than 532.97 to be in the top five percent of sample means (which doesn't look good for our data). To summarize, Figure 1, a sampling distribution, is our null distribution. The rejection region is the top 5% of sample means, and this region starts at 532.97. Our critical value is 532.97, meaning that we will reject the null if our actual sample mean is greater than 532.97. Why? Because such a mean would be very unusual (occurring less than 5% of the time) if the null hypothesis is true, and standard NHST logic (Chapter 4) says that when a result is unusual in the null distribution, then we reject the null hypothesis.

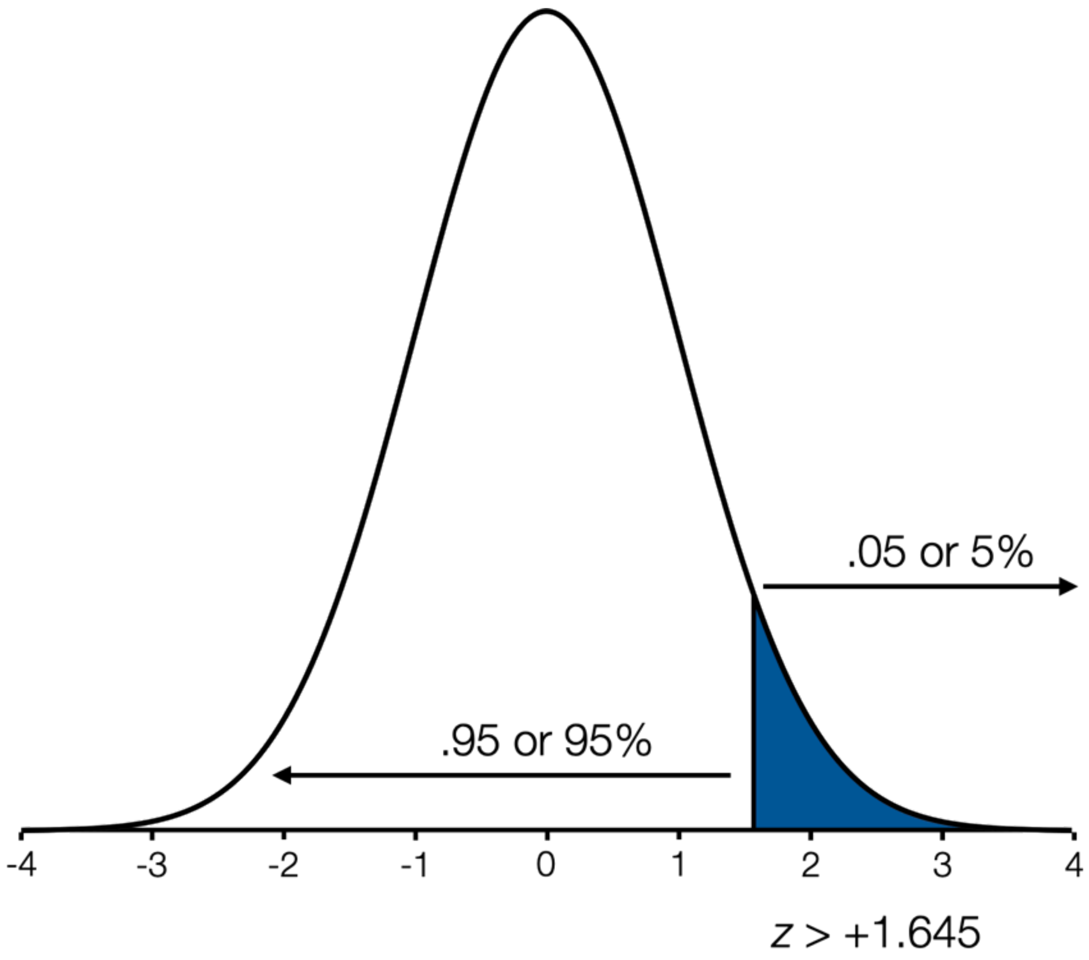
At this point you are probably thinking that this seems like an impossible amount of work every time we conduct a significance test. And you'd be right to think that. We don't do it this way. Due to the magic of the CLT (and some other principles) we can use various pre-existing distributions (e.g., the normal distribution) for significance testing. It's no accident that the test statistic is structured to give us a z value. This structure produces a sampling distribution that is normally distributed with a mean of zero and a standard deviation of one (i.e., the standard normal distribution). The upshot is that we get to simply consult a z table to find the critical value. So if you want to diagram the rejection region for this test, just use a normal distribution, as shown in Figure 2. For the rejection region, find the z table value at the .95 level (z = 1.645 in this case).

So now we have it, a rejection region (Figure 2) and a critical value (1.645). Let's state it for the record.

> We will reject the null hypothesis if z_{obs} is greater than 1.645.

iii . Compute test statistic

FIGURE 2 Upper 5% of a Normal Distribution



$$z_{obs} = \frac{\bar{X} - \mu_c}{\sigma/\sqrt{N}}$$

$$z_{obs} = \frac{505 - 500}{100/\sqrt{25}} = \frac{5}{100/5} = \frac{5}{20}$$

$$z_{obs} = +.25$$

$iv.$ State conclusion

Given that our obtained z of +.25 is not greater than our critical z of +1.65, we are unable to reject the null hypothesis. (Side note: people often say this as "we failed to reject the null." Like it was our fault because we didn't try hard enough or something.)

Why can't we reject the null? A quick check of a full z table indicates that a z score of +.25 is at the 60th percentile ($pr(z > .25) = .60$); samples with means as low as that occur 40% of the time when the null is true meaning these results are not unusual at all. In other words, if we draw a sample of 25 people from a population with a mean of 500 and a standard deviation of 100, we would observe means at that level or higher 40% of the time.

Let's do another example of the same type. Everything is the same as before (sample size = 25, population standard deviation is 100, the population mean to which we compare our sample mean is 500). The differences are in our sample mean, which is now 552, and our hypothesis. Our new hypothesis is the two-tailed variety: This school's test scores are different from national average.

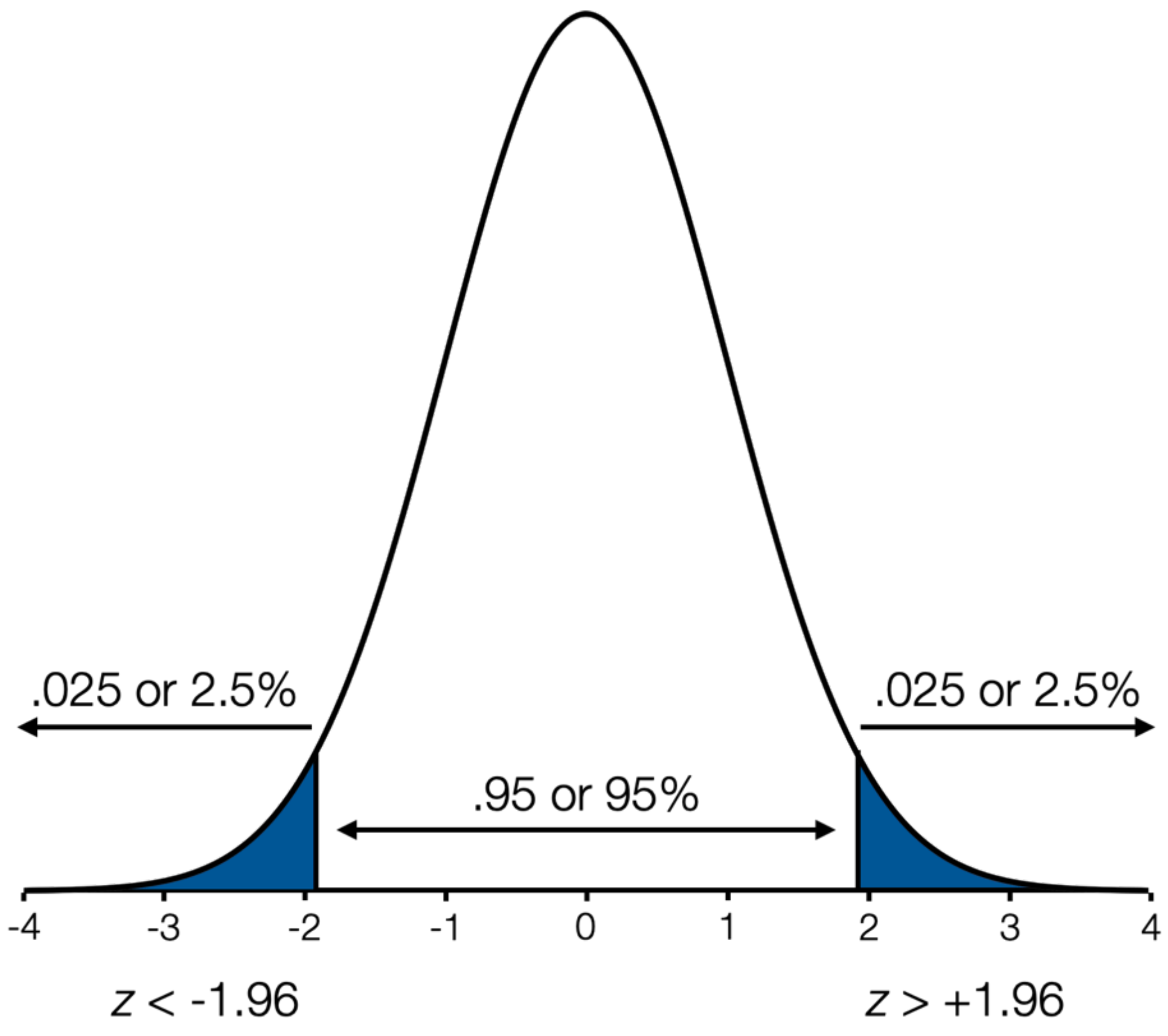
i . List null and alternative hypotheses

H_0 = The school mean is not different than the national average ($\mu_s = 500$).

H_1 = The school mean is different than the national average ($\mu_s \neq 500$).

ii . Determine rejection region and critical value

FIGURE 3 Two-Tailed Rejection Region



The null hypothesis says that the school mean equals the national average of 500. So if we observe a sample mean that is exactly 500, then that's perfectly consistent with the null. What's unusual if the null is true? The unusual values are extremely high or extremely low. And since α = .05, we'll divide that 5% equally between the upper and lower tails of the distribution, giving us the rejection region shown in Figure 3.

As for critical values, a check of the z table shows us that the upper and lower 2.5% of a normal distribution begin at +1.96 and -1.96. Thus, ...

We will reject the null if our test statistic is greater than 1.96 or less than -1.96 (it really helps to think of the diagram for this).

iii . Compute test statistic

This is the easy part. Just plug in and solve.

$$z_{obs} = \frac{\bar{X} - \mu_c}{\sigma/\sqrt{N}}$$

$$z_{obs} = \frac{552 - 500}{100/\sqrt{25}} \qquad = \frac{52}{100/5} \qquad = \frac{52}{20}$$

$$z_{obs} = +2.6$$

iv . State conclusion

Because our z_{obs} is greater than the critical value (2.6 > 1.96), we reject the null hypothesis and accept the alternative hypothesis that the school's mean score is different from the national average of 500.

Thoughts on Null Hypothesis Significance Testing

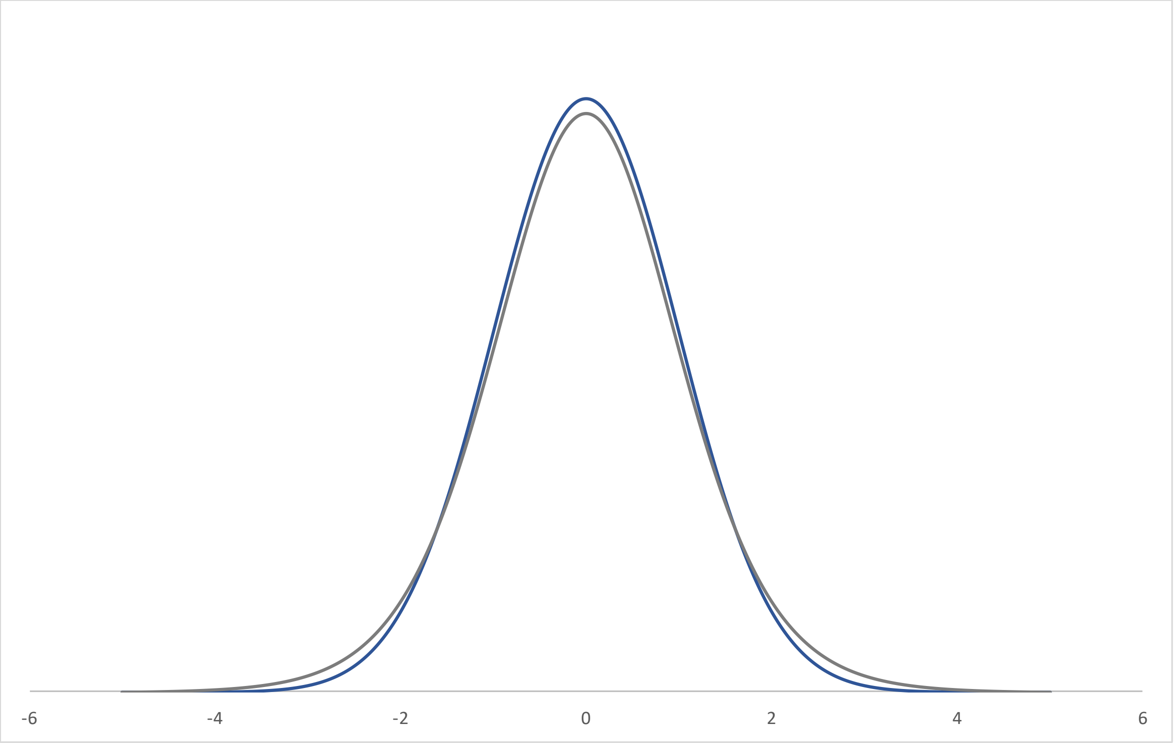
As stated, we are using sample data to make inferences about unmeasured population data. NHST exists to address sampling error only. If we measured the entire population, we wouldn't need to do this strange procedure called NHST.

We are using sample data to choose between H_0 and H_1. We are not computing which one is more likely. Rather, we start with H_0, and, if we judge it to be unlikely, we reject it and conclude H_1 is correct. It's not an H_0 versus H_1 death match – it's a survival game where H_1 simply waits for H_0 to fall in order to be crowned champion.

Student's t Distribution

For no reason whatsoever, let's take a break from this significance testing to discuss a new distribution of data. We already know (and love) the z and the normal distribution. There is a test statistic that we will encounter soon called a *t* statistic. The sampling distribution for this *t* statistic is a family of distributions called Student's *t*, which we'll call the *t* distribution for short. (By the way, there is an interesting story involving a brewery behind the use of the pseudonym "Student".) The easiest way to understand the difference between a normal distribution and a *t* distribution is to examine the tails. Both distributions are unimodal and symmetric (i.e., bell shaped), but the *t* has more scores in the tails (less in the middle) as compared to the normal. See Figure 4 for a comparison of the two distributions.

FIGURE 4 Normal Distribution vs Student's *t* Distribution



The blue curve is the standard normal distribution, and the gray curve is Student's *t* distribution with 10 degrees of freedom. Note the greater probability that a score will be found in the tails of the *t* distribution as compared to the normal.

I mentioned that the *t* distribution is a family of distributions. This is true; it's not just one distribution but a set of distributions that vary on a single parameter called degrees of freedom (*df*). Degrees of freedom is related to *N* size (equations will be given in due time but bigger *N* means bigger *df*). Furthermore, as degrees of freedom in-

Table 1 Student's *t* Table for $\alpha = .05$

df	*t* (1-tail)	*t* (2-tail)
5	2.015	2.571
6	1.943	2.447
7	1.895	2.365
8	1.86	2.306
9	1.833	2.262
10	1.812	2.228
20	1.725	2.086
30	1.697	2.042
40	1.684	2.021
50	1.676	2.009
100	1.66	1.984
200	1.653	1.972

crease, the *t* distribution begins to resemble the normal distribution. They equal each other when $df = \infty$ and are almost indistinguishable when $df > 200$. An abbreviated *t* table is given in Table 1.

One Sample, σ Unknown

Now that we've concluded our discussion of the *t* distribution, back to significance testing. We have covered a test on a mean where the population standard deviation is known. Let's move to a more realistic situation, one in which the population standard deviation is unknown. The setup for the test is otherwise the same as before: one sample in which we use the sample mean to make inferences about an unknown population mean: Is this population mean greater than, less than, or simply not equal to some reference value (a comparison population mean)? Because we don't know the population standard deviation (σ), we will have to use the sample standard deviation (S_X) in its place. This one change in the test statistic causes a new problem (this sort of thing always occurs in statistics): the sampling distribution of the test statistic is no longer normal. This means that we can't use a z table to determine the critical value. You guessed it, we will have to use the *t* distribution.

Characteristics: *t* test; can be one- or two-tailed; designed to use information from a sample mean to draw an inference about population mean; population standard deviation is unknown. Test statistic:

$$t_{obs} = \frac{\bar{X} - \mu_c}{S_X/\sqrt{N}}$$

Distributed as $t_{(N-1)}$

What is this $t_{(N-1)}$ business? That's saying that the test statistic we compute is distributed as a *t* with *N*-1 degrees of freedom.

I think it's clear that aside from sample standard deviation in the denominator and the *t* distribution instead of the normal distribution, everything is the same as before. Same test. A little less known about the population.

Example: Test the hypothesis that a certain aluminum recovery facility achieves recovery rates less than the industry target of 890 pounds for every half ton of scrap aluminum (i.e., $\mu_c = 890$)? Data: 16 runs of aluminum resulted in a mean recovery of 830 pounds (per half ton) with a standard deviation of 96 pounds.

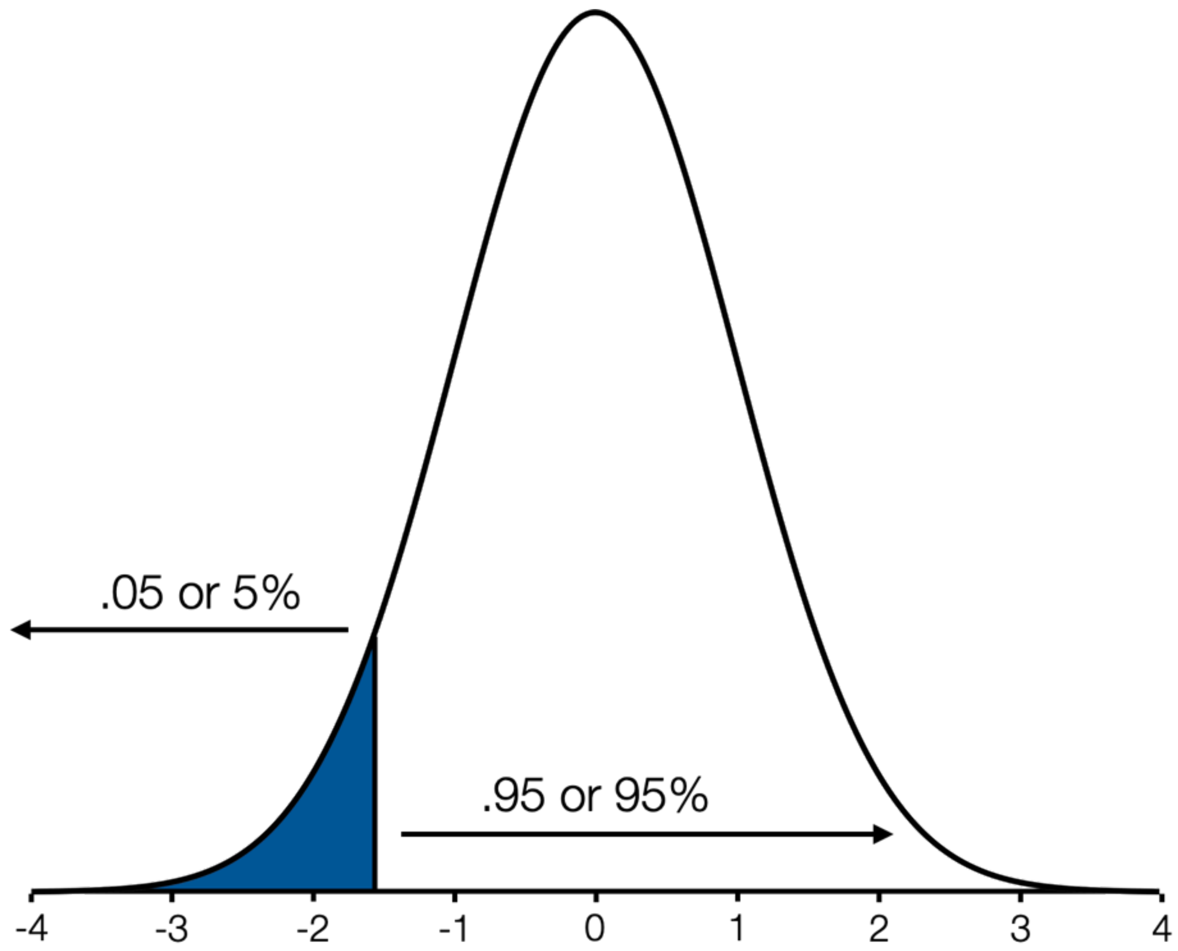
i . State null and alternative hypotheses.

H_0 = The mean rate is not less than the industry standard ($\mu_s \geq 890$).

H_1 = The mean rate is less than the industry standard ($\mu_s < 890$).

ii . Determine rejection region and critical value

FIGURE 5 Rejection Region, Lower Tail



The hypothesis had a direction (*lower* recovery rate), so this is a one-tailed test. But which tail? The null says greater than or equal to 890, so anything equal to the mean or greater than the mean is perfectly consistent with the null. The unusual result if the null is true is the lower tail (see Figure 5). That's our rejection region. What about the critical value? N is 16, so $df = 15$. A check of a proper t table ($\alpha = .05$, 1-tailed, lower tail) yields a value of -1.753 (the tables offer only positive values; make them negative for the lower tail).

Let's state it formally: the rejection region is the lower tail; we will reject the null if the t_{obs} is less than the t_{crit} of -1.753.

iii. Compute test statistic

$$t_{obs} = \frac{\bar{X} - \mu_c}{S_X/\sqrt{N}} = \frac{830 - 890}{96/\sqrt{16}} = \frac{-60}{24}$$

$$t_{obs} = -2.5$$

iv. State conclusion

Because t_{obs} (-2.5) is less than t_{crit} (-1.753), we reject the null hypothesis and conclude that this aluminum recovery facility achieves a recovery rate lower than the industry average of 890. Sounds like they need new management.

Confidence Intervals for a Mean, Population Standard Deviation (σ) Known

Good news: we already covered this in Chapter 3. The equation is listed below.

$$CI = \bar{X} \pm \left(z_{1-\frac{\alpha}{2}}\right) \frac{\sigma_X}{\sqrt{N}}$$

No need to dwell on it. Let's get to something new.

Confidence Interval for Mean, Population Standard Deviation (σ) Unknown

Has it occurred to you that it's odd that we don't know the population mean yet we do know the population standard deviation? I mean, how did we learn the population standard deviation but not the mean? Was someone in the process of telling us both statistics but was interrupted before they could get the mean out? (I'm picturing a spy movie exchange on a foggy, cobblestone street somewhere in Europe. "The population standard deviation is 15 and the mean is..." -poison dart hits person, stopping him in mid-sentence. Yes, my mind goes to strange places.)

The truth is that there are standardized tests that have a set population mean and standard deviation; however, we may wish to know the mean for a differently defined population (i.e., not all test takers, but all test takers from a geographic region). In such a situation, it is reasonable to think that the mean of this particular population will be different from the national average, whereas the standard is likely the same. It's reasonable to think that, but it may not be true.

We need to have a way to compute this confidence interval for situations in which we don't know the population standard deviation, either because there is no national set standard or it is unwise to conclude that this set value holds in our particular population. Given that we just discussed a significance test for this exact situation, you can probably guess what changes we'll make the the confidence interval equation that we already have. That's right, swap out the population standard deviation with the sample standard deviation and the normal distribution with the *t* distribution (with $N - 1$ degrees of freedom).

$$\text{CI} = \bar{X} \pm \left(t_{(N-1),1-\frac{\alpha}{2}}\right) \frac{S_X}{\sqrt{N}}$$

Everything else is the same as before. Same interpretation too (we are 95% confident that the population mean is in the interval of...).

Assumptions of Both Tests and CI

Our significance tests are only as solid as the assumptions supporting them. If the assumptions are not supported, then the results of the tests may be meaningless. To be clear, you can do the test and get a result. You'll always get a result. Maybe it will be significant, maybe not. But that result will not necessary match reality if the assumptions of the test are not supported. So it's fairly important that we discuss assumptions for our tests.

The two tests that we have covered in this chapter (as well as the confidence intervals) have the following assumptions. First, data are *iid* (see Chapter 3 more on *iid*). The big issue with *iid* concerns sampling method. We discussed sampling theory in Chapter 4. For the tests described in this chapter everything (and I mean everything) depends on how the data were sampled. Use a probability sampling technique, and it's possible to justify inferences about a population made with these tests. Fail to use one, and you have forfeited your ability to claim that the results are representative of the population from which they were sampled – which is the entire point of these two types of tests.

Another assumption of these tests is that the distribution of scores on the dependent variable (i.e., X) is normal in the population. There is some flexibility on this issue, meaning that slight departures from normality are not likely to affect the results of these tests.

Finally, t tests also require that a transformation of the variance ($\frac{(N-1)S^2}{\sigma^2}$) is distributed as a chi-square (the chi-square distribution is an entirely different distribution and will be covered later), which may not be the case if the distribution of the dependent variable is not normal. However, as long as the distribution is unimodal and symmetric, then the t test is likely on solid ground.

Other One Sample Tests

6

Sometimes a two sample tests is really just a one sample test in disguise

Other One Sample Tests

The previous chapter covered two types (σ known and σ unknown) of one sample tests on a mean. This chapter offers two more one sample tests, but they differ from the previous ones in important ways. One of them looks like a two sample test even. But it's really a one sample test in disguise. First up is how we handle a standard one sample test on a mean where the dependent variable isn't a continuous variable.

One Sample Tests on a Proportion

In the previous chapter, the dependent variable for our one sample tests on a mean was a continuous variable. As mentioned, a continuous variable is one with an infinite number of possible scores between any two points. There is, however, another very common type of dependent variable that is not continuous but rather is dichotomous. The only two possible scores on this variable are 1 and 0. Think of dependent variables which measure outcomes with only two possible states: success/failure, yes/no, has the disease/doesn't have the disease, graduates/does not graduate, pass/fail, and other unpleasant examples that exist in medical research. These are common dependent variables, and they require an adjustment to the equations used for significance testing.

The mean of a dichotomous variable coded as 1 or 0 is a proportion. Proportions range from 0 to 1. If the mean of a dichotomous variable is .30, then we know that 30% have a score of one and 70% have a score of zero. A study with this type of dependent variable is called a binomial experiment and the hypothesis concerns the proportion of people who are a 1 on the dependent variable.

As for terminology, we will use p (in various forms) to stand for proportion. There will be sample versions of p ($\hat{p}$ = sample proportion, computed as $\hat{p} = r/N$, where r indicates success) and populations versions of p. As with continuous dependent variables, we'll make a distinction between the study-relevant population (p_s = study-relevant population proportion) and the population value against which we compare it (p_c = comparison population proportion value). Just think of $\hat{p}$ as analogous to $\bar{X}$, p_s as analogous to μ_s, and p_c as analogous to μ_c. If that's still confusing, an example will clear it up.

Characteristics: z test; can be one- or two-tailed; designed to use information from a sample proportion to draw inference about the relevant population proportion. Due to the magical properties of dichotomous data, we don't have to worry about whether the population standard deviation is known; if we have hypothesized population proportion, then we know the population standard deviation. Can make all of the usual tests: Is the population proportion greater than some value ($p_s > \ p_c$), less than some value ($p_s < \ p_c$), or not equal to some value ($p_s \neq \ p_c$).

Test statistic:

$$z_{obs} = \frac{\hat{p} - p_c - .5/N}{\sqrt{p_c(1 - p_c)/N}}$$

Note: the $.5/N$ part is called a continuity correction and is necessary. For some reason it is overlooked (i.e., not included) in many published versions of this test. It's important and is required for accuracy. Unimportant but mentioning because it's interesting: Remember how I said that if we have a hypothesized population proportion, then we know the population standard deviation? For a dichotomous variable, variance is $p(1 - p)$, the square root of that gives you the standard deviation; and you know what p is, the mean of a dichotomous variable. Thus, there is a direct relationship between the mean and variance (and standard deviation too, of course) for dichotomous data. If you check the denominator of the test statistic, you'll see the square root of $p(1 - p)$ where the standard deviation goes.

Example: Test the hypothesis that the majority of college students use a tablet computer. (You have no idea how many times I have seen this "majority" type of hypothesis for a dichotomous data. I haven't counted, but it's a lot.) Data: Of 51 students surveyed, 31 use a tablet computer.

i. State null and alternative hypotheses.

The word *majority* in this hypothesis indicates greater than 50%. So we know that this is a one-tailed test. We also know that the hypothesized population value (p_c) is .50.

H_0 = The proportion of students who use a tablet computer is not greater than .50 ($p_s \leq .50$).

H_1 = The proportion of students who use a tablet computer is greater than .50 ($p_s > .50$).

ii . Determine rejection region and critical value

This is a one-tailed test, and we know from the test statistic that this will be distributed normally. And given our null that states the proportion will be less than or equal to .50, the unusual result is a really high proportion (as mentioned, a proportion equal to .50 or less than .50 is completely consistent with the null – not unusual at all). So we want the top 5% of the normal distribution. As we all know by now, the z scores that starts the top 5% of a normal distribution is 1.645.

We will reject the null if the z_{obs} is greater than 1.645.

iii . Compute test statistic

As always, this is easy part. First calculate the sample proportion, $\hat{p}$.

$$\hat{p} = \frac{r}{N} = \frac{31}{51} = .608$$

Worth noting that the sample proportion is great than .50. So the results are in the right direction. But are they unusual enough for us to reject the null? On to the test statistic.

$$z_{obs} = \frac{\hat{p} - p_c - .5/N}{\sqrt{p_c(1 - p_c)/N}}$$

$$z_{obs} = \frac{.608 - .5 - .5/51}{\sqrt{.5(1 - .5)/51}} \quad = \quad \frac{.098}{.07}$$

$$z_{obs} = 1.4$$

iv. State conclusion

Because the z_{obs} (1.4) is not greater than the z_{crit} (1.645), we are unable to reject the null hypothesis.

This one sample test on the mean of a dichotomous variable (i.e., a proportion) is underused in research. Too often, novice statisticians use the standard one sample test on a mean of a continuous variable when their dependent variables are dichotomous. Don't make that mistake. Side note: You should learn this test as I am not aware of its inclusion in any computer stats program.

Dependent Samples Tests on a Mean (i.e., Dependent Samples t Tests)

There is a two sample test that is really a one sample test in disguise. It's a special situation where there are two groups of subjects (more accurately, two sets of dependent variable scores) that are not independent. You could even say that they are dependent. There are two ways for that to happen.

One way is to measure the same group of subjects twice. This design is common in a pre-test/post-test design in which subjects are tested before a treatment (e.g., training program) and then tested again after the treatment. The hypothesis concerns whether the scores changed after the treatment. It should be clear that if you measure the same set of people twice, they are not independent.

The second version of this study actually uses two separate groups of subjects. These subjects are carefully chosen so that they are matched on some variable (e.g., matching on age, for every person in Group 1 with a 28 on the matching variable, there is a person in Group 2 with a 28). Typically, a treatment is given to one group with the matched group serving as a control. Scores on the dependent variable are then compared to see if the means differ by group. This matching is a form of statistical control and is vastly inferior to randomly assigning people to groups. But sometimes our hands are tied by circumstance, and we are unable to randomly assign. Make lemonade out of lemons, right?

In both versions of this study, rather than having two independent groups of subjects (with distinct sample sizes) we have N paired observations (where the observations are *iid*; that is, they are *iid* within groups, but dependent between groups).

The hypotheses that we test are familiar. Using the pre-test/post-test version for our examples, the population mean from what is typically the pre-test (μ_1) vs. the post-test (μ_2) is greater ($\mu_1 > \mu_2$), lesser ($\mu_1 < \mu_2$), or just different ($\mu_1 \neq \mu_2$).

Now, here's the clever part: Because the observations are paired, and because we are interested in whether there is a difference between the scores from the two groups, we can define a new dependent variable:

$$D_i = X_{1i} - X_{2i}$$

Going with the pre- post-test scenario, we subtract the pre-test score from the post-test score, giving us a difference score for each person. Thus, a difference score of zero means that there was no change; positive difference scores mean that the post-test scores are greater than pre-test scores; negative difference scores imply the opposite.

So why is this clever? Instead of having two dependent variables (a pre- and post-test score), we now have just one dependent variable, the difference score, making our analysis a one-sample test. We simply test whether the difference score is greater than zero ($\mu_D > 0$), less than zero ($\mu_D < 0$), or different from zero ($\mu_D \neq 0$).

Characteristics: t test; can be one- or two-tailed; designed to use information from two dependent samples (which are combined to form a single difference score for each observation, thus, becoming a single variable) to draw inferences about the relevant population mean; population standard deviation is unknown. Given that the null hypothesis will always involve some variation of $\mu_D = 0$, the numerator of the test statistic simplifies somewhat.

Test statistic:

$$t_{obs} = \frac{\bar{D}}{S_D\sqrt{N}}$$

With $df = N - 1$

Example: At the start of the semester, five students were timed in the mile run. After a two-week training session, these same five students were timed in the mile again. Test the hypothesis that the training session improved (i.e., lowered) their times. Data for the experiment are listed in Table 1. For the difference ($D = Pre - Post$) scores, the mean and standard deviation are as follows: $\bar{D}$ = 23.4, S_D = 22.

i. State null and alternative hypotheses.

H_0 = The difference in times (pre minus post) will not be greater than zero ($\mu_{Diff} \leq 0$).

H_1 = The difference in times (pre minus post) will be greater than zero ($\mu_{Diff} > 0$).

Table 1 Pre-Test, Post-Test and Difference Score Times (in Seconds) for Running Experiment

Subject	Pre	Post	Difference (Pre-Post)
1	512	497	15
2	697	692	5
3	811	758	53
4	723	683	40
5	603	599	4

ii . Determine rejection region and critical value.

If the mean difference score is zero, that's perfectly consistent with the null being true. So is a negative difference scores. Only a very high mean difference score is inconsistent with the null. Thus, our rejection region will be the top 5% of the distribution. Because this is a *t* test with $N - 1$ degrees of freedom, the critical value is 2.132.

We will reject the null hypothesis if the t_{obs} is greater than 2.132.

iii . Test statistic:

$$t_{obs} = \frac{\bar{D}}{S_D\sqrt{N}} \quad = \quad \frac{23.4}{22\sqrt{5}} \quad = \quad 2.38$$

iv . Because t_{obs} (2.38) is greater than t_{crit} (2.132), we reject the null hypothesis and conclude that the training course did improve times (i.e., the mean difference score is greater than zero in the population; $\mu_{Diff} > 0$).

Problems with Difference Scores as a Variable

Difference scores, although interesting as a dependent variable, have a few issues. First, under certain conditions, difference scores can be unreliable variables even if both the pre- and post-test scores are themselves highly reliable variables. Classical Test Theory reliability is beyond the score of this book (there's probably another book for that), but it is sufficient to say that an unreliable variable is not desirable.

Another problem with difference scores is that they have a nasty habit of being negatively correlated with pre-test scores. Think about what that means for our running study. People who have great times in the pre-test have little room for improvement (their difference score is likely to be low), whereas people with poor times on the pre-test have massive room for improvement (although not all will show big changes). This issue complicates results and may influence how you set up the experiment.

Independent Samples Tests on a Mean

7

We have arrived

Two Independent Samples

We have reached the point where we are discussing the sort of significance tests that are associated with true experiments (which have, at the very least, a treatment group and a control group with random assignment to groups). This chapter introduces the simplest version of this research design. The rest of this book will expand on this tiny kernel of an idea with a series of increasingly complicated twists. So there's something to look forward to.

Tests on a Mean: Two Independent Samples

We are no longer comparing a mean from a group to a real or hypothetical population mean (e.g., the known national average); we are comparing the means of two independent groups to each other. We have two sample means and want to compare them to infer something about their respective population means.

We have data from two *iid* samples, often a treatment and a control. If you want this experiment to have any internal validity, people should be randomly assigned to groups. We wish to test whether the population means are greater ($\mu_1 > \mu_2$), lesser ($\mu_1 < \mu_2$), or just different ($\mu_1 \neq \mu_2$).

(With this research design, you begin to see how our notion of a population starts to become more hypothetical than real. If I have a treatment and control group, I want to generalize from their sample means to an unmeasured population. But how can the entire population be in both the treatment, in which they experience some sort of treatment, and the control, in which they don't? You might attempt to thread this particular needle by saying that we are generalizing to half of the population having the treatment and the other half not. But half of the population isn't a population – it's a very large sample. See the dilemma? The generalization is more hypothetical now. It's best understood as follows. *If* the entire population were in the treatment, this is their mean, and *if* the entire population were in the control, this is their mean. The one sample tests could actually happen. We could actually measure the entire population. Not so here.)

As for how we test the difference between independent samples, there are three versions of this test, all varying on the population standard deviation issue. Population standard deviations are: known, unknown but assumed equal, and unknown not assumed equal.

Two Samples, σ Known

Characteristics: z test; can be one- or two-tailed; designed to use information from two sample means to draw inference about relevant population means; population standard deviations must be known.

Test statistic:

$$z_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example: No example for reasons that will follow.

The big question with this test is: How is that we know the population standard deviations, but we don't know the population means? This is not like the one sample z test days where the population means and standard deviations could be hypothetical values ("the national average is…") – these are real groups in our experiment. The best answer I can offer here is that maybe our dependent variable is a standardized test with a set standard deviation. The problem is that value may hold for the control group, but if our treatment has some crazy strong effect, I lack confidence that it holds for the treatment group. Long story short, you'll never use this test. So we won't either.

Two Samples, σ Unknown But Assumed Equal (i.e., "Pooled" t test; Independent Samples t test)

This is the test. This is the one that you are likely to encounter many times in your research. When people say that you should conduct an independent samples *t* test, this is the one they mean.

In this test, we don't know the population standard deviations for each group, so we use a weighted average of the sample standard deviations as an estimate of the population value (and if you've been paying attention, that makes this a *t* test). We assume that the population standard deviations are the same for both groups, which may or may not be a good idea.

Characteristics: *t* test; can be one- or two-tailed; designed to use information from two sample means to draw inference about relevant population means; population standard deviations are unknown but are assumed to be equal. Test statistic:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Note: S_p^2 is called the pooled variance estimate; S_1^2 and S_2^2 are the sample variances for each group. Also note that we are now making a distinction between the total sample size (N) and the number of people per group (n_1 and n_2).

Finally:

$$df = n_1 + n_2 - 2$$

Example. Test the hypothesis that dependent variable scores are greater for the treatment group than for the control group (shorter version: treatment mean is greater than control mean). Data:

treatment: $\bar{X}_1 = 38, S_1^2 = 14, n_1 = 25$; control: $\bar{X}_2 = 31, S_2^2 = 12, n_2 = 20$.

i . List null and alternative hypotheses

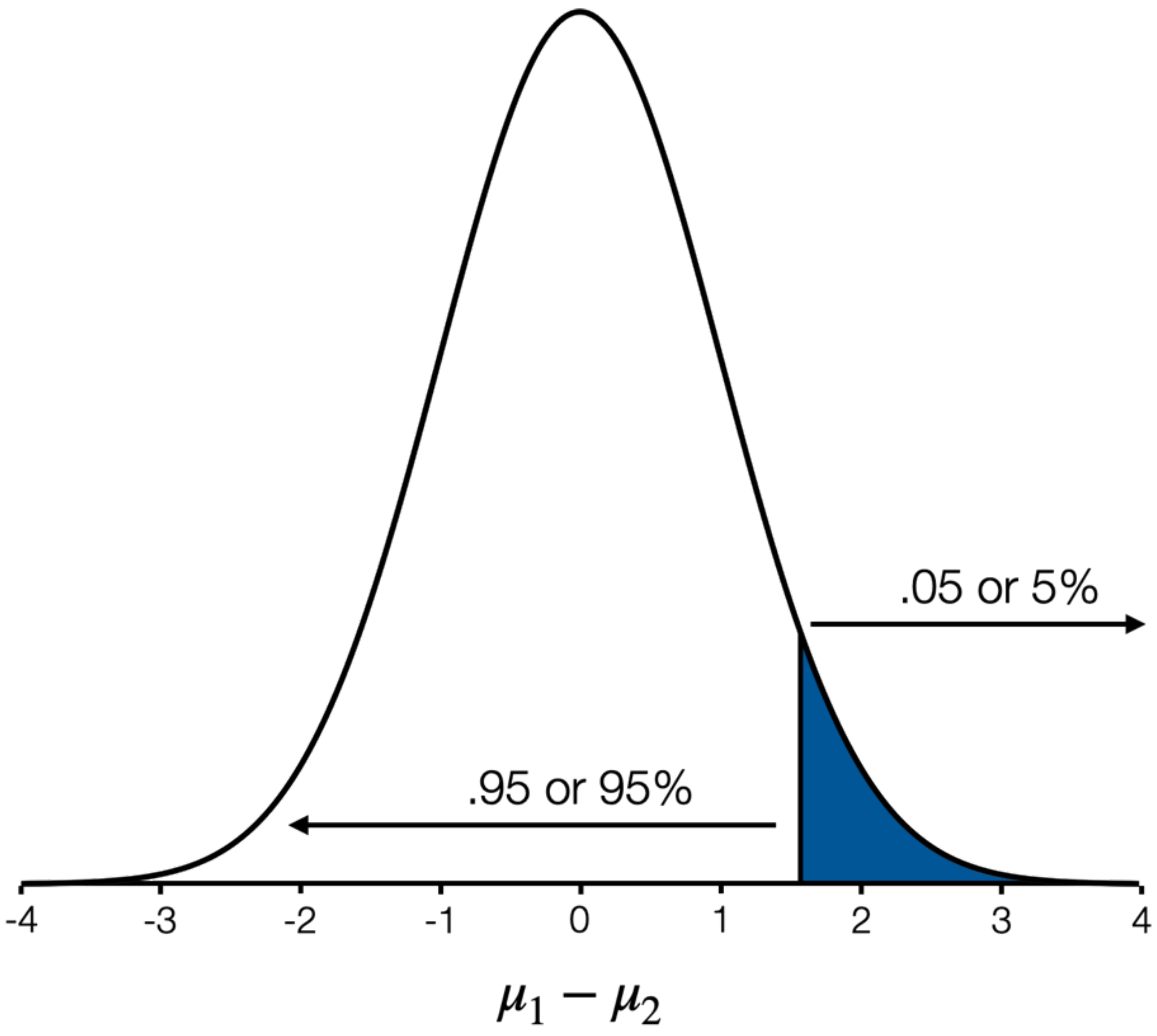
H_0 = The treatment mean is not greater than the control mean $(\mu_1 \leq \mu_2)$.

H_1 = The treatment mean is greater than the control mean $(\mu_1 > \mu_2)$.

ii . Determine rejection region and critical value

To understand the rejection region for this, you first must picture the null distribution. What does it look like here? There are two population means – are there two null distributions? No, we can make this into the same type of null that we know and love by rewriting our null hypothesis. Currently it's $H_0 : \mu_1 \leq \mu_2$. But what if did a little algebra and wrote it like this $H_0 : \mu_1 - \mu_2 \leq 0$? That seems like a legit algebra move. And now we can see the null as a mean difference. If the control population mean is greater than the treatment mean, then the difference between them (treat-

FIGURE 1 Null Distribution and Rejection Region for Independent Samples *t* Test Example



Note that null distribution is for the difference between population means (treatment minus control) in which H_0 is $\mu_1 - \mu_2 \leq 0$

ment minus control) will be negative. If they are equal, then the difference between them will be zero. And if the treatment mean is greater than the control mean, then the difference will be positive. That last case describes the unusual outcome if the null is true. Figure 1 displays the sampling distribution and rejection region for the example.

As for the independent samples *t* test critical value, $df = n_1 + n_2 - 2$. This example has df = 25 + 20 - 2 = 43. A check of a full *t* table shows that the critical value (α = .05, 1-tailed) for 43 degrees of freedom is 1.681.

We will reject the null if t_{obs} is greater than 1.681.

iii . Compute test statistic.

Okay, now it gets annoying. But at least it's rather straightforward. Just make sure you put variance into the S_p^2 equation, because that's what it calls for. If you insert standard deviations, you're in trouble.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$S_p^2 = \frac{(25 - 1)14 + (20 - 1)12}{25 + 20 - 2} = \frac{564}{43}$$

$$S_p^2 = 13.12$$

We said that S_p^2 is a weighted average of the sample variances. Thus it is no surprise that S_p^2 is 13.12, a value in between S_1^2 (14) and S_2^2 (12). You can use this principle as a check on your math, should you ever do this by hand.

On to the actual test statistic.

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{38 - 31}{\sqrt{13.12\left(\frac{1}{25} + \frac{1}{20}\right)}}$$

$$t_{obs} = \frac{7}{1.09} = 6.44$$

iv. State conclusion.

Because t_{obs} (6.44) is greater than t_{crit} (1.681), we reject the null hypothesis and conclude that the treatment mean is greater than the control mean in the population ($\mu_1 > \mu_2$).

Assumptions of Pooled t Test

A quick discussion of the assumptions of the pooled *t* test is in order. Many of these will be familiar. First, as the name *independent* samples *t* test implies, the two samples must be independent random samples. Thus, you cannot use the same subjects in both samples (i.e., a within-subjects design). You already know what test to use for that.

Second, dependent variable scores for each sample should be normally distributed. As before, it is sufficient for the distributions to be merely unimodal and symmetric.

Finally, this pooled *t* test assumes that populations standard deviations, although unknown, are equal. This condition is called homogeneity of variance: $\sigma_1^2 = \sigma_2^2$. Obviously, people can assume anything like they like. But when is it safe to make this assumption? The answer is that it is safe to make this assumption when (a) the sample variances are approximately equal (yes, there is a significance test of the variances that you can use for this) or (b) the sample sizes are approximately equal. If neither of those conditions are met, then you should use the next test.

Two Samples, σ Unknown and Unequal (Welch Approximate t test)

Characteristics: *t* test; can be one- or two-tailed; designed to use information from two sample means to draw inference about relevant population means; population standard deviations are unknown and are not assumed to be equal. This, like the “pooled” *t* test, is also an independent samples *t* test. Test statistic:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

That wasn’t so bad. In fact, it looks a lot like the z test from earlier. But wait until you see the degrees of freedom equation:

$$df = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$

The degrees of freedom is not likely to turn out to be an integer. In that case round down to nearest integer (i.e., 43.2 becomes 43).

The assumptions for the Welch approximate *t* test are the same as the pooled *t* test (other than not having the homogeneity of variance assumption). At least one researcher has argued that the Welch test should be used as the default version of the independent samples *t* test. However, it is worth noting that the Welch approximate *t* test isn’t as sensitive as the pooled *t* test. In other words, you are less likely to obtain a significant result with the Welch test.

Tests on a Proportion (Two Samples)

We introduced a one sample test on a proportion in the previous chapter. It was analogous to our one sample tests on a mean, but set up for a dichotomous dependent variable. You can guess what’s coming, a two sample test for the dichotomous dependent variable. This test allows us to compare two proportions from independent samples.

Characteristics: *z* test; can be one- or two-tailed; designed to use information from two sample proportions to draw an inference about the relevant population proportions. Due to the magical properties of dichotomous data, we don't have to worry about whether the population standard deviation is known. Can make all of the usual tests: Is one population proportion greater than the other ($p_1 > p_2$), less than the other ($p_1 < p_2$), or not equal ($p_1 \neq p_2$).

Test statistic:

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where:

$$\hat{p}_1 = r_1/n_1$$

$$\hat{p}_2 = r_2/n_2$$

$$\hat{p} = \frac{r_1 + r_2}{n_1 + n_2}$$

The symbols should be familiar. p_1 is the population proportion for Group 1. $\hat{p}_1$ is the sample proportion for Group 1 (where r_1 is the number of successes in the group and n_1 is the total number of people in the group). Finally, $\hat{p}$ is the combined sample proportion.

Example. Does a new student advisory program improve 4-year graduation rates? Data: 24 of the 60 students were assigned to the new program graduated in four years; 19 of the 50 students in the control group graduated in four years.

i. List null and alternative hypotheses

H_0 = The graduation rate is not higher for program students than for control group students ($p_1 \leq p_2$).

H_1 = The graduation rate is higher for program students than for control group students ($p_1 > p_2$).

ii. Determine rejection region and critical value

It should be clear by now that when null says less than or equal, the unusual result will be at the high end (i.e., upper tail). As this test statistic is distributed normally, the critical z is the familiar 1.645.

We will reject the null if z_{obs} is greater than 1.645.

iii. Compute test statistic.

Start with the sample proportions.

$$\hat{p}_1 = r_1/n_1 \quad = \quad 24/60 \quad = \quad .40$$

$$\hat{p}_2 = r_2/n_2 \quad = \quad 19/50 \quad = \quad .38$$

(Side note: This does not look good for our program. Yes, .40 is greater than .38, but it's not much better. Likely well within the range of what we would expect to see simply due to sampling error if the null is true.)

$$\hat{p} = \frac{r_1 + r_2}{n_1 + n_2} \quad = \quad \frac{24 + 19}{60 + 50} \quad = \quad \frac{43}{110} \quad = \quad .39$$

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z_{obs} = \frac{.40 - .38}{\sqrt{.39(1 - .39)\left(\frac{1}{60} + \frac{1}{50}\right)}} \quad = \quad \frac{.02}{.0934}$$

$$z_{obs} = .214$$

iv. State conclusion.

Because z_{obs} (.214) is not greater than z_{crit} (1.645), we are unable to reject the null hypothesis.

As with the one sample for tests on a proportion, statistics programs do not compute this two sample test. It's on you to do it.

Final Thoughts on Independent Samples Tests

We have now covered four different tests (three that you would actually use) for testing for differences from independent samples. That's two samples, often a treatment and a control. But what if your study has two treatment groups plus a control group? What do you do then? Run three of these tests in which you compare two groups at a time (Treatment 1 vs Treatment 2, Treatment 1 vs Control, Treatment 2 vs Control)? If you did that, how do you interpret the results if only one or two of them are significant?

Of course that's not the way to handle three independent groups. We need a comprehensive significance test that tests for differences among three or more groups with a single test. Maybe next chapter.

8

One Way ANOVA

The start of something big

Introduction

Many of our studies have three or more independent groups, and we need a way to conduct a single test for differences among these group means. This procedure, known as analysis of variance (ANOVA from now on), allows us to achieve that goal. In this chapter we will introduce the simplest form of ANOVA (one-way, fixed effects model). Future chapters will build on this foundation to introduce increasingly complex designs.

The ANOVA Model

We know how to test whether the means from two groups are different from each other ($H_0 : \mu_1 = \mu_2$). How are we going to handle three or more groups ($H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_a$; where a = number of groups)? To answer that, we need to examine the ANOVA model.

A score on the dependent variable (Y_{ij}) for a given observation (j) within a given group (i) can be broken into three components: an overall mean or grand mean (μ), an effect for group i (α_i), and a random component (e_{ij}). The use of Greek letters tells you that this is a population-level model.

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

The grand mean (μ), the mean across all observations (or across all group means) is a scaling constant and is fundamentally uninteresting. The effect for group i (α_i) is defined as the difference between the mean of that group and the grand mean: $\alpha_i = \mu_i - \mu$. This term describes the difference between groups in the population. If all groups have the same mean (meaning that there is no effect for the independent variables), then all of the group means will equal the grand mean, making $\alpha_i = 0$ for all groups. Bigger differences between the group means result in bigger differences between the group means and the grand mean. So

α_i is an index of the magnitude of the effect of the independent variable in the population.

The random component (e_{ij}) is distributed with a mean of zero and a variance of σ^2. This e_{ij} term describes differences in dependent variable scores within a given group. If all members of the control group have the same score, then $e_{ij} = 0$. Bigger differences within groups result in larger values for e_{ij}.

The fundamental principle of ANOVA is to evaluate the magnitude of between-group differences(α_i), which we hypothesize, against the magnitude of within-group differences (e_{ij}), which we're pretty much not a fan of. It's not a coincidence that the e_{ij} term is given the letter e for error as within-group differences are seen as a type of error. They may not be actual errors, just the result of real differences among people in a group. Because these differences are unrelated to our treatments, we find them annoying at best.

Null Hypotheses Revisited

ANOVA is used to test whether means differ among independent groups. There is no longer the possibility of testing for directional differences. (Imagine that you had eight groups – are you going specify all of the different ways in which the means could go?) In ANOVA-land there is only one issue to be tested: The means are either all the same, or they are not all all the same (i.e., at least one group's mean is different from the rest). Because we always hypothesize differences (hypothesizing sameness is a thread to validity; see Cook and Campbell again), our study hypothesis will always be the same: differences. The null will always be the same: no differences. Thus, the null hypothesis can be written as

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_a.$$

We can revise it to be in terms of the model. If H_0 is true, then $\mu_i = \mu$ for every group, which can be re-written as: $\mu_i - \mu = 0$ for every group. Given that $\mu_i - \mu = \alpha_i$, we can restate as: $\alpha_i = 0$ for all a groups. Thus, saying $H_0\colon \alpha_i = 0$ for all a groups is equivalent to saying that all group means will be the same: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_a$.

The alternative hypothesis will be $H_1\colon \alpha_i \neq 0$ for at least one group. As we said earlier, the means are either all the same or not all the same. Just one group mean different from the rest is enough. (As far how many group means are different, and which groups they happen to be, that's a question that ANOVA in its basic form can't answer. There are ways to address this issue. We'll get there.)

Partition of Variance: Sums of Squares

The key to understanding what ANOVA does rests in understanding how variance is partitioned (i.e., divided up). Because the business end of ANOVA involves manipulating sample data, we'll write things in sample terms (using Y to represent the dependent variable). A score on the dependent variable for subject j in group i is Y_{ij}. Where we had the grand mean of dependent variable scores, μ, we will have the grand mean of dependent variable scores, $\bar{Y}$. Group mean in the population model, μ_i, becomes group mean in our sample, $\bar{Y}_i$. That's just three terms that we need to get the job done. Not bad.

For any Y_{ij} we can view things as:

$$(Y_{ij} - \bar{Y}) = (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)$$

(Algebra fans can simplify the above equation to $Y_{ij} = Y_{ij}$. Not that you need to.)

The above equation means the difference between a given score on Y (Y_{ij}) and the grand mean of Y ($\bar{Y}$) = the difference between the group mean of Y ($\bar{Y}_i$) and the grand mean of Y ($\bar{Y}$) + the difference between that score on Y (Y_{ij}) and the group mean of Y ($\bar{Y}_i$).

Squaring and summing across observations gives us:

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{Y}_i-\bar{Y})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

The subscripts are just saying: sum across all observations in a given group ($\Sigma_{j=1}^{n_i}$) for all of the groups ($\Sigma_{i=1}^{a}$). This is simply the reverent statistician's way of saying: sum across everything.

That's the textbook version of the equation. We can simplify it a bit. Let's get rid of all of the subscripts – do we really need to refer to a score on Y as Y_{ij}? Let's just call it Y. And let's use just one sigma symbol (Σ) to indicate that these calculations should be done for all N observations and summed. This version is a little cleaner.

$$\sum(Y-\bar{Y})^2=\sum(\bar{Y}_i-\bar{Y})^2+\sum(Y-\bar{Y}_i)^2$$

For no reason whatsoever, I'm going to list the sample variance equation from Chapter 3.

$$S_X^2=\frac{\Sigma(X-\bar{X})^2}{N-1}$$

Why don't we rewrite this equation so that it's set up to compute the variance of Y instead of X. Again, for no reason at all.

$$S_Y^2=\frac{\Sigma(Y-\bar{Y})^2}{N-1}$$

Now compare that variance equation with the ANOVA model. Every one of the terms in the ANOVA model looks like a version of the numerator of the sample variance equation: $\Sigma(Y-\bar{Y})^2$. In fact, the first term in the ANOVA model is that very numerator. Which means that all three parts are just chronicling the amount of differences in scores of various types. The first term in the model is the total squared differences in scores (i.e., total squared differences between each score and the overall mean). The second term is the total squared differences between each group mean and the overall mean (differences between groups). The final term is the total squared differences between each score and the group mean (differences within groups). These terms, which are the *sum* of the *squared* differences between something and the mean of something, are called sums of squares.

Using this terminology, we can re-write this equation to be:

$$SS_T = SS_B + SS_W$$

(sum of squares total) = (sum of squares between groups) + (sum of squares within groups).

Sum of Squares Total

$$SS_T = \sum (Y - \bar{Y})^2$$

Addresses the question: How different are the scores? Identical to the numerator of the variance equation because that's what it is: the sum of squared differences between each score and the overall mean score.

Sum of Squares Between

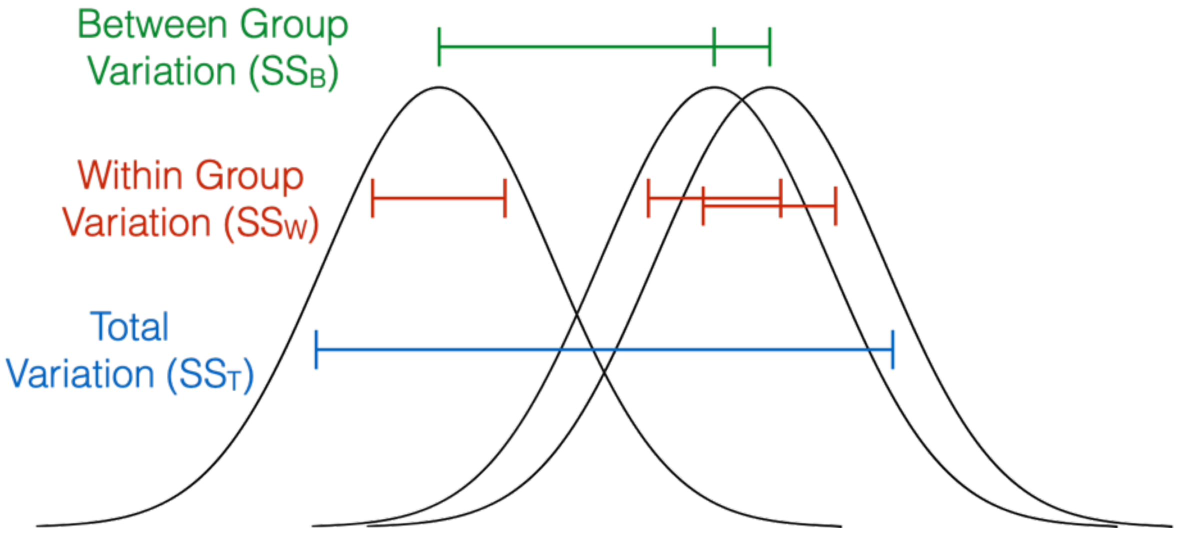
$$SS_B = \sum (\bar{Y}_i - \bar{Y})^2$$

Addresses the question: How different are the group means? Imagine computing the means for the *a* groups (i.e., $\bar{Y}_1,\ \bar{Y}_2,\ \ldots,\ \bar{Y}_a$) and then computing the variance of just these *a* scores. The numerator of that variance calculation would be the sum of the squared difference between each group mean ($\bar{Y}_i$) and the overall mean (Y). SS_B is that, only the squared difference computation $(\left(\bar{Y}_i - \bar{Y}\right)^2)$ is repeated for each observation within a given group (which is as redundant as it sounds).

Sum of Squares Within

$$SS_W = \sum (Y - \bar{Y}_i)^2$$

FIGURE 1 Variance Partitioning in ANOVA


Between Group Variation (SS_B)
Within Group Variation (SS_W)
Total Variation (SS_T)

In the ANOVA model total differences between scores (SS_T) are divided into between group differences (SS_B) and within-group differences (SS_W).

Addresses the question: How different are the scores within a group? Imagine computing the variance of just the observations within a given group. The numerator of that would be the sum of the squared differences between each score in that group and the mean of that group. Repeat for all *a* groups and sum to obtain SS_W.

Thus, ANOVA is an analysis of variance (redundant wording: minus two points) where the total variability is assigned to a between-groups component and a within-groups component (Figure 1).

Computing Sums of Squares

First, I understand that we use computers to do this sort of thing. It is likely that you will never be forced to compute an ANOVA by hand as a practicing researcher. Second, the process we will use to compute this ANOVA is not the most efficient way to do this (the more efficient methods are called computational forms of the equations because they have been optimized for easy computation; computational optimization has the unfortunate side effect of producing equations that don't make any intuitive sense). We are doing these hand calculations this way to help you understand the inner workings of ANOVA. Plus, I made examples to keep the math simple.

Table 1 Sum of Squares Computation

X	Y	Between	Within
1	4	$(5-6)^2 = 1$	$(4-5)^2 = 1$
1	5	$(5-6)^2 = 1$	$(5-5)^2 = 0$
1	6	$(5-6)^2 = 1$	$(6-5)^2 = 1$
2	5	$(6-6)^2 = 0$	$(5-6)^2 = 1$
2	6	$(6-6)^2 = 0$	$(6-6)^2 = 0$
2	7	$(6-6)^2 = 0$	$(7-6)^2 = 1$
3	6	$(7-6)^2 = 1$	$(6-7)^2 = 1$
3	7	$(7-6)^2 = 1$	$(7-7)^2 = 0$
3	8	$(7-6)^2 = 1$	$(8-7)^2 = 1$
		$SS_B = 6$	$SS_W = 6$

An inspection of Table 1 shows that there are three independent groups of subjects with three subjects in each group. The independent variable, *X*, identifies the group. Feel free to think that Groups 1 and 2 are alternate versions of the treatment and Group 3 is the control. To compute sums of squares, we need to know a few means: the grand mean and the mean of each group. The grand mean is 6.0, and the means of Groups 1-3 are 5.0, 6.0, and 7.0, respectively.

The column labeled "between" is where we compute the sum of squares between: $\Sigma(\bar{Y}_i - \bar{Y})^2$. Each entry in that column is the squared difference between the mean of that group and the grand mean. And yes, it's the same for everyone in that group because (yes, this is obvious) the group mean is the same for everyone in a group. The last entry is the sum of these values, the sum of squares between.

The column labeled "within" is for the sum of squares within: $\Sigma(Y - \bar{Y}_i)^2$. Each entry in that column is the squared difference between the actual score on *Y* and the group mean. The final entry is the sum of these values, the sum of squares within.

Partition of Variance: Mean Squares

Recall how the sample variance equation was a sum of squared differences divided by $N - 1$. We will be dividing our ANOVA sums of squares by *something minus something* to get what are called mean squares. The general form of any mean square equation is the sum of squares divided by its degrees of freedom.

$$MS = \frac{SS}{df}$$

The necessity of this step should be clear: we can inflate SS_B simply by having a large number of groups (assuming even trivial differences between means). Thus, we must divide by the number of groups to correct for this possible inflation.

Let's define the degrees of freedom for each of the three components of the model.

$$df_{Total} = N - 1$$

$$df_{Between} = a - 1$$

$$df_{Within} = N - a$$

Where:

N is the total sample size (reminder that n_i is number of observations in a given group)

a is the number of groups (i.e., number of levels or treatments on the independent variable)

Yes, the equations are strange. No, I don't know why. (But if you sum them in the $Total = Between + Within$ fashion, you find that the math works out: $df_{Total} = df_{Between} + df_{Within}$.)

The mean square equations all follow the SS/df structure. First, mean square total:

$$MS_T = \frac{SS_T}{N - 1}$$

Thus, MS_T is the same as the sample variance across all scores.

$$MS_B = \frac{SS_B}{a - 1}$$

Thus, MS_B is the SS_B divided by the *number of groups* – 1.

$$MS_W = \frac{SS_W}{N - a}$$

Thus, MS_W is the SS_W divided by the *total N – number of groups*.

Finally, The Test Statistic

The test statistic for ANOVA isn't distributed normally or as a *t*. The distribution is an *F* distribution, and once again, it's a family of distributions that vary on degrees of freedom. The new twist is that there are two degrees of freedom, called $df_{numerator}$ and $df_{denominator}$. The *F* distribution has the general shape shown in Figure 2. Notice that it is a skewed distribution with only one recognizable tail. Thus, the unusual scores are always the high scores. As mentioned there isn't an option to pick a direction with ANOVA.

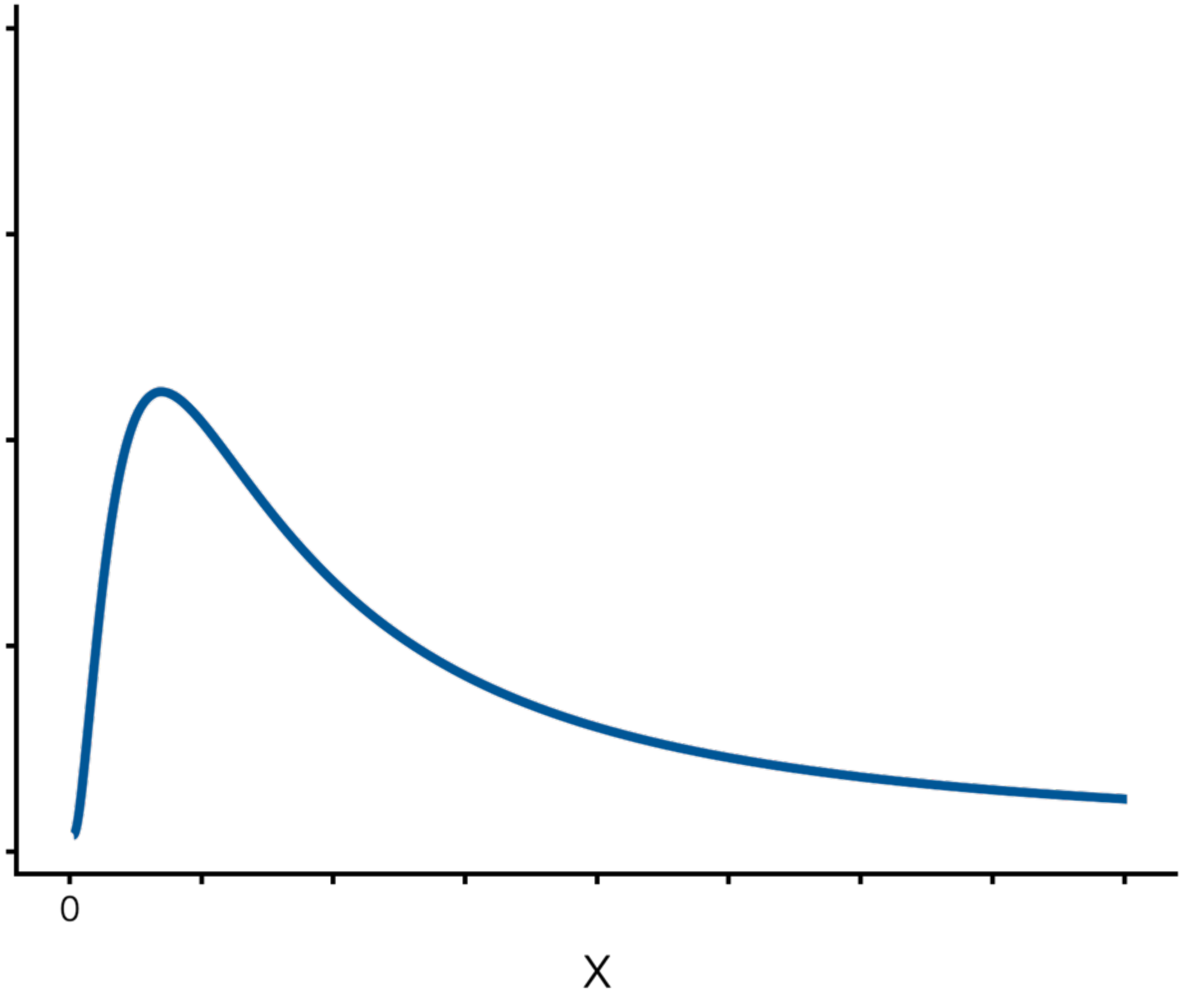
$$F_{obs} = \frac{MS_B}{MS_W}$$

With degrees of freedom: $a - 1$, $N - a$

Degrees of freedom in *F* tests (or *F* distributions) are referred to by their origin: $df_{numerator}$, $df_{denominator}$. The degrees of freedom in the numerator of the *F* test (MS_B) is $a - 1$. And in the denominator of the *F* test (MS_W), the degrees of freedom is $N - a$.

So the *F* test is a ratio of between-group variability to within-group variability. Let's play "What if?" using these equations.

FIGURE 2 *F* Distribution (degrees of freedom = 2, 5)



What if there are no between-group differences in our sample data (i.e., all groups have the same mean)?

> SS_B and MS_B are both zero, making the F test $0/MS_W$, which any mathematician can tell you works out to zero (assuming some within-group differences exist).

What if there are no within-group differences in our sample data (i.e., all members within a group have the same score)?

> Well, you get a division-by-zero error as SS_W and MS_W are both zero. So let's revise the question a bit: What happens to the F test as within-group differences approach zero? As within-group differences approach zero, the denominator of the F test shrinks, causing the F value to increase (as long as there are some between group differences). So bigger F value is the answer.

What if the between-group differences are the same magnitude as the within-group differences in our sample data?

> If MS_B and MS_W are of the same magnitude, then the F ratio works out to 1.0.

What if the between group differences are much bigger than the within group differences in our sample data?

> You get a MS_B that is bigger than MS_W which means you get a big F ratio.

The takeaway from the above Q&A is that when average between-group differences are bigger than average within-group differences, the result is large F values. Thus, the rejection region for all F tests is in the right tail of the distribution; the right tail is where the extreme (or unusual) values are found.

Expected Mean Squares

Let's introduce a concept that starts so small with one-way ANOVA that it's barely noticeable but becomes massively important later: expected mean squares.

Chapter 2 introduced the statistical concept of expectation (or the expected value of a variable). Expectation is defined as *long run average*. For our newfound F test, what is the expected value of each component (numerator and denominator)? These are the expected mean squares. The good news is that they are rather simple. For now.

$$E(MS_B) = \sigma^2 + \frac{1}{a-1}\sum_{i=1}^{a} n_i\alpha_i^2$$

$$E(MS_W) = \sigma^2$$

As terrible as those equations look, they can be reduced to something understandable in two moves. First, the σ^2 term is the pooled variance estimate, which is just another way of saying that it's within-group variance. The only thing in either equation is this ugly thing: $\sum_{i=1}^{a} n_i\alpha_i^2$. That terrible equation simply describes between-group differences. So with these two pieces of information we can re-write the expected mean squares as this.

$E(MS_B)$ = within-group differences + between-group differences

$E(MS_W)$ = within-group differences

Or using W and B as shorthand for within- and between-group differences:

$$E(MS_B) = W + B$$

$$E(MS_W) = W$$

So those are the expected mean squares for one-way ANOVA (fixed effects). Now think about the F test.

$$F = \frac{MS_B}{MS_W}$$

Substituting the expected mean squares, we now see an expected value for F.

$$F = \frac{W + B}{W}$$

Now let's use this to address our what-if questions in terms of expectation. What if there are no between-group differences in the population (i.e., the null hypothesis is true)? The between component (B) of the numerator is zero, and the only MS_B differences are due to sampling error and will be of the same magnitude as within-group variation. The expected value of the F test is 1.0. Expected values for F will only be greater than 1.0 when the null is false (i.e., there are actual differences between groups in the population).

The point of studying expected mean squares is to understand that the F test in ANOVA is designed to yield an expected value of 1.0 when the null is true and to yield an expected value greater than 1.0 when the null is false. Later, when the concept of expected mean squares takes an unpleasant turn, we will use this concept to construct the proper F test given the conditions of the study.

Back to Our Example

In our example we computed a SS_B of 6 and a SS_W of 6. Let's turn those into mean squares and finish this. Let's see, to compute mean squares we need to know degrees of freedom. We had nine total subjects, so $N = 9$. And we had three treatments, or groups, so $a = 3$. That means:

$$df_{Between} = a - 1 \quad = 3 - 1 = 2$$

$$df_{Within} = N - a \quad = 9 - 3 = 6$$

As for mean squares, we simply divide the sums of squares by their respective degrees of freedom.

$$MS_B = \frac{SS_B}{a-1} \quad = \frac{6}{2} \quad = 3$$

$$MS_W = \frac{SS_W}{N-a} \quad = \frac{6}{6} \quad = 1$$

And, finally, our F_{obs}:

$$F_{obs} = \frac{MS_B}{MS_W} \quad = \frac{3}{1} \quad = \ 3.0$$

We need to consult an F table (Table 2) to determine the critical value. We already calculated the degrees of freedom as 2 for the numerator and 6 for the denominator. So need $\alpha = .05$ value for $F_{(2,6)}$. According to the table, $F_{(2,6)} = 5.14$. Because F_{obs} (3.0) is not greater than F_{crit} (5.14), we are unable to reject the null that all of the means are the same in the population.

Comments

Remember: A significant F test means that at least one of the group means is different from the rest (we are rejecting null: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_a$). It does not mean that all of the group means are different from each other. It also doesn't tell us how many or which group means are different. There are follow-up tests (call post hoc tests) that

TABLE 2 Selected *F* Table Values (α = .05)

	df_b				
df_w	1	2	3	4	5
5	6.61	5.79	5.41	5.19	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48
10	4.96	4.10	3.71	3.48	3.33
15	4.54	3.68	3.29	3.06	2.90
20	4.35	3.49	3.10	2.87	2.71
25	4.24	3.39	2.99	2.76	2.60
40	4.08	3.23	2.84	2.61	2.45
50	4.03	3.18	2.79	2.56	2.40
100	3.94	3.09	2.70	2.46	2.31

we can use to determine that. These post hoc tests are structured differently from the ANOVA *F* test and thus, are not testing the same hypothesis.

In addition, there is no directionality to the ANOVA *F* test. Differences are differences and the *F* test doesn't know or care what direction the differences go. If you want to test a directional hypothesis, you need to do an independent samples *t* test (wherein only two groups are compared) or a planned comparison post hoc-type test (which, in an ironic note, technically isn't post hoc because it was planned).

The ANOVA F Test Is an Enhanced Version of the Independent Samples (Pooled) t Test

The ANOVA *F* test is really a super version of the pooled *t* test. Why do I say this? First, consider MS_W. In the days of independent samples *t* test, part of the denominator of the *t* statistic was a pooled variance estimate (i.e., S_p^2). Pooled variance is just an average variance of the two groups (look at the equation – it's an actual weighted average). Well, MS_W is a pooled variance for any number of groups (like an average of the SS_W for each group). If you're not convinced, set up a two-group study and compute both; MS_W will equal the pooled variance estimate (S_p^2).

Second, for the test statistic itself, $F = t^2$ when there are only two groups (and $n_1 = n_2$). A study of the equations for each would reveal this, but due to the arrangement of the independent samples t equation, this takes a fair bit of algebraic manipulation to see. And I don't think anyone wants that. Suffice to say that upon re-arrangement of the squared independent samples t statistic, the numerator (when squared) equals the MS_B of the F statistic and the denominator (when squared) equals the MS_W.

Long story short, the F test is a fancy version of the t test modified to handle any number of groups. Because the F test is a squared version of the t test ($F = t^2$), the squaredness of the F of the test precludes any possibility of negative values, meaning that there can't be a directional hypothesis with the F test (i.e., in other words, all F tests are similar to two-tailed t tests – there is no one-tailed possibility with the F test).

Assumptions

The assumptions for the F test in ANOVA are the same as with pooled t test for independent samples: independent random samples, normality, and homogeneity of variance ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_a^2$).

As before, the normality assumption is satisfied as long as the distribution is unimodal and symmetric. As for homogeneity of variance, this also matches what we saw with the pooled t test – as long as the n sizes are close to equal, this assumption are not likely to adversely impact the results.

Final Thoughts on One-Way ANOVA (Fixed Effects Model)

One-way ANOVA is a powerful tool for testing for differences between multiple independent groups. We said at the beginning that this is the version for fixed effects. Fixed effects are independent variables whose values were chosen by the researcher. There will be a different model for the random effects case. There will also be higher-order ANOVA models (two-way and beyond) in our future. For now, let's appreciate the simplicity of the one-way model.

Effect Size and Statistical Power

9

The full name of this chapter is “Statistical Significance, NHST Decisions, and Statistical Power.”

But that wouldn’t fit in the available space.

Statistical Significance vs Practical Importance

Statistical significance addresses whether sampling error is a likely cause of the observed results. It's a yes/no question (the Fisher inferential model): We can rule out sampling error as a cause of the results (i.e., reject the null) or not (fail to reject the null). Practical importance is whether the observed results are impressive (i.e., large differences between the means for an independent sample *t* test), moderate, or trivial (small). These are two different, albeit related, issues.

Here's how they are related: Other things being equal, larger effects are more likely to be significant.

Here's how they are different: When sample sizes are small, even large effects may not be significant (and nobody cares how big the effects are when you fail to reject the null because these "big effects" that you claim to have could be just sampling error). Or when sample sizes are huge, we may reject the null even though the effects are trivial in magnitude (yes, the results are unlikely due to sampling error, but given how small they are, who cares?).

Remember, no one cares about how big your effects are if your results are not significant. Can't rule out sampling error as a cause of your results? Then, the conversation is over.

Indices of Effect Size: Cohen's d

We need a way to quantify practical importance. Because we don't want our indices of practical importance to be affected by sample size, *z* statistics, *t* statistics, and *F* statistics are out.

For comparisons of two group means (the independent samples *t* test scenario) we will use a statistic called Cohen's *d*. sCohen's *d* allows us to describe the difference in a standardized metric. Like z scores, Cohen's *d* is in standard deviation units; means that are one standard deviation apart have a Cohen's *d* of 1.0 (or -1.0 if the direction is reversed).

Statistic:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$$

Sharp-eyed equation fans will recognize the denominator; it's the square root of the pooled variance estimate (i.e., S_p^2) from the pooled *t* test that you know and love from Chapter 5. The square root of variance is standard deviation. So this equation is really just a matter of dividing the difference between two means by their (average) standard deviation. That's the essence of Cohen's *d*: the difference between group means divided by a standard deviation (that is formed from a weighted average of their sample variances).

There are a number of twists to Cohen's *d* – it really is quite flexible. I'll list them below.

1. If σ is known (and is assumed the same for both groups), then the denominator above is simply σ.

2. If this is a one-sample situation, then replace $\bar{X}_2$ with the comparison population mean (μ_c). The denominator is σ, if known, or S_X, if not.

3. Note 1 and Note 2 will probably never be relevant to you. Cohen's *d* is almost always used for the two-group independent samples "pooled" *t* test situation.

4. Cohen's *d* is related to the *t* statistic from the independent samples pooled *t* test in the following fashion (which can be a major labor-saving move):

$$d = t_{obs}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

This is nice as it means that every t_{obs} can be turned into a Cohen's *d* with a simple multiplication involving the group *n* sizes.

5. If you like using online calculators for various statistical applications, don't use them for Cohen's *d* as every one that I have seen uses an incorrect (oversimplified) form of the equation. Do not trust them. Trust me.

Standards for Cohen's *d*. According to (Cohen, 1988, p. 40), the following standards should be used to interpret the magnitude of *d*:

Small: +/- .20

Medium: +/- .50

Large: +/- .80

Anything greater than +/- 1.0 isn't just large; it's enormous. A *d* of 1.0 is an entire standard deviation. Effects that big are quite rare.

Example. Let's use our example from Chapter 7. Data: treatment: $\bar{X}_1 = 38$, $S_1^2 = 14$, $n_1 = 25$; control: $\bar{X}_2 = 31$, $S_2^2 = 12$, $n_2 = 20$. The first thing to note is that we rejected the null ($t_{obs} = 6.44$, $t_{crit} = 1.681$). There isn't anything to discuss regarding effect size if we hadn't. Also note that I provided variance statistics for each group; had I given standard deviations, squaring would be necessary (do not overlook this detail in life). Let's compute *d*.

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$$

$$d = \frac{38 - 31}{\sqrt{\frac{(24)14 + (19)12}{24 + 19 - 2}}} = \frac{7}{\sqrt{13.12}} = 1.93$$

Cohen's $d = 1.93$, which is a massive difference. The means of these two aren't just a little different – they are different by almost two standard deviations. These results would be very impressive if this were not just made up data.

You may have noticed that we have a t_{obs} (6.44). Why don't we compute d from that?

$$d = t_{obs}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$d = 6.44\sqrt{\frac{1}{25} + \frac{1}{20}} \quad = \quad 6.44\sqrt{.09} \quad = \quad 1.93$$

Told you this way was easier.

Indices of Effect Size: Eta-squared

Cohen's d is our effect size solution for two-group situations. What about the ANOVA case in which we have three or more groups? Are we to compute Cohen's d for every pairing of groups? Well we could, but it would be better to have a single effect size statistic that captures all of the groups. That statistic is called eta-squared (η^2).

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

Note: It's that simple.

Here's how it works. In Chapter 8 we mentioned that the ANOVA model sees the within-group variability (SS_W) as error. (Side note: one of my undergraduate professors viewed the field of psychology as having two philosophies. One side sees things just like the ANOVA model: within-group differences are uninteresting. The other side sees within-group differences as the interesting part – the entire purpose of the field is to understand why people within groups are different on variables of interest.) In the ANOVA model between-group variability (SS_B) is the interesting part. Because $SS_T = SS_B + SS_W$, dividing SS_B by SS_T gives us the percent of total variance that is due to differences between groups (presumably a treatment effect). If all group means are the same (making $SS_B = 0$), then all variability is due to variability within groups and $\eta^2 = 0$. If the group means are different and all scores within a given group are identical, then all variability is due to between-group differences and $\eta^2 = 1.0$. If you're a fan of regression analysis (and what decent person isn't?), η^2 is analogous to R^2. Some stat programs even report η^2 as R^2.

Example. I would say let's use our Chapter 8 one-way ANOVA example, but that analysis was non-significant (i.e., we didn't reject the null), meaning that there's no reason to discuss effect size. I made a slight edit to the Chapter 8 dataset to give us a new example to work with (I lowered everyone's score in Group 1 by a point and raised Group 3 scores by a point). The dataset and sums of squares are shown in Table 1. If you complete the ANOVA, and you should, you will end up rejecting the null. Let's get to the eta-squared calculation.

TABLE 1 Sum of Squares Computation

X	Y	Between	Within
1	3	$(4-6)^2 = 4$	$(3-4)^2 = 1$
1	4	$(4-6)^2 = 4$	$(4-4)^2 = 0$
1	5	$(4-6)^2 = 4$	$(5-4)^2 = 1$
2	5	$(6-6)^2 = 0$	$(5-6)^2 = 1$
2	6	$(6-6)^2 = 0$	$(6-6)^2 = 0$
2	7	$(6-6)^2 = 0$	$(7-6)^2 = 1$
3	7	$(8-6)^2 = 4$	$(7-8)^2 = 1$
3	8	$(8-6)^2 = 4$	$(8-8)^2 = 0$
3	9	$(8-6)^2 = 4$	$(9-8)^2 = 1$
		$SS_B = 24$	$SS_W = 6$

$$\eta^2 = \frac{SS_B}{SS_T} \qquad = \frac{24}{SS_T}$$

What is the SS_T? We could compute it from the raw data with the $SS_T = \Sigma(Y - \bar{Y})^2$ equation. Or we could recall that $SS_T = SS_B + SS_W$. Since we already have SS_B and SS_W, that seems easier. So SS_T = 24 + 6 = 30.

$$\eta^2 = \frac{SS_B}{SS_T} \quad = \quad \frac{24}{30} \quad = \quad .80$$

We have our answer: 80% of the total variance is between-group variance, an incredibly high amount.

NHST Decisions

If the null hypothesis is in reality false, and we reject it, that's good.

If the null hypothesis is in reality true, and we fail to reject it, that's also good (although we're probably none too happy about it). Why is this good? Because we learned the truth.

But what of the other outcomes? Table 2 lists the various outcomes that can occur with hypothesis testing. Please understand that "accept H_0" is just shorthand for "fail to reject H_0."

Type I error is rejecting the null when the null is true, and the probability (at maximum) of this error equals alpha (α). Think of the logic of NHST here: Assume the null is true and ask how unusual this result would be in a null distribution; if

TABLE 2 NHST Outcomes

		Truth/Reality	
		H_0 (null is true)	H_1 (alt is true)
Study Decision	Accept H_0	Correct Decision	Type II Error
	Reject H_0	Type I Error	Correct Decision

the answer is very unusual (a standard that corresponds to α), then we reject the null. So, if the null is true, there is still an α probability that we will reject it.

Many researchers believe that when we have rejected the null, the probability that the null is true (and thus made a Type I Error) equals alpha (α). Those people could not be more wrong. They could try to be more wrong. But they would fail. The probability of a Type I error is not the probability that the null is true given that we rejected the null. The probability of a Type I error is the probability that we reject the null given that it's true. I know this sounds confusing (which is why this misunderstanding is so common), so I'll say it a few more ways.

> Correct Version: "If the null is true (this means that we are assuming that the null is true), then the probability that I made a Type I error (i.e., rejected the null when true) is .05 (i.e., alpha)."
>
> Incorrect Version: "I rejected the null. Thus, the probability that the null is true (meaning I made a Type I error) is .05 (i.e., alpha)."

The second version is incorrect because given the limited knowledge we have when we conduct research, we simply do not have enough information to estimate the probability that the null is true (thus make a Type I error when we reject it). (If we did have this info, then our approach to NHST would be very different.) I shouldn't be too hard on researchers holding to this incorrect version – I was one of them for many years.

Finally, Beta (β) is the probability of Type II error:

$$\beta = Pr(\text{fail to reject } H_0 \text{ when } H_1 \text{ is true*})$$

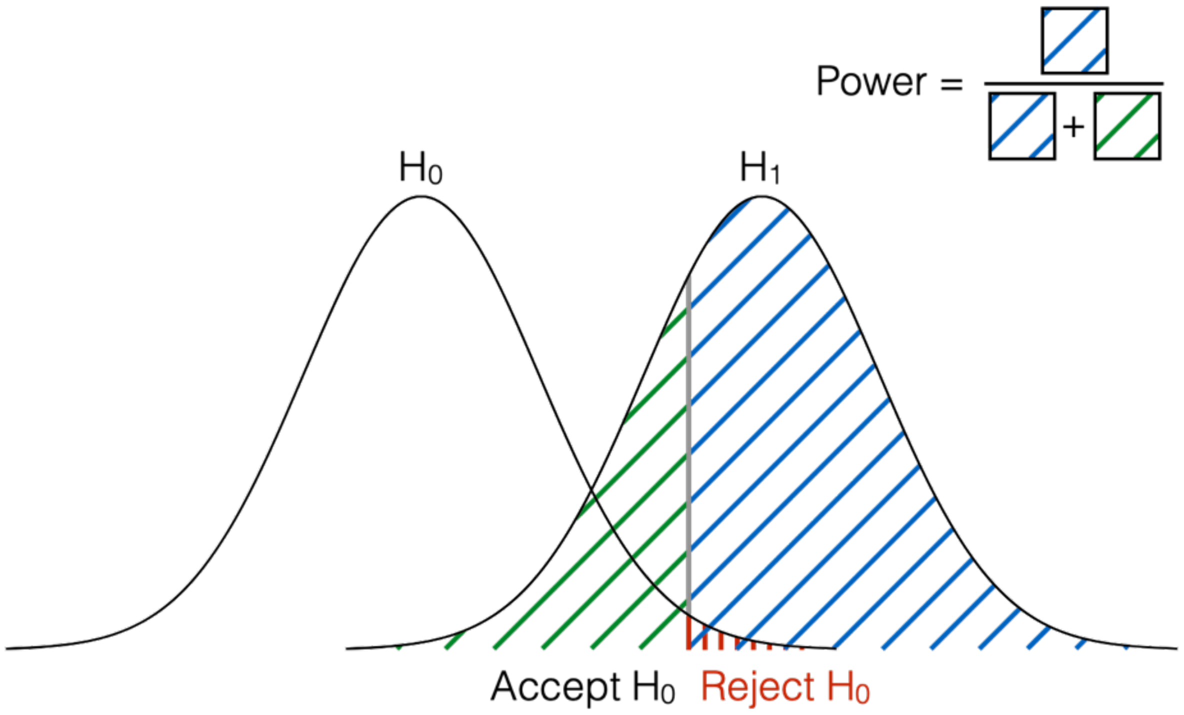
*I could have said "H_0 is false." Same thing.

We don't get to set this one – it depends on many factors (see Power section). Let's understand Type II error. If H_0 is false, we should reject it every time. But sometimes the results of our study do not rise to the level that allows us to reject the null (i.e., not in the 5% of the null distribution that is the rejection region for that type of hypothesis). This is quite bad as it means that effective treatments (e.g., medical, educational, job training) are thought to be ineffective. Why does these Type II errors occur? The answer is almost always a lack of sufficient statistical power.

Statistical Power

Power is defined as the probability that we reject the null hypothesis given that the null is false. This means that we are obtaining a significant result when we should have a significant result. Here are some more ways of saying the same thing.

FIGURE 1 Power Diagram



$$\text{Power} = 1 - Pr(\text{accept } H_0 \text{ given that the null is false})$$

$$\text{Power} = 1 - Pr(\text{Type II Error})$$

$$\text{Power} = 1 - \beta$$

We want power. Lots of power. We want to be able to reject the null every time our hypothesis is true (which, of course, means that the null is false). Notice that I didn't say that we want to reject the null every time – we want to reject it *every time it is false*. NHST is a (imperfect) tool we use to discover the truth.

The key to understanding a power diagram (Figure 1) is to keep in mind that these are sampling distributions and that the H_0 and H_1 distributions represent two different possibilities: H_0 for a "no effect" situation and H_1 for the "there is an effect" outcome. These are alternate realities – they can't both be true. We can construct the H_0 distribution for every hypothesis we test, but the H_1 distribution you see in the diagram is unknown in practice; it represents the "if there is an effect that of this magnitude, then its sampling distribution looks like this" scenario.

An inspection of Figure 1 will help us understand power. First, focus on the null (H_0) distribution. As we have discussed, the null distribution is a sampling distribution that reflects the characteristics of the null hypothesis. When the null is true some sample means will be low, most will be average, and some will be high. The vertical line in the graph represents the critical value; sample means greater than this value are in the top 5%, and if observed, allow us to reject the null. If the null is actually true and we reject it because we obtained a sample mean that was in the top 5%, then we made a Type I error (red area).

That's if the null is true. Now let's examine the distribution for the alternative hypothesis (H_1) (which makes the null false). If we obtain a sample mean greater than the critical value (i.e., falls in the top 5% of the null distribution), then we correctly reject the null (blue area). Note that about 2/3 of the H_1 distribution is above that critical value. Thus, 2/3 of the time when the null is false (making the alternative hypothesis is true), we are rejecting the null. That's power. A great power level is .9; .8 is still very good. Anything less than .8 (e.g., the .67 in) is not very good.

As to why low power is a problem, consider that when the null is false, we should be rejecting it every time. In Figure 1, when the alternative hypothesis is true we will obtain a sample mean that is the below the critical value about 1/3 of the time, leaving us unable to reject a false null. That is a Type II error (green area). Consider that even a good power of .8 means that the Type II error rate is .20, four times the rate of a Type I error.

One criticism of power analysis that you might have is that the analysis involves information that we don't know and can't know. That criticism is correct; power isn't computed – it's estimated. And the quality of the estimated power is dependent on the quality of the assumptions that go into the analysis. Power can't be directly computed as we don't know whether H_1 is true, and we don't know its parameters (and, of course, we don't know β). Power can be estimated, but that's only because we estimate various parameters (e.g., population effect size). Thus, the accuracy of the power estimate is dependent on the quality of our parameter estimates.

Ways to Increase Statistical Power

Increase Alpha. What if we set α to .10 instead of .05? This change would make it easier to reject the null. But this option is not really an option. And trades one error for another (i.e., increased alpha means greater risk of Type I error). If you want to go this route, good luck.

Conduct a One-Tailed Test Instead of Two-Tailed. Of course this option is only appropriate when we have a directional hypothesis. You would be surprised how often people conduct a two-tailed test when they could have or should have conducted a one-tailed test. There are two common reasons for this oversight. First, there are many times when a researcher has enough information to make a directional hypothesis but for some reason fails to state it that way. Thus, researchers making this error perform two-tailed tests when they could have performed the one-tailed variety.* Think of every z and t table – other things being equal, the two-tailed critical value is always greater than the one-tailed value. Another cause of this error lies in the stat software we use. These programs always default to two-tailed tests. Many times there isn't even an option for a one-tailed test. Thus, careless researchers end up conducting two-tailed tests even when they intended to run a one-tailed test of their directional hypothesis.

(* I once heard a researcher defend this practice on the grounds that it results in a more conservative – reduced chance of a Type I error – test. To that, I say: If you want to lower α to something less than .05, then lower α; there's little to be gained by doing it in a backhanded manner such as this.)

Decrease σ. Before we address the *how*, let's start with the *what*. Decreasing the standard deviation makes the sampling distributions (both null and alternative) narrower, leaving less of the H_1 distribution below the critical value. (Picture the Figure 1 distributions only skinnier.) How can we make this happen? The answer is in within-group variance (think SS_W). Differences within groups are seen as error; ideally, everyone receiving the same treatment would perform the same way. Of course, this line of thinking is a fantasy, completely disconnected from reality. But follow it backwards: the more irrelevant influences (e.g., distracting half of the subjects within a given treatment) that we introduce into our studies, the greater amount of irrelevant variance we introduce into their scores; that's a bigger standard deviation (σ). If it's possible to make things worse (introduce irrelevant influences to increase σ, then it should be possible to make things better by removing irrelevant sources of variance from your study (i.e., exercise more experimental control). Of course there are limits to how much we can reduce σ, but some of this is under the control of the researcher. It is possible to do this with experimental research and very difficult to do with non-experimental research.

Increase *N*. Bigger *N* means more power. The denominator of every standard error contains *N*. Increase *N* and you decrease standard errors, the standard deviation of the sampling distribution – making those distributions skinnier. Increasing *N* has the same effect that decreasing σ has. The advantage here is that *N* is a factor that is always under the direct control of the researcher. Just collect bigger samples already.

Post Hoc Tests

10

Things to do with a significant ANOVA

What ANOVA Can't Tell You

A significant *F* test in an ANOVA tells you that at least one of the means is different from the other means. But we don't know which one or how many. We must do follow up tests to find out (if we care about that sort of thing – and we may not). These tests are called post hoc (Latin for *I know a few words of Latin*) tests. As the plural "tests" implies, there are many of these post hoc tests. We will discuss two in this chapter.

Fisher's Least Significant Difference

Fisher's Least Significant Difference (i.e., Fisher's LSD, a name that engendered far fewer giggles when it was introduced in the early 20th century) allows us to compare any pair of group means, hence the term for this sort of thing: pairwise comparisons.

Setup: Compare any two group means ($H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$). Assumes equal *n* size per group. Also assumes equal variance (like pooled *t* test). This is basically a hopped up independent samples *t* test.

Start with independent samples "pooled" *t* test. Simplify with equal *n* size per group. And substitute the pooled variance estimate with its ANOVA equivalent (MS_W), and you get this:

$$t_{obs} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{MS_W\left(\frac{2}{n}\right)}}$$

Where the degrees of freedom are the degrees of freedom from the MS_W (i.e., the denominator *df*; i.e., $N - a$). And where *n* is the number of people per group (*N* is the total sample size)

Since we don't have a direction hypothesized, either one is fine (making this a 2-tailed test); so we'll take the absolute value of the numerator. The result will be significant if:

$$\frac{|\bar{Y}_1 - \bar{Y}_2|}{\sqrt{MS_W\left(\frac{2}{n}\right)}} > t_{(df_W),1-\frac{\alpha}{2}}$$

Where $t_{(df_W),1-\frac{\alpha}{2}}$ is the critical value of t (two-tailed).

If we move the denominator over to the other side (i.e., basic algebra), we get this:

$$|\bar{Y}_1 - \bar{Y}_2| > \left(t_{(df_W),1-\frac{\alpha}{2}}\right)\sqrt{MS_W\left(\frac{2}{n}\right)}$$

In other words, if the absolute value of the mean difference is greater than what's on the right side of the equation, then we have a significant difference. Thus, the right side of the equation represents the minimum difference needed to be significant, or (who could see this coming?) the *least significant difference*.

Example. Table 1 lists the data for 15 subjects in three groups (n = 5 for each group). Due to space limitations, I wasn't able to show the sums of squares, but I think you can handle that one yourself. (Go ahead, and finish off the ANOVA – the tedious work is already done.) The means for each group are as follows: Group 1 = 5.0, Group 2 = 6.0, Group 3 = 7.0.

Degrees of freedom are 2, 12 (a – 1, N – a), giving us an F_{crit} of 3.89. Because F_{obs} (5.0) is greater than F_{crit}, we can reject the null and conclude that the group means are not all the same. (Recall that I mentioned stat programs report an exact p-value and leave it up to you to conclude significance. In this case, the exact is p-value .026. Because .026 is less than .05 – and this is an ANOVA in which there is neither a direction nor tail option – we can reject the null.) Now it's time for a post hoc test to identify which groups are different. At least one of them should be.

TABLE 1 One-Way ANOVA Dataset

X	Y	Between	Within
1	4	$(5\text{-}6)^2 = 1$	$(4\text{-}5)^2 = 1$
1	4	$(5\text{-}6)^2 = 1$	$(4\text{-}5)^2 = 1$
1	5	$(5\text{-}6)^2 = 1$	$(5\text{-}5)^2 = 0$
1	6	$(5\text{-}6)^2 = 1$	$(6\text{-}5)^2 = 1$
1	6	$(5\text{-}6)^2 = 1$	$(6\text{-}5)^2 = 1$
2	5	$(6\text{-}6)^2 = 0$	$(5\text{-}6)^2 = 1$
2	5	$(6\text{-}6)^2 = 0$	$(5\text{-}6)^2 = 1$
2	6	$(6\text{-}6)^2 = 0$	$(6\text{-}6)^2 = 0$
2	7	$(6\text{-}6)^2 = 0$	$(7\text{-}6)^2 = 1$
2	7	$(6\text{-}6)^2 = 0$	$(7\text{-}6)^2 = 1$
3	6	$(7\text{-}6)^2 = 1$	$(6\text{-}7)^2 = 1$
3	6	$(7\text{-}6)^2 = 1$	$(6\text{-}7)^2 = 1$
3	7	$(7\text{-}6)^2 = 1$	$(7\text{-}7)^2 = 0$
3	8	$(7\text{-}6)^2 = 1$	$(8\text{-}7)^2 = 1$
3	8	$(7\text{-}6)^2 = 1$	$(8\text{-}7)^2 = 1$

Here's the equation for Fisher's LSD:

$$|\bar{Y}_1 - \bar{Y}_2| > \left(t_{(df_W),1-\frac{\alpha}{2}}\right)\sqrt{MS_W\left(\frac{2}{n}\right)}$$

The right side is the LSD part. We need to compute that before we start comparing means. To start, we'll need the t statistic. Let's see, the df_W is 12 ($N – a$: 15 - 3 = 12). A quick check of a t table (α = .05, two-tailed) tells us that t = 2.179. If you computed the F_{obs} for this ANOVA, then you know that the MS_W is 1.0. Finally, we know that n = 5. Plugging all of that info in:

$$|\bar{Y}_1 - \bar{Y}_2| > (2.179)\sqrt{1.0\left(\frac{2}{5}\right)}$$

$$|\bar{Y}_1 - \bar{Y}_2| > 1.378$$

So, if the difference between any pair of means is greater than 1.378, then those means are significantly different. Now let's get to comparing the means.

First, let's examine the mean of Group 1 versus 2. That's 5.0 and 6.0, respectively.

$$|\bar{Y}_1 - \bar{Y}_2|$$

$$|5.0 - 6.0| = 1.0$$

Is a mean difference of 1.0 greater than the LSD of 1.378? No, so the difference between 1 and 2 is not significant. Group 3 has a mean of 7.0, so the comparison of Groups 2 with 3 is also nonsignificant. But Group 1 versus 3 is different.

$$|\bar{Y}_1 - \bar{Y}_3|$$

$$|5.0 - 7.0| = 2.0$$

The difference between Groups 1 and 3 is greater than the LSD, so it is these two groups that are significantly different. Now we know.

As we said at the start, there are things ANOVA just can't tell us. Post hoc tests are employed to get that additional information.

The Trouble with Fisher

There is just one problem with Fisher's LSD: α is set to .05 for each comparison. We had three comparisons in the example. This means that, if the null is true for all three, the probability is greater than .05 that we will reject the null for at least one test when the null is actually true – a Type I error. (Beware of the typical misinterpretation of NHST and Type I errors. See the previous chapter for more on that.). If there are, say, 10 groups, then this probability is quite high (assuming the null is true for all). You may have heard of the difference between test-wise alpha and experiment-wise alpha. That is what we're dealing with here. This problem is known as probability pyramiding (best name ever). In spite of the cute name, some people can't sleep at night because of issues like this (I am not one of those people – think about what we've said about how often the null is true) and have proposed solutions. Which brings us to our next post hoc test.

Tukey's Honestly Significant Difference

Tukey's Honestly Significant Difference procedure (i.e., Tukey's HSD – more three letter acronyms?) is designed to control Type I error rate so that it is 5% for all the tests combined. That means that each individual comparison will have an α that is less than .05 or 5% (i.e., test-wise α for these tests will be less than .05 so that experiment-wise α will be .05). Tukey's test is not distributed as a z or a t (or even an F) distribution but as a Studentized Range distribution (q). The Studentized Range distribution is somewhat like the F distribution in that it has two degrees of freedom-type things: one is the number of groups (a), and the other is df_W. As with Fisher's LSD, this test assumes equal n per group.

$$|\bar{Y}_1 - \bar{Y}_2| > \left(q_{(a,df_W)}\right)\sqrt{MS_W/n}$$

Where a is the number of groups in the original ANOVA. MS_W and df_w are also from the original ANOVA.

Note that many tables use k in place of a and v in place of df_w for reasons that don't matter.

Example. We'll use the same Table 1 example from before. First, we compute the HSD, and for that we need to consult a Studentized Range distribution table of critical values. Because there are three groups and $df_W = 12$, we'll look for the 3, 12 value. A quick check of the table tells us that for $\alpha = .05$, the critical value is 3.77. Solving for HSD,

$$|\bar{Y}_1 - \bar{Y}_2| > (3.77)\sqrt{1.0/5}$$

$$|\bar{Y}_1 - \bar{Y}_2| > 1.685$$

Now, we know. If the difference between any two means is greater than 1.685, then they are significantly different (i.e., we reject the null that they are the same in the population).

Because Groups 1 and 2 as well as Groups 2 and 3 differ by only a point, they are not significantly different. But Groups 1 and 3 differ by two points, which is greater than 1.685, which makes them different.

Thus, these results are the same as with Fisher's LSD test. But you may have noticed that the minimum difference between group means required for significance increased with Tukey's HSD (1.685 for Tukey's vs 1.378 for Fisher's). In short, you can see that there will be situations in which groups that were significantly different with Fisher's test will not be significantly different with Tukey's. Choose wisely.

What About the Equal n per Group Thing?

Both Fisher's LSD and Tukey's HSD are approximately valid in the unequal n per group situation. For the unequal n scenario we must replace n in equation with an average n (called a harmonic average):

$$n_h = \frac{a}{\frac{1}{n_1} + \frac{1}{n_2} + \ldots + \frac{1}{n_a}}$$

If nothing else, harmonic average is an objectively cool name.

Summary of Pairwise Comparisons

In terms of Type I error, Fisher's LSD is the least conservative test (most likely to make a Type I error) and Tukey's the most conservative test (least likely) with other tests (e.g., Duncan's) somewhere in between. Duncan's Multiple Range Test is quite nice.

Don't Be That Guy

The ANOVA + post hoc model of significance testing is appropriate for one and only one situation: where you don't have any hypotheses about specific groups (e.g., Group 1 will be different from Group 3). For the ANOVA + post hoc model to be appropriate, your hypothesis must be of the form there will be differences (of an unspecified nature) between the groups. The ANOVA is the test of that hypothesis and the post hoc offers you supplemental information about the nature of the differences

If you have specific hypotheses about individual groups, test those hypotheses. At the most extreme, let's say that you had a hypothesis about Group 1 vs Group 3. Running an ANOVA just so that you can run the post hoc to check this is the wrong way to go. You don't run one significance test just so that you can run some other test that you really wanted to run (which is how this model is often treated). Just imagine that the ANOVA is non-significant. According to the rules for post hoc testing, you aren't even supposed to run the post hoc. Are you really not going to do that? And if you do, you haven't even checked your hypotheses properly as the MS_W for both analyses will be affected by these other groups that you didn't even care about.

In short, don't be that guy. Test your hypothesis as written. If your hypothesis only concerns two groups, but you also collected data on some other groups (maybe due to poor planning or factors outside of your control – ahem, thesis advisors), just ignore the other groups. Test your hypothesis with an independent samples *t* test. If your hypothesis had a direction, it should be obvious that this is the only way to go. But even if it didn't, the idea that you would let data from irrelevant groups affect your hypothesis test is plain daft.

Here's another "don't be that guy scenario": you have hypotheses about all the groups (e.g., Group 1 will have a greater mean than Groups 2 and 3). This is also inappropriate for the ANOVA + post hoc model. What you need is something called a planned comparison. A planned comparison is not a post hoc test. You are looking for something specific, not just which groups are different from which other groups. We’ll get to planned comparisons soon. Very soon.

Planned Comparisons

11

They're not post hoc tests

Planned Comparisons

Planned comparisons are comparisons of specific groups (in specific ways) that are *planned* – they were hypothesized in advance of data analysis. It's great when the name actually means something.

A Planned Comparison is Not a Post Hoc Test

Although planned comparisons are often treated as a fancier version of a post hoc test, there is a key difference – planned comparisons are planned, that is, hypothesized. Post hoc tests are a follow up to a significant ANOVA. With a true post hoc (Fisher's, Tukey's, and the rest) there are no hypotheses stating which specific groups are different from other specific groups, which is why there are no directional tests with post hocs. The planned part of a planned comparison means that there is a hypothesized comparison. We're not testing all of the groups against each other in some sort of free for all (that's ANOVA) – we're testing specific groups against specific groups (with directional tests, if desired). In fact, the flexibility inherent to planned comparisons can get overly complicated in a hurry.

Now here's the controversial part, although it shouldn't be controversial at all. A common question about planned comparisons is: Do you need to have a significant *F* test from an ANOVA before you can conduct a planned comparison? Let's revise that question: Do you need to conduct an ANOVA at all before you conduct a planned comparison? It should be obvious that the answer is no. (It should be, but it isn't to everyone.) The ANOVA and the planned comparison are testing two different types of hypotheses.

> ANOVA: There are differences among the group means (with optional post hoc tests to determine which means are different).

Planned Comparison: Group X has a higher/lower/different mean than the other groups (to offer but one possibility).

Although the planned comparison example probably won't be significant if the ANOVA is not significant (it shouldn't be, but strange things do happen), you don't have to run the ANOVA to have permission to run the planned comparison – they are testing different hypotheses. In short, your only obligation is to test the hypothesis you made. No more. If you're still not convinced, let's say you do the ANOVA *F* test, and it's non-significant. Are you not going to run the planned comparison? Of course you are going to run it. Like I said, strange things happen. (Here's how: In close cases, one degree of freedom can make the difference.)

The point of all of this is that the significance of a planned comparison (and the propriety of conducting one) does not depend upon the significance of an ANOVA.

What Can You Do with a Planned Comparison?

You can test just about anything you want (within only a few limits).

If there are three groups, you can test whether:

Group 1's mean is greater/lesser/different from the means of Groups 2 and 3 (note that with this hypothesis, we don't care whether Group 2 is different from Group 3).

Or some other variation of that.

If there are four groups, you can test whether:

Groups 1 and 2 are greater/lesser/different than Groups 3 and 4.

Group 1 is greater/lesser/different than Groups 2, 3 and 4.

Or some other variation of that

You get the idea. The flexibility of the planned comparison model allows to test a wide variety of hypotheses. To address the first example of the four group scenario (Groups 1 and 2 are greater than Groups 3 and 4), you might ask why someone would want to do that. The answer might be that Groups 1 and 2 are two treatment groups and Groups 3 and 4 are two control groups, and you want to find out if the treatments in general have an effect as compared to the control groups in general. Maybe this hypothesis sounds strange to you. That's fine. The point is that the planned comparison process is flexible enough to accommodate a variety of comparisons.

Did you notice the other big advantage of planned comparisons? You can test directional hypotheses (Group 1 is greater than Groups 2 and 3). Good luck doing that with the ANOVA + post hoc model.

Conducting a Planned Comparison

First, figure out what you want to test (i.e., your hypothesis). Let's do one of the hypotheses from the four group thing in the previous section: Group 1 and 2 are different than Group 3 and 4. Unlike the simple days of one-way ANOVA in which the null and alternative hypotheses were always the same, we're going to have to specify them for planned comparisons. For the example, the alternative hypothesis would look like this:

$H_1 : \mu_1 + \mu_2 \neq \mu_3 + \mu_4$

The null would look like this:

$H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4$

Which we could re-write as:

$H_0 : \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$

Thus, our null can be conceived as a linear combination (l) of means

$$l = (1)\mu_1 + (1)\mu_2 + (-1)\mu_3 + (-1)\mu_4$$

Notice how weights sum to zero. They have to sum to zero.

The general form of that linear combination thing is this:

$$l = c_1\mu_1 + c_2\mu_2 + \ldots + c_a\mu_a = \Sigma c_i\mu_i$$

Where $\Sigma c_i = 0$. Note that restriction: you can do whatever you want regarding weights as long as they sum to zero.

Thus for all of these planned comparisons, the null is really $H_0 : l = 0$ (or $H_0 : l \leq 0$ or $H_0 : l \geq 0$ if your hypothesis has a direction). Think of l as a summary statement about the population means. In this case, the null states that these two group means (and we don't care about them individually) are equal (because this is the null) to these other two group means (which we also don't care about individually). The alternative hypothesis is of course that they are unequal in the population.

The sample equivalent of l is $\hat{l}$ and is defined as $\hat{l} = \Sigma c_i \bar{Y}_i$. Thus, $\hat{l}$ is just the sum of the sample means weighted according to our hypothesis. My suggestion to make this easy: Specify the null in the $H_0 : \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$ format, turn it into the linear combination format:

$$l = (1)\mu_1 + (1)\mu_2 + (-1)\mu_3 + (-1)\mu_4$$

Use that to set up your $\hat{l}$ computation. In this case, $\hat{l}$ would look like this:

$$\hat{l} = (1)\bar{Y}_1 + (1)\bar{Y}_2 + (-1)\bar{Y}_3 + (-1)\bar{Y}_4$$

Let's do an example with the data listed in Table 1. The means for Groups 1-4 are 5.0, 6.0, 7.0, 8.0, respectively. For this hypothesis, $\hat{l}$ is:

$$\hat{l} = (1)5.0 + (1)6.0 + (-1)7.0 + (-1)8.0 = \text{-4.0}$$

TABLE 1 Planned Comparison Dataset

X	Y	Between	Within
1	4	$(5\text{-}6.5)^2 = 2.25$	$(4\text{-}5)^2 = 1$
1	4	$(5\text{-}6.5)^2 = 2.25$	$(4\text{-}5)^2 = 1$
1	6	$(5\text{-}6.5)^2 = 2.25$	$(6\text{-}5)^2 = 1$
1	6	$(5\text{-}6.5)^2 = 2.25$	$(6\text{-}5)^2 = 1$
2	5	$(6\text{-}6.5)^2 = .25$	$(5\text{-}6)^2 = 1$
2	5	$(6\text{-}6.5)^2 = .25$	$(5\text{-}6)^2 = 1$
2	7	$(6\text{-}6.5)^2 = .25$	$(7\text{-}6)^2 = 1$
2	7	$(6\text{-}6.5)^2 = .25$	$(7\text{-}6)^2 = 1$
3	6	$(7\text{-}6.5)^2 = .25$	$(6\text{-}7)^2 = 1$
3	6	$(7\text{-}6.5)^2 = .25$	$(6\text{-}7)^2 = 1$
3	8	$(7\text{-}6.5)^2 = .25$	$(8\text{-}7)^2 = 1$
3	8	$(7\text{-}6.5)^2 = .25$	$(8\text{-}7)^2 = 1$
4	7	$(8\text{-}6.5)^2 = 2.25$	$(7\text{-}8)^2 = 1$
4	7	$(8\text{-}6.5)^2 = 2.25$	$(7\text{-}8)^2 = 1$
4	9	$(8\text{-}6.5)^2 = 2.25$	$(9\text{-}8)^2 = 1$
4	9	$(8\text{-}6.5)^2 = 2.25$	$(9\text{-}8)^2 = 1$

This means that in our sample, the means of Groups 1 & 2 differed from the means of Groups 3 & 4 by -4.0 points. Is that a significant difference? We don't know yet; that's why we do the *t* test. The equal *n* test statistic is:

$$t_{obs} = \frac{\hat{l}}{\sqrt{\frac{MS_W}{n}\Sigma c_i^2}}$$

With df_w (i.e., $N - a$) degrees of freedom.

If *n*s are not equal, then it's this:

$$t_{obs} = \frac{\hat{l}}{\sqrt{MS_W \Sigma \frac{c_i^2}{n_i}}}$$

Where n_i is the *n* size for each group.

Back to our example. We computed $\hat{l}$ as -4.0. You should do the work to compute MS_W on your own, but I'll just tell you that $MS_W = 1.333$ and $df_W = 12$ ($df_W = N - a = 16 - 4$). With this info, we can compute our test statistic. But first, we need to address this Σc_i^2 thing. Σc_i^2 is just the sum of the squared weights. In the case of our example data the weights were all +1 or -1. Squaring and summing is pretty easy.

$$\Sigma c_i^2 = (1)^2 + (1)^2 + (-1)^2 + (-1)^2 = 4$$

Finally:

$$t_{obs} = \frac{\hat{l}}{\sqrt{\frac{MS_W}{n}\Sigma c_i^2}} \qquad = \frac{-4}{\sqrt{\frac{1.333}{4}(4)}} \qquad = \frac{-4}{1.15}$$

t_{obs} = -3.46

To complete the test we need a critical value. This planned comparison did not specify a direction, so we need the two-tailed (α = .05) value. A check of the t table tell is that t_{crit} = ±2.179.

The conclusion of our test of the hypothesis that Groups 1 and 2 are different from Groups 3 and 4 is as follows. Because the t_{obs} (-3.46) is less than the at t_{crit} (-2.179 for this situation), we reject that the null that the Groups 1 and 2 are the same as 3 and 4 and conclude that they are different.

Final Thoughts on Planned Comparisons

That may have seemed fairly pointless to you, but let me remind you that you have almost complete freedom on what you test in a planned comparison. Here's a version of the example that might be more interesting. Let's say that Groups 1-3 are three different treatments for something and Group 4 is a control. We might be curious is the treatments, taken together, have lower scores than the control on the dependent variable. The issue isn't which treatment but the treatments as a whole. The alternative hypothesis would look like this:

$$H_1 : \mu_1 + \mu_2 + \mu_3 < \mu_4$$

The null would look like this:

$H_0 : \mu_1 + \mu_2 + \mu_3 \geq \mu_4$

Let's do the null rewrite so that we can get the weights (remember, they have to sum to zero). First, a little algebra to get it to equal zero.

$H_0 : \mu_1 + \mu_2 + \mu_3 - \mu_4 \geq 0$

Now to translate into l form:

$l = c_1\mu_1 + c_2\mu_2 + c_3\mu_3 - c_4\mu_4$

The one rule for weights is that they must sum to zero. So if we go with unit (i.e., 1.0) weights, we have a problem: 1 + 1+ 1 - 1 ≠ 0. But there are solutions. One is this:

$l = (1)\mu_1 + (1)\mu_2 + (1)\mu_3 + (-3)\mu_4$

Or if you prefer fractions, there is this option:

$l = (1/3)\mu_1 + (1/3)\mu_2 + (1/3)\mu_3 + (-1)\mu_4$

That works. If you want to get creative, you could do this:

$l = (10)\mu_1 + (10)\mu_2 + (10)\mu_3 + (-30)\mu_4$

Whatever weights you choose, just make sure they sum to zero. You'll then use those weights when computing $\hat{l}$ and Σc_i^2 for the t statistic.

Maybe we should have done that version of the example. It looks more interesting. You know, there's nothing stopping you from doing it now.

Random Effects ANOVA

12

Something old, something new

Fixed vs Random Effects ANOVA

A quick review of fixed effects versus random effects independent variables is necessary. Way back in Chapter 1 we said:

> A fixed variable is one whose values are determined by the researcher. This can only apply to an independent variable in a true or quasi experiment. In the case of a fixed (independent) variable, the researcher decides what values the variable will have in the study. For example, if the study involves time spent studying a new language, the researcher might assign one group of subjects to study for zero minutes, another group to study for 10 minutes, and a third group to study for 20 minutes. Why these values and not some other values? Ask the experimenters – they're the ones who chose them.

The fixed variable should be very familiar to us. Everything we've done to this point has dealt with fixed variables for independent variables. The ANOVA model that we covered in Chapter 8 was the fixed effects (FE) model.

We need to address random variables. Our earlier definition of a random variable stated that it is a variable whose values were not the result of a choice made by the experimenter. When the independent variable is a random variable the values were chosen by a random process. An example may help to highlight the difference. Suppose there are twenty possible treatments (e.g., 20 techniques for improving performance on a test). We probably don't want to test all 20 conditions as the number of subjects required would be prohibitive (i.e., a real pain). We could pick the most promising five or so treatments and just test them; that version of the study uses the FE ANOVA model (that we know and love). That study might be a fine study, but it wouldn't tell us anything about the 15 treatments that we didn't test. However, if we pick 5 treatments at random from the 20, we are using the random effects (RE) model of ANOVA. The RE model has a surprise benefit that may justify the decision.

Random Effects ANOVA

First, the good news. For a one-way ANOVA, RE ANOVA calculations are the same as FE. Same sums of squares, mean squares, degrees of freedom, and *F* test. Bad news: What is different are the hypotheses and the interpretation of the results. With random effects (RE) ANOVA, the values of the IV chosen are just some of the possible values we could have chosen.

The big difference between the FE and RE models is that with FE ANOVA we are testing for differences between the values of the independent variable that we chose ($H_0: \alpha_i = 0$ for all a groups), whereas for RE ANOVA we are testing for differences between all possible levels of the independent variable ($H_0: \sigma_\alpha^2 = 0$). It may help to remember that $\alpha_i = \mu_i - \mu$ (i.e., the difference between the mean of group i and the grand mean). Thus, by saying $\sigma_\alpha^2 = 0$, we're saying that the means of all possible levels are the same (i.e., when variance equals zero, everything is the same).

That one difference in the null hypothesis may not sound like much, but it carries some big implications. To restate, the FE null states that the means of the treatments in the study are the same; the RE null states that the means of *all possible* treatments (including those not in the actual study) are the same. The RE version of the null is a much bigger statement.

To summarize how FE case and RE case are different, consider the various conclusions that occur with each model (Table 1). The interesting case is when the we fail to reject the null. The random effects model allows for much greater generalizability. Because the treatments studies were selected at random from a domain of possible treat-

Table 1 Fixed vs Random Effects ANOVA Conclusions

		ANOVA Model	
		Fixed Effects	Random Effects
Study Decision	Accept H_0	The a treatments are equivalent	All possible treatments are equivalent
	Reject H_0	At least one treatment is different	Some treatments that exist are different

ments, we are able to generalize to the entire domain (no differences in the ones we picked at random means no differences among all of them).

Expected Mean Squares

In the FE ANOVA model we introduced expected mean squares. It was fairly simple and reduced to this.

$$E(MS_B) = W + B$$

$$E(MS_W) = W$$

The real equations were more complicated than this, but we simply labeled the parts as describing within-group stuff (W) or between-group stuff (B). Given the structure of the F test ($F = MS_B/MS_W$), the expected value of the F test is 1.0 if the null is true because the expected value for MS_B equals the expected value for MS_W (i.e., when the null is true $E(MS_B) = W + 0$ because there are no between-group differences).

We have now reached a good news/bad news situation with RE ANOVA. The good news is that for one-way RE ANOVA, the expected mean squares end up looking like the expected mean squares for the FE one-way ANOVA model.

$$E(MS_B) = W + B$$

$$E(MS_W) = W$$

The actual equation is different the expected mean square between ($E(MS_B) = \sigma^2 + n\sigma_\alpha^2$) but the net result is the same.

The bad news is that when we get to two-way ANOVA and beyond, expected mean squares go completely sideways for RE ANOVA. But that's a problem for another day. For today, expected mean squares along with sums of squares, degrees of freedom, mean squares, and the *F* test are the same for FE and RE one-way ANOVA.

Concluding Thoughts on Random Effects Model

This all seems like a big waste of time. Why introduce the RE effects model when hardly anything changes (only the scope of the conclusions change)? The answer is that although nothing changes in one-way ANOVA, things will change in a big way for not just expected mean squares but also for the *F* test itself with multi-way ANOVA (i.e., two-way, three-way, etc.). Even worse, there will be mixed models in which some independent variables are FE and others are RE.

Two-Way ANOVA

13

Twice the fun

Why Two-Way ANOVA?

To this point our discussion of ANOVA has been limited to a single factor (i.e., one independent variable). Sometimes we desire to conduct research on multiple independent variables at once. To do that, we will need an ANOVA that can accommodate these multiple independent variables. That's where two-way (and beyond) ANOVA steps in. Two-way ANOVA allows researchers, in a single analysis, to analyze the effect that two independent variables have on a single dependent variable.

The big question you might have is why? Why analyze two independent variables in a single analysis? Why not simply perform the analysis as two one-way ANOVAs? There are multiple good answers to this question. First, if the same set of subjects are used for both independent variables, then a two-way ANOVA can be more sensitive to the effects of each independent variable than two one-way ANOVAs. We'll demonstrate this later. Second, a two-way ANOVA allows us to examine more than just the effects for each independent variable in isolation; it allows us to examine whether the two independent variables interact with the dependent variable. Testing for interactions is very important, and is not something to be ignored. We will also address this later. For starters we need to lay the foundation for the two-way ANOVA (fixed effects) model.

Factors and Cells

With one-way ANOVA, we had one independent variable with various levels (or treatments or groups). With two-way ANOVA, we have two independent variables (which we will now call factors), each with various levels/ treatments/groups (we'll use those terms interchangeably). A fully-crossed design (which we definitely want) has every combination of the two factors (i.e., all possible combinations of treatments). If there are three treatments on the first factor and four treatments on the second, there are 3 × 4 = 12 possible combinations, and all combinations are included in the study. A given combination (say, the group with Treatment 2 on the first factor and Treatment 3 on the second factor) is called a cell. A cell is like the group of the one-way ANOVA days. Each cell in a multifactor ANOVA consist of a different group of subjects, necessary to satisfy the independence requirement of *iid*.

One more weird thing, in two-way ANOVA you often hear talk about row effects and column effects. These are just shorthand terms for the effect at a given treatment of Factor A (irrespective of Factor B) and vice versa. Table 1 displays an example of the row/column structure of two-way ANOVA.

The effect for the entire set of rows (or columns) for a given factor is called a main effect. A main effect for Factor A is the same (in terms of the numerator of the *F* test) as a one-way ANOVA on Factor A (the main effect for Factor A is obtained by collapsing the data across the groups of Factor B – as if we didn't know there was a Factor B).

TABLE 1 Two-Way ANOVA Data Structure

		Factor B (with b levels)				
		1	2	3	4	
Factor A (with a levels)	1	n obs	n obs	n obs	n obs	bn obs
	2	n obs	n obs	n obs	n obs	bn obs
	3	n obs	n obs	n obs	n obs	bn obs
		an obs	an obs	an obs	an obs	abn (or N) obs

Note how in Table 1 each cell has a sample size of size n (we're assuming equal sample sizes per cell). Also note the concept of collapsing across a factor: there are n subjects in cell (1,1), but if we collapse across Factor B, there are bn

(where b is the number of levels on Factor B) in Level 1 of Factor A. Similarly we can collapse across Factor A, giving us an subjects in Level 1 of Factor B. The total number of subjects is abn (number of levels of A × number of levels of B × number of subjects per cell) or just N.

A Recap of the One-Way Model

In the one-way days Y_{ij} was the score on Y for j^{th} observation of group/level/treatment i, and a was the number of groups/levels/treatments in our independent variable. The difference between the group/level/treatment mean and grand mean was α_i (i.e., $\alpha_i = \mu_i - \mu$). Finally, the model was given as:

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

Which means a given score on Y (i.e., Y_{ij}) can be decomposed into the overall mean of Y (i.e., μ), the effect for group (α_i), and a random component (e_{ij}, distributed as $\sim N(0, \sigma^2)$).

The Two-Way ANOVA Model

In two-way ANOVA dependent variable scores are Y_{ijk} (score on Y for the k^{th} observation within level i of Factor A and level j of Factor B). As before, a is number of levels/treatments for Factor A. To that we can add b, the number of levels/treatments for Factor B. We also had $\alpha_i = \mu_i - \mu$, which is now specific to Factor A. To that we add the Factor B equivalent, $\beta_j = \mu_j - \mu$.

The two-way ANOVA model has the following structure:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$$

Which means a given score on Y (i.e., Y_{ijk}) can be decomposed into: the overall mean of Y (i.e., μ), the effect for the level of Factor A (α_i), the effect for the level of Factor B (β_j), the effect for the interaction of Factor A and B ($\alpha\beta_{ij}$), and a random component (e_{ijk}, distributed as $\sim N(0, \sigma^2)$).

In summary, the two-way ANOVA model is like the one-way but with effects for two independent variables plus an interaction term.

Partition of Variance

In one-way ANOVA total variability was divided into a between-groups component (which will now be labeled as $SS_{Between}$ to eliminate ambiguity) and a within-groups component:

$$SS_T = SS_{Between} + SS_W$$

In two-way ANOVA we will be breaking what was a single between-group component into multiple parts. Figure 1 shows the division of the sums of squares in two-way ANOVA. What was once a single term for between-group differences is now three terms: a main effect for Factor A (SS_A), a main effect for Factor B (SS_B), and an interaction term (SS_{AB}).

Figure 1 Sums of Squares in One-Way ANOVA vs Two Way ANOVA

$$\begin{tikzcd}
SS_T = & & SS_{Between} \arrow[dl] \arrow[d] \arrow[dr] & + SS_{Within} \\
SS_T = & SS_A + & SS_B + & SS_{AB} + SS_{Within}
\end{tikzcd}$$

Degrees of Freedom

As always, *N* is the total number of observations, and *n* is number of observations per cell; if equal number of observations per cell (as we as-

Table 2 Degrees of Freedom and Mean Squares in Two-Way ANOVA

Source	*df*	Mean Square
A	$a - 1$	SS_A/df_A
B	$b - 1$	SS_B/df_B
AB	$(a - 1)(b - 1)$	SS_{AB}/df_{AB}
W	$ab(n - 1)$	SS_W/df_W
T	$N - 1$	

Note: AB is the interaction term.

sumed), then $N = abn$. Because we have so many sources of variance in two-way ANOVA we will make a table (Table 2) for the various degrees of freedom.

Mean Squares

As always, we obtain mean squares by dividing the sum of squares by its relevant degrees of freedom. The various mean squares are listed below in full detail (this is really the same information presented in Table 2).

$$MS_A = SS_A/(a - 1)$$

$$MS_B = SS_B/\left(b - 1\right)$$

$$MS_{AB} = SS_{AB}/(a - 1)(b - 1)$$

$$MS_W = SS_W/ab(n - 1)$$

Hypothesis Testing in Two-Way ANOVA

In the one-way days, life was simple. All we could test was whether there were differences in groups/treatments. In two-way ANOVA, we can test three things: Are there differences among groups/treatments of Factor A (null hypothesis: the means of the *a* groups of Factor A are the same; $\alpha_i = 0$ for all groups), are there differences among groups/treatments of Factor B (null hypothesis: the means of the *b* groups of Factor B are

TABLE 3 *F* Tests in Two-Way ANOVA

Source	F_{obs}	df_{num}	df_{denom}
A	MS_A/MS_W	$a - 1$	$ab(n - 1)$
B	MS_B/MS_W	$b - 1$	$ab(n - 1)$
AB	MS_{AB}/MS_W	$(a - 1)(b - 1)$	$ab(n - 1)$

the same; $\beta_i = 0$ for all groups), and is there an interaction of Factor A and B (null hypothesis: after controlling for main effects, are the means of the *ab* cells the same; $\alpha\beta_{ij} = 0$ for all cells)? That's three significance tests. It's not as bad as it sounds.

Table 3 lists the three *F* tests in two-way ANOVA as well as their degrees of freedom (which can be deduced easily enough by examining the mean squares in the numerator and denominator of the *F* tests). One of the nice things about the fixed effects model for multi-way ANOVA is that the denominator mean square for all *F* tests is MS_W. That won't be the case with two-way (and beyond) random effects ANOVA.

Expected Mean Squares

Expected mean squares in two-way fixed effects ANOVA aren't much worse than in one-way. There are just the obvious additions. First, the ugly equations:

$$E(MS_A) = \sigma^2 + \frac{bn}{a-1}\sum_{i=1}^{a}\alpha_i^2$$

$$E(MS_B) = \sigma^2 + \frac{an}{b-1}\sum_{j=1}^{b}\beta_i^2$$

$$E(MS_{AB}) = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{i=1}^{b}(\alpha\beta)_{ii}^2$$

$$E(MS_W) = \sigma^2$$

That was unpleasant. Let's see the easy versions.

$$E(MS_A) = W + A$$

$$E(MS_B) = W + B$$

$$E(MS_{AB}) = W + AB$$

$$E(MS_W) = W$$

Because MS_W is used as the denominator, the expected value for the F test of Factors A and B and their interaction will be 1.0 when the null is true (as was the case with the one-way fixed effects model).

Computing Sums of Squares

As a refresher, the relevant sums of squares in one-way ANOVA were computed as follows.

$$SS_{Between} = \sum \left(\bar{Y}_i - \bar{Y}\right)^2$$

$SS_{Between}$ was computed as the difference between the group mean from the overall mean.

TABLE 4 One-Way ANOVA Sum of Squares

A	Y	Between	Within
1	6	$(15\text{-}17)^2 = 4$	$(6\text{-}15)^2 = 81$
1	8	$(15\text{-}17)^2 = 4$	$(8\text{-}15)^2 = 49$
1	12	$(15\text{-}17)^2 = 4$	$(12\text{-}15)^2 = 9$
1	14	$(15\text{-}17)^2 = 4$	$(14\text{-}15)^2 = 1$
1	16	$(15\text{-}17)^2 = 4$	$(16\text{-}15)^2 = 1$
1	18	$(15\text{-}17)^2 = 4$	$(18\text{-}15)^2 = 9$
1	22	$(15\text{-}17)^2 = 4$	$(22\text{-}15)^2 = 49$
1	24	$(15\text{-}17)^2 = 4$	$(24\text{-}15)^2 = 81$
2	8	$(19\text{-}17)^2 = 4$	$(8\text{-}19)^2 = 121$
2	10	$(19\text{-}17)^2 = 4$	$(10\text{-}19)^2 = 81$
2	14	$(19\text{-}17)^2 = 4$	$(14\text{-}19)^2 = 25$
2	16	$(19\text{-}17)^2 = 4$	$(16\text{-}19)^2 = 9$
2	22	$(19\text{-}17)^2 = 4$	$(22\text{-}19)^2 = 9$
2	24	$(19\text{-}17)^2 = 4$	$(24\text{-}19)^2 = 25$
2	28	$(19\text{-}17)^2 = 4$	$(28\text{-}19)^2 = 81$
2	30	$(19\text{-}17)^2 = 4$	$(30\text{-}19)^2 = 121$

$$SS_W = \sum \left(Y - \bar{Y}_i\right)^2$$

SS_W was computed as the difference between the score on Y from the group mean.

Let's review an example to refresh our memories. In the Table 4 dataset we have a single independent variable (A) with two groups or levels (n = 8 for each). The groups means are 15 for Group 1 and 19 for Group 2. The grand mean is 17. The $SS_{Between}$ = 64, and the SS_W = 752. With that process established, let's explore how sums of squares are calculated in two-way ANOVA.

Two-Way ANVOA

An understanding how sums of squares are calculated is crucial for a full understanding of two-way ANOVA. Everything after the sums of squares, including mean squares and F tests, is simply a matter of following directions (Table 3 is essentially a recipe card). The sum of squares calculation is where the magic happens.

The sum of squares for Factor A is computed as follows:

$$SS_A = \sum \left(\bar{Y}_i - \bar{Y}\right)^2$$

SS_A is computed as the difference between the mean of the group (or level) of Factor A from the overall mean. Unlike the $SS_{Between}$ computation we are specifying group mean to be the mean of a group of a certain factor (Factor A in this case).

The sum of squares for Factor B is similar, only with Factor B level means.

$$SS_B = \sum \left(\bar{Y}_j - \bar{Y}\right)^2$$

SS_B is computed as the difference between the mean of the group (or level) of Factor B from the overall mean.

There is a sum of squares $SS_{Between}$ in two-way ANOVA. We need to expand our idea of what $SS_{Between}$ means. In the one-way days, it was just how different the group means were. In two-way land, $SS_{Between}$ indicates the difference between the groups are at the smallest possible grouping, the cells. Thus, we compute $SS_{Between}$ as the difference between the cell mean and the overall mean. As a reminder, $SS_{Between}$ in two-way ANOVA is the sum of three sources of variance.

$$SS_{Between} = SS_A + SS_B + SS_{AB}.$$

We need the $SS_{Between}$ to compute interaction sum of squares (SS_{AB}). With that in mind, we compute $SS_{Between}$ as follows.

$$SS_{Between} = \sum \left(\bar{Y}_{ij} - \bar{Y}\right)^2$$

The key term in the above equation is $\bar{Y}_{ij}$, which is the cell mean. In multi-way ANOVA a cell is the specific combination of the independent variables. For example, cell (2,3) is Level 2 of Factor A and Level 3 of Factor B. So the equation for $SS_{Between}$ is computing the difference between the cell mean and the grand mean. These are between-group differences at the smallest grouping possible.

Because $SS_{Between} = SS_A + SS_B + SS_{AB}$ we can rearrange to solve for SS_{AB} once we know the other terms. In other words, compute between-cell variability and subtract from it the main effect variability to get interaction variability:

$$SS_{AB} = SS_{Between} - SS_A - SS_B$$

Finally, there is the sum of squares within to compute. In the one-way days, SS_W indicated the difference between the actual score on Y and the group mean. In multi-way ANOVA, it's almost the same but with cell means instead of group means.

$$SS_W = \sum \left(Y - \bar{Y}_{ij}\right)^2$$

Thus SS_W is calculated as the difference between the score on Y and the cell mean. It's within-cell

TABLE 5 Two-Way ANOVA Sum of Squares Part 1

A	B	Y	SS_A	SS_B
1	1	6	$(15-17)^2 = 4$	$(11-17)^2 = 36$
1	1	8	$(15-17)^2 = 4$	$(11-17)^2 = 36$
1	1	12	$(15-17)^2 = 4$	$(11-17)^2 = 36$
1	1	14	$(15-17)^2 = 4$	$(11-17)^2 = 36$
1	2	16	$(15-17)^2 = 4$	$(23-17)^2 = 36$
1	2	18	$(15-17)^2 = 4$	$(23-17)^2 = 36$
1	2	22	$(15-17)^2 = 4$	$(23-17)^2 = 36$
1	2	24	$(15-17)^2 = 4$	$(23-17)^2 = 36$
2	1	8	$(19-17)^2 = 4$	$(11-17)^2 = 36$
2	1	10	$(19-17)^2 = 4$	$(11-17)^2 = 36$
2	1	14	$(19-17)^2 = 4$	$(11-17)^2 = 36$
2	1	16	$(19-17)^2 = 4$	$(11-17)^2 = 36$
2	2	22	$(19-17)^2 = 4$	$(23-17)^2 = 36$
2	2	24	$(19-17)^2 = 4$	$(23-17)^2 = 36$
2	2	28	$(19-17)^2 = 4$	$(23-17)^2 = 36$
2	2	30	$(19-17)^2 = 4$	$(23-17)^2 = 36$

TABLE 6 Two-Way ANOVA Sum of Squares Part 2

A	B	Y	$SS_{Between}$	SS_{Within}
1	1	6	$(10-17)^2 = 49$	$(6-10)^2 = 16$
1	1	8	$(10-17)^2 = 49$	$(8-10)^2 = 4$
1	1	12	$(10-17)^2 = 49$	$(12-10)^2 = 4$
1	1	14	$(10-17)^2 = 49$	$(14-10)^2 = 16$
1	2	16	$(20-17)^2 = 9$	$(16-20)^2 = 16$
1	2	18	$(20-17)^2 = 9$	$(18-20)^2 = 4$
1	2	22	$(20-17)^2 = 9$	$(22-20)^2 = 4$
1	2	24	$(20-17)^2 = 9$	$(24-20)^2 = 16$
2	1	8	$(12-17)^2 = 25$	$(8-12)^2 = 16$
2	1	10	$(12-17)^2 = 25$	$(10-12)^2 = 4$
2	1	14	$(12-17)^2 = 25$	$(14-12)^2 = 4$
2	1	16	$(12-17)^2 = 25$	$(16-12)^2 = 16$
2	2	22	$(26-17)^2 = 81$	$(22-26)^2 = 16$
2	2	24	$(26-17)^2 = 81$	$(24-26)^2 = 4$
2	2	28	$(26-17)^2 = 81$	$(28-26)^2 = 4$
2	2	30	$(26-17)^2 = 81$	$(30-26)^2 = 16$

differences instead of within-group differences.

Example. Now it's time for a two-way ANOVA example. Much of the data will look familiar. In fact, it's much the same dataset as before, now with two independent variables. Factor A is still there. We have a Factor B now.

Table 5 lists the data for our example as well as the sums of squares calculation for the main effects, Factors A and B. As always, we need means. The grand mean is 17, the means for Levels 1 and 2 for Factor A are 15 and 19, respectively. The means for Levels 1 and 2 for Factor B are 11 and 23, respectively.

For Factor A the sum of squares is 64. For Factor B the sum of squares is 576. At the point, it is clear that there are much bigger differences between the levels on Factor B than on Factor A. Of course, you could have spotted this simply by noting that the Factor A level means (15 and 19) are much closer to the grand mean of 17 than the Factor B level means (11 and 23). That's exactly what the sum of squares for the various independent variables calculates. Treatment (or level) means that are further from the grand mean = a bigger sum of squares for that factor.

As for the sum of squares between and sum of squares within, I couldn't fit all of the column on one table, so I had to list them in a new table, Table 6 along with the raw data. Both sums of squares ($SS_{Betweeen}$ and SS_W) use cell means, so we need to list them now. The means for cell (1,1) is 10. For cell (1,2) the mean is 20. Cell (2,1) is 12. Cell (2,2) is 26.

The $SS_{Betweeen}$ is the difference between the cell mean and the grand mean. The sum of all of these squared differences is 656. The SS_W is the difference between the actual score on the dependent variable and the cell mean. This is cataloging within-cell differences in scores. The SS_W is 160 for the Table 6 data.

Finally, we can compute the sum of squares for the interaction of Factors A and B (SS_{AB}). Because $SS_{Between} = SS_A + SS_B + SS_{AB}$, we can compute SS_{AB} with a bit of algebraic rearrangement:

$$SS_{AB} = SS_{Between} - SS_A - SS_B$$

$$SS_{AB} = 656 - 64 - 576 = 16$$

Which doesn't sound like much.

Finally, we can compute some F tests. We need degrees of freedom first. Using the Table 2 information we find that both Factors A and B have 1 degree of freedom (both factors have two levels, so it's 2 - 1 = 1 for both). The interaction also has one degree of freedom, (2 - 1)(2 - 1) = 1. Finally, degrees of freedom for within deserve some attention. The equation is $ab(n - 1)$; a and b are 2, and n, the number of people per cell, is 4. Thus, we have $2 \times 2(4 - 1) = 12$ degrees of freedom within.

As for mean squares, this is where everything starts to become incredibly simple. Just divide the sum of squares by its df, and you have your mean squares.

$$MS_A = SS_A / df_A \quad = \quad 64/1 = 64$$

$$MS_B = SS_B / df_B \quad = \quad 576/1 = 576$$

$$MS_{AB} = SS_{AB} / df_{AB} \quad = \quad 16/1 = 16$$

$$MS_W = SS_W / df_W \quad = \quad 160/12 = 13.333$$

Table 3 shows us the way from here. Recall that one of the great things about fixed effects ANOVA is that the denominator mean square is MS_W for all three tests. Let's get the critical values first. So our critical F will have 1 degree of freedom for the numerator and 12 degrees of freedom for the denominator for all three tests. A check of the F table ($\alpha = .05$) indicates that F_{crit} (1, 12) = 4.75.

As for the F values, for Factor A we have:

$F_{A.obs} = MS_A/MS_W = 64/13.333 = 4.8$

For Factor B:

$F_{B.obs} = MS_B/MS_W = 576/13.333 = 43.2$

And for the interaction:

$F_{AB.obs} = MS_{AB}/MS_W = 16/13.333 = 1.2$

We can now draw our conclusions. Because F_{obs} for Factor A (4.8) is greater than F_{crit} (4.75), we reject the null and conclude that there are differences on Factor A. Because F_{obs} for Factor B (43.2) is greater than F_{crit} (4.75), we reject the null and conclude that there are differences on Factor B. Finally Because F_{obs} for the interaction (1.2) is not greater than F_{crit} (4.75), we do not reject the null, and we are unable to conclude that there is an interaction.

So that's three significance tests. Both main effects were significant, but the interaction was not. Speaking of interactions, further discussion is warranted. But first, let's compare the results of our two-way ANOVA test of Factor A with a one-way ANOVA. I think you'll find something interesting.

Two-Way ANOVA vs One-Way on Factor A

As we saw with Table 4 (one-way ANVOA) and Table 5 (two-way ANVOA), the sum of squares for Factor A is computed in the same fashion with the same result. In both cases, we are squaring (and summing) the difference between the group mean and the grand mean. Because it's the same dataset in both tables, the sum of squares was 64 for both. Lesson: computing the sum of squares for a main effect in two-way ANVOA is comparable to ignoring the second factor and treating it like it was a one-way ANOVA. (To that point, when we did the sum of squares for Factor B, we ignored Factor A.)

So computing the sum of squares for Factor A in a two way ANOVA is just like computing the sum of squares between for that same factor in a one-way ANOVA. We just act like Factor B isn't there.

Here's a question: If we are only interested in determining whether there is an effect for Factor A, is there any reason to conduct this analysis as a two-way ANOVA? The answer is yes. As for why, it's in the the sum of squares within. Let's compare the one-way and two-way ANVOAs again. In the one-way version, SS_W 752. In the two-way version $SS_W = 160$. Why the massive difference? Bear in mind that a bigger SS_W means a bigger MS_W, and given the nature of the F test ($F = MS_A/MS_W$), a smaller F_{obs}. All in all, we don't want the SS_W to be big.

So why is SS_W bigger in the one-way vs a two-way ANOVA of the same dataset. Let's remember that the total sum of squares (SS_T) is the same for both since they are the same data. The answer to that question can be found by looking at the equations for each version. In one-way ANOVA, SS_T is broken into two components ($SS_{Between}$ is the sum of squares for Factor A).

$$SS_T = SS_{Between} + SS_W$$

Two-way ANOVA breaks the $SS_{Between}$ into three parts ($SS_{Between} = SS_A + SS_B + SS_{AB}$). So the full equation looks like this:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_W$$

Another reminder, the data in Table 4 and Table 5 are the same. SS_T is the same, and SS_A is the same (64 for both). What do you think happens to the SS_B and SS_{AB} variance when you analyze it as a one-way? It goes into SS_W. And that's what we see here.

I'll fill in the values for the two-way

$$SS_T = 64 + 752$$

Now I'll fill in the values for the one-way

$SS_T = 64 + 576 + 16 + 160$

See what happened? All of the variance associated with Factor B (SS_B) and the interaction (SS_{AB}) in the two-way ANOVA shifted over to the SS_W in the one-way (576 + 16 +160 = 752). (This wouldn't be an issue if they had zero variance, but that's incredibly rare.) Let's compare the *F* tests of Factor A in both cases to take this comparison to its conclusion. In the one-way ANVOA F_{obs} is 1.19, and the F_{crit} is 4.6. We do not reject the null. In the two-way ANOVA the F_{obs} for Factor A is 4.8 with an F_{crit} of 4.75, allowing us to reject the null. Same dataset, different conclusions.

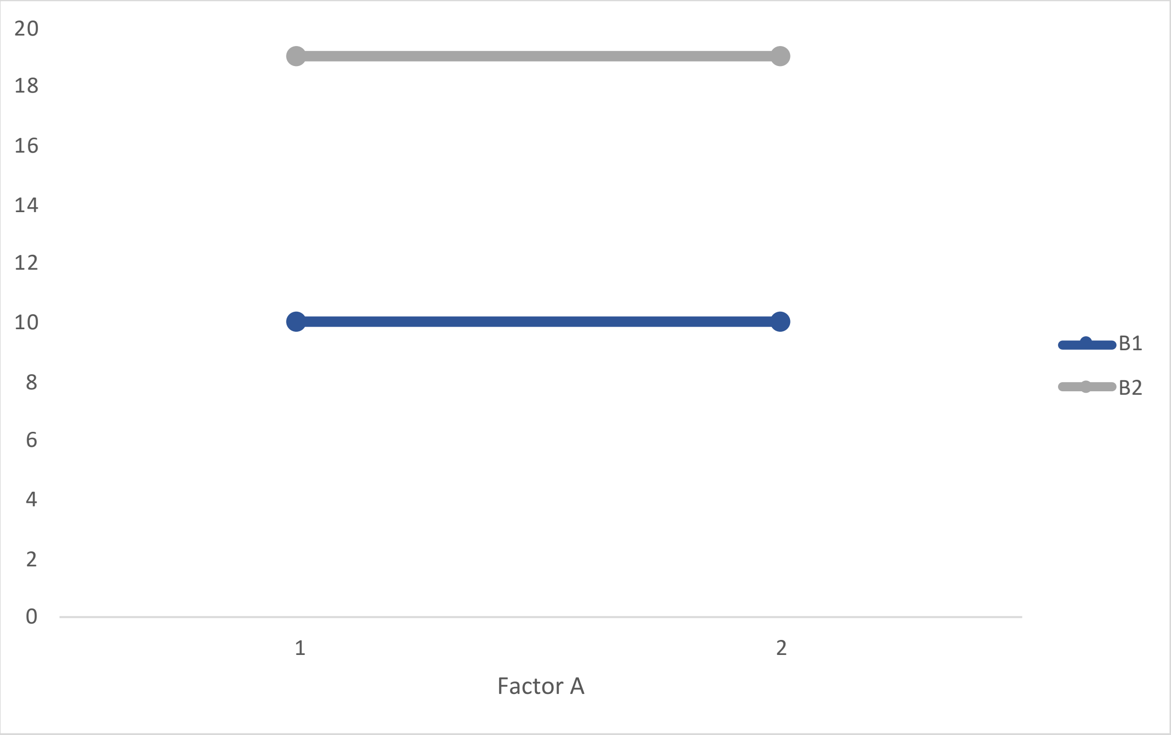
Long story short, if a variable is manipulated in any sort of multi-factor study (and this variable has a non-zero effect), you do not want the variance associated with that variable shoved into SS_W. You may not be all that interested in the second factor, but the proper way to test for the main effect of Factor A is in a two-way ANOVA. (Why is that second variable in your study if you're not interested in it? Probably someone, your advisor maybe, told you to include it even though it doesn't directly pertain to your hypotheses. These things happen in life.)

Interactions vs Main Effects

Let's use some graphs to understand main effects. After that we'll move on to interactions.

Example 1: One Main Effect (Factor B). This first example is a 2 × 2 ANOVA with one main effect. In this case, the effect is on Factor B ($p < .05$); there are no differences on Factor A ($p > .05$). The means for each of the four groups are displayed in Figure 2. In all of the graphs in this section, Factor A will be on the *x*-axis. Factor B will be represented by the separate lines. The dots indicate the means of the different cells. In this example, the mean of cell (1,1) is represented by the left dot on the lower (dark blue) line and is 10.0.

FIGURE 2 One Main Effect (Factor B)



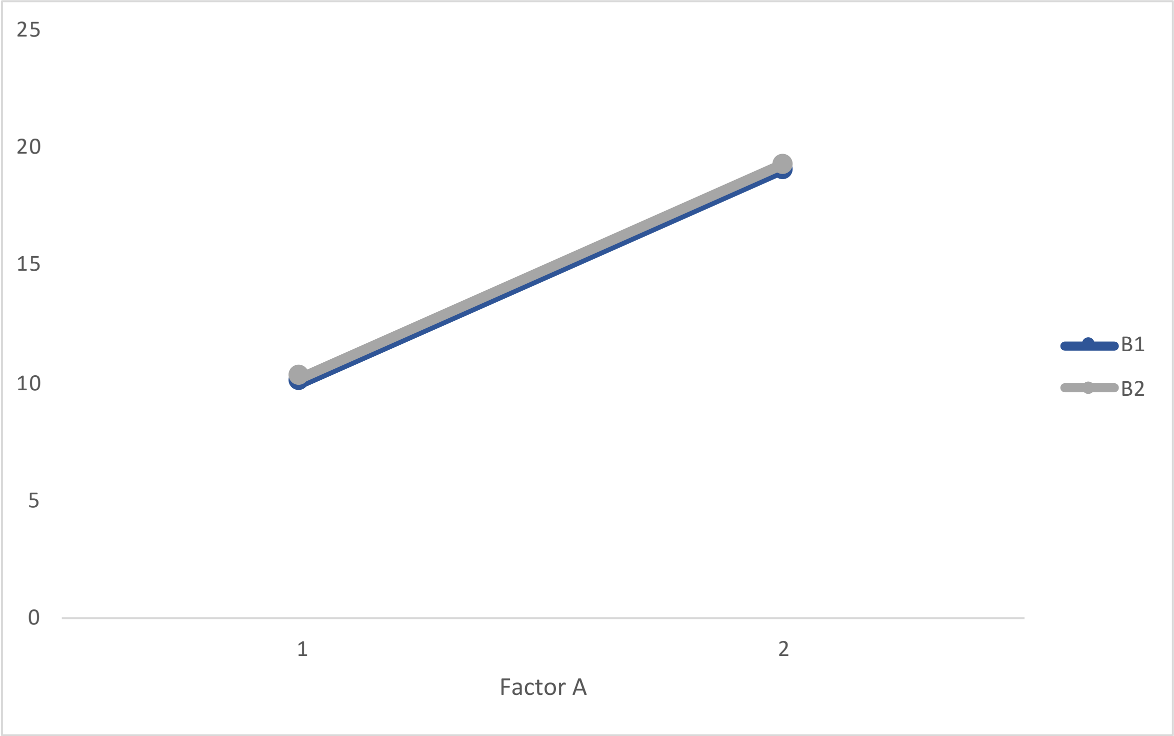
Let's understand why there is an effect for Factor B but not for Factor A. For Factor B, the means are very difference for subjects in Level 1 versus Level 2. For Factor A, the level matters not – the means are the same. As for the interaction, there is none (p > .05). This can be seen in that the lines are perfectly parallel. More on interactions later. Finally, if you want to inspect the ANOVA table that a stats program produces, the table of results for this analysis is given in Table 7.

TABLE 7 One Main Effect (Factor B)

Source	*SS*	*df*	*MS*	*F*	*p-value*	F_{crit}
A	0	1	0	0	1	4.75
B	324	1	324	17.4	0.001	4.75
AB	0	1	0	0	1	4.75
Within	224	12	18.67			
Total	548	15				

Example 2: One Main Effect (Factor A). In this new example, the effect is for Factor A (p < .05) not for Factor B (p > .05). The means for each of the four groups are displayed in Figure 3. The ANOVA table is shown in Table 8. In this example, there is no effect for Factor B – note how the lines are almost on top of each other. The means tell us why there is an effect for A but not one for B. Means are much higher for people in

FIGURE 3 One Main Effect (Factor A)



Level 2 of Factor A than for Level 1, but there is almost no difference between levels 1 and 2 of Factor B. The F tests bear this out. Also note that the lines are once again parallel, no interaction here.

Example 3: Two Main Effects. In this third example (Figure 4) both Factor A ($p < .05$) and Factor B ($p < .05$) have significant main effects (Table 9 lists the ANOVA results). In this example we

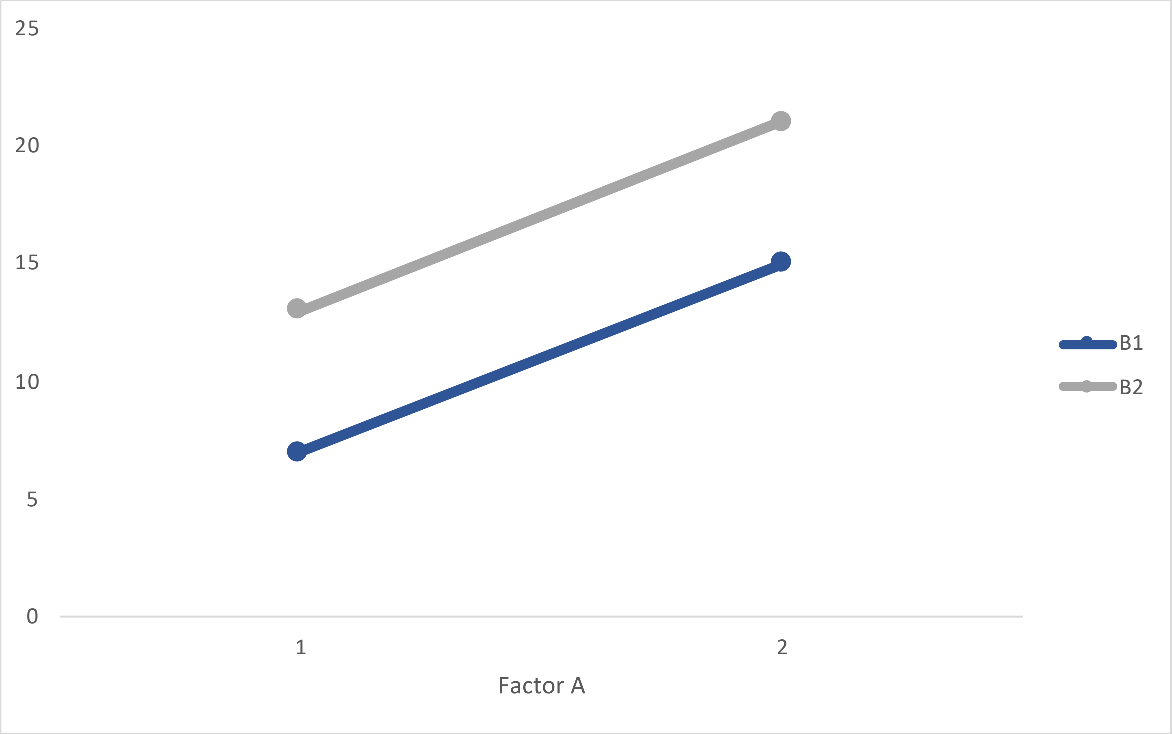
TABLE 8 One Main Effect (Factor A)

Source	*SS*	*df*	*MS*	*F*	*p-value*	F_{crit}
A	324	1	324	19.01	0.001	4.75
B	0.25	1	0.25	0.01	0.906	4.75
AB	0	1	0	0	1	4.75
Within	204.5	12	17.04			
Total	528.8	15				

can see that Level 2 means are greater than Level 1 means for both factors. Once again, the lines are parallel, indicating no interaction.

Example 4: No Main Effects with an Interaction. There are an infinite number of types of interactions, but let's look at the coolest one, the interaction that renders any talk of main effects pointless. To reiterate, interactions are observed when the lines are non-parallel; they do not have to cross. The easiest way to understand an interaction is that an interaction occurs when there are

FIGURE 4 Two Main Effects



differences in cell means that can't be explained by main effects alone.

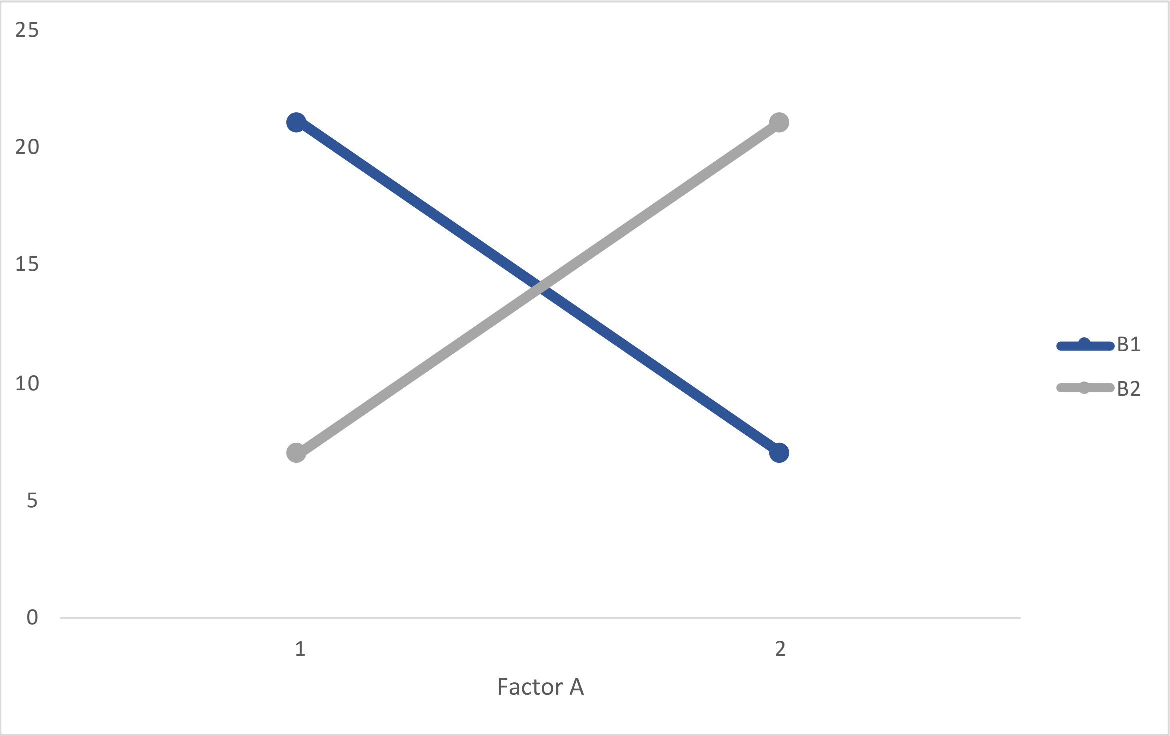
Figure 5 displays the means of a dataset with a significant interaction (see Table 10 for the ANOVA results). As shown in Figure 5 the cell means are 21 and 7 for the four cells. So there are differences among the cell means. Yet the main effects are all non-significant. Why? The reason is that when you collapse across cells, the mean for each level of Factor A is 14. The is also true of Factor B (mean = 14 for both levels). That's why there is no main effect. Each factor when examined in isolation appears to have zero differences on the dependent variable.

TABLE 9 Two Main Effects

Source	*SS*	*df*	*MS*	*F*	*p-value*	F_{crit}
A	256	1	256	11.82	0.005	4.75
B	144	1	144	6.65	0.002	4.75
AB	0	1	0	0	1	4.75
Within	260	12	21.67			
Total	660	15				

Now about that interaction. A significant *F* test for the interaction means that the lines are not parallel. In this case the slopes of the lines are completely inverted. This is a prime example of the principle that interactions can hide main ef-

FIGURE 5 No Main Effects, Significant Interaction



fects. I don't know if that's a great way to say it though. I think this example is showing us that interactions can make main effects a moot point. The important point is that even if you didn't hypothesize an interaction, you should always check for one as failing to do so in this case would give you the erroneous impression that these independent variables had no effect on the dependent variable. They do have an effect. It's just complicated and can't be described with a "Level 2 of Factor B has higher scores than Level 1" type statement.

TABLE 10 No Main Effects, Significant Interaction

Source	*SS*	*df*	*MS*	*F*	*p-value*	F_{crit}
A	0	1	0	0	1	4.75
B	0	1	0	0	1	4.75
AB	784	1	784	42	0.0001	4.75
Within	224	12	18.67			
Total	1008	15				

Example 5: Two Main Effects and an Interaction. As I mentioned at the beginning of this example, there are an infinite number of forms of the interaction. Interactions can exist with main effects as well.

Figure 6 displays a case in which both factors and the interaction are significant ($p < .05$ for all, see Table 11). This case is a perfect example of the

FIGURE 6 Two Main Effects and an Interaction

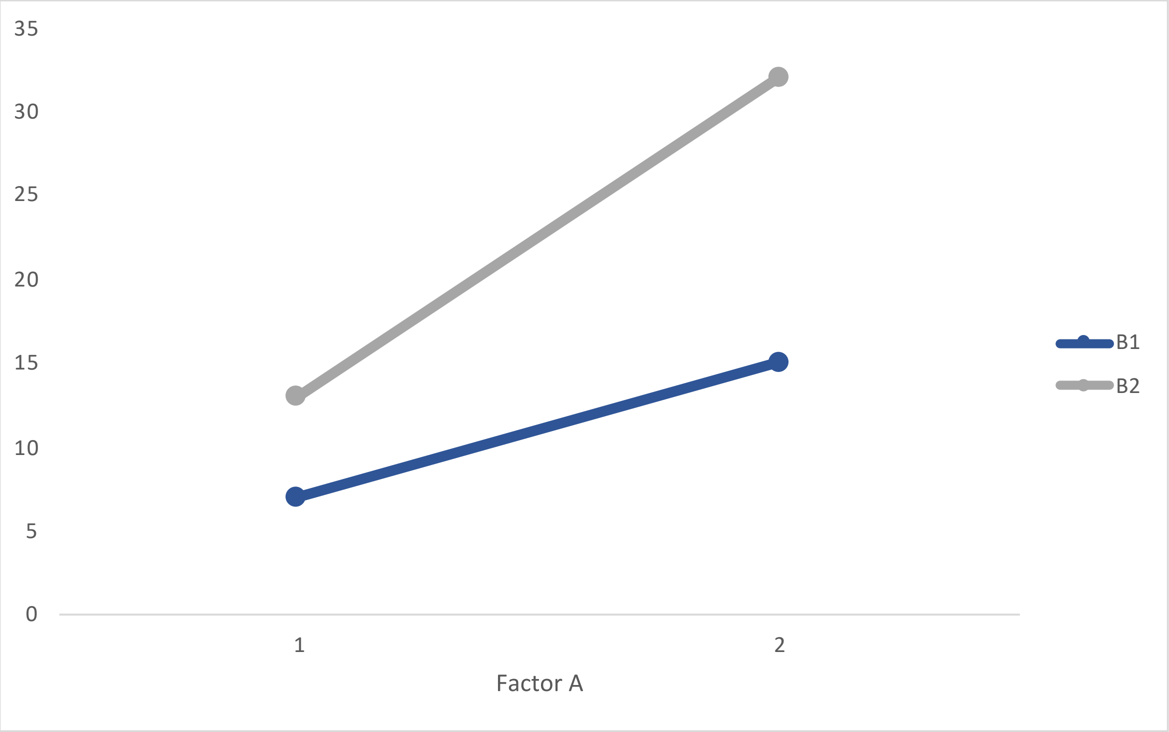


TABLE 11 Two Main Effects and an Interaction

Source	*SS*	*df*	*MS*	*F*	*p-value*	F_{crit}
A	729	1	729	33.65	0.0001	4.75
B	529	1	529	24.41	0.0001	4.75
AB	121	1	121	5.58	0.04	4.75
Within	260	12	21.67			
Total	1639	15				

nature of an interaction: differences among cell means that can't be explained by main effects alone. There are main effects for both factors; Level 2 means are greater than Level 1 means for both; however, the mean for cell (2, 2) is greater than the pattern suggested by the main effects. In short, it's as if the combination of Level 2 on Factor A and Level 2 on Factor B combines to form some sort of extra bonus. Also notice that the lines are not parallel.

Three-Way ANOVA and Beyond

There is no limit to the number of independent variables we can have in our studies. Three? Four? Fifty? Sure. As many as you want. The good news is that fixed effects ANOVA handles all of this with ease. It's just more of the same stuff. There is one piece of bad news: You'll have more interaction terms to test than anyone would ever want. In just a three-way ANOVA, there are four interaction terms: the three-way interaction between Factors A, B, and C plus three two-way interactions (AB, AC, and BC). And of course you still have your main effects to test. That's seven significance tests in all.

Now imagine what a five-way ANOVA would be like.

The Chi-Square Test of Independence

14

This really belongs in a different textbook

Maybe one covering correlation and regression

The Way It Was

To this point we have been looking for differences in dependent variable scores by level of the independent variable. The fundamental question always was: Are the dependent variable means different by level of the independent variable? Now we are going to address a different question: Is there a relationship between the independent variable and the dependent variable? This is a question that is most commonly addressed with a correlation, but when the variables are categorical, there is a more intuitive way to conduct the analysis. In this context, the terms dependent and independent are used instead of related and unrelated. If the variables are independent, then there isn't a relationship. To the converse, if the variables are dependent, then there is a relationship.

Chi-Square Test of Independence

The chi-square test of independence is our much easier correlation type test for a relationship between variables. As mentioned, both variables are categorical. Categorical not only means *not continuous* – it also means that there is no order to the levels (i.e., these are not selected levels of a continuous variable, we'll get to that later). Each variable can have any number of categories, but let's start with the simplest case, the 2 × 2 design.

The null hypothesis is that the variables are independent. More on this in bit. Table 1 displays the structure of the frequency table used for a chi-square analysis.

Each entry in the table is simply the number of cases in that condition. For example n_{11} is the number of people who are in Level 1 of row variable and Level 1 of the column variable. As to which variable, row or column, is the dependent variable, that's up to the researcher (because it

Table 1 Chi-Square Frequency Table Structure

		Column Variable		
		1	2	
Row Variable	1	n_{11}	n_{12}	$n_{1.}$
	2	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$n_{..}$

really doesn't matter). The convention for the various frequency terms is as follows.

n_{ij} = number observed for row_i, $column_j$

$n_{i.}$ = number observed for row_i

$n_{.j}$ = number observed for $column_j$

$n_{..}$ = total number of subjects (i.e., N)

An example will help illustrate things. Let's make our example about whether student level (undergraduate vs graduate student) is associated with having a job at the time of graduation. We collect data from a sample of 100 students (because we're making this up and having 100 total students makes for some easy math). At this point, let's say that we don't know the actual cell frequencies, but we know the category frequencies (i.e., row and column totals). Table 2 lists what we know so far.

So half of the sample has a job and half doesn't. And 80% of the sample is undergrad and 20% graduate. If these variables were independent (i.e., no association), then we would expect to see

Table 2 Example with Totals Only

		Student Level		
		Under	Grad	
Job?	No			50
	Yes			50
		80	20	100

that half of the undergraduates have jobs and half of the graduate students have jobs. (Or, we could state that 80% of the job holders will be undergrads and 80% of the non-job people will be undergrads, but that's less interesting). Let's fill in the cells with values to fit these expectations (Table 3).

How did I get these values? Because the math is so easy in this case, you can just reason it out. But we need an actual equation to do this for those times where the numbers aren't as friendly.

TABLE 3 Expected Frequencies If Independent

		Student Level		
		Under	Grad	
Job?	No	40	10	50
	Yes	40	10	50
		80	20	100

Just multiply the respective column and row means and divide by the total *N*. For example, for Cell (1,1) the expected frequency (e_{11}) is:

$$e_{11} = (80 \times 50)/100 = 40$$

Even without the math, it should be clear that if there are 80 undergrads, and we expect half of them to have jobs, then that's 40 people. Here's the general form of the equation for an expected frequency (expected frequency means *expected if the variables are independent*):

$$e_{ij} = \frac{(n_i\,.)(n\,._j)}{n\,.\,.}$$

Which is really just:

$$e_{ij} = \frac{(\mathit{row\ total})(\mathit{column\ total})}{\mathit{grand\ total}}$$

So, those are the expected frequencies. You have probably guessed that we will be comparing the expected frequencies (e_{ij}) to the observed fre-

Table 4 Observed Frequencies

		Student Level		
		Under	Grad	
Job?	No	46	4	50
	Yes	34	16	50
		80	20	100

quencies (n_{ij}) to determine if there is a relationship. (Table 4 lists the observed frequencies.) If the observed frequencies deviate from the expected frequencies, then the variables are not independent. Of course, a small deviation is to be expected given sampling error. Thus, we'll need to construct a significance test to determine how big of a difference is necessary to allow us to reject the null that there are no differences.

Before we get to that, let's talk null and alternative hypotheses (note that we lack proper symbols to differentiate between sample and population values, but rest assured, these hypotheses concern population values).

H_0: Level and Job are independent ($n_{ij} = e_{ij}$ for all cells)

H_1: Level and Job are dependent ($n_{ij} \neq e_{ij}$ for at least one cell)

Now, on to the observed frequencies:

No surprise, the observed frequencies depart from the expected values. We need a statistic to index the magnitude of the departure. The obvious approach is to compute the sum of the difference between observed and expected frequencies for each cell (i.e., $\sum (n_{ij} - e_{ij})$). However, as we have seen before (cf. variance) this will always sum to zero. As with variance, we square these differences so they can sum to something other than zero. One more issue: if we go this route, the biggest cells will have a disproportionate effect on the result. So, let's adjust this squared difference by the expected frequency. This gives us something approximately distributed as a chi-square (χ^2).

The chi-square distribution is similar to the *F* distribution in shape; it's skewed with the unusual scores in the right tail. Unlike the *F*, the chi-square has a single degree of freedom. Selected chi-square critical values are shown in Table 5.

As for the test statistic, it is listed below.

$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Where: r = the number of rows and c = the number of columns

With $(r - 1)(c - 1)$ degrees of freedom

This equation simply tell us to (a) divide the squared difference between the observed and expected values by the expected value for each cell and (b) sum that value across all cells. It's as easy as it sounds.

TABLE 5 Selected Chi-Square Critical Values (α = .05)

df	Chi-Square	*df*	Chi-Square
1	3.841	9	16.919
2	5.991	10	18.307
3	7.815	15	24.996
4	9.488	20	31.410
5	11.07	25	37.652
6	12.592	30	43.773
7	14.067	40	55.758
8	15.507	50	67.505

As for our example, the observed and expected frequencies (expected frequencies in parentheses) are shown in Table 6. None of the values in this table are new to us. I just listed it again this way because this is the traditional method of

TABLE 6 Observed and Expected Frequencies

		Student Level		
		Under	Grad	
Job?	No	46 (40)	4 (10)	50
	Yes	34 (40)	16 (10)	50
		80	20	100

showing the observed and expected values for a chi-square test of independence. The calculation of χ^2_{obs} is shown in Table 7. According to Table 7, the χ^2_{obs} for our example dataset is 9.0 with 1 degree of freedom, $df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$.

Or so you would think. Remember how I said that this test statistic was *approximately* distributed as a chi-square? Well, I did. You can check. This approximation will be on solid ground if:

TABLE 7 Chi-Square Computation (Part 1)

Cell	$n_{ij} - e_{ij}$	$(n_{ij} - e_{ij})^2$	$(n_{ij} - e_{ij})^2/e_{ij}$
1, 1	46 - 40 = 6	$(6)^2 = 36$	36/40 = .9
1, 2	4 - 10 = -6	$(-6)^2 = 36$	36/10 = 3.6
2, 1	34 - 40 = -6	$(-6)^2 = 36$	36/40 = .9
2, 2	16 - 10 = 6	$(6)^2 = 36$	36/10 = 3.6
			$\sum = 9.0$

(a) e_{ij} for each cell is greater than or equal to 5 for any analysis with designs greater than 2 × 2 (greater than 10 for a 2 × 2 design).

(b) and a continuity correction is used for the 2 × 2 design:

$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|n_{ij} - e_{ij}| - .5)^2}{e_{ij}}$$

Where the numerator is 0 for cells in which $n_{ij} = e_{ij}$.

This being a 2 × 2 design, we need to re-compute our previous example to include the continuity correction. Table 8 lists the new version of the calculation for a 2 × 2 design, yielding a χ^2_{obs} of 7.5625.

According to the table of chi-square critical values (Table 5), $\chi^2_{obs}(1) = 3.841$, allowing us to reject the null that level and job status are independent. We conclude that they are dependent, meaning that there is a relationship between student level and having a job at graduation (the nature of the relationship is unknown, but they are related).

Table 8 Chi-Square Computation (Part 2)

Cell	$\lvert n_{ij} - e_{ij} \rvert$	$(\lvert n_{ij} - e_{ij} \rvert - .5)^2$	$(\lvert n_{ij} - e_{ij} \rvert - .5)^2/e_{ij}$
1, 1	$\lvert 46 - 40 \rvert = 6$	$(6 - .5)^2 = 30.25$	30.25/40 = .756
1, 2	$\lvert 4 - 10 \rvert = 6$	$(6 - .5)^2 = 30.25$	30.25/10 = 3.025
2, 1	$\lvert 34 - 40 \rvert = 6$	$(6 - .5)^2 = 30.25$	30.25/40 = .756
2, 2	$\lvert 16 - 10 \rvert = 6$	$(6 - .5)^2 = 30.25$	30.25/10 = 3.025
			$\sum = 7.5625$

Measures of Association Based On Chi-Square

The chi-square test is a significance test. Like all significance tests, it is heavily dependent on sample size and does not convey magnitude of association, or effect size. We'll need a different statistic for that. This will sound very familiar to fans of ANOVA where the *F* test just told us if there were differences but did not indicate the magnitude of the differences (we used eta-squared for that). Or for the two-group scenario, think of the relationship between the *t* test and Cohen's *d*. As with the ANOVA situation, we'll be using the same basic building blocks to get what we want here. Unlike ANOVA, we will have two measures to enjoy

Phi Coefficient

$$\phi = \sqrt{\chi^2_{obs}/N}$$

Nice and simple. Note that the χ^2_{obs} used is the one without the continuity correction. The only problem is that phi ranges from 0 to $\sqrt{\text{min}(r-1, c-1)}$. That's right, the square root of $r-1$ or $c-1$, whichever is smaller

This is not what we want. We want something that ranges from 0 to 1. That's where our next index steps in

Cramer's V

$$V = \frac{\phi}{\sqrt{min(r-1, c-1)}}$$

Or,

$$V = \frac{\sqrt{\chi^2_{obs}/N}}{\sqrt{min(r-1, c-1)}}$$

For our example, phi = $\sqrt{9/100}$ = .30 and V = .30/1 = .30. That's right, for a 2 × 2 design $\phi = V$.

Directional Hypotheses with a 2 × 2 Design

You may have noticed something about the null and alternative hypotheses to the chi-square test: there aren't a lot of options. Either the variables are independent or they aren't. My point is that there is no direction to the chi-square test. Well, with a 2 × 2 design, we might want to test the direction of the association. There are a few ways to do this, but the simplest is to compute a correlation.

But wait, you say. I thought the correlation was only for continuous data – these data are decidedly not continuous. Well, there is a special correlation for two dichotomous variables, the phi correlation. Better news: it's really the same correlation that we use for two continuous variables (the Pearson correlation). The phi correlation is actually a simplified version of the Pearson equation – simplified for dichotomous data. Note that the version derived from a chi-square statistic (previous section) can only yield a positive value – there are other versions that indicate direction. In the case of our previous example, I coded Job as 0 for no and 1 for yes. I coded level as 1 for undergrad and 2 for grad. With these codes phi correlation is +.3. Had I coded undergrad and grad in the opposite fashion, the correlation would have been -.3. Note that the coding scheme made no difference to the chi-square test because there is no direction to the chi-square test – differences are differences, they get squared.

As for a significance test for a correlation, it's a fairly simple *t* test:

$$t_{obs} = \frac{r}{\sqrt{\frac{1 - r^2}{N - 2}}}$$

With $N - 2$ degrees of freedom

Final Thoughts on the Chi-Square Test

The chi-square test is wonderfully simple and intuitive. But this simplicity causes limitations. We know that it's insensitive to direction in a 2 × 2 design – you'll need to use a real correlation to test for a direction. What about ordered categories when there are more than two levels? As you may have guessed, the chi-square test is insensitive to this as well. It can't tell if things are falling in a certain order. It only knows that there are differences between the observed and expected frequencies to the point we conclude that the variables are not independent. If you don't believe me, find any chi-square with a design more complicated than a 2 × 2 and switch up the order of the categories. You will get the same result every time. There is no order to the categories on a chi-square test of independence.

How to test if there is an association with more than two ordered categories? There are a few options, but I favor correlations. If both variables are continuous measured at the interval level, use the Pearson correlation. Otherwise (i.e., with ordinal data), use Spearman's correlation.

References

15

The responsible parties

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Orlando, FL: Harcourt Brace.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.