Significantly Significant

What It Means for You and Me

Difficulty Scale 😊😊 (somewhat thought provoking and key to it all!)

How much Excel? 🐬 (just a mention)

WHAT YOU’LL LEARN ABOUT IN THIS CHAPTER

✦ What the concept of significance is and why it is important
✦ The importance of and difference between Type I and Type II errors
✦ How inferential statistics works
✦ How to select the appropriate statistical test for your purpose

THE CONCEPT OF SIGNIFICANCE

Probably no term or concept causes the beginning statistics student more confusion than the concept of statistical significance. But it doesn’t have to be that way for you. Although it’s a powerful idea, it is also relatively straightforward and can be understood by anyone in a basic statistics class.

We need an example of a study to illustrate the points we want to make. Let’s take E. Duckett and M. Richards’s “Maternal Employment and Young Adolescents’ Daily Experiences in Single Mother Families” (paper presented at the Society for Research in Child Development, Kansas City, MO, 1989—a long time ago in a galaxy far, far away . . . ). These two authors examined the attitudes of 436 fifth- through ninth-grade adolescents toward maternal employment. Even though the presentation took place some years ago, as an example it’s perfect for illustrating many of the important ideas at the heart of this chapter.
Specifically, the two researchers investigated whether differences are present between the attitudes of adolescents whose mothers work and the attitudes of adolescents whose mothers do not work. They also examined some other factors, but for this example, we'll stick with the mothers-who-work and mothers-who-don't-work groups. One more thing. Let's add the word significant to our discussion of differences, and we have a research hypothesis something like this:

There is a significant difference in attitude toward maternal employment between adolescents whose mothers work and adolescents whose mothers do not work, as measured by a test of emotional state.

What we mean by the word significant is that any difference between the attitudes of the two groups is due to some systematic influence and not due to chance. In this example, that influence is whether or not mothers work. We assume that all of the other factors that might account for any differences between groups were controlled. Thus, the only thing left to account for the differences between adolescents' attitudes is whether or not their mothers work. Right? Yes. Finished? Not quite.

**If Only We Were Perfect**

Because our world is not a perfect one, we must allow for some leeway in how confident we are that only those factors we identify could cause any difference in our measure (or measures) of interest between groups. In other words, you need to be able to say that although you are pretty sure the difference between the two groups of adolescents is due to maternal employment, you cannot be absolutely, 100%, positively, unequivocally, indisputably (get the picture?) sure. There's always a chance, no matter how small, that you are wrong.

Why? Many reasons. For example, you could (horrors) just be plain ol' wrong. Maybe during this one experiment, differences between adolescents' attitudes were not due to whether mothers worked or didn't work but were due to some other factor that was inadvertently not accounted for, such as a speech given by the local Mothers Who Work Club that several students attended. How about if the people in one group were mostly adolescent males and the people in the other group were mostly adolescent females? That could be the source of a difference as well.
If you are a good researcher and do your homework, you can account for such differences, but it's always possible that you can't. And as a good researcher, you have to take that possibility into account.

So what do you do? In most scientific endeavors that involve testing hypotheses (such as the group differences example here), there is bound to be a certain amount of error that cannot be controlled—this is the chance factor that we have been talking about in the past few chapters. The level of chance or risk you are willing to take is expressed as a **significance level**, a term that unnecessarily strikes fear in the hearts of even strong men and women. Significance level (here's the quick-and-dirty definition) is the risk associated with not being 100% confident that what you observe in an experiment is due to the treatment or what was being tested—in our example, whether or not mothers worked. If you read that significant findings occurred at the .05 level (or $p < .05$ in the tech talk you regularly see in professional journals), the translation is that there is 1 chance in 20 (or .05 or 5%) that any differences found were not due to the hypothesized reason (whether mom works) but to some other, unknown reason or reasons. Your job is to reduce this likelihood as much as possible by removing all of the competing reasons for any differences that you observed. Because you cannot fully eliminate the likelihood (because no one can control every potential factor), you assign some level of probability and report your results with that caveat.

In sum (and in practice), the researcher defines a level of risk that he or she is willing to take. If the results fall within the region that says, “This could not have occurred by chance alone—something else is going on,” the researcher knows that the null hypothesis (which states an equality) is not the most attractive explanation for the observed outcomes. Instead, the research hypothesis (that there is an inequality or a difference) is the favored explanation.

Let’s take a look at another example, and this one is hypothetical.

A researcher is interested in whether there is a difference in the academic achievement of children who participated in a preschool program and children who did not participate. The null hypothesis is that the two groups are equal to each other on some measure of achievement.

The research hypothesis is that the mean score for the group of children who participated in the program is higher than the mean score for the group of children who did not participate in the program. As a good researcher, your job is to show (as best you can—and no one is so perfect that he or she can account for everything) that any difference that exists between the two groups is due only to the effects of the preschool experience and no other factor or combination of factors.
However, through a variety of techniques (that you'll learn about in your Stats II class!), you control or eliminate all the possible sources of difference, such as the influence of parents' education, number of children in the family, and so on. Once these other potential explanatory variables are removed, the only remaining alternative explanation for differences is the effect of the preschool experience itself.

But can you be absolutely (which is pretty darn) sure? No, you cannot. Why? First, because you can never be sure that you are testing a sample that reflects the profile of the population in every single respect. And even if the sample perfectly represents the population, there could always be other influences that might affect the outcome that you inadvertently missed when designing the experiment. There's always the possibility of error (sort of another word for chance).

By concluding that the differences in test scores are due to differences in treatment, you accept some risk. This degree of risk is, in effect (drumroll, please), the level of statistical significance at which you are willing to operate. 

**Statistical significance** (here's the formal definition) is the degree of risk you are willing to take that you will reject a null hypothesis when it is actually true. For our example above, the null says that there is no difference between the two sample groups (remember, the null is always a statement of equality). In your data, however, you did find a difference. That is, given the evidence you have so far, group membership seems to have an effect on achievement scores. In reality, however, maybe there is no difference. If you reject the null you stated, you would be making an error. The risk you take in making this kind of error (or the level of significance) is also known as a Type I error.

---

**The World's Most Important Table (for This Semester Only)**

Here's what it all boils down to.

A null hypothesis can be true or false. Either there really is no difference between groups, or there really and truly is an inequality (such as the difference between two groups). But remember, you'll never know this true state because the null cannot be tested directly (remember that the null applies only to the population, and for a variety of reasons we have talked about, the population cannot be directly tested).
And, as a crackerjack statistician, you can choose to either reject or accept the null hypothesis. Right? These four different conditions create the table you see here in Table 9.1.

Let’s look at each cell.

**More About Table 9.1**

Table 9.1 has four important cells that describe the relationship between the nature of the null (whether it’s true or not) and your action (fail to reject or reject the null hypothesis). As you can see, the null can be either true or false, and you can either reject or fail to reject it.

The most important thing about understanding this table is the fact that the researcher never really knows the true nature of the null hypothesis and whether there really is or is not a difference between groups. Why? Because the population (which the null

<table>
<thead>
<tr>
<th>Table 9.1 Different Types of Errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Action You Take</strong></td>
<td><strong>Accept the Null Hypothesis</strong></td>
</tr>
<tr>
<td>True nature of the null hypothesis</td>
<td>1 🎉 Bingo, you accepted a null when it is true and there is really no difference between groups.</td>
</tr>
<tr>
<td></td>
<td>3 😞 Uh-oh—you made a Type II error and accepted a false null hypothesis. Type II errors are also represented by the Greek letter beta, or β.</td>
</tr>
</tbody>
</table>
represents) is never tested directly. Why? Because it's impractical to do so. That's why we have inferential statistics.

- 😎 Cell 1 in Table 9.1 represents a situation in which the null hypothesis is really true (there's no difference between groups) and the researcher made the correct decision accepting it. No problem here. In our example, our results would show that there is no difference between the two groups of children, and we have acted correctly by accepting the null that there is no difference.

- 😎 Oops. Cell 2 represents a serious error. Here, we have rejected the null hypothesis (that there is no difference) when it is really true (and there is no difference). Even though there is no difference between the two groups of children, we will conclude there is, and that's an error—clearly a boo-boo called a **Type I error**, also known as the level of significance.

- 😎 Uh-oh, another type of error. Cell 3 represents a serious error as well. Here, we have accepted the null hypothesis (that there is no difference) when it is really false (and, indeed, there is a difference). We have said that even though there is a difference between the two groups of children, we will conclude there is not—also clearly a boo-boo, this time known as a **Type II error**.

- 😎 Cell 4 in Table 9.1 represents a situation where the null hypothesis is really false and the researcher made the correct decision in rejecting it. No problem here. In our example, our results show that there is a difference between the two groups of children, and we have acted correctly by rejecting the null that states there is no difference.

So, if .05 is good and .01 is even “better,” why not set your Type I level of risk at .000001? For the very good reason that you would be so rigorous in your rejection of false null hypotheses that you might reject a null when it was actually true. Such a stringent Type I error rate allows for little leeway—indeed, the research hypothesis might be true but the associated probability might be .015—still quite rare and probably very useful information, but missed with the too-rigid Type I level of error.

**Back to Type I Errors**

Let's focus a bit more on Cell 2 in Table 9.1, where a Type I error was made, because this is the focus of our discussion.
This Type I error, or level of significance, has certain values associated with it that define the risk you are willing to take in any test of the null hypothesis. The conventional levels set are between .01 and .05.

For example, if the level of significance is .01, then on any one test of the null hypothesis, there is a 1% chance you will reject the null hypothesis when the null is true and conclude that there is a group difference when there really is no group difference at all.

If the level of significance is .05, it means that on any one test of the null hypothesis, there is a 5% chance you will reject it when the null is true (and conclude that there is a group difference) when there really is no group difference at all. Notice that the level of significance is associated with an independent test of the null. Therefore, it is not appropriate to say that “on 100 tests of the null hypothesis, I will make an error on only 5, or 5% of the time.”

In a research report, statistical significance is usually represented as \( p < .05 \), read as “the probability of observing that outcome is less than .05,” often expressed in a report or journal article simply as “significant at the .05 level.”

With the introduction of fancy-schmancy software such as SPSS and Excel that can do statistical analysis, there’s no longer the worry about the imprecision of such statements as \( p < .05 \) or \( p < .01 \). For example, \( p < .05 \) can mean anything from .000 to .049999, right? Instead, software such as Excel gives you the exact probability, such as .013 or .158, of the risk you are willing to take that you will commit a Type I error. So, when you see in a research article the statement that \( p < .05 \), it means that the value of \( p \) is equal to anything from .00 to .04999999999 (you get the picture). Likewise, when you see \( p > .05 \) or \( p = \text{n.s.} \) (for nonsignificant), it means that the probability of rejecting a true null exceeds .05 and, in fact, can range from .0500001 to 1.00. So, it’s actually terrific when we know the exact probability of an outcome because we can measure more precisely the risk we are willing to take.

But what to do if the \( p \) value is exactly .05? Well, given what you’ve already read, if you want to play by the rules, then the outcome is not significant. A result either is, or is not. So, .04999999999 is and .05 is not. Now, if Excel (or any other program) generates a value of .05, extend the number of decimal places—it may really be .04999999999.

There is another kind of error you can make, which, along with the Type I error, is shown in Table 9.1. A Type II error (Cell 3 in the chart) occurs when you inadvertently accept a false null hypothesis.
When talking about the significance of a finding, you might hear the word power used. Power is a construct that has to do with how well a statistical test can detect and reject a null hypothesis when it is false. The value is calculated by subtracting the value of the Type II error from 1, or $1 - \beta$. A more powerful test is always more desirable than a less powerful test, because the more powerful one lets you get to the heart of what’s false and what’s not.

For example, there may really be differences between the populations represented by the sample groups, but you mistakenly conclude there are not.

Ideally, you want to minimize both Type I and Type II errors, but doing so is not always easy or under your control. You have complete control over the Type I error level or the amount of risk that you are willing to take (because you actually set the level itself). Type II errors are not as directly controlled but, instead, are related to factors such as sample size. Type II errors are particularly sensitive to the number of subjects in a sample; as that number increases, Type II error decreases. In other words, as the sample characteristics more closely match those of the population (achieved by increasing the sample size), the likelihood that you will accept a false null hypothesis decreases. Just as you would expect, right?

**SIGNIFICANCE VERSUS MEANINGFULNESS**

What an interesting situation for the researcher when he or she discovers that the results of an experiment indeed are statistically significant. You know technically what statistical significance means—that the research was a technical success and the null hypothesis is not a reasonable explanation for what was observed. Now, if your experimental design and other considerations were well taken care of, statistically significant results are unquestionably the first step toward making a contribution to the literature in your field. However, the value of statistical significance and its importance or meaningfulness must be kept in perspective.

For example, let’s take the case where a very large sample of illiterate adults (say, 10,000) is divided into two groups. One group receives intensive training to read using classroom teaching, and the other receives intensive training to read using computers. The average score for Group 1 (which learned in the classroom) is 75.6 on a reading test, the outcome variable. The average score on the reading test for Group 2 (which learned using the computer) is 75.7.
The amount of variance in both groups is about equal. As you can see, the difference in score is only one tenth of 1 point (75.6 vs. 75.7), but when a t-test for the significance between independent means is applied, the results are significant at the .01 level, indicating that computers do work better than classroom teaching on the variable of interest. (Chapters 11 and 12 discuss t-tests.)

The difference of 0.1 is indeed statistically significant, but is it meaningful? Does the improvement in test scores (by such a small margin) provide sufficient rationale for the $300,000 it costs to fund the program? Or is the difference negligible enough that it can be ignored, even if it is statistically significant?

Here are some conclusions about the importance of statistical significance that we can reach, given this and the countless other possible examples:

- Statistical significance, in and of itself, is not very meaningful unless the study has a sound conceptual base that lends some meaning to the significance of the outcome.
- Statistical significance cannot be interpreted independently of the context within which it occurs. For example, if you are the superintendent in a school system, are you willing to retain children in Grade 1 if the retention program significantly raises their standardized test scores by one half point?
- Although statistical significance is important as a concept, it is not the end-all and certainly should not be the only goal of scientific research. That is the reason why we set out to test hypotheses rather than prove them. If your study is designed correctly, then even null results tell you something very important. If a particular treatment does not work, that's important information that others need to know about. If your study is designed well, then you should know why the treatment does not work, and the next person down the line can design his or her study taking into account the valuable information you provided.

AN INTRODUCTION TO INFERENTIAL STATISTICS

Whereas descriptive statistics are used to describe a sample’s characteristics, inferential statistics are used to infer something about the population based on the sample’s characteristics.

At several points throughout the first half of Statistics for People Who (Think They) Hate Statistics, Excel 2016 Edition, we have emphasized that a hallmark of good scientific research is choosing
a sample in such a way that it is representative of the population from which it was selected. The process then becomes an inferential one, in which you infer from the smaller sample to the larger population based on the results of tests (and experiments) conducted using the sample.

Before we start discussing individual inferential tests, let’s go through the logic of how the inferential method works.

**How Inference Works**

Here are the general steps to see how the process of inference might work. We’ll stay with adolescents’ attitudes toward mothers working as an example.

Here’s the sequence of events that might happen:

1. The researcher selects representative samples of adolescents who have mothers who work and adolescents who have mothers who do not work. These are selected in such a way that the samples represent the populations from which they are drawn.

2. Each adolescent is administered a test to assess his or her attitude. The mean scores for groups are computed and compared using some statistical test.

3. A conclusion is reached as to whether the difference between the scores is the result of chance (meaning some factor other than moms working is responsible for the difference) or the result of “true” and statistically significant differences between the two groups (meaning the results are due to moms working).

4. A conclusion is reached as to the relationship between maternal employment and adolescents’ attitudes in the population from which the sample was originally drawn. In other words, an inference, based on the results of an analysis of the sample data, is made about the population of all adolescents.

**How to Select What Test to Use**

Step 3 above brings us to ask the question “How do I select the appropriate statistical test to determine whether a difference between groups exists?” Heaven knows, there are plenty of them (many hundreds), and you have to decide which one to use and when to use it. Well, the best way to learn which test to use is to be
an experienced statistician who has taken lots of courses in this area and participated in lots of research. Experience is still the greatest teacher. In fact, there’s no way you can really learn what to use and when to use it unless you’ve had the real-life, applied opportunity to use these tools. And as a result of taking this course, you are learning how to use these very tools.

So, for our purposes and to get started, we’ve created this nice little flowchart (aka cheat sheet) of sorts that you see in Figure 9.1. You have to have some idea of what you’re doing so that selecting the correct statistical test is not done entirely on autopilot, but this flowchart certainly is a good place to start.

Don’t think for a second that Figure 9.1 takes the place of your need to learn about when these different tests are appropriate. The flowchart is here only to help you get started.

This is really important. We just wrote that selecting the appropriate statistical test is not necessarily an easy thing to do. And the best way to learn how to do it is to do it, and that means practicing and even taking more statistics courses. The simple flowchart we present here works, but use it with caution. When you make a decision, check with your professor or some other person who has been through this stuff and feels more confident than you might (who also knows more!). You may also find the neat tool named Statistical Navigator at http://rimarcik.com/en/navigator/ to be interesting and of some help.

---

*Here’s How to Use the Chart*

1. Assume that you’re very new to this statistics stuff (which you are) and that you have some idea of what these tests of significance are, but you’re pretty lost as far as deciding which one to use when.

2. Answer the questions at the top of the flowchart.

3. Proceed down the chart by answering each of the questions until you get to the end of the chart. That’s the statistical test you should use. This is not rocket science, and with some practice (which you will get throughout this part of *Statistics for People* . . . ), you’ll be able to quickly and reliably select the appropriate test. Each of the chapters in this part of the book will begin with a chart like the one you see in Figure 9.1 and take you through the specific steps for the test statistic you should use.
Figure 9.1  A Quick (but Not Always Great) Approach to Determining What Statistical Test to Use

Are you examining differences between one sample and a population?

I’m examining relationships between variables.

How many variables are you dealing with?

One-sample Z-test

Two variables

t-test for the significance of the correlation coefficient

More than two variables

Regression, factor analysis, or canonical analysis

Are you examining relationships between variables or examining the difference between groups on one or more variables?

I’m examining differences between groups on one or more variables.

How many groups are you dealing with?

Two groups

t-test for dependent samples

More than two groups

Repeated measures analysis of variance

Are the same participants being tested more than once?

Yes

No

How many groups are you dealing with?

Two groups

t-test for independent samples

More than two groups

How many factors are you dealing with?

One

Simple analysis of variance

More than one

Factorial analysis of variance
Does the cute flowchart in Figure 9.1 contain all the statistical tests there are? Not by a long shot. There are hundreds, but those in Figure 9.1 are the ones used most often (and we’ll discuss most of them in coming chapters). And if you are going to become familiar with the research in your own field, you are bound to run into these.

AN INTRODUCTION TO TESTS OF SIGNIFICANCE

What inferential statistics does best is allow decisions to be made about populations based on information about samples. One of the most useful tools for doing this is a test of statistical significance that can be applied to different types of situations depending on the nature of the question being asked and the form of the null hypothesis.

For example, do you want to look at the difference between two groups, such as whether boys score significantly differently than girls on some test? Or the relationship between two variables, such as number of children in a family and average score on intelligence tests? The two cases call for different approaches, but both will result in a test of a null hypothesis using a specific test of statistical significance.

How a Test of Significance Works: The Plan

Tests of significance are based on the fact that each type of null hypothesis has associated with it a particular type of statistic. And each of the statistics has associated with it a special distribution that you compare with the data you obtain from a sample. A comparison between the characteristics of your sample and the characteristics of the test distribution allows you to conclude whether the sample characteristics are different from what you would expect by chance.

Here are the general steps to take in the application of a statistical test to any null hypothesis. These steps will serve as a model for each of the chapters that follow in Part IV.

1. A statement of the null hypothesis. Do you remember that the null hypothesis is a statement of equality? The null hypothesis is the “true” state of affairs given no other information on which to make a judgment.
2. Setting the level of risk (or the level of significance or Type I error) associated with the null hypothesis. With any research hypothesis comes a certain degree of risk that you are wrong. The smaller this error is (such as .01 compared with .05), the less risk you are willing to take. No test of a hypothesis is completely risk-free because you never really know the “true” relationship between variables. Remember that it is traditional to set the Type I error rate at .01 or .05; Excel and other programs specify the exact level.

3. Selection of the appropriate test statistic. Each null hypothesis has associated with it a particular test statistic. You can learn what test is related to what type of question in this part of Statistics for People . . .

4. Computation of the test statistic value. The test statistic value (also called the obtained value) is the result or product of a specific statistical test. For example, there are test statistics for the significance of the difference between the averages of two groups, for the significance of the difference of a correlation coefficient from zero, and for the significance of the difference between two proportions. You’ll actually compute the test statistic and come up with a numerical value.

5. Determination of the value needed for rejection of the null hypothesis using the appropriate table of critical values for the particular statistic. Each test statistic (along with group size and the risk you are willing to take) has a critical value associated with it. This is the value you would expect the test statistic to yield if the null hypothesis is indeed true.

6. Comparison of the obtained value with the critical value. This is the crucial step. Here, the value you obtained from the test statistic (the one you computed) is compared with the value (the critical value) you would expect to find by chance alone.

7. If the obtained value is more extreme than the critical value, the null hypothesis cannot be accepted. That is, the null hypothesis’s statement of equality (reflecting chance) is not the most attractive explanation for differences that were found. Here is where the real beauty of the inferential method shines through. Only if your obtained value is more extreme than chance (meaning that the result of the test statistic is not a result of some chance fluctuation) can you say that any differences you obtained are not due to chance and that the equality stated by the null hypothesis is not the most attractive explanation for any differences you might have found.
Instead, the differences must be due to the treatment. What if the two values are equal (and you were about to ask your instructor that question, right?)? Nope—due to chance.

8. If the obtained value does not exceed the critical value, the null hypothesis is the most attractive explanation. If you cannot show that the difference you obtained is due to something other than chance (such as the treatment), then the difference must be due to chance or something you have no control over. In other words, the null is the best explanation.

Here’s the Picture That’s Worth a Thousand Words

What you see in Figure 9.2 represents the eight steps that we just went through. This is a visual representation of what happens when the obtained and critical values are compared. In this example, the significance level is set at .05, or 5%. It could have been set at .01, or 1%.

In examining Figure 9.2, note the following:

1. The entire curve represents all the possible outcomes based on a specific null hypothesis, such as the difference between two groups or the significance of a correlation coefficient.
2. The critical value is the point beyond which the obtained outcomes are judged to be so rare that we conclude the obtained outcome is not due to chance but to some other factor. In this example, we define “rare” as having a less than 5% chance of occurring.

3. If the outcome representing the obtained value falls to the left of the critical value (it is less extreme), we conclude that the null hypothesis is the most attractive explanation for any differences that are observed. In other words, the obtained value falls in the region (95% of the area under the curve) where we expect only outcomes due to chance to occur.

4. If the obtained value falls to the right of the critical value (it is more extreme), the conclusion is that the research hypothesis is the most attractive explanation of any differences that are observed. In other words, the obtained value falls in the region (5% of the area under the curve) where we would expect only outcomes due to something other than chance to occur.

**CONFIDENCE INTERVALS—BE EVEN MORE CONFIDENT**

You now know that probabilities can be associated with outcomes—that's been an ongoing theme for this and the last chapter. Now we are going to say the same thing in a slightly different way and introduce a new idea called confidence intervals.

A **confidence interval** (or c.i.) is the best estimate of the range of a population value (or population parameter) that we can come up with given the sample value (or sample statistic representing the population parameter). Say we knew the mean spelling score for a sample of 20 third graders (of all the third graders in a school district). How much confidence could we have that the population mean will fall between two scores? A 95% confidence interval would be correct 95% of the time.

You already know that the probability of a raw score falling within ±1.96 $z$ scores or standard deviations is 95%, right? (See page 204 in Chapter 8 if you need some review.) And you already know that the probability of a raw score falling within ±2.56 $z$ scores or standard deviations is 99%. If we use the positive or negative raw scores equivalent to those $z$ scores, we have ourselves a confidence interval.
While we are using the standard deviation to compute confidence intervals, many people choose to use the standard error of the mean, or SEM (see Chapter 10). The standard error of the mean is the standard deviation of all the sample means that could, in theory, be selected from the population. Remember that both the standard deviation and the standard error of the mean are “errors” in measurement that surround a certain “true” point (which in our case would be the true mean and the true amount of variability). The use of the SEM is a bit more complex, but it is an alternative way of computing, and understanding, confidence intervals.

Let’s fool around with some real numbers.

Let’s say that the mean spelling score for a random sample of 100 sixth graders is 64 (out of 75 words) and the standard deviation is 5. What confidence can we have in predicting the population mean for the average spelling score for the entire population of sixth graders?

The 95% confidence interval is equal to . . .

$$64 \pm 1.96(5)$$

In other words, it is a range from 54.2 to 73.8, so at the least you can say with 95% confidence that the population mean for the average spelling score for all sixth graders falls between those two scores.

Want to be more confident? The 99% confidence interval would be computed as . . .

$$64 \pm 2.56(5)$$

This is a range from 51.2 to 76.8, so you can conclude with 99% confidence that the population mean falls between those two scores.

Why does the confidence interval itself get larger as the probability of your being correct increases (from, say, 95% to 99%)? Because the larger range of the confidence interval (in this case from 73.8 – 54.2 = 19.6 for a 95% confidence interval to 76.8 – 51.2 = 25.6 for a 99% confidence interval) encompasses a larger number of possible outcomes. You can thereby be more confident. Ha! Isn’t this stuff cool?
REAL-WORLD STATS

It’s really interesting how different disciplines can learn from one another when they share. And it’s a shame that this does not happen more often, which is one of the reasons why interdisciplinary studies are so vital to creating an environment where new and old ideas can be used in new and old settings.

One such discussion took place in a medical journal that devotes itself to articles on anesthesia. The focus was a discussion of the relative merits of statistical versus clinical significance. Drs. Timothy Houle and David Stump point out how many large clinical trials obtain a high level of statistical significance with minuscule differences between groups (just as we talked about earlier in the chapter in our hypothetical example of teaching adults to read). In this case, statistically significant results are clinically irrelevant. However, the authors point out that with proper marketing, billions can be made from results of dubious clinical importance. This is really a caveat emptor, or buyer beware, state of affairs. Clearly, there a few very good lessons here to learn about how the significance of an outcome may or may not be meaningful. How to know? Look at the context of the research and substance of the outcomes.

Want to know more? Go online or go to the library, and look for . . .


SUMMARY

So, now you know exactly how the concept of significance works, and all that is left is applying it to a variety of research questions. That’s what we’ll start with in the next chapter and continue with through most of this part of the book.

TIME TO PRACTICE

1. Why is significance an important construct in the study and use of inferential statistics?

2. What’s wrong with the following statements?
   a. A Type I error of .05 means that in 5 out of 100 tests of the research hypothesis, I will reject a true null hypothesis.
b. It is possible to set the Type I error rate to zero.
   c. The smaller the Type I error rate, the better the results.

3. What does chance have to do with the testing of the research hypothesis for significance?

4. Given the following information, would your decision be to reject or fail to reject the null hypothesis? (Set the level of significance at .05 for decision making.) Provide an explanation for your conclusion.
   a. The null hypothesis that there is no relationship between the type of music a person listens to and the number of crimes he or she commits. $p < .05$.
   b. The null hypothesis that there is no relationship between the amount of coffee a student consumes and GPA. $p = .62$.
   c. The research hypothesis that there is a negative relationship between the number of hours worked and level of job satisfaction. $p = .51$.

5. Why is it “harder” to find a significant outcome (all other things being equal) when the research hypothesis is being tested at the .01 rather than the .05 level of significance?

6. Why should we think in terms of “failing to reject” the null rather than just accepting it?

7. What does the actual critical value represent?

8. What does the obtained value represent?

9. If you were to look at differences between two independent groups, what steps would you take using Figure 9.1?

10. In Figure 9.2, there is a striped area in the right-hand part of the curve.
    a. What does that area represent?
    b. If you tested the research hypothesis at a more rigorous level (say at .01 rather than .05), would that area get bigger or smaller, and why?