

Chapter 16 Inferential Statistics

(REMINDER: as you read the lectures, it's a good idea to also look at the concept map for each chapter. The concept maps help to give you the big picture and see how the concepts are related. Here is the link to all of the concept maps; just select the one for this chapter: http://www.southalabama.edu/coe/bset/johnson/dr_johnson/2conceptmaps.htm)

This is probably the most challenging chapter in your book. However, you can understand it. It just takes attention and effort. After you carefully study the material, it will become clear to you. I will also be available to answer any questions you have.

Please start this chapter by taking a look (again) at the divisions in the field of statistics that were shown in Figure 15.1 (p. 434) and also shown in the previous lecture.

- This shows the "big picture."
- As you can see, inferential statistics is divided into estimation and hypothesis testing, and estimation is further divided into point and interval estimation.

Inferential statistics is defined as the branch of statistics that is used to make inferences about the characteristics of a populations based on sample data.

- The goal is to go beyond the data at hand and make inferences about population parameters.
- In order to use inferential statistics, it is assumed that either random selection or random assignment was carried out (i.e., some form of randomization must is assumed).

Looking at Table 16.1 (p.464 and shown below) you can see that statisticians use Greek letters to symbolize population parameters (i.e., numerical characteristics of populations, such as means and correlations) and English letters to symbolize sample statistics (i.e., numerical characteristics of samples, such as means and correlations).

■ **TABLE 16.1**

A List of Symbols Used
for Statistics and
Parameters

Name	Sample Statistic	Population Parameter
Mean	\bar{X}	μ (mu)
Variance	SD^2	σ^2 (sigma squared)
Standard deviation	SD	σ (sigma)
Correlation	r	ρ (rho)
Proportion	p	π (pi)
Regression coefficient	b	β (beta)

Note: Statistics are usually symbolized with Roman letters and parameters with Greek letters.

For example, we use the Greek letter mu (i.e., μ) to symbolize the population mean and the Roman/English letter X with a bar over it, \bar{X} (called X bar), to symbolize the sample mean.

Sampling Distributions

One of the most important concepts in inferential statistics is that of the sampling distribution. That's because the use of a sampling distributions is what allows us to make "probability" statements in inferential statistics.

- A sampling distribution is defined as "The theoretical probability distribution of the values of a statistic that results when all possible random samples of a particular size are drawn from a population." (For simplicity you can view the idea of "all possible samples" as taking a million random samples. That is, just view it as taking a whole lot of samples!)
- A one specific type of sampling distribution is called the sampling distribution of the mean. If you wanted to generate this distribution through the laborious process of doing it by hand (which you would NOT need to do in practice), you would randomly select a sample, calculate the mean, randomly select another sample, calculate the mean, and continue this process until you have calculated the means for all possible samples. This process will give you a lot of means, and you can construct a line graph to depict your sampling distribution of the mean (e.g., see Figure 16.1 on page 468).
- The sampling distribution of the mean is normally distributed (as long as your sample size is about 30 or more for your sampling).
- Also, note that the mean of the sampling distribution of the mean is equal to the population mean! That tells you that repeated sampling will, over the long run, produce the correct mean. The spread or variance shows you that sample means will tend to be somewhat different from the true population mean in most particular samples.

Although I just described the sampling distribution of the mean, it is important to remember that a sampling distribution can be obtained for any statistic. For example, you could also obtain the following sampling distributions:

- Sampling distribution of the percentage (or proportion).
- Sampling distribution of the variance.
- Sampling distribution of the correlation.
- Sampling distribution of the regression coefficient.
- Sampling distribution of the difference between two means.

The standard deviation of a sampling distribution is called the standard error. In other words, the standard error is just a special kind of standard deviation and you learned what a standard deviation was in the last chapter.

- The smaller the standard error, the less the amount of variability present in a sampling distribution.

It is important to understand that researchers do not actually empirically construct sampling distributions! When conducting research, researchers typically select only one sample from the population of interest; they do not collect all possible samples.

- The computer program that a researcher uses (e.g., SPSS and SAS) uses the appropriate sampling distribution for you.

- The computer program will look at the type of statistical analysis you select (and also consider certain additional information that you have provided, such as the sample size in your study), and then the statistical program selects the appropriate sampling distribution.
- (It's kind of like the Greyhound Bus analogy: Leave the driving to us...SPSS will take care of generating the appropriate sampling distribution for you if you give it the information it needs.)

So please remember that the idea of sampling distributions (i.e., the idea of probability distributions obtained from repeated sampling) underlies our ability to make probability statements in inferential statistics.

Now, I'm going to cover the two branches of inferential statistics (i.e., estimation and hypothesis testing) that were shown in Figure 15.1: estimation and hypothesis testing.

Estimation

The key estimation question is "Based on my random sample, what is my estimate of the population parameter?"

- The basic idea is that you are going to use your sample data to provide information about the population.

There are actually two types of estimation.

- They can be first understood through the following analogy: Let's say that you take your car to your local car dealer's service department and you ask the service manager how much it will cost to repair your car. If the manager says it will cost you \$500 then she is providing a point estimate. If the manager says it will cost somewhere between \$400 and \$600 then she is providing an interval estimate.

In other words, a point estimate is a single number, and an interval estimate is a range of numbers.

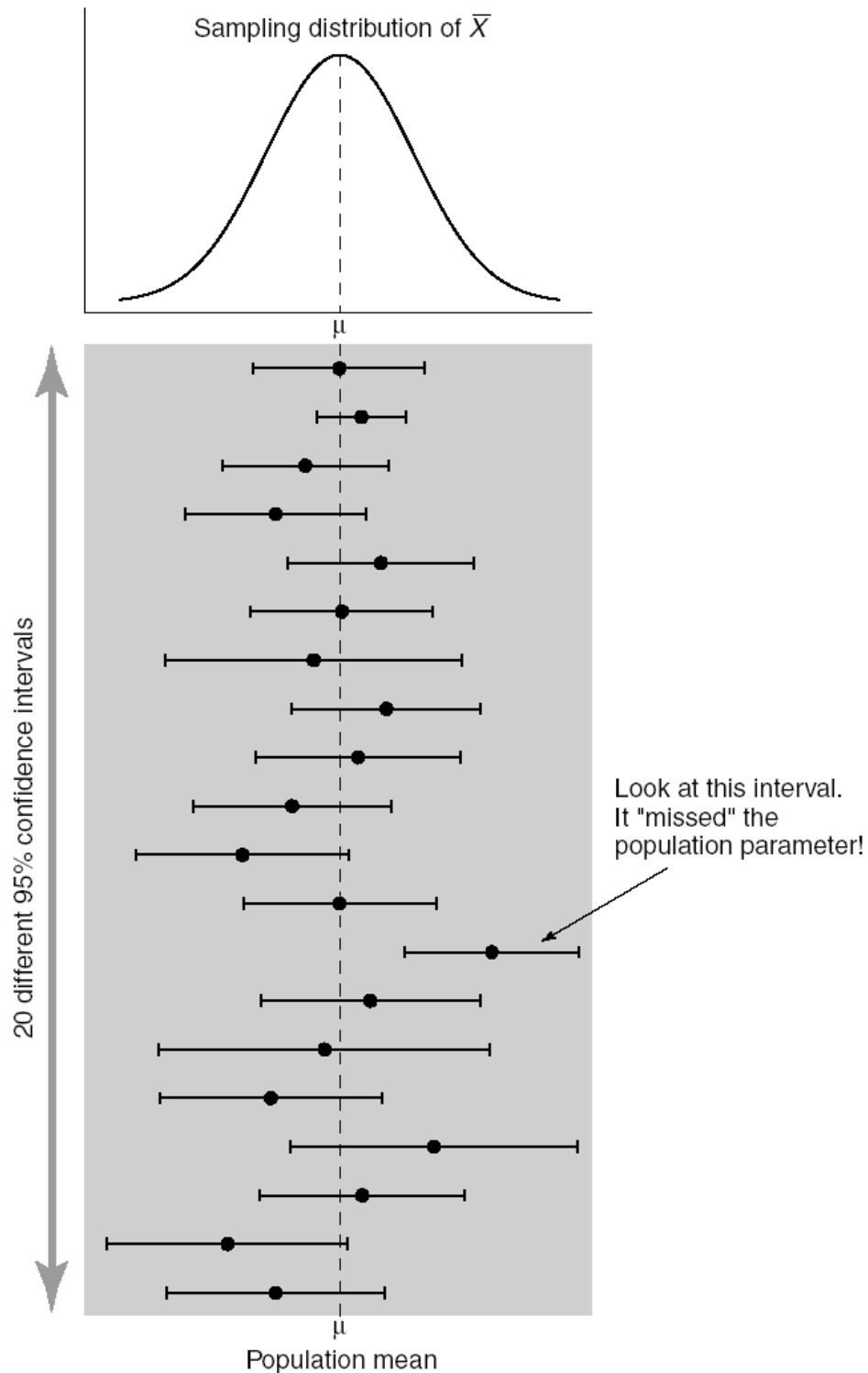
- A point estimate is the value of your sample statistic (e.g., your sample mean or sample correlation), and it is used to estimate the population parameter (e.g., the population mean or the population correlation).
- For example, if you take a random sample from adults living in the United States and you find that the average income for the people in your sample is \$45,000, then your best guess or your point estimate for the population of adults in the U.S. will be \$45,000.

In the above example, you used the value of the sample mean as the estimate of the population mean.

- Again, whenever you engage in point estimation, all you need to do is to use the value of your sample statistic as your "best guess" (i.e., as your estimate) of the (unknown) population parameter.

Oftentimes, we like to put an interval around our point estimates so that we realize that the actual population value is somewhat different from our point estimate because sampling error is always present in sampling.

- An interval estimate (also called a confidence interval) is a range of numbers inferred from the sample that has a known probability of capturing the population parameter over the long run (i.e., over repeated sampling).
- See Figure 16.2, p.471, for a picture of twenty different confidence intervals randomly jumping around the population mean from sample to sample.) Here it is for your convenience:



■ **FIGURE 16.2** A sampling distribution of the mean (based on all possible samples of size 100) and an illustration of the 95 percent confidence intervals for twenty possible samples. The width of the intervals will be slightly different because they are estimated from different random samples. In the long run, 95 percent of confidence intervals will capture the population mean.

- The "beauty" of confidence intervals is that we know their probability (over the long run) of including the true population parameter. (You can't do this with a point estimate.)
- Specifically, if you have the computer provide you with a 95 percent confidence interval (based on your data), then you will be able to be "95% confident" that it will include the population parameter. That is, your "level of confidence" is 95%.
- For example, you might take the point estimate of annual income of U.S. adults of \$45,000 (used earlier as a point estimate) and surround it by a 95% confidence interval. You might find that the confidence interval is \$43,000 to \$47,000. In this case, you can be "95% confident" that the average income is somewhere between \$43,000 and \$47,000.
- If you have the computer program give you a 99% confidence interval, then you can be "99% confident" that the confidence interval provided will include the population parameter (i.e., it will capture the true parameter 99% of the time in the long run).

You might ask: So why don't we just use 99% confidence intervals rather than 95% intervals, since you will make fewer mistakes?

- The answer is that for a given sample size, the 99% confidence interval will be wider (i.e., less precise) than a 95% confidence interval. For example, the interval \$40,000 to 50,000 is wider than the interval \$43,000 to \$47,000.
- 95% confidence intervals are popular with many researchers. However, you may, at times, want to use other confidence intervals (e.g., 90% confidence intervals or 99% confidence intervals).

Hypothesis Testing

Hypothesis testing is the branch of inferential statistics that is concerned with how well the sample data support a null hypothesis and when the null hypothesis can be rejected in favor of the alternative hypothesis.

- First note that the null hypothesis is usually the prediction that there is no relationship in the population.
- The alternative hypothesis is the logical opposite of the null hypothesis and says there is a relationship in the population.
- We use hypothesis testing when we expect a relationship to be present; in other words, we usually hope to "nullify" the null hypothesis and tentatively accept the alternative hypothesis. (Note: if you expect the null to be true, you can use the estimation approach described in this chapter; several additional procedures for this special case are discussed in Shadish, Cook, and Campbell's book *Experimental and Quasi-Experimental Designs*, 2002, pp. 52-53)
- Here is the key question that is answered in hypothesis testing: "*Is the value of my sample statistic unlikely enough (assuming that the null hypothesis is true) for me to reject the null hypothesis and tentatively accept the alternative hypothesis?*"
- Note that it is the null hypothesis that is directly tested in hypothesis testing (not the alternative hypothesis).

To get the idea of null hypothesis testing in your head, reread Exhibit 16.1 (p. 473 and shown below).

Exhibit 16.1 An Analogy From Jurisprudence

The United States criminal justice system operates on the assumption that the defendant is innocent until proven guilty beyond a reasonable doubt. In hypothesis testing, this assumption is called the null hypothesis. That is, researchers assume that the null hypothesis is true until the evidence suggests that it is not likely to be true. The researcher's null hypothesis might be that a technique of counseling does not work any better than no counseling. The researcher is kind of like a prosecuting attorney. The prosecuting attorney brings someone to trial when he or she believes there is some evidence against the accused, and the researcher brings a null hypothesis to "trial" when he or she believes there is some evidence against the null hypothesis (i.e., the researcher actually believes that the counseling technique does work better than no counseling). In the courtroom, the jury decides what constitutes reasonable doubt, and they make a decision about guilt or innocence. The researcher uses inferential statistics to determine the probability of the evidence under the assumption that the null hypothesis is true. If this probability is low, the researcher is able to reject the null hypothesis and accept the alternative hypothesis. If this probability is not low, the researcher is not able to reject the null hypothesis. No matter what decision is made, things are still not completely settled because a mistake could have been made. In the courtroom, decisions of guilt or innocence are sometimes overturned or found to be incorrect. Similarly, in research, the decision to reject or not reject the null hypothesis is based on probability, so researchers sometimes make a mistake. However, inferential statistics gives researchers the probability of their making a mistake.

- Here is the main point: In the United States System of Jurisprudence, a defendant is "presumed innocent" until evidence calls this assumption into question. That is, the jury is told to assume that a person is innocent until they have heard all of the evidence and can make a decision. Likewise, in hypothesis testing, the null hypothesis is assumed to be true (i.e., it is assumed that there is no relationship) until evidence clearly calls this assumption into question.
- In jurisprudence, the jury rejects the claim of innocence (rejects the null) in the face of strong evidence to the contrary and makes the opposite conclusion that the defendant is guilty. Likewise, in hypothesis testing, the researcher rejects the null hypothesis in the face of strong evidence to the contrary.
- In hypothesis testing, "strong evidence to the contrary" is found in a small probability value, which says the research result is unlikely if the null hypothesis is true. When the researcher rejects the null hypothesis (i.e., rejects the assumption of no relationship), he or she tentatively accepts the alternative hypothesis (i.e., which says there *is* a relationship in the population).
- In short . . . *in the procedure called hypothesis testing the researcher states the null and alternative hypotheses. Then if the probability value is small, the researcher rejects the null hypothesis and goes with the alternative hypothesis and makes the claim that statistical significance has been found.*

Now take a look at the research questions and the null and alternative hypotheses shown below and in Table 16.2 (p.474).

- When you look at the table be sure to notice that the null hypothesis has the equality sign in it and the alternative hypothesis has the "not equals" sign in it.
- You can also see in the table that hypotheses can be tested for many different kinds of research questions such as questions about means, correlations, and regression coefficients.

■ **TABLE 16.2** Examples of Null and Alternative Hypotheses in Inferential Statistics

Research Question	Verbal Null (H_0) Hypothesis	Symbolic H_0 Hypothesis	Verbal Alternative (H_1) Hypothesis	Symbolic H_1 Hypothesis
Do teachers score higher on the GRE verbal than the national average?	The teacher population GRE verbal mean is equal to the national average of 476.	$H_0: \mu_{\text{GRE V}} = 476$	The teacher population GRE verbal mean is different from the national average of 476.	$H_1: \mu_{\text{GRE V}} \neq 476$
Do males or females tend to score better on the GRE verbal?	The male and female population means are not different.	$H_0: \mu_M = \mu_F$	The male and female population means are different.	$H_1: \mu_M \neq \mu_F$
Do education, arts and sciences, and business students have different starting incomes?	The education, arts and sciences, and business student populations have the same mean starting incomes.	$H_0: \mu_E = \mu_{\text{A\&S}} = \mu_B$	At least two of the three population means are different.	$H_1: \text{Not all equal}$
Is there a correlation between GPA (X), and starting salary (Y)?	The population correlation between GPA and starting salary is equal to zero.	$H_0: \rho_{XY} = 0$	The population correlation between GPA and starting salary is not equal to zero.	$H_1: \rho_{XY} \neq 0$
Is there a relationship between GRE verbal (X_1), and starting salary (Y), controlling for GPA (X_2)?	The population regression coefficient is equal to zero.	$H_0: \beta_{YX_1 \cdot X_2} = 0$	The population regression coefficient is not equal to zero.	$H_1: \beta_{YX_1 \cdot X_2} \neq 0$

You may be wondering, when do you actually reject the null hypothesis and make the decision to tentatively accept the alternative hypothesis?

- Earlier I mentioned that you reject the null hypothesis when the *probability* of your result assuming a true null is very small. That is, you reject the null when the evidence would be unlikely under the assumption of the null.
- In particular, you set a significance level (also called the alpha level) to use in your research study, which is the point at which you would consider a result to be very unlikely. Then, if your probability value is less than or equal to your significance level, you reject the null hypothesis.
- It is essential that you understand the difference between the probability value (also called the p-value) and the significance level (also called the alpha level).

- The **probability value** is a number that is obtained from the SPSS computer printout. It is based on your empirical data, and *it tells you the probability of your result or a more extreme result when it is assumed that there is no relationship in the population (i.e., when you are assuming that the null hypothesis is true which is what we do in hypothesis testing and in jurisprudence).*
- The **significance level** is just that point at which you would consider a result to be "rare." *You are the one who decides on the significance level to use in your research study.* A significance level is not an empirical result; it is the level that you set so that you will know what probability value will be small enough for you to reject the null hypothesis.
- The significance level that is usually used in education is .05.
- It boils down to this: *if your probability value is less than or equal to the significance level (e.g., .05) then you will reject the null hypothesis and tentatively accept the alternative hypothesis. If not (i.e., if it is $> .05$) then you will fail to reject the null.* You just compare your probability value with your significance level.
- You must memorize the definitions of probability value and significance level right away because they are at the heart of hypothesis testing. At the most simple level, the process just boils down to seeing whether your probability value is less than (or equal to) your significance level. If it is, you are happy because you can reject the null hypothesis and make the claim of statistical significance. (Still don't forget the last step of determining practical significance.)

This full process of hypothesis testing is summarized in Table 16.3 (p.480) and shown below.

- Be sure to note the final step shown in the table, because after conducting a hypothesis test, you must interpret your results, make a substantive, real-world decision, and determine the practical significance of your result.

Here is Table 16.3, in case you don't have your book handy.

■ **TABLE 16.3** Steps in Hypothesis Testing

1. State the null and alternative hypotheses.
2. Set the *significance level* before the research study.
(Most educational researchers use .05 as the significance level. Note that the significance level is also called the *alpha level* or, more simply, *alpha*.)
3. Obtain the *probability value* using a computer program such as SPSS.
4. Compare the probability value to the significance level and make the statistical decision.

Step 4 includes two decision-making rules:

Rule 1:

If: Probability value \leq significance level (i.e., probability value \leq alpha)

Then: Reject the null hypothesis.

And: Conclude that the research finding is statistically significant.

In practice, this usually means:

If: Probability value \leq .05

Then: Reject the null hypothesis.

And: Conclude that the finding is statistically significant.

Rule 2:

If: Probability value $>$ significance level (i.e., probability value $>$ alpha)

Then: Fail to reject the null hypothesis.

And: Conclude that the research finding is not statistically significant.

In practice this usually means:

If: Probability value $>$.05

Then: Fail to reject the null hypothesis.

And: Conclude that the research finding is not statistically significant.

5. Interpret the results. That is, make a substantive, real-world decision and determine practical significance.
This means that you must decide what the results of your research study actually mean. Statistics are only a tool for determining statistical significance. If you obtain statistical significance, you must now interpret your results in terms of the variables used in your research study. For example, you might decide that females perform better, on average, than males on the GRE verbal test or that client-centered therapy works better than rational emotive therapy, or that phonics and whole language in combination work better than phonics only.

You must also determine the *practical significance* of your findings. A finding is practically significant when the difference between the means or the size of the relationship is big enough, in your opinion, to be of practical use. For example, a correlation of .15 would probably not be practically significant, even if it was statistically significant. On the other hand, a correlation of .85 would probably be practically significant. *Effect-size indicators* (p. 484) are important aids when you are making a judgment about practical significance.

Step 5 shows that you must decide what the results of your research study actually mean.

- Statistical significance does not tell you whether you have practical significance. At the end of step four you will know whether your result is statistically significant.
- If a finding is statistically significant then you can claim that the evidence suggests that the observed result (e.g., your observed correlation or your observed difference between two means) was probably *not* just due to chance. That is, there probably is some non-zero relation present in the population.
- An effect size indicator can aid in your determination of practical significance and should always be examined to help interpret the strength of a statistically significant relationship. An effect size indicator is defined as a measure of the strength of a relationship.
- A finding is practically significant when the difference between the means or the size of the correlation is big enough, in your opinion, to be of practical use. For example, a correlation of .15 would probably not be practically significant, even if it was statistically significant. On the other hand, a correlation of .85 would probably be practically significant.
- Practical significance requires you to make a non-quantitative decision and to think about many different factors such as the size of the relationship, whether an intervention would transfer well to the real world, the costs of using a statistically significant intervention in the real world, etc. It is a decision that YOU make.

The next idea is for you to realize that you will either make a correct decision about statistical significance or you will make an error whenever you conduct a hypothesis test.

- This idea is shown below and in Table 16.5 (p. 482) and here for your convenience.

■ **TABLE 16.5** The Four Possible Outcomes in Hypothesis Testing

		<u>The True (but Unknown) Status of the Null Hypothesis</u>	
		The null hypothesis is true (It should not be rejected.)	The null hypothesis is false (It should be rejected.)
Your Decision*	Fail to reject the null hypothesis	Type A Correct decision!	Type II Error (false negative)
	Reject the null hypothesis	Type I Error (false positive)	Type B Correct decision!

*Remember that if the null hypothesis is true, it should *not* be rejected, but if the null hypothesis is false, it *should* be rejected. The problem is that you will not know if the null hypothesis is true or false. You only have the probabilistic evidence obtained from your sample data.

- Looking at the top of the table (i.e., above the two columns) you will see that the null hypothesis is either *true* or *not true* in the empirical world.
- If you look at the side of the table (i.e., beside the two rows) you will see that you must make a decision to either *fail to reject* or to *reject* the null hypothesis.
- When the null is false you want to reject it, but when it is true you do not want to reject it.
- The four logical possibilities of hypothesis testing are shown in the table.
- When the null hypothesis is true you can make the correct decision (i.e., fail to reject the null) or you can make the incorrect decision (rejecting the true null). The incorrect decision is called a Type I error or a "false positive" because you have erroneously concluded that there is an effect or relationship in the population.
- When the null hypothesis is false you can also make the correct decision (i.e., rejecting the false null) or you can make the incorrect decision (failure to reject the false null). The incorrect decision is called a Type II error or a "false negative" because you have erroneously concluded that there is no effect or relationship in the population.
- You need to memorize the definitions of Type I and Type II errors, and after working with many examples of hypothesis testing they will become easier to ponder.
- Exercise: In law, a person is presumed to be innocent (i.e., that is the null hypothesis). Explain the idea of Type I and Type II errors here. Which error has occurred when an innocent person is found guilty? Which error has occurred when a guilty person is found innocent by the jury? (The answers are below.)

Hypothesis Testing in Practice

In this last section of the chapter, I apply the process of hypothesis testing (which is also called "significance testing") to the data set given in Table 15.1 (p. 435) and shown again here (below).

■ **TABLE 15.1** Hypothetical Set of Data for Twenty-Five Recent College Graduates

Person	Starting Salary	GPA	College Major	Gender	GRE Verbal
1	31,000	2.9	2	1	520
2	32,000	3.6	1	1	600
3	33,000	3.7	1	1	610
4	28,000	2.4	2	0	450
5	37,000	3.4	3	0	500
6	32,000	3.0	3	0	500
7	33,000	3.1	2	0	520
8	25,000	2.5	1	1	450
9	38,000	3.0	3	0	650
10	33,000	2.7	2	0	490
11	30,000	3.0	2	1	500
12	32,000	2.6	1	0	580
13	32,000	3.1	2	1	480
14	31,000	3.1	1	1	530
15	24,000	2.5	1	1	460
16	40,000	3.3	3	0	630
17	31,000	3.3	1	1	510
18	38,000	3.2	2	1	620
19	35,000	3.1	3	1	680
20	32,000	3.2	2	0	550
21	41,000	3.5	3	0	680
22	34,000	3.0	3	1	590
23	28,000	3.0	1	1	650
24	30,000	2.9	2	0	480
25	36,000	3.5	2	0	570

Note: For the categorical variable college major, the value labels are 1 = education, 2 = arts and sciences, 3 = business. For the categorical variable gender, the value labels are 0 = male and 1 = female.

- Since we are now using this data set for inferential statistics, we will assume that the 25 people were randomly selected.
- Note that there are three quantitative variables and two categorical variables (can you list them?).
- Also note that I will use the significance level of .05 for all of my statistical tests below.

(The answers to the earlier questions about the two types of errors are in the first case a Type I error was made and in the second case a Type II error was made.)

- Before I test some hypotheses, I want to point out the reason WHY we use hypothesis or significance testing: We do it because researchers do not want to interpret findings that are not statistically significant because these findings are probably nothing but a reflection of chance fluctuations.

Note that in all of the following examples I will be doing the same thing. I will get the p-value and compare it to my preset significance level of .05 to see if the relationship is statistically significant. And then I will also interpret the results by looking at the data, looking at an effect size indicator, and by thinking about the practical importance of the result.

- Again, after practice, significance becomes very easy because you do the same procedure every single time. Determining the practical significance is probably the hardest part.

t-Test for Independent Samples

One frequently used statistical test is called the t-test for independent samples. We do this when we want to determine if the difference between two groups is statistically significant.

Here is an example of the t-test for independent samples using our recent college graduate data set:

- Research Question: Is the difference between average starting salary for males and the average starting salary for females significantly different?
- Here the hypotheses (note that they are stated in terms of population parameters):
 - Null Hypothesis $H_0: \mu_M = \mu_F$ (i.e., the population mean for males equals the population mean for females)
 - Alternative Hypothesis $H_1: \mu_M \neq \mu_F$ (i.e., the population mean for males does not equal the population mean for females)

The probability value was .048 (I got this off of my SPSS printout).

- Since my probability value of .049 is less than my significance level of .05, I reject the null hypothesis and accept the alternative.
- I conclude that the difference between the two means is statistically significant.
- Now I would need to look at the actual means and interpret them for substantive and practical significance.
- The males' mean is \$34,333.33 and the females' mean is \$31,076.92.
- I can simply look at these means and see how different they are.
- To help in judging how different the means are, I also calculated an effect size indicator called eta-squared which was equal to .16. This tells me that gender explains 16% of the variance in starting salary in my data set.
- I conclude that males earn more than females, and because this is an important issue in society, I also conclude that this difference is practically significant.

One-Way Analysis of Variance

One-way analysis of variance is used to compare two or more group means for statistical significance.

Here is an example using our “recent college graduate” data set:

- Research Question: Is there a statistically significant difference in the starting salaries of education majors, arts and sciences majors, and engineering majors?
- Here the hypotheses (note that they are stated in terms of population parameters):
- Null Hypothesis. $H_0: \mu_E = \mu_{A\&S} = \mu_B$ (i.e., the population means for education students, arts and sciences students, and business students are all the same)
- Alternative Hypothesis. $H_1: \text{Not all equal}$ (i.e., the population means are not all the same)

The probability value was .001 (I got this off of my SPSS printout).

- Since .001 is less than .05, I reject the null hypothesis and accept the alternative. I conclude that at least two of the means are significantly different.
- The effect size indicator, eta-squared, was equal to .467 which say that almost 47 percent in the variance of starting salary was explained or accounted for by differences in college major.
- Now I need to find out which of the three means are different.
- In order to decide which of these three means are significantly different, I must follow the “post hoc testing” procedure explained in the next. Notice that is I had done an ANOVA with an independent variable that was composed of only two groups, I would not need follow-up tests (which are only needed when there are three or more groups).

Post Hoc Tests in Analysis of Variance

Here are the three average starting salaries for the three groups examined in the previous analysis of variance (i.e., these are the three sample means):

- Education: \$29,500
- Arts and Sciences: \$32,300
- Business: \$36,714.29

The question in post hoc testing is "Which pairs of means are significantly different?"

In this case that results in three post hoc tests that need to be conducted:

1. First, is the difference between education and arts and sciences significantly different"
 - Here are the null and alternative hypotheses for this first post hoc test:

- Null Hypothesis $H_0: \mu_E = \mu_{A\&S}$ (i.e., the population mean for education majors equals the population mean for arts and sciences majors)
 - Alternative Hypothesis $H_1: \mu_E \neq \mu_{A\&S}$ (i.e., the population mean for education majors does not equal the population mean for arts and sciences majors)
 - The Bonferroni "adjusted" p-value, which I got off the SPSS printout, was .233.
 - Since .233 is $> .05$, I fail to reject the null that the population means for education and arts and sciences are equal.
 - In short, this difference was not statistically significant.
2. Second, is the difference between education and business significantly different?
- Here are the null and alternative hypotheses for this first post hoc test:
 - Null Hypothesis $H_0: \mu_E = \mu_B$ (i.e., the population mean for education majors equals the population mean for business majors)
 - Alternative Hypothesis $H_1: \mu_E \neq \mu_B$ (i.e., the population mean for education majors does not equal the population mean for business majors)
 - The adjusted p-value was .001.
 - Since .001 is $< .05$, I reject the null that the two population means are equal.
 - I make the claim that the difference between the means is statistically significant.
 - I also claim that the salaries are higher for business than for education students in the populations from which they were randomly selected.
 - Because this finding could affect many students' choices about majors and because it may also reflect the nature of salary setting by the private versus public sectors, I also conclude that this difference is practically significant.
3. Third, is the difference between arts and sciences and business significantly different?
- Here are the null and alternative hypotheses for this first post hoc test:
 - Null Hypothesis $H_0: \mu_B = \mu_{A\&S}$ (i.e., the population mean for business majors equals the population mean for arts and sciences majors)
 - Alternative Hypothesis $H_1: \mu_B \neq \mu_{A\&S}$ (i.e., the population mean for business majors does not equal the population mean for arts and sciences majors)
 - The adjusted p-value was .031.
 - Since .031 is $< .05$, I reject the null hypothesis that the two population means are significantly different.
 - I make the claim that this difference between the means is statistically significant.

- I also claim that the salaries are higher form arts and sciences than for education students in the populations from which they were randomly selected.
- Because this finding could affect students' choices about majoring in business versus arts and sciences, I believe that this finding is practically significant.

In short, based on my post hoc tests, I have found that two of the differences in starting salary were statistically significant, and, in my view, these differences were also practically significant.

The t-Test for Correlation Coefficients

This test is used to determine whether an observed correlation coefficient is statistically significant.

Here is an example using our “recent college graduate” data set:

- Research Question: Is there a statistically significant correlation between GPA (X) and starting salary (Y)?
- Here are the hypotheses:
 - Null Hypothesis. $H_0: P_{XY} = 0$ (i.e., there is no correlation in the population)
 - Alternative Hypothesis. $H_1: P_{XY} \neq 0$ (i.e., there is a correlation in the population)
- The observed correlation in the sample was .63.
- The probability value was .001.
- Since .001 is $< .05$, I reject the null hypothesis.
- The observed correlation was statistically significant.
- I conclude that GPA and starting salary are correlated in the population.
- If you square the correlation coefficient you obtain a “variance accounted for” effect size indicator: .63 squared is .397 which means that almost 40 percent of the variance in starting salary is explained or accounted for by GPA
- Because the effect size is large and because GPA is something that students can control through studying, I conclude that this statistically significant correlation is also practically significant.

The t-Test for Regression Coefficients

This test is used to determine whether a regression coefficient is statistically significant.

The multiple regression equation analyzed in the last chapter is shown here again, but this time we will test each of the two regression coefficients for statistical significance.

$$\hat{Y} = 3,890.05 + 4,675.41 (X_1) + 26.13(X_2)$$

where,

\hat{Y} is predicted starting salary

3,890.05 is the Y intercept (or predicted starting salary when GPA and GRE Verbal are zero)

4,675.41 is the regression coefficient for grade point average

X_1 is grade point average (GPA)

X_2 is GRE Verbal

Research Question One: Is there a statistically significant relationship between starting salary (Y) and GPA (X_1) controlling for GRE Verbal (X_2)? That is, is the first regression coefficient statistically significant?

- Here are the hypotheses:
- Null Hypothesis. $H_0: \beta_{YX_1.X_2} = 0$ (i.e., the population regression coefficient expressing the relationship between starting salary and GPA, controlling for GRE Verbal is equal to zero; that is, there is no relationship)
- Alternative Hypothesis. $H_1: \beta_{YX_1.X_2} \neq 0$ (i.e., the population regression coefficient expressing the relationship between starting salary and GPA, controlling for GRE Verbal is NOT equal to zero; that is, there IS a relationship)
- The observed regression coefficient was 4,496.45.
- The probability value was .035
- Since .035 is $< .05$, I conclude that the relationship expressed by this regression coefficient is statistically significant.
- A good measure of effect size for regression coefficients is the semi-partial correlation squared (sr^2). In this case it is equal to .10, which means that 10% of the variance in starting salary is uniquely explained by GPA
- Because GPA is something we can control and because the effect explains a good amount of variance in starting salary, I conclude that the relationship expressed by this regression coefficients is practically significant.

Research Question Two: Is there a statistically significant relationship between starting salary (Y) and GRE Verbal (X_2), controlling for GPA (X_1)? That is, is the second regression coefficient statistically significant?

- Here are the hypotheses:
- Null Hypothesis. $H_0: \beta_{YX_2.X_1} = 0$ (i.e., the population regression coefficient expressing the relationship between starting salary and GRE Verbal, controlling for GPA is equal to zero; that is, there is no relationship)

- Alternative Hypothesis. $H_1 : \beta_{YX2.X1} \neq 0$ (i.e., the population regression coefficient expressing the relationship between starting salary and GRE Verbal, controlling for GPA is NOT equal to zero; that is, there IS a relationship)
- The observed regression coefficient was 26.13.
- The probability value was .014
- Since .014 is $< .05$, I conclude that the relationship expressed by this regression coefficient is statistically significant.
- A good measure of effect size for regression coefficients is the semi-partial correlation squared (sr^2). In this case it is equal to .15, which means that 15% of the variance in starting salary is uniquely explained by GRE Verbal
- Because GRE Verbal is also something we can work at (as well as take preparation programs for) and because the effect explains 15% of the variance in starting salary, I conclude that the relationship expressed by this regression coefficient is practically significant.

The Chi-Square Test for Contingency Tables

This test is used to determine whether a relationship observed in a contingency table is statistically significant.

- Research Question: Is the observed relationship between college major and gender statistically significant?
- The probability value was .046.
- Since .046 is $< .05$, I conclude that the observed relationship in the contingency table shown in Table 16.6 (p.492) is statistically significant.
- The effect size indicator used for this contingency table is Cramer's V. It was equal to .496, which tells us that the relationship is moderately large.
- Because the effect size indicator suggested a moderately large relationship and because of the importance of these variables in real world politics, I would also conclude that this relationship is practically significant.

Believe it or not, we are done. My goal in this last section was to show that every single time we do one of these tests, you do the same thing. You get your probably value, compare it to your significance level, and, finally, you make a decision.

You have now come a long way toward understanding the logic of significance testing. Remember, when reading journal articles look out for those probability values (to see if they are less than .05), and also look for effect sizes and statements about whether a finding is practically significant

Congratulations!