14 Inferential Statistics

> The great tragedy of science - the slaying of a beautiful hypothesis by an ugly fact. —Thomas Huxley

> Truth in science can be defined as the working hypothesis best suited to open the way to the next better one. —Konrad Lorenz

Recall that Matthias Mehl and his colleagues, in their study of sex differences in talkativeness, found that the women in their sample spoke a mean of 16,215 words per day and the men a mean of 15,669 words per day (Mehl et al. 2007). But despite this sex difference in their sample, they concluded that there was no evidence of a sex difference in talkativeness in the population. Recall also that Allen Kanner and his colleagues, in their study of the relationship between daily hassles and symptoms, found a correlation of +.60 in their sample (Kanner et al. 1981). But they concluded that this finding means there is a relationship between hassles and symptoms in the population. This assertion raises the question of how researchers can say whether their sample result reflects something that is true of the population.

The answer to this question is that they use a set of techniques called inferential statistics, which is what this chapter is about. We focus, in particular, on null hypothesis testing, the most common approach to inferential statistics in psychological research. We begin with a conceptual overview of null hypothesis testing, including its purpose and basic logic. Then we look at several null hypothesis testing techniques for drawing conclusions about differences between means and about correlations between quantitative variables. Finally, we consider a few other important ideas related to null hypothesis testing, including some that can be helpful in planning new studies and interpreting results. We also look at some long-standing criticisms of null hypothesis testing and some ways of dealing with these criticisms.

Understanding Null Hypothesis Testing

The Purpose of Null Hypothesis Testing

Learning Objectives

1. Explain the purpose of null hypothesis testing, including the role of sampling error.
2. Describe the basic logic of null hypothesis testing.
3. Describe the role of relationship strength and sample size in determining statistical significance and make reasonable judgments about statistical significance based on these two factors.

As we have seen, psychological research typically involves measuring one or more variables for a sample and computing descriptive statistics for that sample. In general, however, the researcher's goal is not to draw conclusions about that sample but to draw conclusions about the population that the sample was selected from. Thus researchers must use sample statistics to draw conclusions about the corresponding values in the population. These corresponding values in the population are called parameters. Imagine, for example, that a researcher measures the number of depressive symptoms exhibited by each of 50 clinically depressed adults and computes the mean number of symptoms. The researcher probably wants to use this sample statistic (the mean number of symptoms for the sample) to draw conclusions about the corresponding population parameter (the mean number of symptoms for clinically depressed adults).

Unfortunately, sample statistics are not perfect estimates of their corresponding population parameters. This is because there is a certain amount of random variability in any statistic from sample to sample. The mean number of depressive symptoms might be 8.73 in one sample of clinically depressed adults, 6.45 in a second sample, and 9.44 in a third—even though these samples are selected randomly from the same population. Similarly, the correlation (Pearson's r) between two variables might be +.24 in one sample, -.04 in a second sample, and +.15 in a third—again, even though these samples are selected randomly from the same population. This random variability in a statistic from sample to sample is called sampling error. (Note that the term error here refers to random variability and does not imply that anyone has made a mistake. No one "commits a sampling error.")

One implication of this is that when there is a statistical relationship in a sample, it is not always clear that there is a statistical relationship in the population. A small difference between two group means in a sample might indicate that there

is a small difference between the two group means in the population. But it could also be that there is no difference between the means in the population and that the difference in the sample is just a matter of sampling error. Similarly, a Pearson's r value of -.29 in a sample might mean that there is a negative relationship in the population. But it could also be that there is no relationship in the population and that the relationship in the sample is just a matter of sampling error.

In fact, any statistical relationship in a sample can be interpreted in two ways:

- There is a relationship in the population, and the relationship in the sample reflects this.
- There is no relationship in the population, and the relationship in the sample reflects only sampling error.

The purpose of null hypothesis testing is simply to help researchers decide between these two interpretations.

The Logic of Null Hypothesis Testing

Null hypothesis testing is a formal approach to deciding between two interpretations of a statistical relationship in a sample. One interpretation is called the null hypothesis (often symbolized H0 and read as "H-naught"). This is the idea that there is no relationship in the population and that the relationship in the sample reflects only sampling error. Informally, the null hypothesis is that the sample relationship "occurred by chance." The other interpretation is called the alternative hypothesis (often symbolized as H1). This is the idea that there is a relationship in the population and that the relationship in the sample reflects this relationship in the population.

Again, every statistical relationship in a sample can be interpreted in either of these two ways: It might have occurred by chance, or it might reflect a relationship in the population. So researchers need a way to decide between them. Although there are many specific null hypothesis testing techniques, they are all based on the same general logic. The steps are as follows:

- Assume for the moment that the null hypothesis is true. There is no relationship between the variables in the population.
- Determine how likely the sample relationship would be if the null hypothesis were true.
- If the sample relationship would be extremely unlikely, then reject the null hypothesis in favor of the alternative hypothesis. If it would not be extremely unlikely, then retain the null hypothesis.

Following this logic, we can begin to understand why Mehl and his colleagues concluded that there is no difference in talkativeness between women and men in the population. In essence, they asked the following question: "If there were no difference in the population, how likely is it that we would find a small difference of d = 0.06 in our sample?" Their answer to this question was that this sample relationship would be fairly likely if the null hypothesis were true. Therefore, they retained the null hypothesis—concluding that there is no evidence of a sex difference in the population. We can also see why Kanner and his colleagues concluded that there is a correlation between hassles and symptoms in the population. They asked, "If the null hypothesis were true, how likely is it that we would find a strong correlation of +.60 in our sample?" Their answer to this question was that this sample relationship would be fairly unlikely if the null hypothesis were true. Therefore, they rejected the null hypothesis in favor of the alternative hypothesis—concluding that there is a positive correlation between these variables in the population.

A crucial step in null hypothesis testing is finding the likelihood of the sample result if the null hypothesis were true. This probability is called the p value. A low p value means that the sample result would be unlikely if the null hypothesis were true and leads to the rejection of the null hypothesis. A high p value means that the sample result would be likely if the null hypothesis were true and leads to the retention of the null hypothesis. But how low must the p value be before the sample result is considered unlikely enough to reject the null hypothesis? In null hypothesis testing, this criterion is called α (alpha) and is almost always set to .05. If there is less than a 5% chance of

a result as extreme as the sample result if the null hypothesis were true, then the null hypothesis is rejected. When this happens, the result is said to be statistically significant. If there is greater than a 5% chance of a result as extreme as the sample result when the null hypothesis is true, then the null hypothesis is retained. This does not necessarily mean that the researcher accepts the null hypothesis as true—only that there is not currently enough evidence to conclude that it is true. Researchers often use the expression "fail to reject the null hypothesis" rather than "retain the null hypothesis," but they never use the expression "accept the null hypothesis."

The Misunderstood p Value

The p value is one of the most misunderstood quantities in psychological research (Cohen 1994). Even professional researchers misinterpret it, and it is not unusual for such misinterpretations to appear in statistics textbooks!

The most common misinterpretation is that the p value is the probability that the null hypothesis is true—that the sample result occurred by chance. For example, a misguided researcher might say that because the p value is .02, there is only a 2% chance that the result is due to chance and a 98% chance that it reflects a real relationship in the population. But this is incorrect. The p value is really the probability of a result at least as extreme as the sample result if the null hypothesis were true. So a p value of .02 means that if the null hypothesis were true, a sample result this extreme would occur only 2% of the time.

You can avoid this misunderstanding by remembering that the p value is not the probability that any particular hypothesis is true or false. Instead, it is the probability of obtaining the sample result if the null hypothesis were true.

Role of Sample Size and Relationship Strength

Recall that null hypothesis testing involves answering the question, "If the null hypothesis were true, what is the probability of

a sample result as extreme as this one?" In other words, "What is the p value?" It can be helpful to see that the answer to this question depends on just two considerations: the strength of the relationship and the size of the sample. Specifically, the stronger the sample relationship and the larger the sample, the less likely the result would be if the null hypothesis were true. That is, the lower the p value. This should make sense. Imagine a study in which a sample of 500 women is compared with a sample of 500 men in terms of some psychological characteristic, and Cohen's d is a strong 0.50. If there were really no sex difference in the population, then a result this strong based on such a large sample should seem highly unlikely. Now imagine a similar study in which a sample of three women is compared with a sample of three men, and Cohen's d is a weak 0.10. If there were no sex difference in the population, then a relationship this weak based on such a small sample should seem likely. And this is precisely why the null hypothesis would be rejected in the first example and retained in the second.

Of course, sometimes the result can be weak and the sample large, or the result can be strong and the sample small. In these cases, the two considerations trade off against each other so that a weak result can be statistically significant if the sample is large enough and a strong relationship can be statistically significant even if the sample is small. Figure @ref(fig:IS1) shows roughly how relationship strength and sample size combine to determine whether a sample result is statistically significant. The columns of the table represent the three levels of relationship strength: weak, medium, and strong. The rows represent four sample sizes that can be considered small, medium, large, and extra large in the context of psychological research. Thus each cell in the table represents a combination of relationship strength and sample size. If a cell contains the word Yes, then this combination would be statistically significant for both Cohen's d and Pearson's r. If it contains the word No, then it would not be statistically significant for either. There is one cell where the decision for d and r would be different and another where it might be different depending on some additional considerations, which are discussed in Section "Some Basic Null Hypothesis Tests"

Although Figure Figure 1 provides only a rough guideline, it

	Relationship strength		
Sample Size	Weak	Medium	Strong
Small ($N = 20$)	No	No	d = Maybe r = Yes
Medium ($N = 50$)	No	Yes	Yes
Large ($N = 100$)	d = Yes r = No	Yes	Yes
Extra large ($N = 500$)	Yes	Yes	Yes

Figure 1: How Relationship Strength and Sample Size Combine to Determine Whether a Result Is Statistically Significant

shows very clearly that weak relationships based on medium or small samples are never statistically significant and that strong relationships based on medium or larger samples are always statistically significant. If you keep this lesson in mind, you will often know whether a result is statistically significant based on the descriptive statistics alone. It is extremely useful to be able to develop this kind of intuitive judgment. One reason is that it allows you to develop expectations about how your formal null hypothesis tests are going to come out, which in turn allows you to detect problems in your analyses. For example, if your sample relationship is strong and your sample is medium, then you would expect to reject the null hypothesis. If for some reason your formal null hypothesis test indicates otherwise, then you need to double-check your computations and interpretations. A second reason is that the ability to make this kind of intuitive judgment is an indication that you understand the basic logic of this approach in addition to being able to do the computations.

Statistical Significance Versus Practical Significance

Figure Figure 1) illustrates another extremely important point. A statistically significant result is not necessarily a strong one. Even a very weak result can be statistically significant if it is based on a large enough sample. This is closely related to Janet Shibley Hyde's argument about sex differences (Hyde 2007). The differences between women and men in mathematical problem solving and leadership ability are statistically significant. But the word significant can cause people to interpret these differences as strong and important—perhaps even important enough to influence the college courses they take or even who they vote for. As we have seen, however, these statistically significant differences are actually quite weak—perhaps even "trivial."

This is why it is important to distinguish between the statistical significance of a result and the practical significance of that result. Practical significance refers to the importance or usefulness of the result in some real-world context. Many sex differences are statistically significant—and may even be interesting for purely scientific reasons—but they are not practically significant. In clinical practice, this same concept is often referred to as "clinical significance." For example, a study on a new treatment for social phobia might show that it produces a statistically significant positive effect. Yet this effect still might not be strong enough to justify the time, effort, and other costs of putting it into practice—especially if easier and cheaper treatments that work almost as well already exist. Although statistically significant, this result would be said to lack practical or clinical significance.

Key Takeaways

- Null hypothesis testing is a formal approach to deciding whether a statistical relationship in a sample reflects a real relationship in the population or is just due to chance.
- The logic of null hypothesis testing involves assuming that the null hypothesis is true, finding how likely the sample result would be if this assumption were correct, and then

making a decision. If the sample result would be unlikely if the null hypothesis were true, then it is rejected in favor of the alternative hypothesis. If it would not be unlikely, then the null hypothesis is retained.

- The probability of obtaining the sample result if the null hypothesis were true (the p value) is based on two considerations: relationship strength and sample size. Reasonable judgments about whether a sample relationship is statistically significant can often be made by quickly considering these two factors.
- Statistical significance is not the same as relationship strength or importance. Even weak relationships can be statistically significant if the sample size is large enough. It is important to consider relationship strength and the practical significance of a result in addition to its statistical significance.

Exercises

1. Discussion: Imagine a study showing that people who eat more broccoli tend to be happier. Explain for someone who knows nothing about statistics why the researchers would conduct a null hypothesis test.
2. Practice: Use Figure @ref(fig:IS1) to decide whether each of the following results is statistically significant.

- The correlation between two variables is r = -.78 based on a sample size of 137.
- The mean score on a psychological characteristic for women is 25 (SD = 5) and the mean score for men is 24 (SD = 5). There were 12 women and 10 men in this study.
- In a memory experiment, the mean number of items recalled by the 40 participants in Condition A was 0.50 standard deviations greater than the mean number recalled by the 40 participants in Condition B.
- In another memory experiment, the mean scores for participants in Condition A and Condition B came out exactly the same!

- A student finds a correlation of r = .04 between the number of units the students in his research methods class are taking and the students' level of stress.

Some Basic Null Hypothesis Tests

In this section, we look at several common null hypothesis testing procedures. The emphasis here is on providing enough information to allow you to conduct and interpret the most basic versions. In most cases, the online statistical analysis tools mentioned in Chapter 12 will handle the computations—as will programs such as Microsoft Excel and SPSS.

Learning Objectives

1. Conduct and interpret one-sample, dependent-samples, and independent-samples t tests.
2. Interpret the results of one-way, repeated measures, and factorial ANOVAs.
3. Conduct and interpret null hypothesis tests of Pearson's r.

The t Test

As we have seen throughout this book, many studies in psychology focus on the difference between two means. The most common null hypothesis test for this type of statistical relationship is the t test. In this section, we look at three types of t tests that are used for slightly different research designs: the one-sample t test, the dependent- samples t test, and the independent-samples t test.

One-Sample t Test

The one-sample t test is used to compare a sample mean (M) with a hypothetical population mean (\$ 0)*thatprovidessomeinterestingstandardofcomparison.Thenullhypothesi* is equal to the hypothetical population mean: $\mu =$ \$ \$0. The alternative hypothesis is that the mean for the population is different from the hypothetical population mean: $\mu \neq$ \$ \$0. To decide between these two hypotheses, we need to find the probability of obtaining the sample mean (or one more extreme) if the null hypothesis were true. But finding this p value requires first computing a test statistic called t. (A test statistic is a statistic that is computed only to help find the p value.) The formula for t is as follows:

$$t = \frac{M - u0}{\left(\frac{SD}{\sqrt{N}}\right)}$$

Again, M is the sample mean and $ $0 is the hypothetical population mean of interest. SD is the sample standard deviation and N is the sample size.

The reason the t statistic (or any test statistic) is useful is that we know how it is distributed when the null hypothesis is true. As shown in Figure Figure 2, this distribution is unimodal and symmetrical, and it has a mean of 0. Its precise shape depends on a statistical concept called the degrees of freedom, which for a one-sample t test is N - 1. (There are 24 degrees of freedom for the distribution shown in Figure Figure 2.) The important point is that knowing this distribution makes it possible to find the p value for any t score. Consider, for example, a t score of +1.50 based on a sample of 25. The probability of a t score at least this extreme is given by the proportion of t scores in the distribution that are at least this extreme. For now, let us define extreme as being far from zero in either direction. Thus the p value is the proportion of t scores that are +1.50 or above or that are -1.50 or below—a value that turns out to be .14.

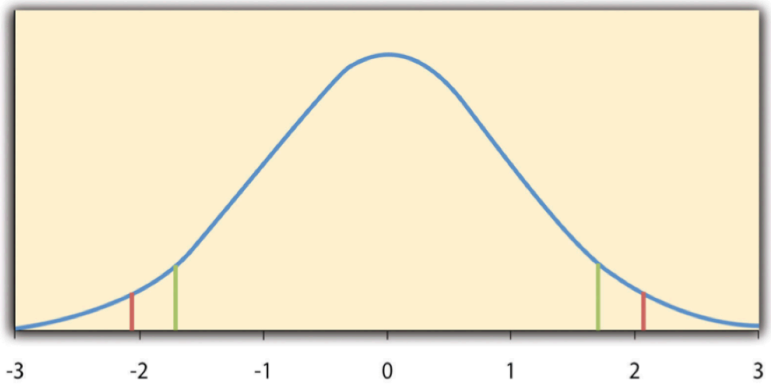

Figure 2: Distribution of t Scores (With 24 Degrees of Freedom) When the Null Hypothesis Is True. The red vertical lines represent the two-tailed critical values, and the green vertical lines the one-tailed critical values when alpha = .05.

Figure Figure 2 Distribution of t Scores (With 24 Degrees of Freedom) When the Null Hypothesis Is True. The red vertical lines represent the two-tailed critical values, and the green vertical lines the one-tailed critical values when $\alpha = .05$.

Fortunately, we do not have to deal directly with the distribution of t scores. If we were to enter our sample data and hypothetical mean of interest into one of the online statistical tools in Chapter 12 or into a program like SPSS (Excel does not have a one-sample t test function), the output would include both the t score and the p value. At this point, the rest of the procedure is simple. If p is less than .05, we reject the null hypothesis and conclude that the population mean differs from the hypothetical mean of interest. If p is greater than .05, we retain the null hypothesis and conclude that there is not enough evidence to say that the population mean differs from the hypothetical mean of interest. (Again, technically, we conclude only that we do not have enough evidence to conclude that it does differ.)

If we were to compute the t score by hand, we could use a table like Table 13.2 to make the decision. This table does not provide actual p values. Instead, it provides the critical values of t for different degrees of freedom (df) when α is .05. For now, let us focus on the two-tailed critical values in the last column of the table. Each of these values should be interpreted as a pair of values: one positive and one negative. For example, the two-tailed critical values when there are 24 degrees of freedom are +2.064 and -2.064. These are represented by the red vertical lines in Figure Figure 2. The idea is that any t score below the lower critical value (the left-hand red line in Figure Figure 2) is in the lowest 2.5% of the distribution, while any t score above the upper critical value (the right-hand red line) is in the highest 2.5% of the distribution. Therefore any t score beyond the critical value in either direction is in the most extreme 5% of t scores when the null hypothesis is true and has a p value less than .05. Thus if the t score we compute is beyond the critical value in either direction, then we reject the null hypothesis. If the t score we compute is between the upper and lower critical values, then we retain the null hypothesis.

Thus far, we have considered what is called a two-tailed test,

where we reject the null hypothesis if the t score for the sample is extreme in either direction. This test makes sense when we believe that the sample mean might differ from the hypothetical population mean but we do not have good reason to expect the difference to go in a particular direction. But it is also possible to do a one-tailed test, where we reject the null hypothesis only if the t score for the sample is extreme in one direction that we specify before collecting the data. This test makes sense when we have good reason to expect the sample mean will differ from the hypothetical population mean in a particular direction.

Here is how it works. Each one-tailed critical value in Table 13.2 can again be interpreted as a pair of values: one positive and one negative. A t score below the lower critical value is in the lowest 5% of the distribution, and a t score above the upper critical value is in the highest 5% of the distribution. For 24 degrees of freedom, these values are -1.711 and +1.711. (These are represented by the green vertical lines in Figure Figure 2.) However, for a one-tailed test, we must decide before collecting data whether we expect the sample mean to be lower than the hypothetical population mean, in which case we would use only the lower critical value, or we expect the sample mean to be greater than the hypothetical population mean, in which case we would use only the upper critical value. Notice that we still reject the null hypothesis when the t score for our sample is in the most extreme 5% of the t scores we would expect if the null hypothesis were true—so α remains at .05. We have simply redefined extreme to refer only to one tail of the distribution. The advantage of the one-tailed test is that critical values are less extreme. If the sample mean differs from the hypothetical population mean in the expected direction, then we have a better chance of rejecting the null hypothesis. The disadvantage is that if the sample mean differs from the hypothetical population mean in the unexpected direction, then there is no chance at all of rejecting the null hypothesis.

Example One-Sample t Test

Imagine that a health psychologist is interested in the accuracy of university students' estimates of the number of calories in a

chocolate chip cookie. He shows the cookie to a sample of 10 students and asks each one to estimate the number of calories in it. Because the actual number of calories in the cookie is 250, this is the hypothetical population mean of interest ($ 0). *The null hypothesis is is that the mean estimate for the population* ($) is 250. Because he has no real sense of whether the students will underestimate or overestimate the number of calories, he decides to do a two-tailed test. Now imagine further that the participants' actual estimates are as follows:

250, 280, 200, 150, 175, 200, 200, 220, 180, 250

The mean estimate for the sample (M) is 212.00 calories and the standard deviation (SD) is 39.17. The health psychologist can now compute the t score for his sample:

$$t = \frac{212 - 250}{\left(\frac{39.17}{\sqrt{10}}\right)} = -3.07$$

If he enters the data into one of the online analysis tools or uses SPSS, it would also tell him that the two- tailed p value for this t score (with 10 - 1 = 9 degrees of freedom) is .013. Because this is less than .05, the health psychologist would reject the null hypothesis and conclude that university students tend to underestimate the number of calories in a chocolate chip cookie. If he computes the t score by hand, he could look at Table 13.2 and see that the critical value of t for a two-tailed test with 9 degrees of freedom is ±2.262. The fact that his t score was more extreme than this critical value would tell him that his p value is less than .05 and that he should reject the null hypothesis.

Finally, if this researcher had gone into this study with good reason to expect that university students underestimate the number of calories, then he could have done a one-tailed test instead of a two-tailed test. The only thing this decision would change is the critical value, which would be -1.833. This slightly less extreme value would make it a bit easier to reject the null hypothesis. However, if it turned out that university students overestimate the number of calories—no matter how much they overestimate it—the researcher would not have been able to reject the null hypothesis.

The Dependent-Samples t Test

The dependent-samples t test (sometimes called the paired-samples t test) is used to compare two means for the same sample tested at two different times or under two different conditions. This comparison is appropriate for pretest-posttest designs or within-subjects experiments. The null hypothesis is that the means at the two times or under the two conditions are the same in the population. The alternative hypothesis is that they are not the same. This test can also be one-tailed if the researcher has good reason to expect the difference goes in a particular direction.

It helps to think of the dependent-samples t test as a special case of the one-sample t test. However, the first step in the dependent-samples t test is to reduce the two scores for each participant to a single difference score by taking the difference between them. At this point, the dependent-samples t test becomes a one-sample t test on the difference scores. The hypothetical population mean ($ 0) of interest is 0 because this is what the mean difference score would be if there were no difference on average between
0) and the alternative hypothesis as being that the mean difference score in the population is not 0 ($0
≠ 0).

Example Dependent-Samples t Test

Imagine that the health psychologist now knows that people tend to underestimate the number of calories in junk food and has developed a short training program to improve their estimates. To test the effectiveness of this program, he conducts a pretest-posttest study in which 10 participants estimate the number of calories in a chocolate chip cookie before the training program and then again afterward. Because he expects the program to increase the participants' estimates, he decides to do a one-tailed test. Now imagine further that the pretest estimates are

230, 250, 280, 175, 150, 200, 180, 210, 220, 190

and that the posttest estimates (for the same participants in the same order) are

250, 260, 250, 200, 160, 200, 200, 180, 230, 240.

The difference scores, then, are as follows:

+20, +10, -30, +25, +10, 0, +20, -30, +10, +50.

Note that it does not matter whether the first set of scores is subtracted from the second or the second from the first as long as it is done the same way for all participants. In this example, it makes sense to subtract the pretest estimates from the posttest estimates so that positive difference scores mean that the estimates went up after the training and negative difference scores mean the estimates went down.

The mean of the difference scores is 8.50 with a standard deviation of 27.27. The health psychologist can now compute the t score for his sample as follows:

$$t = \frac{8.5 - 0}{\left(\frac{27.27}{\sqrt{10}}\right)} = 1.11$$

If he enters the data into one of the online analysis tools or uses Excel or SPSS, it would tell him that the one- tailed p value for this t score (again with 10 - 1 = 9 degrees of freedom) is .148. Because this is greater than .05, he would retain the null hypothesis and conclude that the training program does not increase people's calorie estimates. If he were to compute the t score by hand, he could look at Table 13.2 and see that the critical value of t for a one- tailed test with 9 degrees of freedom is +1.833. (It is positive this time because he was expecting a positive mean difference score.) The fact that his t score was less extreme than this critical value would tell him that his p value is greater than .05 and that he should fail to reject the null hypothesis.

The Independent-Samples t Test

The independent-samples t test is used to compare the means of two separate samples (M1 and M2). The two samples might have been tested under different conditions in a between-subjects experiment, or they could be preexisting groups in a correlational design (e.g., women and men, extraverts and

introverts). The null hypothesis is that the means of the two populations are the same: $ $1 = $ $2. The alternative hypothesis is that they are not the same: $ $1 ≠ $ $2. Again, the test can be one-tailed if the researcher has good reason to expect the difference goes in a particular direction.

The t statistic here is a bit more complicated because it must take into account two sample means, two standard deviations, and two sample sizes. The formula is as follows:

Notice that this formula includes squared standard deviations (the variances) that appear inside the square root symbol. Also, lowercase n1 and n2 refer to the sample sizes in the two groups or condition (as opposed to capital N, which generally refers to the total sample size). The only additional thing to know here is that there are N - 2 degrees of freedom for the independent-samples t test.

Example Independent-Samples t test

Now the health psychologist wants to compare the calorie estimates of people who regularly eat junk food with the estimates of people who rarely eat junk food. He believes the difference could come out in either direction so he decides to conduct a two-tailed test. He collects data from a sample of eight participants who eat junk food regularly and seven participants who rarely eat junk food. The data are as follows:

Junk food eaters: 180, 220, 150, 85, 200, 170, 150, 190

Non–junk food eaters: 200, 240, 190, 175, 200, 300, 240

The mean for the junk food eaters is 220.71 with a standard deviation of 41.23. The mean for the non–junk food eaters is 168.12 with a standard deviation of 42.66. He can now compute his t score as follows:

$$t = \frac{220.71 - 168.12}{\sqrt{\frac{41.28^2}{8} + \frac{42.66^2}{7}}} = 2.42$$

If he enters the data into one of the online analysis tools or uses Excel or SPSS, it would tell him that the two- tailed p

value for this t score (with 15 - 2 = 13 degrees of freedom) is .015. Because this p value is less than .05, the health psychologist would reject the null hypothesis and conclude that people who eat junk food regularly make lower calorie estimates than people who eat it rarely. If he were to compute the t score by hand, he could look at Table 13.2 and see that the critical value of t for a two-tailed test with 13 degrees of freedom is ±2.160. The fact that his t score was more extreme than this critical value would tell him that his p value is less than .05 and that he should fail to retain the null hypothesis.

The Analysis of Variance

When there are more than two groups or condition means to be compared, the most common null hypothesis test is the analysis of variance (ANOVA). In this section, we look primarily at the one-way ANOVA, which is used for between-subjects designs with a single independent variable. We then briefly consider some other versions of the ANOVA that are used for within-subjects and factorial research designs.

One-Way ANOVA

The one-way ANOVA is used to compare the means of more than two samples (M1, M2 …MG) in a between- subjects design. The null hypothesis is that all the means are equal in the population: $ $1= $ $2 =…= μG. The alternative hypothesis is that not all the means in the population are equal.

The test statistic for the ANOVA is called F. It is a ratio of two estimates of the population variance based on the sample data. One estimate of the population variance is called the mean squares between groups (MSB) and is based on the differences among the sample means. The other is called the mean squares within groups (MSW) and is based on the differences among the scores within each group. The F statistic is the ratio of the MSB to the MSW and can therefore be expressed as follows:

$$F = \frac{MSW}{MSW}$$

Again, the reason that F is useful is that we know how it is distributed when the null hypothesis is true. As shown in Figure Figure 3, this distribution is unimodal and positively skewed with values that cluster around 1. The precise shape of the distribution depends on both the number of groups and the sample size, and there is a degrees of freedom value associated with each of these. The between-groups degrees of freedom is the number of groups minus one: dfB = (G - 1). The within-groups degrees of freedom is the total sample size minus the number of groups: dfW = N - G. Again, knowing the distribution of F when the null hypothesis is true allows us to find the p value.

The online tools in Chapter 12 and statistical software such as Excel and SPSS will compute F and find the p value. If p is less than .05, then we reject the null hypothesis and conclude that there are differences among the group means in the population.

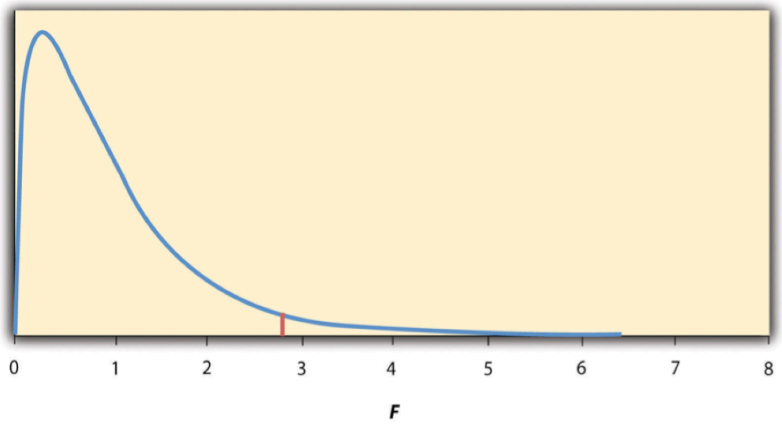

Figure 3: Distribution of the F Ratio With 2 and 37 Degrees of Freedom When the Null Hypothesis Is True. The red vertical line represents the critical value when alpha is .05

Figure Figure 3 Distribution of the F Ratio With 2 and 37 Degrees of Freedom When the Null Hypothesis Is True. The red vertical line represents the critical value when α is .05.

If p is greater than .05, then we retain the null hypothesis and

conclude that there is not enough evidence to say that there are differences. In the unlikely event that we would compute F by hand, we can use a table of critical values like Table 13.3 "Table of Critical Values of" to make the decision. The idea is that any F ratio greater than the critical value has a p value of less than .05. Thus if the F ratio we compute is beyond the critical value, then we reject the null hypothesis. If the F ratio we compute is less than the critical value, then we retain the null hypothesis.

Example One-Way ANOVA

Imagine that the health psychologist wants to compare the calorie estimates of psychology majors, nutrition majors, and professional dieticians. He collects the following data:

Psych majors: 200, 180, 220, 160, 150, 200, 190, 200

Nutrition majors: 190, 220, 200, 230, 160, 150, 200, 210, 195

Dieticians: 220, 250, 240, 275, 250, 230, 200, 240

The means are 187.50 (SD = 23.14), 195.00 (SD = 27.77), and 238.13 (SD = 22.35), respectively. So it appears that dieticians made substantially more accurate estimates on average. The researcher would almost certainly enter these data into a program such as Excel or SPSS, which would compute F for him and find the p value. Figure Figure 4 shows the output of the one-way ANOVA function in Excel for these data.

ANOVA						
Source of variation	*SS*	*df*	*MS*	*F*	*p-value*	F_{crit}
Between groups	11,943.75	2	5,971.875	9.916234	0.000928	3.4668
Within groups	12,646.88	21	602.2321			
Total	24,590.63	23				

Figure 4: ANOVA table output

This table is referred to as an ANOVA table. It shows that MSB is 5,971.88, MSW is 602.23, and their ratio, F, is 9.92. The p

value is .0009. Because this value is below .05, the researcher would reject the null hypothesis and conclude that the mean calorie estimates for the three groups are not the same in the population. Notice that the ANOVA table also includes the "sum of squares" (SS) for between groups and for within groups. These values are computed on the way to finding MSB and MSW but are not typically reported by the researcher. Finally, if the researcher were to compute the F ratio by hand, he could look at Table 13.3 and see that the critical value of F with 2 and 21 degrees of freedom is 3.467 (the same value in Figure Figure 4 under Fcrit). The fact that his F score was more extreme than this critical value would tell him that his p value is less than .05 and that he should reject the null hypothesis.

ANOVA Elaborations

Post Hoc Comparisons

When we reject the null hypothesis in a one-way ANOVA, we conclude that the group means are not all the same in the population. But this can indicate different things. With three groups, it can indicate that all three means are significantly different from each other. Or it can indicate that one of the means is significantly different from the other two, but the other two are not significantly different from each other. It could be, for example, that the mean calorie estimates of psychology majors, nutrition majors, and dieticians are all significantly different from each other. Or it could be that the mean for dieticians is significantly different from the means for psychology and nutrition majors, but the means for psychology and nutrition majors are not significantly different from each other. For this reason, statistically significant one-way ANOVA results are typically followed up with a series of post hoc comparisons of selected pairs of group means to determine which are different from which others.

One approach to post hoc comparisons would be to conduct a series of independent-samples t tests comparing each group mean to each of the other group means. But there is a problem with this approach. In general, if we conduct a t test when

the null hypothesis is true, we have a 5% chance of mistakenly rejecting the null hypothesis (see Section "Additional Considerations" for more on such Type I errors). If we conduct several t tests when the null hypothesis is true, the chance of mistakenly rejecting at least one null hypothesis increases with each test we conduct. Thus researchers do not usually make post hoc comparisons using standard t tests because there is too great a chance that they will mistakenly reject at least one null hypothesis. Instead, they use one of several modified t test procedures—among them the Bonferonni procedure, Fisher's least significant difference (LSD) test, and Tukey's honestly significant difference (HSD) test. The details of these approaches are beyond the scope of this book, but it is important to understand their purpose. It is to keep the risk of mistakenly rejecting a true null hypothesis to an acceptable level (close to 5%).

Repeated-Measures ANOVA

Recall that the one-way ANOVA is appropriate for between-subjects designs in which the means being compared come from separate groups of participants. It is not appropriate for within-subjects designs in which the means being compared come from the same participants tested under different conditions or at different times. This requires a slightly different approach, called the repeated-measures ANOVA. The basics of the repeated-measures ANOVA are the same as for the one-way ANOVA. The main difference is that measuring the dependent variable multiple times for each participant allows for a more refined measure of MSW. Imagine, for example, that the dependent variable in a study is a measure of reaction time. Some participants will be faster or slower than others because of stable individual differences in their nervous systems, muscles, and other factors. In a between-subjects design, these stable individual differences would simply add to the variability within the groups and increase the value of MSW. In a within-subjects design, however, these stable individual differences can be measured and subtracted from the value of MSW. This lower value of MSW means a higher value of F and a more sensitive test.

Factorial ANOVA

When more than one independent variable is included in a factorial design, the appropriate approach is the factorial ANOVA. Again, the basics of the factorial ANOVA are the same as for the one-way and repeated- measures ANOVAs. The main difference is that it produces an F ratio and p value for each main effect and for each interaction. Returning to our calorie estimation example, imagine that the health psychologist tests the effect of participant major (psychology vs. nutrition) and food type (cookie vs. hamburger) in a factorial design. A factorial ANOVA would produce separate F ratios and p values for the main effect of major, the main effect of food type, and the interaction between major and food. Appropriate modifications must be made depending on whether the design is between subjects, within subjects, or mixed.

Testing Pearson's r

For relationships between quantitative variables, where Pearson's r is used to describe the strength of those relationships, the appropriate null hypothesis test is a test of Pearson's r. The basic logic is exactly the same as for other null hypothesis tests. In this case, the null hypothesis is that there is no relationship in the population. We can use the Greek lowercase rho (ρ) to represent the relevant parameter: $\rho = 0$. The alternative hypothesis is that there is a relationship in the population: $\rho \neq 0$. As with the t test, this test can be two-tailed if the researcher has no expectation about the direction of the relationship or one-tailed if the researcher expects the relationship to go in a particular direction.

It is possible to use Pearson's r for the sample to compute a t score with N - 2 degrees of freedom and then to proceed as for a t test. However, because of the way it is computed, Pearson's r can also be treated as its own test statistic. The online statistical tools and statistical software such as Excel and SPSS generally compute Pearson's r and provide the p value associated with that value of Pearson's r. As always, if the p value is less than .05, we reject the null hypothesis and

conclude that there is a relationship between the variables in the population. If the p value is greater than .05, we retain the null hypothesis and conclude that there is not enough evidence to say there is a relationship in the population. If we compute Pearson's r by hand, we can use a table like Table 13.5, which shows the critical values of r for various samples sizes when α is .05. A sample value of Pearson's r that is more extreme than the critical value is statistically significant.

Example Test of Pearson's r

Imagine that the health psychologist is interested in the correlation between people's calorie estimates and their weight. He has no expectation about the direction of the relationship, so he decides to conduct a two-tailed test. He computes the correlation for a sample of 22 university students and finds that Pearson's r is -.21. The statistical software he uses tells him that the p value is .348. It is greater than .05, so he retains the null hypothesis and concludes that there is no relationship between people's calorie estimates and their weight. If he were to compute Pearson's r by hand, he could look at Table 13.5 and see that the critical value for 22 - 2 = 20 degrees of freedom is .444. The fact that Pearson's r for the sample is less extreme than this critical value tells him that the p value is greater than .05 and that he should retain the null hypothesis.

Key Takeaways

- To compare two means, the most common null hypothesis test is the t test. The one-sample t test is used for comparing one sample mean with a hypothetical population mean of interest, the dependent-samples t test is used to compare two means in a within-subjects design, and the independent-samples t test is used to compare two means in a between-subjects design.
- To compare more than two means, the most common null hypothesis test is the analysis of variance (ANOVA). The one-way ANOVA is used for between-subjects designs with one independent variable, the repeated-measures

ANOVA is used for within-subjects designs, and the factorial ANOVA is used for factorial designs.

- A null hypothesis test of Pearson's r is used to compare a sample value of Pearson's r with a hypothetical population value of 0.

Exercises

1. Practice: Use one of the online tools, Excel, or SPSS to reproduce the one-sample t test, dependent-samples t test, independent-samples t test, and one-way ANOVA for the four sets of calorie estimation data presented in this section.
2. Practice: A sample of 25 university students rated their friendliness on a scale of 1 (Much Lower Than Average) to 7 (Much Higher Than Average). Their mean rating was 5.30 with a standard deviation of 1.50. Conduct a one-sample t test comparing their mean rating with a hypothetical mean rating of 4 (Average). The question is whether university students have a tendency to rate themselves as friendlier than average.
3. Practice: Decide whether each of the following Pearson's r values is statistically significant for both a one-tailed and a two-tailed test.

- The correlation between height and IQ is +.13 in a sample of 35.
- For a sample of 88 university students, the correlation between how disgusted they felt and the harshness of their moral judgments was +.23.
- The correlation between the number of daily hassles and positive mood is -.43 for a sample of 30 middle-aged adults.

Additional Considerations

In this section, we consider a few other issues related to null hypothesis testing, including some that are useful in

Learning Objectives

1. Define Type I and Type II errors, explain why they occur, and identify some steps that can be taken to minimize their likelihood.
2. Define statistical power, explain its role in the planning of new studies, and use online tools to compute the statistical power of simple research designs.
3. List some criticisms of conventional null hypothesis

planning studies and interpreting results. We even consider some long-standing criticisms of null hypothesis testing, along with some steps that researchers in psychology have taken to address them.

Errors in Null Hypothesis Testing

In null hypothesis testing, the researcher tries to draw a reasonable conclusion about the population based on the sample. Unfortunately, this conclusion is not guaranteed to be correct. This discrepancy is illustrated by Figure Figure 5. The rows of this table represent the two possible decisions that we can make in null hypothesis testing: to reject or retain the null hypothesis. The columns represent the two possible states of the world: The null hypothesis is false or it is true. The four cells of the table, then, represent the four distinct outcomes of a null hypothesis test. Two of the outcomes—rejecting the null hypothesis when it is false and retaining it when it is true—are correct decisions. The other two—rejecting the null hypothesis when it is true and retaining it when it is false—are errors.

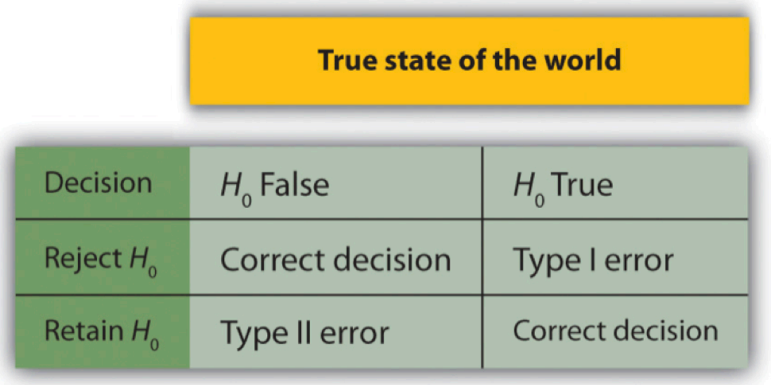

Figure 5: Two Types of Correct Decisions and Two Types of Errors in Null Hypothesis Testing

Rejecting the null hypothesis when it is true is called a Type I error. This error means that we have concluded that there is a relationship in the population when in fact there is not. Type I errors occur because even when there is no relationship in the population, sampling error alone will occasionally produce an

extreme result. In fact, when the null hypothesis is true and α is .05, we will mistakenly reject the null hypothesis 5% of the time. (This possibility is why α is sometimes referred to as the "Type I error rate.") Retaining the null hypothesis when it is false is called a Type II error. This error means that we have concluded that there is no relationship in the population when in fact there is. In practice, Type II errors occur primarily because the research design lacks adequate statistical power to detect the relationship (e.g., the sample is too small). We will have more to say about statistical power shortly.

In principle, it is possible to reduce the chance of a Type I error by setting α to something less than .05. Setting it to .01, for example, would mean that if the null hypothesis is true, then there is only a 1% chance of mistakenly rejecting it. But making it harder to reject true null hypotheses also makes it harder to reject false ones and therefore increases the chance of a Type II error. Similarly, it is possible to reduce the chance of a Type II error by setting α to something greater than .05 (e.g., .10). But making it easier to reject false null hypotheses also makes it easier to reject true ones and therefore increases the chance of a Type I error. This provides some insight into why the convention is to set α to .05. There is some agreement among researchers that level of α keeps the rates of both Type I and Type II errors at acceptable levels.

The possibility of committing Type I and Type II errors has several important implications for interpreting the results of our own and others' research. One is that we should be cautious about interpreting the results of any individual study because there is a chance that it reflects a Type I or Type II error. This possibility is why researchers consider it important to replicate their studies. Each time researchers replicate a study and find a similar result, they rightly become more confident that the result represents a real phenomenon and not just a Type I or Type II error.

Another issue related to Type I errors is the so-called file drawer problem (Rosenthal 1979). The idea is that when researchers obtain statistically significant results, they tend to submit them for publication, and journal editors and reviewers tend to accept them. But when researchers obtain non-significant results,

they tend not to submit them for publication, or if they do submit them, journal editors and reviewers tend not to accept them. Researchers end up putting these non-significant results away in a file drawer (or nowadays, in a folder on their hard drive). One effect of this tendency is that the published literature probably contains a higher proportion of Type I errors than we might expect on the basis of statistical considerations alone. Even when there is a relationship between two variables in the population, the published research literature is likely to overstate the strength of that relationship. Imagine, for example, that the relationship between two variables in the population is positive but weak (e.g., $\rho = +.10$). If several researchers conduct studies on this relationship, sampling error is likely to produce results ranging from weak negative relationships (e.g., r = -.10) to moderately strong positive ones (e.g., r = +.40). But because of the file drawer problem, it is likely that only those studies producing moderate to strong positive relationships are published. The result is that the effect reported in the published literature tends to be stronger than it really is in the population.

Figure 6: An Example of How Type I and Type II Errors Could Play out in Pregnancy Exams.

The file drawer problem is a difficult one because it is a product of the way scientific research has traditionally been conducted and published. One solution might be for journal editors and reviewers to evaluate research submitted for publication without knowing the results of that research. The idea is that if the

research question is judged to be interesting and the method judged to be sound, then a non-significant result should be just as important and worthy of publication as a significant one. Short of such a radical change in how research is evaluated for publication, researchers can still take pains to keep their non-significant results and share them as widely as possible (e.g., at professional conferences). Many scientific disciplines now have journals devoted to publishing non-significant results. In psychology, for example, there is the Journal of Articles in Support of the Null Hypothesis (http://www.jasnh.com).

In 2014, Uri Simonsohn, Leif Nelson, and Joseph Simmons published a leveling article at the field of psychology accusing researchers of creating too many Type I errors in psychology by chasing a significant p value through what they called p-hacking (Simonsohn, Nelson, and Simmons 2014). Researchers are trained in many sophisticated statistical techniques for analyzing data that will yield a desirable p value. They propose using a p-curve to determine whether the data set with a certain p value is credible or not. They also propose this p-curve as a way to unlock the file drawer because we can only understand the finding if we know the true effect size and the likelihood of a result was found after multiple attempts at not finding a result. Their groundbreaking paper contributed to a major conversation in the field about publishing standards and the reliability of our results.

Statistical Power

The statistical power of a research design is the probability of rejecting the null hypothesis given the sample size and expected relationship strength. For example, the statistical power of a study with 50 participants and an expected Pearson's r of +.30 in the population is .59. That is, there is a 59% chance of rejecting the null hypothesis if indeed the population correlation is +.30. Statistical power is the complement of the probability of committing a Type II error. So in this example, the probability of committing a Type II error would be 1 - .59 = .41. Clearly, researchers should be interested in the power of their research designs if they want to avoid making Type II errors.

In particular, they should make sure their research design has adequate power before collecting data. A common guideline is that a power of .80 is adequate. This guideline means that there is an 80% chance of rejecting the null hypothesis for the expected relationship strength.

The topic of how to compute power for various research designs and null hypothesis tests is beyond the scope of this book. However, there are online tools that allow you to do this by entering your sample size, expected relationship strength, and α level for various hypothesis tests (see "Computing Power Online"). In addition, Figure Figure 7 shows the sample size needed to achieve a power of .80 for weak, medium, and strong relationships for a two- tailed independent-samples t test and for a two-tailed test of Pearson's r. Notice that this table amplifies the point made earlier about relationship strength, sample size, and statistical significance. In particular, weak relationships require very large samples to provide adequate statistical power.

Null Hypothesis Test		
Relationship Strength	Independent-Samples *t* Test	Test of Pearson's *r*
Strong (*d* = .80, *r* = .50)	52	28
Medium (*d* = .50, *r* = .30)	128	84
Weak (*d* = .20, *r* = .10)	788	782

Figure 7: Sample Sizes Needed to Achieve Statistical Power of .80 for Different Expected Relationship Strengths for an Independent- Samples t Test and a Test of Pearson's r Null Hypothesis Test

What should you do if you discover that your research design does not have adequate power? Imagine, for example, that you are conducting a between-subjects experiment with 20 participants in each of two conditions and that you expect a medium difference (d = .50) in the population. The statistical power of this design is only .34. That is, even if there is a medium difference in the population, there is only about a one in three chance of rejecting the null hypothesis and about a two in three

chance of committing a Type II error.

Given the time and effort involved in conducting the study, this probably seems like an unacceptably low chance of rejecting the null hypothesis and an unacceptably high chance of committing a Type II error. Given that statistical power depends primarily on relationship strength and sample size, there are essentially two steps you can take to increase statistical power: increase the strength of the relationship or increase the sample size. Increasing the strength of the relationship can sometimes be accomplished by using a stronger manipulation or by more carefully controlling extraneous variables to reduce the amount of noise in the data (e.g., by using a within- subjects design rather than a between-subjects design). The usual strategy, however, is to increase the sample size. For any expected relationship strength, there will always be some sample large enough to achieve adequate power.

Computing Power Online

The following links are to tools that allow you to compute statistical power for various research designs and null hypothesis tests by entering information about the expected relationship strength, the sample size, and the α level. They also allow you to compute the sample size necessary to achieve your desired level of power (e.g., .80). The first is an online tool. The second is a free downloadable program called G*Power.

- Russ Lenth's Power and Sample Size Page: http://www.stat.uiowa.edu/~rlenth/Power/index.html
- G*Power: http://www.gpower.hhu.de

Problems With Null Hypothesis Testing, and Some Solutions

Again, null hypothesis testing is the most common approach to inferential statistics in psychology. It is not without its critics, however. In fact, in recent years the criticisms have become so prominent that the American Psychological Association convened a task force to make recommendations about how to deal

with them (Wilkinson 1999). In this section, we consider some of the criticisms and some of the recommendations.

Criticisms of Null Hypothesis Testing

Some criticisms of null hypothesis testing focus on researchers' misunderstanding of it. We have already seen, for example, that the p value is widely misinterpreted as the probability that the null hypothesis is true. (Recall that it is really the probability of the sample result if the null hypothesis were true.) A closely related misinterpretation is that 1 - p is the probability of replicating a statistically significant result. In one study, 60% of a sample of professional researchers thought that a p value of .01—for an independent-samples t test with 20 participants in each sample—meant there was a 99% chance of replicating the statistically significant result (Oaks 1986). Our earlier discussion of power should make it clear that this figure is far too optimistic. As Table 13.5 shows, even if there were a large difference between means in the population, it would require 26 participants per sample to achieve a power of .80. And the program G*Power shows that it would require 59 participants per sample to achieve a power of .99.

Another set of criticisms focuses on the logic of null hypothesis testing. To many, the strict convention of rejecting the null hypothesis when p is less than .05 and retaining it when p is greater than .05 makes little sense. This criticism does not have to do with the specific value of .05 but with the idea that there should be any rigid dividing line between results that are considered significant and results that are not. Imagine two studies on the same statistical relationship with similar sample sizes. One has a p value of .04 and the other a p value of .06. Although the two studies have produced essentially the same result, the former is likely to be considered interesting and worthy of publication and the latter simply not significant. This convention is likely to prevent good research from being published and to contribute to the file drawer problem.

Yet another set of criticisms focus on the idea that null hypothesis testing—even when understood and carried out correctly—is simply not very informative. Recall that the null hypothesis

is that there is no relationship between variables in the population (e.g., Cohen's d or Pearson's r is precisely 0). So to reject the null hypothesis is simply to say that there is some nonzero relationship in the population. But this assertion is not really saying very much. Imagine if chemistry could tell us only that there is some relationship between the temperature of a gas and its volume—as opposed to providing a precise equation to describe that relationship. Some critics even argue that the relationship between two variables in the population is never precisely 0 if it is carried out to enough decimal places. In other words, the null hypothesis is never literally true. So rejecting it does not tell us anything we did not already know!

To be fair, many researchers have come to the defense of null hypothesis testing. One of them, Robert Abelson, has argued that when it is correctly understood and carried out, null hypothesis testing does serve an important purpose (Abelson 2012). Especially when dealing with new phenomena, it gives researchers a principled way to convince others that their results should not be dismissed as mere chance occurrences.

The end of p-values?

In 2015, the editors of Basic and Applied Social Psychology announced6 a ban on the use of null hypothesis testing and related statistical procedures. Authors are welcome to submit papers with p-values, but the editors will remove them before publication. Although they did not propose a better statistical test to replace null hypothesis testing, the editors emphasized the importance of descriptive statistics and effect sizes. This rejection of the "gold standard" of statistical validity has continued the conversation in psychology of questioning exactly what we know.

What to Do?

Even those who defend null hypothesis testing recognize many of the problems with it. But what should be done? Some suggestions now appear in the Publication Manual. One is that each null hypothesis test should be accompanied by an effect

size measure such as Cohen's d or Pearson's r. By doing so, the researcher provides an estimate of how strong the relationship in the population is—not just whether there is one or not. (Remember that the p value cannot substitute as a measure of relationship strength because it also depends on the sample size. Even a very weak result can be statistically significant if the sample is large enough.)

Another suggestion is to use confidence intervals rather than null hypothesis tests. A confidence interval around a statistic is a range of values that is computed in such a way that some percentage of the time (usually 95%) the population parameter will lie within that range. For example, a sample of 20 university students might have a mean calorie estimate for a chocolate chip cookie of 200 with a 95% confidence interval of 160 to 240. In other words, there is a very good chance that the mean calorie estimate for the population of university students lies between 160 and 240. Advocates of confidence intervals argue that they are much easier to interpret than null hypothesis tests. Another advantage of confidence intervals is that they provide the information necessary to do null hypothesis tests should anyone want to. In this example, the sample mean of 200 is significantly different at the .05 level from any hypothetical population mean that lies outside the confidence interval. So the confidence interval of 160 to 240 tells us that the sample mean is statistically significantly different from a hypothetical population mean of 250.

Finally, there are more radical solutions to the problems of null hypothesis testing that involve using very different approaches to inferential statistics. Bayesian statistics, for example, is an approach in which the researcher specifies the probability that the null hypothesis and any important alternative hypotheses are true before conducting the study, conducts the study, and then updates the probabilities based on the data. It is too early to say whether this approach will become common in psychological research. For now, null hypothesis testing—supported by effect size measures and confidence intervals—remains the dominant approach.

Key Takeaways

- The decision to reject or retain the null hypothesis is not guaranteed to be correct. A Type I error occurs when one rejects the null hypothesis when it is true. A Type II error occurs when one fails to reject the null hypothesis when it is false.
- The statistical power of a research design is the probability of rejecting the null hypothesis given the expected relationship strength in the population and the sample size. Researchers should make sure that their studies have adequate statistical power before conducting them.
- Null hypothesis testing has been criticized on the grounds that researchers misunderstand it, that it is illogical, and that it is uninformative. Others argue that it serves an important purpose—especially when used with effect size measures, confidence intervals, and other techniques. It remains the dominant approach to inferential statistics in psychology.

Exercises

1. Discussion: A researcher compares the effectiveness of two forms of psychotherapy for social phobia using an independent-samples t test. a. Explain what it would mean for the researcher to commit a Type I error. b. Explain what it would mean for the researcher to commit a Type II error.
2. Discussion: Imagine that you conduct a t test and the p value is .02. How could you explain what this p value means to someone who is not already familiar with null hypothesis testing? Be sure to avoid the common misinterpretations of the p value.
3. For additional practice with Type I and Type II errors, try these problems from Carnegie Mellon's Open Learning Initiative.

From the "Replicability Crisis" to Open Science Practices

Learning Objectives

1. Describe what is meant by the "replicability crisis" in psychology.
2. Describe some questionable research practices.
3. Identify some ways in which scientific rigor may be increased.
4. Understand the importance of openness in psychological science.

At the start of this book we discussed the "Many Labs Replication Project," which failed to replicate the original finding by Simone Schnall and her colleagues that washing one's hands leads people to view moral transgressions as less wrong (Schnall, Benton, & Harvey, 2008)1. Although this project is a good illustration of the collaborative and self-correcting nature of science, it also represents one specific response to psychology's recent "replicability crisis," a phrase that refers to the inability of researchers to replicate earlier research findings. Consider for example the results of the Reproducibility Project, which involved over 270 psychologists around the world coordinating their efforts to test the reliability of 100 previously published psychological experiments (Aarts et al., 2015). Although 97 of the original 100 studies had found statistically significant effects, only 36 of the replications did! Moreover, even the effect sizes of the replications were, on average, half of those found in the original studies (see Figure 13.5). Of course, a failure to replicate a result by itself does not necessarily discredit the original study as differences in the statistical power, populations sampled, and procedures used, or even the effects of moderating variables could explain the different results (Yong, 2015).

Although many believe that the failure to replicate research results is an expected characteristic of cumulative scientific progress, others have interpreted this situation as evidence of systematic problems with conventional scholarship in psychology, including a publication bias that favors the discovery and publication of counter-intuitive but statistically significant findings instead of the duller (but incredibly vital) process of replicating previous findings to test their robustness (Aschwanden, 2015; Frank, 2015; Pashler & Harris, 2012; Scherer, 2015).

Worse still is the suggestion that the low replicability of many studies is evidence of the widespread use of questionable research practices by psychological researchers. These may include:

1. The selective deletion of outliers in order to influence

(usually by artificially inflating) statistical relationships among the measured variables.

2. The selective reporting of results, cherry-picking only those findings that support one's hypotheses.

3. Mining the data without an a priori hypothesis, only to claim that a statistically significant result had been originally predicted, a practice referred to as "HARKing" or hypothesizing after the results are known (Kerr 1998).

4. A practice colloquially known as "p-hacking" (briefly discussed in the previous section), in which a researcher might perform inferential statistical calculations to see if a result was significant before deciding whether to recruit additional participants and collect more data (Head et al. 2015). As you have learned, the probability of finding a statistically significant result is influenced by the number of participants in the study.

5. Outright fabrication of data (as in the case of Diederik Stapel, described at the start of Chapter 3), although this would be a case of fraud rather than a "research practice."

It is important to shed light on these questionable research practices to ensure that current and future researchers (such as yourself) understand the damage they wreak to the integrity and reputation of our discipline (see, for example, the "Replication Index," a statistical "doping test" developed by Ulrich Schimmack in 2014 for estimating the replicability of studies, journals, and even specific researchers). However, in addition to highlighting what not to do, this so-called "crisis" has also highlighted the importance of enhancing scientific rigor by:

1. Designing and conducting studies that have sufficient statistical power, in order to increase the reliability of findings.

2. Publishing both null and significant findings (thereby counteracting the publication bias and reducing the file drawer problem).

3. Describing one's research designs in sufficient detail to enable other researchers to replicate your study using an identical or at least very similar procedure.
4. Conducting high-quality replications and publishing these results (Brandt et al. 2014).

One particularly promising response to the replicability crisis has been the emergence of open science practices that increase the transparency and openness of the scientific enterprise. For example, Psychological Science (the flagship journal of the Association for Psychological Science) and other journals now issue digital badges to researchers who pre-registered their hypotheses and data analysis plans, openly shared their research materials with other researchers (e.g., to enable attempts at replication), or made available their raw data with other researchers (see Figure 13.6).

These initiatives, which have been spearheaded by the Center for Open Science, have led to the development of "Transparency and Openness Promotion guidelines" (see Table 13.7) that have since been formally adopted by more than 500 journals and 50 organizations, a list that grows each week. When you add to this the requirements recently imposed by federal funding agencies in Canada (the Tri-Council) and the United States (National Science Foundation) concerning the publication of publicly-funded research in open access journals, it certainly appears that the future of science and psychology will be one that embraces greater "openness" (Nosek et al. 2015).

Key Takeaways

- In recent years psychology has grappled with a failure to replicate research findings. Some have interpreted this as a normal aspect of science but others have suggested that this is highlights problems stemming from questionable research practices.
- One response to this "replicability crisis" has been the emergence of open science practices, which increase the transparency and openness of the research process. These

open practices include digital badges to encourage pre-registration of hypotheses and the sharing of raw data and research materials.

Exercises

1. Discussion: What do you think are some of the key benefits of the adoption of open science practices such as pre-registration and the sharing of raw data and research materials? Can you identify any drawbacks of these practices?

2. Practice: Read the online article "Science isn't broken: It's just a hell of a lot harder than we give it credit for" and use the interactive tool entitled "Hack your way to scientific glory" in order to better understand the data malpractice of "p-hacking."

References

Abelson, Robert P. 2012. *Statistics as Principled Argument*. Psychology Press.

Brandt, Mark J., Hans IJzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James A. Grange, Marco Perugini, Jeffrey R. Spies, and Anna Van't Veer. 2014. "The Replication Recipe: What Makes for a Convincing Replication?" *Journal of Experimental Social Psychology* 50: 217–24.

Cohen, J. 1994. "The World Is Round: P $<$ .05." *American Psychologist* 49: 997–1003.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of p-Hacking in Science." *PLoS Biology* 13 (3): e1002106.

Hyde, Janet Shibley. 2007. "New Directions in the Study of Gender Similarities and Differences." *Current Directions in Psychological Science* 16 (5): 259–63.

Kanner, Allen D., James C. Coyne, Catherine Schaefer, and Richard S. Lazarus. 1981. "Comparison of Two Modes of Stress Measurement: Daily Hassles and Uplifts Versus

Major Life Events." *Journal of Behavioral Medicine* 4 (1): 1–39.

Kerr, Norbert L. 1998. "HARKing: Hypothesizing After the Results Are Known." *Personality and Social Psychology Review* 2 (3): 196–217.

Mehl, Matthias R., Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher, and James W. Pennebaker. 2007. "Are Women Really More Talkative Than Men?" *Science* 317 (5834): 82–82.

Nosek, Brian A., George Alter, George C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. "Promoting an Open Research Culture." *Science* 348 (6242): 1422–25.

Oaks, M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143 (2): 534.

Wilkinson, Leland. 1999. "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist* 54 (8): 594.