6

GENERAL PRINCIPLES OF HYPOTHESIS TESTING

In Chapter 1, we described an experiment by Barlett (2015) in which he attempted to investigate whether there is a difference in hostility between those who receive insulting or nice online messages by conducting an experiment in which participants received messages that were either insulting or nice and then measuring the participants' levels of hostility. We presented the results of this experiment at the beginning of Chapter 2. In this chapter, we will apply the concepts discussed in preceding chapters to describe the basic principles for testing statistical hypotheses. To make it easier to see those basic principles, we will assume for the moment that we know the population variances. We will postpone the actual analysis of Barlett's data until Chapter 7, where we will use estimates of the population variance in the application of Student's *t*-test.

As we saw in Chapter 1, we start with a research question and generate mutually exclusive and exhaustive experimental hypotheses as possible answers to our research question. Then we design a research study based on our research hypotheses and collect data. By making certain assumptions about the data, we can use a statistical model to assess whether the obtained results reflect real experimental effects or merely random (chance) factors. With the classical statistical model, this assessment is carried out by making assumptions about the shape of the populations from which the data were obtained, setting up statistical hypotheses about the parameters of these populations, and evaluating which hypothesis is best supported by the data. The results of our statistical hypothesis test are then generalized back to our experimental hypotheses to hopefully answer the question originally posed. In this chapter, we will examine the principles involved in testing statistical hypotheses with the classical statistical model, and in Chapter 9, we will do the same with the randomization/permutation model.

EXPERIMENTAL AND STATISTICAL HYPOTHESES

> *Experimental hypotheses are statements about the relationship between the independent and dependent variables in our experiment. In the classical statistical model, the parallel statistical hypotheses are about unknown parameters in the populations from which our data were sampled.*

The independent variable in Barlett's (2015) experiment was the message participants received from their "partner" after failing to solve an unsolvable Sudoku puzzle. This message was either an insulting message ("you suck") or a nice message ("it's OK"). There were a number of dependent variables; the one we discussed was the participants' scores on a scale measuring state hostility (scores could range from 34 to 170, with higher scores indicating more hostility) after they received their "partner's" message. (More details of this experiment are given in Chapter 1.)

Experimental Hypotheses

The experimental hypotheses for Barlett's experiment were:

Cybervictimization (message content) *does not* affect state hostility.
Cybervictimization (message content) *does* affect state hostility.

These hypotheses are stated in terms of the independent and dependent variables in the experiment. To analyze his data using the classical statistical model, Barlett had to construct a parallel set of statistical hypotheses about the means of the populations from which the data were presumably sampled.

Statistical Hypotheses

If we assume that the experimental participants were sampled randomly from a normal distribution[1] with a mean of $\mu_{\text{Insulting message}}$ and variance $\sigma^2_{\text{Insulting message}}$, and that the control participants were sampled from a normal distribution with a mean $\mu_{\text{Nice message}}$ and a variance $\sigma^2_{\text{Nice message}}$, then the following statistical hypotheses can be generated to correspond to the experimental hypotheses in the experiment:

$$\mu_{\text{Insulting message}} = \mu_{\text{Nice message}}$$

or

$$\mu_{\text{Insulting message}} - \mu_{\text{Nice message}} = 0$$

[1] Here we are using the principle of potential populations we discussed in Chapter 1.

$$\mu_{\text{Insulting message}} \neq \mu_{\text{Nice message}}$$

or

$$\mu_{\text{Insulting message}} - \mu_{\text{Nice message}} \neq 0$$

In words, the first statistical hypothesis reads: "The mean of the population from which the participants in the Insulting message group were sampled is equal to the mean of the population from which the participants in the Nice message group were sampled." This hypothesis can also be stated as "The mean of the population from which the participants in the Insulting message group were sampled *minus* the mean of the population from which the participants in the Nice message group were sampled equals 0." Both versions convey the same information. We will see shortly that the second version fits better with the test statistic we use to test this hypothesis.

The second hypothesis is read in a similar manner, except that the means of the two populations are stated as not being equal; in other words, the difference between them is stated as not being equal to 0. As noted in Chapter 1, with the classical statistical model, statistical hypotheses are statements about the parameters of potential populations created for the purpose of using that model.

The two experimental hypotheses are *mutually exclusive* and *exhaustive*. That is, one of them must be true, but they both cannot be true, and they are the only possibilities: Either being a victim of cyberbullying (through having received an insulting message online) does or does not affect whether someone responds by experiencing hostility. The same logic holds for the statistical hypotheses; either $\mu_{\text{Insulting message}} = \mu_{\text{Nice message}}$ or $\mu_{\text{Insulting message}} \neq \mu_{\text{Nice message}}$.

For ease of exposition and to frame the discussion in this chapter in general terms, the groups in this experiment will be referred to as E and C, denoting experimental and control groups, respectively.

ESTIMATING PARAMETERS

Our statistical hypotheses tell us what parameters we need to estimate to construct a test statistic to help us decide which hypothesis our data support.

Statistical hypotheses are statements about the parameters of the populations from which our samples were obtained. To test these hypotheses, we use the information in our samples to obtain estimates of the unknown parameters. From Chapter 4, we know that the best estimate of the mean of a population is the mean of a random sample from that population. In a two-group experiment, we are concerned with two population means (in this case, μ_E and μ_C); therefore, we must use two estimates, one from the control group ($\overline{X}_C$) and one from the experimental group ($\overline{X}_E$). Since our statistical hypotheses are about the difference between these population means, the value of interest is the difference between the sample means $\bar{X}_E - \bar{X}_C$. Assuming that the sampling is random, we expect that our estimates $\bar{X}_E$ and $\bar{X}_C$ will be close to the parameters μ_E and μ_C.[2] Therefore, if the hypothesis $\mu_E - \mu_C = 0$ is indeed true, we expect that the quantity $\bar{X}_E - \bar{X}_C$ to be close to zero. On the other hand, if the other hypothesis ($\mu_E - \mu_C \neq 0$) is true, then we expect $\bar{X}_E - \bar{X}_C$ to be a non-zero value that is somewhere close to the value of $\mu_E - \mu_C$. Since this other hypothesis says nothing about how far μ_E is from μ_C, the only thing we can say is that a large positive or negative value of $\bar{X}_E - \bar{X}_C$ is more consistent with $\mu_E - \mu_C \neq 0$ than with $\mu_E - \mu_C = 0$.

Therefore, we are faced with trying to answer the following questions: How far can $\bar{X}_E - \bar{X}_C$ be from zero before we conclude that the value is consistent with the hypothesis $\mu_E - \mu_C \neq 0$? How close can $\bar{X}_E - \bar{X}_C$ be to zero before we conclude that it is consistent with the hypothesis $\mu_E - \mu_C = 0$?

THE CRITERION FOR EVALUATING OUR STATISTICAL HYPOTHESES

> *When we test statistical hypotheses, we select one of the hypotheses and calculate the probability of obtaining (in a future research study) our results, or results even more extreme, assuming that the hypothesis is true. When that probability is below a predetermined level, we reject the hypothesis and accept the other one.*

All statistical hypothesis tests are carried out by selecting one of our mutually exclusive and exhaustive hypotheses and calculating the *probability of obtaining (in a future experiment) a similar result, or one even more extreme, assuming that the hypothesis being considered is true*. If the hypothesis we select to test is true, we expect that probability to be high. If our chosen hypothesis is false, we expect that probability to be low (and the probability calculated from the other hypothesis to be higher). To calculate the probability of obtaining our observed difference $\bar{X}_E - \bar{X}_C$ based on the statistical hypothesis we choose to test, we need to create a test statistic to determine the probability of obtaining our data, assuming the statistical hypothesis we chose to test is correct. From the distribution of that test statistic, we can find the probability of obtaining our observed data.

CREATING OUR TEST STATISTIC

> *Of interest here is the difference between the two sample means, and therefore we need to create a test statistic that allows us to easily determine the probability of*

[2] As noted in Chapter 1, we rarely expect our estimates to equal our parameters.

obtaining our observed difference between those two sample means based on one of our statistical hypotheses. In the situation under consideration, we apply the standard score transformation to $\bar{\mathrm{X}}_{E} - \bar{\mathrm{X}}_{C}$ *to create our test statistic, z.*

To calculate the probability of obtaining our observed value of $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$, we need to know the distribution of that difference. From the central limit theorem, we know that when the original populations are normal distributions, the sample means $\bar{X}_{\mathrm{E}}$ and $\bar{X}_{\mathrm{C}}$ will both have normal distributions with means μ_{E} and μ_{C} and variances $\sigma^2_{\bar{X}_{\mathrm{E}}} = \frac{\sigma^2_{\mathrm{E}}}{n_{\mathrm{E}}}$ and $\sigma^2_{\bar{X}_{\mathrm{C}}} = \frac{\sigma^2_{\mathrm{C}}}{n_{\mathrm{C}}} \sigma^2_{\bar{X}_C} = \frac{\sigma^2_C}{n_C}$, respectively. Theorem 4 in Chapter 3 can be extended to the present situation in which the numbers under consideration are sample means. Therefore, according to Theorem 4, the distribution of $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$ has a mean $\mu_{\mathrm{E}} - \mu_{\mathrm{C}}$ and a variance $\sigma^2_{\bar{X}_{\mathrm{E}}} + \sigma^2_{\bar{X}_{\mathrm{C}}}$, and when the original populations are normal distributions, the distribution of $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$ is also a normal distribution.[3]

Knowing the mean and variance of this normal distribution, we can calculate the probability of obtaining values of $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$ that occur in any particular region of the distribution by using the normal distribution table in Appendix E. However, to use that table, we must transform the raw scores to a standard score distribution:

$$\text{standard score} = \frac{\text{score} - \text{mean}}{\text{standard deviation}} \tag{6.1}$$

Using Equation 6.1 and substituting $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$ for the score, $\mu_{\mathrm{E}} - \mu_{\mathrm{C}}$ for the mean, and $\sqrt{\sigma^2_{\bar{X}_{\mathrm{E}}} + \sigma^2_{\bar{X}_{\mathrm{C}}}}$ for the standard deviation, we get the following formula:

$$z = \frac{\left(\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}\right) - \left(\mu_{\mathrm{E}} - \mu_{\mathrm{C}}\right)}{\sqrt{\sigma^2_{\bar{X}_{\mathrm{E}}} + \sigma^2_{\bar{X}_{\mathrm{C}}}}} \tag{6.2}$$

The final step is to select a hypothesis to test. In the present situation, only one of our two statistical hypotheses allows us to calculate the probability of obtaining a particular result, and that is the hypothesis $\mu_{\mathrm{E}} - \mu_{\mathrm{C}} = 0$. The other hypothesis, $\mu_{\mathrm{E}} - \mu_{\mathrm{C}} \neq 0$, cannot be used for direct calculation because it does not give a value for the difference between the population means. If that hypothesis were stated with a definite value (such as $\mu_{\mathrm{E}} - \mu_{\mathrm{C}} = 10$), then calculating the probability of that result would be possible.

[3] The distribution of $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$ is a theoretical distribution that could be generated if we perform our research study an infinite number of times. Each time we perform the experiment, we can obtain a value for $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$. Because the $\bar{X}$'s are estimates of parameters, neither they, nor the values of $\bar{X}_{\mathrm{E}} - \bar{X}_{\mathrm{C}}$, will be the same from study to study. After an infinite number of these replications of the study, we will have a distribution that is normal in shape with a mean $\mu_{\mathrm{E}} - \mu_{\mathrm{C}}$ and a variance $\sigma^2_{\bar{X}_{\mathrm{E}}} + \sigma^2_{\bar{X}_{\mathrm{C}}}$.

The hypothesis that allows us to calculate this probability is called the **null hypothesis** and in this case is written as H_0: $\mu_E = \mu_C$ or H_0: $\mu_E - \mu_C = 0$. *The other hypothesis is called the* **alternative hypothesis** and in this case is written as H_1: $\mu_E \neq \mu_C$ or H_1: $\mu_E - \mu_C \neq 0$.

By substituting $\mu_E - \mu_C = 0$ into Equation 6.2, we arrive at our **test statistic**:

$$z = \frac{\left(\bar{X}_E - \bar{X}_C\right)}{\sqrt{\sigma^2_{\bar{X}_E} + \sigma^2_{\bar{X}_C}}} \tag{6.3}$$

Equation 6.3 gives us the value we use to find the probability of obtaining our observed results, assuming the null hypothesis (in this case $\mu_E - \mu_C = 0$) is true.

Because the form of a distribution is not changed by the standard score transformation, the distribution of *z* will be normal when the distribution of $\bar{X}_E - \bar{X}_C$ is normal (which it is when the original populations from which the random samples were taken were normal distributions). Furthermore, when H_0 is true, $\mu_z = 0$ and $\sigma^2_z = 1$.

DRAWING CONCLUSIONS ABOUT OUR NULL HYPOTHESIS

We can divide the distribution of our test statistic into two regions. Values of our test statistic in one of these regions (called the critical region*) leads us to reject our null hypothesis. The size of the critical region represents the probability of getting a result judged unlikely to have occurred when the null hypothesis is true. The dividing line between these two regions is called the* critical value.

The *p*-Value

The probability of obtaining a particular experimental result or one more extreme, assuming that H_0 is true, is the ***p*-value**. According to Fisher (1925), we can use the *p*-value to assess the correctness of H_0. For Fisher, the *p*-value is a measure of the implausibility of the null hypothesis; that is, the lower the *p*-value, the more implausible the null hypothesis. Therefore, when the *p*-value is lower than a certain value, we should take that as evidence that the null hypothesis is false. Fisher recognized that the null hypothesis might be true even when the *p*-value is a low number. According to Fisher (1956), a low *p*-value means that "either an exceptionally rare event has occurred, or the theory of random distribution [the null hypothesis] is not true" (p. 42). Nevertheless, Fisher took the position that a low p-value should be taken as evidence that the null hypothesis is false.

How small does the p*-value have to be for us to conclude that H_0 is false?* Unfortunately, there is no straightforward answer to that question. Should that probability be .1 (1 chance in 10), or .05 (1 chance in 20), or .01 (1 chance in 100), or .001 (1 chance in 1,000)? While, as we will see later, the choice can depend in part on other considerations, Fisher proposed that we use $p < .05$ as our definition of "a rare event when H_0 is true." In other words, anytime the calculated value of p is less than or equal to .05, we reject H_0 and accept the alternative hypothesis H_1. Fisher's proposal has become the standard by which we decide whether to reject H_0. We do not reject H_0 when $p > .05$. As we will see later in this chapter, the latter case does *not* lead us to conclude that H_0 is true.

The value of z that divides those results that lead to rejection of H_0 from those that lead to non-rejection is called the **critical value**, and the set of all values of z that lead to rejection of H_0 is called the **critical region** (see Figure 6.1).

We use the symbol α to represent the probability of getting a result in the critical region when the null hypothesis is true. We can represent this situation symbolically as follows:

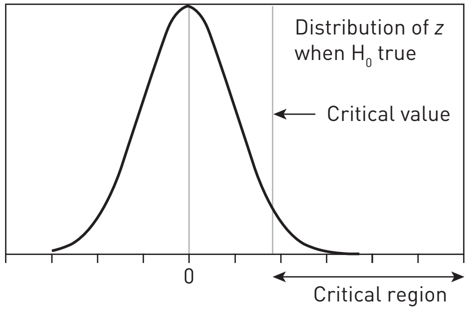
$$\alpha = \Pr(\text{test statistic in critical region}|\text{H}_0 \text{ true}) \tag{6.4a}$$

or

$$\alpha = \Pr(\text{reject H}_0 \text{ when it is true}) \tag{6.4b}$$

Therefore, any value of the test statistic for which $p < \alpha$ leads us to reject H_0.

FIGURE 6.1 ■ Location of the critical value and critical region. Any calculated z that is in the critical region leads to rejection of H_0. Calculated values of z that are outside the critical region do not lead to rejection of H_0.



BUT SUPPOSE H$_0$ IS FALSE?

> *When H_0 is false, our test statistic does not come from the distribution based on H_0 being true; it comes from another normal distribution that has a mean ≠ 0. This other distribution is called the non-central distribution. While the* p*-value for our test statistic calculated from the distribution based on H_0 being true may be low, the probability of getting our data from the non-central distribution will be higher when H_0 is false.*

The strategy for testing statistical hypotheses described above requires us to assume that the hypothesis we test (the null hypothesis) is true and to use the distribution of our test statistic (in this case z) to calculate the probability of obtaining our data. But suppose H$_0$ is false?

When H$_0$ is false, the distribution of z will also be a normal distribution with variance equal to 1, but the mean will *not* be equal to zero. The formulas for μ_z and σ_z^2 when H$_0$ is both true and false are derived in Box 6.1.

BOX 6.1

ALGEBRAIC DERIVATION OF THE MEAN AND VARIANCE OF THE *z*-TEST STATISTIC

The mean of the test statistic $z = \frac{\bar{X}_E - \bar{X}_C}{\sqrt{\frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C}}}$ is found by repeatedly drawing samples from the two populations, finding $\bar{X}_E - \bar{X}_C$ for each set of samples, converting to z using Equation 6.3, and then averaging all of the resulting z-statistics. Using the definition of the mean, we find that

$$\mu_z = \frac{\sum_{i=1}^{K} z_i}{K}$$

where K is the number of z-scores generated by this procedure. Substituting Equation 6.3 for z produces

$$\mu_z = \frac{\sum_{i=1}^{K}\left(\frac{\bar{X}_E - \bar{X}_C}{\sqrt{\frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C}}}\right)}{K}$$

Because the term under the square root sign is a constant with respect to the summation,

$$\mu_z = \frac{1}{\sqrt{\frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C}}} \cdot \left[\frac{\sum_{i=1}^{K}\left(\bar{X}_E - \bar{X}_C\right)}{K}\right]$$

We can distribute the summation sign to get the following:

$$\mu_z = \frac{1}{\sqrt{\frac{\sigma_E^2}{n_E}+\frac{\sigma_C^2}{n_C}}} \bullet \left[\frac{\sum_{i=1}^{K}\bar{X}_E}{K} - \frac{\sum_{i=1}^{K}\bar{X}_C}{K}\right]$$

The values in the brackets represent $\mu_E - \mu_C$. Therefore,

$$\mu_z = \frac{\mu_E - \mu_C}{\sqrt{\frac{\sigma_E^2}{n_E}+\frac{\sigma_C^2}{n_C}}}$$

When H_0 is true, $\mu_E - \mu_C = 0$, and $\mu_z = 0$. When H_0 is false, the value of μ_z depends on the actual difference between μ_E and μ_C, the population variances, and the sample sizes.

The variance of the *z*-test statistic can be found in a similar manner. Start with the definition of the variance as the mean of the squared deviations of scores from their means:

$$\sigma_z^2 = \frac{\sum_{i=1}^{K}\left(z_i - \mu_z\right)^2}{K}$$

We can substitute $z = \frac{\bar{X}_E - \bar{X}_C}{\sigma_{\bar{X}_E - \bar{X}_C}}$ and $\mu_z = \frac{\mu_E - \mu_C}{\sigma_{\bar{X}_E - \bar{X}_C}}$ to get this:

$$\sigma_z^2 = \frac{\sum_{i=1}^{K}\left[\left(\frac{\bar{X}_E - \bar{X}_C}{\sigma_{\bar{X}_E - \bar{X}_C}}\right) - \left(\frac{\mu_E - \mu_C}{\sigma_{\bar{X}_E - \bar{X}_C}}\right)\right]^2}{K}$$

Because the term $\sigma^2_{\bar{X}_E - \bar{X}_C}$ occurs in both parts and is a constant with respect to the summation,

$$\sigma_z^2 = \frac{1}{\sigma^2_{\bar{X}_E - \bar{X}_C}} \bullet \frac{\sum_{i=1}^{K}\left[\left(\bar{X}_E - \bar{X}_C\right) - \left(\mu_E - \mu_C\right)\right]^2}{K}$$

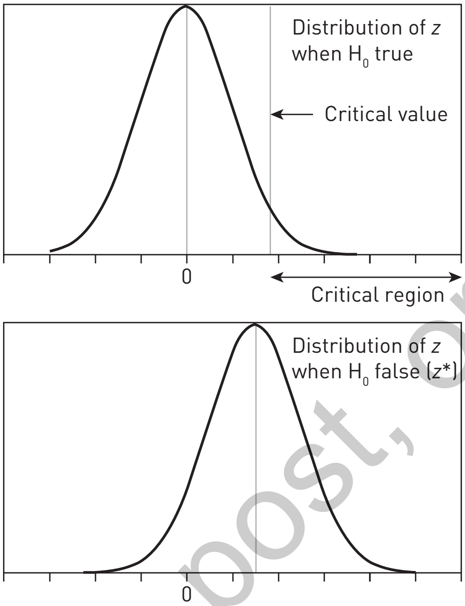
(*Note:* $\sigma^2_{\bar{X}_E - \bar{X}_C}$ is squared because it comes from a squared expression.)

The summation term is, by definition of variance, $\sigma^2_{\bar{X}_E - \bar{X}_C}$. Therefore, $\sigma_z^2 = 1$ both when H_0 is true and when H_0 is false.

The distributions of our test statistic when H_0 is true and when H_1 is true are represented in Figure 6.2. The two curves are presented separately because they are two *mutually exclusive possibilities*; that is, they cannot both happen at the same time: Either H_0 is true, in which case our calculated *z* (Equation 6.3) comes from the top distribution, or H_1 is true, in which case our calculated *z* (Equation 6.3) comes from the bottom distribution.[4]

[4] Although our alternative hypothesis is two tailed (that is, $\mu_E - \mu_C \neq 0$), this difference can only be on one side when the null hypothesis is false: Either $\mu_E - \mu_C > 0$ or $\mu_E - \mu_C < 0$. In the first situation, *z* is greater than 0, and in the second situation, *z* is less than zero. For the discussion here, we will assume the true state of affairs is that $\mu_E - \mu_C > 0$ and $z > 0$.

FIGURE 6.2 ■ Location of the critical value and critical region. Any calculated *z* that is in the critical region leads to the rejection of H_0. Calculated values of *z* that are outside the critical region do not lead to the rejection of H_0.



We decide whether to reject or not reject H_0 by calculating the probability of obtaining our observed value of *z* from the H_0 (top) distribution. When H_0 is true, the probability of getting a result in the critical region is low. On the other hand, when H_0 is false and H_1 is true, our test statistic comes from the bottom distribution, and our getting a result in the critical region is more likely to occur. The bottom distribution in Figure 6.2 is called the **non-central *z* distribution** and is symbolized as z^* (*z*-star).

ERRORS IN HYPOTHESIS TESTING

There are two types of errors we might make when testing statistical hypotheses. One error is to reject H_0 when it is true, which is called a type I or alpha error. The other error is to not reject H_0 when it is false, which is called a type II or beta error. We can only make a type I error when H_0 is true and our test statistic is in the critical region. We can only make a type II error when H_0 is false and our test statistic is not in the critical region.

Although we reject H_0 when the *p*-value is lower than our predetermined definition of an improbable event, there is the possibility that H_0 is indeed true. In this case, our rejection of H_0 leads to an error.[5] Following Neyman and Pearson (1928a, 1928b), this error is called a **type I** or **alpha error** and is represented by the part of the distribution based on H_0 being true that is in the critical region. *We can only make a type I error (or an alpha error) when H_0 is true and we obtain a result that falls in the critical region. The error is that we will decide to reject H_0 when it is indeed true.* The probability that we might make a type I or alpha error is α (see Equations 6.4a and 6.4b).

There is another kind of error we can make, namely, when H_0 is false and our data do not lead us to reject it; that error will occur when our test statistic does *not* fall in the critical region. We use the symbol β to represent *the probability of getting a result that is not in the critical region when the null hypothesis is false*. We can represent this situation symbolically as follows:

$$\beta = \Pr(\text{test statistic is not in the critical region}|\text{H}_0 \text{ false}) \tag{6.5a}$$

or

$$\beta = \Pr(\text{do not reject H}_0 \text{ when it is false}) \tag{6.5b}$$

This probability is represented by the portion of the z^* distribution that is *outside* the critical region. This error is called a **type II** or **beta error**. Since the probability of making a type II error (or beta error) is dependent on how much of the z^* curve is outside the critical region, which in turn is dependent on the unknown mean of z^*. It is not possible to calculate the probability of making a type II error. Nevertheless, there are ways to minimize this error, and these will be discussed in the next section.

When H_0 is false and the calculated value of z is in the critical region, then H_0 will be correctly rejected. This is not an error; it is a correct decision. The probability that one will correctly reject H_0 when it is false is represented by the area of the z^* curve that is in the critical region. *The probability of correctly rejecting H_0 when it is false* is called **power**; that is,

$$\text{Power} = \Pr(\text{test statistic in critical region}|\text{H}_0 \text{ false}) \tag{6.6a}$$

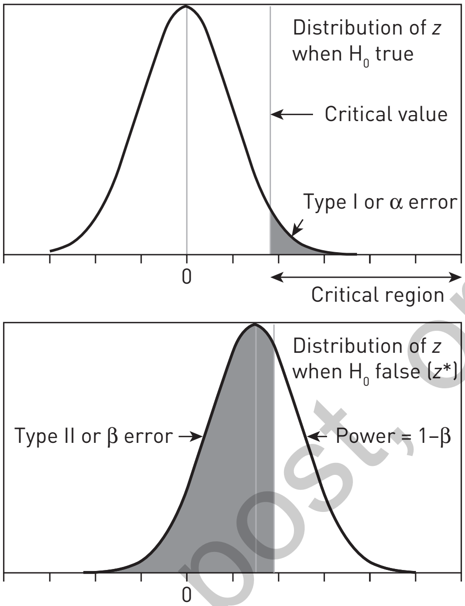
or

$$\text{Power} = \Pr(\text{reject H}_0 \text{ when it is false}) \tag{6.6b}$$

Since power and beta (probability of a beta or type II error) represent those parts of the z^* distribution that are inside and outside the critical region, respectively, they are related by the

[5] It is easy to make the mistake of stating that "there is such and such a probability that H_0 is true or not true." H_0 is either true or not, there is no probability attached to the truth or falseness of H_0; the only probability here is associated with obtaining a given result (in a future experiment) assuming H_0 is true.

FIGURE 6.3 ■ Relationship between the critical region, the probability of making a type I error (alpha), the probability of making a type II error (beta), and power (the probability of rejecting H_0 when it is false).



formula, power = 1 − beta. Therefore, those things that increase power also decrease beta, and vice versa. The relationships described here are represented in Figure 6.3.

Finally, when H_0 is true and the calculated z does not fall in the critical region, one has no reason to reject H_0. Does this outcome mean that H_0 can be accepted? The answer is no, since logically one cannot prove a hypothesis true by finding evidence consistent with it. It may be that the hypothesis is actually false, but the data are still technically consistent with that hypothesis (or just not inconsistent enough with the hypothesis to warrant its rejection). We will discuss the reasons why one cannot prove H_0 to be true and what one can do with results that do not lead to the rejection of H_0 (see the section in this chapter titled "What Should We Do (or Not Do) When Our Data Do Not Allow Us to Reject the Null Hypothesis?").

POWER AND POWER FUNCTIONS

If our experimental question involves trying to determine whether two treatments have different effects on behavior, then we want to design an experiment in which the probability of rejecting a false null hypothesis (the power of our test) is high. Although we *cannot calculate the power of our statistical hypothesis test, we know what factors affect the power, and we can use that information to design our experiment to maximize its power. Power functions provide us with a way to decide how large a sample size to use when we design our experiment.*

Power is defined as the probability of rejecting H_0 when it is false. It is represented in Figure 6.3 as that area of the z^* distribution that lies in the critical region. To determine the power of a hypothesis test, we would have to know the form and location of the z^* distribution. Because z^* is a normal distribution with variance = 1 (see Box 6.1), the location of z^* is determined by the mean of the distribution, which is shown in Box 6.1 to be

$$\mu_{z^*} = \frac{\mu_E - \mu_C}{\sqrt{\frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C}}} \tag{6.7}$$

From Equation 6.7, it can be seen that the location of the z^* distribution (and hence the power of the test) is dependent on

1. the difference between the population means ($\mu_E - \mu_C$),
2. the population variances (σ_E^2 and σ_C^2), and
3. the sample sizes (n_E and n_C).

In addition, power will depend on

4. the critical value for the test that is based on the value of α we select as defining an improbable result when H_0 is true.

Of these four things, all but the value $\mu_E - \mu_C$ are known for any given experiment for which the z-test statistic can be used. Therefore, while we do not know what μ_{z^*} is, we certainly know how to change its value and to influence the power of the test.

The Population Variances

We tend to think of the population variances as fixed values not under our control, but that belief is not entirely true. The scores under consideration are measurements of behavior. From classical measurement theory, we know that an observed score is made up of two components, a true score and an error component. Although we cannot directly measure the true score, we can estimate it as the expected value of an infinite number of the measurements. The error component is assumed to be randomly distributed around the true score so that the mean of the errors = 0. The observed score can thus be represented by the following formula:

$$Y = T + Error, \tag{6.8}$$

where $\mu_{Error} = 0$.

From Theorem 3 in Chapter 3, the mean of a sum is the sum of the means, and the variance of a sum is the sum of the variances. Therefore,

$$\mu_Y = \mu_T + \mu_{Error} = \mu_T \tag{6.9}$$

$$\sigma_Y^2 = \sigma_T^2 + \sigma_{Error}^2 \tag{6.10}$$

The variance of the true scores is a function of the variation of the individuals being measured. Homogeneous groups (such as when all members of the group share the same gender, same age, same background, and so forth) will have lower values for σ_T^2 . Heterogeneous groups will have larger values. For this reason, we try to use homogeneous groups in our experiments. We will see in Chapter 14 how we can apply analysis of variance to the possible sources of variance mentioned above to reduce the variances of the observed scores.

The variance of the errors is affected by the reliability of our measuring instrument and how consistently the participants in a given group are treated. *Reliability* refers to getting the same value each time we take the measurement. A perfectly reliable measuring instrument yields the same value every time we use it (assuming the true score does not change over time). An unreliable measuring instrument yields different values each time the measurement is taken. An extreme example of using an unreliable measuring instrument is using a rubber band to measure the length of an object. Therefore, it is important to choose measuring instruments that have high reliability.

The other factor that affects σ_{Error}^2 involves whether the conditions under which participants are treated and tested are held constant. Sloppy experimental procedures (not administering the treatments or instructions to our participants in the same way, or otherwise handling the participants in different ways that are not directly connected to the manipulation of the independent variables in our experiment) increase σ_{Error}^2 , and thus in turn decrease the power of our statistical test. To improve the power of our statistical test, all experiments should be conducted in a way that ensures all participants in a given condition are treated the same.

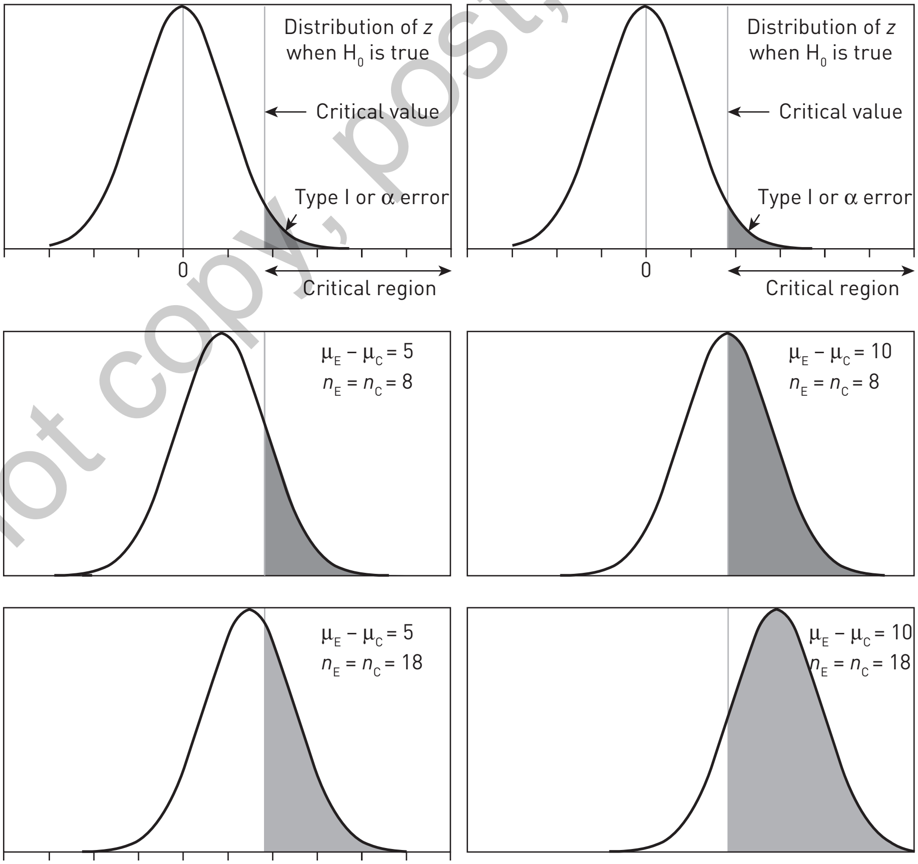
It is clear from Equation 6.7 that anything we can do to *decrease* σ_E^2 and σ_C^2 will *increase* the μ_{z^*} and therefore move the non-central distribution to the right. Doing so will increase the power of our statistical test, which uses the *z* in Equation 6.3 as the test statistic.

The Sample Sizes

It is clear in Equation 6.7 that μ_{z^*} is also a function of the sample size. The larger the sample size (for any given difference $\mu_E - \mu_C$ and any given values of σ_E^2 and σ_C^2), the larger the value of μ_{z^*} and the greater the power of our *z*-test. This relationship is illustrated in Figure 6.4. Power is represented by the shaded area of the distributions on which H_0 is assumed to be false.

Clearly, as sample size increases, more and more of the distribution of the test statistic when H_0 is false falls in the critical region of the distribution based on H_0 being true, thus increasing the power. In this case, because the form and variance of the H_1 distribution do not change, the location of that distribution must shift further into the critical region. This concept is illustrated in both sides of Figure 6.4. The difference between the two sides of Figure 6.4 reflects the size of the difference between μ_E and μ_C. In these examples, the difference between μ_E and μ_C on the right is twice as large as that on the left; hence, the power of the test at any given sample size is greater for detecting the larger difference. It should be noted, however, that if the sample sizes were large enough, the

FIGURE 6.4 ■ In the top panels, there are distributions of the test statistic (Equation 6.3) when H_0 is true, and in the middle and bottom panels, when the null hypothesis is false. In the middle panels, $n_E = n_C = 8$, and in the bottom panels, $n_E = n_C = 18$. On the left, $\mu_E - \mu_C = 5$, and on the right, $\mu_E - \mu_C = 10$. In all cases, $\alpha = .05$ and $\sigma_E^2 = \sigma_C^2 = 10$.



H_1 distribution would shift far enough into the critical region to allow detection of the difference between μ_E and μ_C, even if that difference were small. Therefore, any test can be made powerful enough to detect the smallest differences by increasing the sample size. We will have much more to say about this in the section "The Use of Power Functions" later in this chapter.

The Value of α Selected for Our Test

Finally, if the critical value were shifted to the right by decreasing the probability of making a type I error from $\alpha = .05$, where the critical value is 1.64, to $\alpha = .01$, where the critical value increases to 2.33, the power of the statistical test would decrease because less of the H_1 distribution would be in the critical region.

Power Functions

A **power function** is a *functional relationship between the power of a test and various possible values of the alternative hypothesis for a given sample size*. It is conventional to express the alternative hypothesis $\mu_E - \mu_C \neq 0$ in terms of the number of standard deviations μ_E is from μ_C. This index is the **standardized effect size**, where

$$ES = \frac{\mu_E - \mu_C}{\sigma} \tag{6.11}$$

The horizontal axis of the power function graph is in units of the standardized effect size $\left(\frac{\mu_E - \mu_C}{\sigma}\right)$, and the power is on the vertical axis.

To construct a power function, we must find the area of the alternative distribution curve (in this case, z^*) that lies in the critical region for various possible values of $\mu_E - \mu_C$. Thus, if we adopt a certain critical value (based on our choice of alpha) and particular sample sizes (n_E and n_C), we can examine the probability that our test would reject H_0 for different values of $\mu_E - \mu_C$. If we chose a different sample size and again looked at different values of $\mu_E - \mu_C$, we would construct another curve that would reflect the power at that sample size. In fact, if we chose many different values of n, we could construct a *family* of power functions. Similarly, we could hold sample size constant and vary alpha to construct another family of power functions that reflect the effects of that factor. Examples of these power functions are given below.

In the case of the z-test statistic under consideration, the power of the test can easily be found because we know that the distribution of z^* is normal with a known mean (see Equation 6.8) and variance ($\sigma_z^2 = 1$). Therefore, we can find the area of this normal curve that lies in the critical region by using the standard score transformation so that we can use the unit normal curve tables. The method for finding the power for the two-sample z-test is given in Box 6.2. The derivation in Box 6.2 is provided to illustrate how various factors affect the power of a test. All of these factors have similar effects in

BOX 6.2

FORMULA FOR THE POWER FUNCTION FOR THE TEST OF THE NULL HYPOTHESIS

H_0: $\mu_E - \mu_C = 0$ with the *z*-test statistic.

The power of a test is defined as the probability of rejecting H_0 when it is false. In this case, it is represented by the area of the *z** distribution that lies in the critical region (see Figure 6.4). Because *z** is a normal distribution, we can find the power by converting to standard scores again and finding the area of the curve from the critical value to $+\infty$. To do this, we substitute into the following formula:

$$\text{standard score} = \frac{\text{score} - \text{mean}}{\text{standard deviation}}$$

$$z' = \frac{z - \mu_{z^*}}{\sigma_{z^*}}$$

We will designate this standard score as *z′* (*z*-prime) to differentiate it from the other *z*-values used in this section. (The symbol *z* represents the test statistic in Equation 6.3 and the distribution of Equation 6.3 when H_0 is true; *z** represents the distribution of the *z*-test statistic when H_0 is false.) From Box 6.1 we know that

$$\mu_z = \frac{\mu_E - \mu_C}{\sqrt{\frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C}}}$$

and $\sigma_z^2 = 1$.

This deviation can be simplified if the following assumptions are made:

1. $\sigma_E^2 = \sigma_C^2$. (The two population variances are equal; this is called homogeneity of variance.)

2. $n_E = n_C$. (The sample sizes are equal.)

If we drop the subscripts on σ^2 and *n* to denote that they are equal,

$$\mu_{z^*} = \frac{\mu_E - \mu_C}{\sqrt{\frac{2\sigma^2}{n}}} = \frac{\mu_E - \mu_C}{\sigma} \bullet \sqrt{\frac{n}{2}}$$

Because we want to find the area to the right of the critical value (*c.v.*),

$$z' = \frac{c.v. - \mu_{z^*}}{\sigma_{z^*}}$$

$$z' = \frac{c.v. - \frac{\mu_E - \mu_C}{\sigma} \bullet \sqrt{\frac{n}{2}}}{1}$$

$$z' = c.v. - \frac{\mu_E - \mu_C}{\sigma} \bullet \sqrt{\frac{n}{2}}$$

The value of *z´* determines the area of the *z** curve that lies in the critical region. The smaller the value of *z´*, the greater the area of *z** in the critical region and the greater the power. Note that when *z´* is a negative number, the power is greater than .5.

all hypothesis tests. This derivation applies only to situations in which the distribution where H_0 is false is normal. When that distribution is not normal, other methods must be used. Some of these strategies will be described in subsequent chapters.

From the formula derived in Box 6.2, we can readily construct a set of power functions for any *z*-test of the hypothesis H_0: $\mu_E - \mu_C = 0$. Two examples of such functions are given below. In the first example, the probability of a type I (α) error is set at .05 and sample size is varied (see Table 6.1 and Figure 6.5). In the second example, sample size is held constant at 8 participants in a group and the value of α is varied (see Table 6.2 and Figure 6.6).

TABLE 6.1 ■ Power for the Test of H_0: $\mu_E - \mu_C = 0$ Against H_1: $\mu_E - \mu_C > 0$ With a *z*-Statistic for Various Standardized Effect Sizes. $\alpha = .05$, and Sample Size (*n*) Is Varied

n	**Standardized Effect Size $\frac{\mu_E-\mu_C}{\sigma}$**						
	0	0.5	1.0	1.5	2.0	2.5	3.0
2	.05	.13	.26	.44	.63	.80	.91
8	.05	.26	.63	.91	.99	.99	.99
18	.05	.44	.91	.99	.99	.99	.99

FIGURE 6.5 ■ Power functions for the *z*-test of H_0: $\mu_E - \mu_C = 0$ against H_1: $\mu_E - \mu_C > 0$ with $\alpha = .05$.

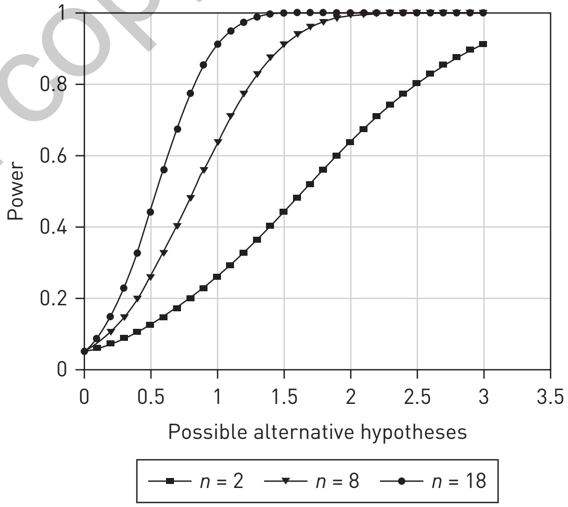
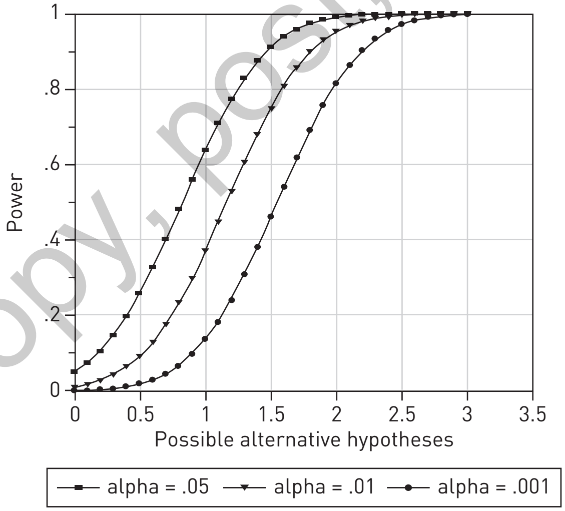


TABLE 6.2 ■ Power for the Test of H_0: $\mu_E - \mu_C = 0$ Against H_1: $\mu_E - \mu_C > 0$ With a *z*-Statistic for Various Standardized Effect Sizes (Sample Size = 8; α Is Varied)

α	Standardized Effect Size $\frac{\mu_E - \mu_C}{\sigma}$						
	0	0.5	1.0	1.5	2.0	2.5	3.0
.05	.05	.26	.63	.91	.99	.99	.99
.01	.01	.09	.37	.75	.95	.99	.99
.001	.001	.02	.14	.46	.82	.97	.99

FIGURE 6.6 ■ Power functions for the *z*-test of H_0: $\mu_E - \mu_C = 0$ against H_1: $\mu_E - \mu_C > 0$ with sample size = 8.



THE USE OF POWER FUNCTIONS

When used properly, power functions can help us decide how large a sample size to use to find a treatment effect of a certain size at a given level of α*. Too large a sample size can be wasteful of resources, while too small a sample size will not result in a test with enough power to consistently detect a real difference between the population means.*

If we are going to commit time and resources to an experiment, we want to do so in a way that maximizes our chances of finding an effect that we predicted. In view of the proceeding discussion about power, it is apparent that the sample size one chooses has a great deal to do with this.

Too small a sample size will lead to an experiment in which the power of the test (probability of rejecting H_0 when it is false) is also too small. For example, if the effect size $\frac{\mu_E - \mu_C}{\sigma} = 1$, then an experiment with a sample size of 8 in a group will have a power of .63 when alpha = .05 (see Table 6.1 and Figure 6.5). Thus, if the experiment were repeated a large number of times, over *one third* of these replications would not lead us to reject H_0 even though μ_E and μ_C are one standard deviation apart (a rather large difference). If, however, the sample size is increased to 18 in a group, the power increases to .91, which means that less than 1 in 10 replications would *not* lead us to reject H_0. In this case, the risk of making a type II error (not rejecting H_0 when it is false) is at an acceptably small level.

On the other hand, too large a sample size can be wasteful of resources. For example, if the sample size above were increased from 18 per group to 25 (that is, 50 participants in all), the power would increase from .91 to .97, a gain in power that, while useful, might not be worth the added expense of the additional participants. In general, when the power of a test is around .9, adding more participants does little to increase it because there is so little room for further increase.

Clearly, the choice of sample size is an important matter. Too small a sample can lead to a test that possesses little power and consequently will most likely not lead to rejection of H_0 even when it is false (type II error). This situation is most apt to occur when the difference between μ_E and μ_C is small (see Figures 5.5 and 5.6 where $\frac{\mu_E - \mu_C}{\sigma}$ is a small number). Too large a sample size, on the other hand, creates a different problem. Note that in Figure 6.5, where $\frac{\mu_E - \mu_C}{\sigma}$ is a large number, the power is very high and almost the same for a wide range of sample sizes. Thus, the smaller sample sizes lead to tests as powerful as the larger ones.

The major problem with using power functions to aid in the determination of sample size for an experiment is that one needs to have some estimate of the size of the difference between the population means. Because this unknown value is what the experimenter is attempting to find, power functions can only be useful if some thought is given to how small a difference one wants to look for. For example, would a difference of 0.1 standard deviation $\left(\frac{\mu_E - \mu_C}{\sigma} = 0.1\right)$ be worth looking for? If so, it would be possible to choose a sample size large enough to have a test with power = .9 when that is the difference between μ_E and μ_C. In this situation, one would need 1,717 participants in a group. (As an exercise, you should try to verify this value.) Whether such a difference is worth looking for can only be decided in the context of a particular research area. Such a choice is not a statistical decision. The statistical model merely tells us what sample size to use to obtain a test with a certain power when the actual difference between μ_E and μ_C is of a particular magnitude. After we decide how large a difference between μ_E and μ_C to look for, the power function can be consulted to determine the smallest sample size necessary to achieve a test of a certain power. Fortunately, there are power functions available for all of the test statistics used in the social sciences that researchers can consult.

p-VALUES, α, AND ALPHA (TYPE I) ERRORS: WHAT THEY DO AND DO NOT MEAN

> *The p-value, α, and a type I (α error) are not the same thing. Unfortunately, many people inappropriately use these terms interchangeably. The problem is that the* p*-value was introduced by R. A. Fisher in his hypothesis-testing paradigm. Alpha and alpha error were introduced by Jerzy Neyman and Egon Pearson to test statistical hypotheses. Although we use concepts from both models, it is important for us to understand where they come from and what they mean and do not mean.*

The model for testing statistical hypotheses presented here is a combination of two competing and perhaps contradictory paradigms, one by R. A. Fisher (1925, 1935) and the other by Jerzy Neyman and Egon Pearson (1928a, 1928b). Fisher's paradigm is called *significance testing*, and Neyman and Pearson's paradigm is called *hypothesis testing*. The concepts of significance, null hypothesis, and *p*-values come from Fisher, and the concepts of alternative hypotheses, α and β errors, and power come from Neyman and Pearson.

With Fisher's significance-testing paradigm, we propose a null hypothesis that the sample comes from a specific infinite population, and the *p*-value is a measure of how unlikely it is that our data came from that population. Fisher adopted $p < .05$ as his criterion for an unlikely event when the null hypothesis is true. He called his paradigm a significance test and argued that, although obtaining a result where $p < .05$ could mean we have observed an unlikely event, such rare events can be taken as evidence that the null hypothesis is false. For Fisher, the smaller the observed *p*-value, the stronger the evidence against the null hypothesis. Fisher said nothing about alternative hypotheses, α, β, or power.

On the other hand, with Neyman and Pearson's hypothesis-testing paradigm, we set up two competing hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_1), and test H_0 against H_1. These researchers introduced the probabilities of committing two types of errors: type I or alpha errors (rejecting true null hypotheses) and type II or beta errors (retaining false null hypotheses). Rather than conceptualizing an experiment as sampling from an infinite population, they built their paradigm on repeated random sampling from a population. Therefore, α is the probability of getting results that fall in the critical region when H_0 is true if we repeat the same experiment a large number of times. The same is the case for β with respect to getting results that do not fall in the critical region when the H_0 is false. Neyman and Pearson also introduced the concept of the power of the test, or the probability of rejecting a false null hypothesis. Because the power of a test depends of a number of factors, such as sample size and effect size, they argued that we can improve our experiments by considering what the power of our test might be to detect a particular effect size in the population.

Because the concepts we use to describe statistical hypothesis testing come from these two competing paradigms, there is a lot of confusion about the proper understanding of *p*-values, α, β, and power (Hubbard, 2004; Hubbard & Bayarri, 2003; Huberty, 1993).

Differentiating *p*-Values From α Levels

With the Neyman–Pearson paradigm, we select an α-level (the probability of getting results in the critical region and rejecting H_0) prior to conducting the experiment. Therefore, α is a fixed value. On the other hand, the *p*-value is not fixed; instead, it is a value that varies depending on our data. It is the probability of getting the value of our test statistic (or one more extreme), assuming the null hypothesis is true. As noted in this chapter, when the null hypothesis is false, we can make the *p*-value as small as we want by increasing the sample size. The *p*-value is *not* the probability of making, or having made, a type I error. We connect these two concepts by saying that we will reject the null hypothesis when the *p*-value is less than α (the value we selected to define the critical region), a statement that neither Fisher nor Neyman and Pearson would have made.

Can we calculate the probability that we might make a type I or α error? *The answer to this question is no!* What we *can* determine is the probability that we might make a type I or α error when the null hypothesis is true. That probability is α, the probability of getting a result in the critical region when the null hypothesis is true (see Equation 6.5). Alpha is *not* the probability that a result in the critical region is a type I or α error. It is *not* the probability that we *might* make a type I or α error, because we can only make a type I or α error when the null hypothesis is true *and* our test statistic is in the critical region. As Pollard and Richardson (1987) noted, the probability that the null hypothesis is true and our test statistic is in the critical region equals the probability that the null hypothesis is true times the probability of getting a test statistic in the critical region when the null hypothesis is true.[6] Using the symbol H_0 for the null hypothesis is true and D^* for data in the critical region, Pollard and Richardson (1987) expressed this idea as follows:

$$\Pr(H_0 \text{ and } D^*) = \Pr(H_0) \times \Pr(D^*|H_0) \tag{6.12}$$

$\Pr(D^*|H_0)$ is α, and $\Pr(H_0)$ is an unknown value between 0 and 1 because we have no idea of how often the null hypothesis is true across all possible experiments that can be performed; therefore, the probability that we might make a type I or α error is less than α, but we have no way of calculating that value.

Can we calculate the probability that we did make a type I or α error when we rejected the null hypothesis? *The answer to this question is also no!* As Pollard and Richardson (1987) noted, "When the null hypothesis is rejected, the probability of *having made* a Type I error is a probability about the null hypothesis because a Type I error has been made if and only if the null hypothesis is in fact true" (p. 160; emphasis added). This question is really about estimating the proportion of experiments in which the null hypothesis is true *and* the test statistic is in the critical region, that is, $\Pr(H_0|D^*)$. The formula for calculating a posterior (after-the-fact) probability is Bayes' theorem:

$$\Pr\left(\mu_0|D^*\right) = \frac{\Pr\left(D^*|\mu_0\right) \times \Pr\left(\mu_0\right)}{\Pr\left(D^*\right)} \tag{6.13}$$

It is clear from Equation 6.13 that to calculate the probability that the null hypothesis is true when our test statistic falls in the critical region, we need to know the probability that the null hypothesis is true (H_0) and the probability of getting data that fall in the critical region in any experiment. Both of these values are unknown; therefore, there is no way to know after the fact whether our result that fell in the critical region occurred because the null hypothesis is false or because we made a type I or α error.

A WORD OF CAUTION ABOUT ATTEMPTING TO ESTIMATE THE POWER OF A HYPOTHESIS TEST AFTER THE DATA HAVE BEEN COLLECTED

Power is the probability of correctly rejecting the null hypothesis when that hypothesis is false. Because probability refers to future events, we use power functions to help us

[6] This logic is based on the addition law of probability: $\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B|A)$.

> *design our experiments so that when the null hypothesis is false, the probability that we will reject it will be high. Therefore, trying to calculate the power of a statistical test from the sample data gives us no more information than we already have.*

The words of caution are simple: Do not do this! As noted above, the power of a hypothesis test is the probability of rejecting the null hypothesis when it is false. Although we cannot know the power of our test because we do not know the location of the distribution of our test statistic (the non-central distribution) when the null hypothesis is false, we do know the factors that affect the power of our test (the true difference between the population means; the sample size; the value for α we select; whether we are performing a one- or two-tailed test; and, in the case of hypothesis tests on population means, the population variance). We demonstrated that we can use power functions to select the sample sizes for an experiment when we are attempting to look for a given effect size (or one larger). *We estimate the power of a hypothesis test before we perform the experiment, not afterward*. See Lenth (2001) for some practical advice for selecting the appropriate sample sizes.

The issue of attempting to estimate the power of our statistical hypothesis test arises when the results of an experiment have been analyzed and the test statistic did not fall in the critical region. The failure to reject the null hypothesis could have been because of low power (due to a variety of factors noted above) even though there actually is an effect. On the other hand, if we could show that the power is indeed high, this information might be taken as evidence that the null hypothesis is true. A number of statistical software packages provide retrospective or post hoc (after-the-fact) power analyses based on the observed data. These analyses are based on the premise that the observed differences between sample means and the observed estimate of variance are *perfect* estimates of the parameters of the populations. However, as Hoenig and Heisey (1991) and Lenth (2007) observed, this type of analysis is doomed to failure. The reason is simple: The observed power (that is, the power calculated from the observed data) is a monotonic function of the *p*-value for that experiment; that is, the larger the *p*-value, the lower the observed power, and vice versa. Therefore, when $p > .05$, power will always be less than .5. In fact, .5 is the maximum value of the observed power when the null hypothesis is not rejected.

One possible use of a retrospective or post hoc power analysis is to try to determine what sample size we would need to reject the null hypothesis for the observed differences between sample means and the observed estimate of variance, assuming they are perfect estimates of the population effect size and variance, an assumption that is almost always false. While we could calculate how large a sample size we would need to find that effect size *in a future experiment,* it is *not appropriate* to use that information to add more participants to our sample. In fact, using the classical statistical model, it is *never appropriate* to keep collecting data until we have enough data to reject the null hypothesis!

IS IT EVER APPROPRIATE TO USE A ONE-TAILED HYPOTHESIS TEST?

> *One-tailed tests are always more powerful than two-tailed tests when the difference between the sample means is in the predicted direction, but what will you do when there is a large difference in the opposite direction? Changing your theory and statistical hypothesis to accommodate that situation increases the probability of making a type I error. To discourage that practice, journal editors often advise authors to not use one-tailed tests. However, there are situations in which one-tailed tests are appropriate.*

The answer to this question is "it depends." One-tailed hypothesis tests are designed to test a directional hypothesis, for instance, that the mean of Group 1 is greater than the mean of Group 2. In such a situation, the null hypothesis would only be rejected if the mean for Group 1 were larger than the mean of Group 2 and large enough to result in the test statistic falling in the critical region. In this situation, the critical region would reside on only one side of the distribution such that, conventionally, 5% of the distribution on one side, but 0% of the distribution on the other side, would comprise the critical region. This circumstance would provide the most powerful test of the directional hypothesis, which was likely based on good theoretical foundations.

Such an outcome seems to be the optimal situation, with researchers being rewarded for their good theoretical foresight in predicting the direction of the mean difference with increased power to reject the null hypothesis in detecting it. However, researchers are often unsuccessful in their predictions (as researchers will begrudgingly admit). Sometimes the effect manifests differently than extant theory would suggest, such as when the mean of Group 1 is substantially less than the mean of Group 2 in our example above. Were the researchers using a one-tailed test, they would have no opportunity (that is, no power) to reject the null hypothesis in this case because there would be no critical region on that side of the distribution for their test statistic.

In thinking about their counterintuitive results, the researchers may, being bright and creative individuals, create a reasonable (nay, compelling!) explanation for why they would get results opposite to that which they predicted. Further, they may even convince themselves that they should have known that the results would turn out that way. Consequently, they may change their analytical approach to a two-tailed test in an effort to reject the null hypothesis, or they may even adopt a one-tailed test with a directional hypothesis opposite to their original directional hypothesis.

The problem with this hypothesizing after the results are known, or "HARKing" (Kerr, 1998), is that the probability of making a type I or alpha error has increased. The original one-tailed test in the wrong direction carried with it a 5% chance of making a type I error if the null hypothesis were true. Following this one-tailed test with a two-tailed test raises that to a 7.5% chance of making a type I error if the null hypothesis were true (with the initial 5% critical region on one side of the distribution added to the subsequent 2.5% critical region on the other side of the distribution). Following the initial one-tailed test with a one-tailed test in the other direction similarly raises the probability of making a type I error if the null hypothesis were true to 10%. Given the severity of the consequences of making type I errors, it should be obvious that practices that may be exploited to increase these error rates are not things to be considered casually. And given the bias for the publication of significant effects and the "publish or perish" mentality that researchers too often face, the researchers may decide that the downsides of making type I errors are temporarily overshadowed by the short-term rewards of finding significant results.

For these reasons, it has become standard practice for many psychology journals to require authors to use two-tailed tests almost exclusively, thereby more explicitly controlling the probability of making type I errors when the null hypothesis is true. It is our opinion that one-tailed tests are valuable, particularly when an effect in the opposite direction is as meaningful as no effect (such as when a new therapy has either no effect or makes the clients' situations worse—in either case, the therapy would be discontinued). However, researchers who intend to use them should be well prepared to justify why they were appropriate to use, and we recommend that researchers do so proactively.

WHAT SHOULD WE MEAN WHEN WE SAY OUR RESULTS ARE STATISTICALLY SIGNIFICANT?

When R. A. Fisher described the results of an experiment as significant, he meant that the result was unlikely to occur by chance if the tested hypothesis were true. That does not mean that the observed difference is important or of consequence. Furthermore, statistical significance is not the same as practical significance. What is the correct way to deal with situations in which our p*-value is slightly larger than the criterion we adopted for judging the result of our experiment as unlikely?*

We say our results are "statistically significant" when our test statistic falls in the critical region and we reject our null hypothesis in favor of our alternative hypothesis. What should we mean when we say our results are statistically significant?

It is unfortunate that Fisher used the term *significance testing* to describe his paradigm for statistical inference because in everyday discourse, *significance* and *significant* mean something very different than they do in statistics. The dictionary definitions of these terms include the language "important" and "of consequence." Such a definition is not what Fisher meant when he applied those terms to experiments. As noted above, Fisher used those terms to describe the situation in which the obtained *p*-value is below a certain level judged to be improbable when the null hypothesis is true, which is all *significance* means in the context of statistical inference and hypothesis testing. Unfortunately, we tend to drop the adjective *statistically* and just say a result is "significant," adding to the confusion.

Fisher went on to argue that the smaller the observed *p*-value, the stronger the evidence against the null hypothesis, and some people take that to mean that the smaller the *p*-value, the more significant the results are in terms of their importance. But as we have seen in this chapter, as long as the null hypothesis is false, the *p*-value can be made as small as we want by increasing the sample size. David Bakan (1966) made a persuasive argument that the null hypothesis is never true; that is, anything we do that treats the participants in our experiments differently will have some effect, even if that effect is small. Furthermore, in the Neyman and Pearson hypothesis-testing paradigm, we reject the null hypothesis whenever our test statistic falls in the critical region, but it does not matter where in that region the test statistic falls. Therefore, it is not correct to say that a result is "more significant" when the *p*-value is small or that a small *p*-value is a reflection of a large treatment effect.

Can a Result Be "Marginally Significant"?

As noted above, Fisher used the *p*-value as a measure of the strength of the evidence against the null hypothesis, and he proposed the use of $p < .05$ as the threshold beyond which the result of an experiment is considered to be unlikely when the null hypothesis is true. He would have treated $p < .049$ and $p < .051$ as similar results. On the other hand, Neyman and Pearson set the value of $\alpha = .05$ to differentiate those results in the critical region that lead us to reject the null hypothesis from those that do not. Clearly, the decision to adopt Fisher's $p < .05$ as the threshold is arbitrary. As Rosnow and Rosenthal (1989) commented, "Surely, God loves the .06 nearly as much as the .05." So what does it mean, and what should we do, when the *p*-value in our experiment is slightly greater than .05? Such results are sometimes referred to as *marginally significant*, *approaching significance*, *nearly significant*, or *trending*. It has become more common over the past 40 years for psychologists to use these terms to describe those situations (Pritschet, Powell, & Horne, 2016).

One argument for using such terms is to convey to the reader that the researcher is not confident that the null hypothesis is false but still thinks there is something worth reporting, and by using terms like *marginally significant*, the researcher can highlight findings that do not fall in the region of rejection so that the reader can decide how to interpret those findings. Researchers who use these terms need to be careful about how they characterize these findings, because, despite the arbitrary nature of the .05 threshold, these findings do not result in rejection of the null hypothesis, and researchers should not make strong conclusions as if they did.

Statistical Significance Versus Practical Significance

Statistical significance is not the same as practical significance, or importance. The importance of the results of an experiment depends on the context and other nonstatistical aspects of an experiment, even when the treatment effects are small (Prentice & Miller, 1992). In Chapter 7, we will look at how to estimate the effect size, and in Chapter 13 we will look at another way to estimate the effect size. In both cases, the measure of effect size is not affected by the sample size. It has long been the recommendation of the American Psychological Association, and it is being required by more and more journals, that researchers accompany their significance tests with effect sizes when reporting the results of their studies. But even then, researchers should be careful not to overstate the implications of their "significant" findings and the effect sizes they compute.

What Should We Do (or Not Do) When Our Data Do Not Allow Us to Reject the Null Hypothesis?

This situation happens to all of us. We carefully consider how many participants to use when we design our experiment to find an effect size of at least a certain magnitude, and we conduct our study carefully. But our obtained results are not in the critical region, and therefore we cannot reject our statistical null hypothesis. Where do we go from here?

When we have not rejected our null hypothesis, we have *not* provided evidence that the null hypothesis is true. As we have discussed earlier, when our null hypothesis is false, our statistical test result is not always in the critical region. That is, we may have made a type II (or beta) error.

There are a number of perspectives to consider in determining what this finding allows us to conclude. Bakan (1966) argued that the null hypothesis is never true because it is hard to conceive of a situation in which different treatments will have *exactly* the same effects on behavior. And logically it makes sense that a firm stance on any prediction that a predetermined exact value is true is unlikely to be verified by our findings. Therefore, it should be obvious that non-rejection of a null hypothesis, which does not allow us to say that predetermined exact value is true, is not evidence that the null hypothesis is true. We can only conclude that we do not have sufficient evidence to demonstrate that it is false.

In Fisher's (1925, 1928) significance-testing model, obtaining a result with a $p < .05$ can be taken as evidence that the null hypothesis is false, but obtaining a result with $p > .05$ cannot be taken as evidence that the null hypothesis is true. In this model, a non-significant result could be due to a number of possibilities, from the null hypothesis being true to the treatment effect being small. Under these circumstances, we cannot draw a firm conclusion about the null hypothesis. At best, we can say the results of our experiment are inconclusive.

On the other hand, Neyman and Pearson (1928a, 1928b) viewed the situation as a test between two competing hypotheses, H_0 and H_1. According to their perspective, we can reject the null hypothesis and accept the alternative when our test statistic is in the critical region. On the other hand, we "accept" the null hypothesis when the data are not in the critical region. However, according to them, "accepting" the null hypothesis does not mean that we are concluding that the null hypothesis is true. Their position is that we should "act" as if the null hypothesis is true until we get more data to indicate otherwise.

Bakan (1966) provided a justification for Neyman and Pearson's approach to non-rejection of the null hypothesis by distinguishing between "sharp" and "loose" null hypotheses. A sharp null hypothesis is that there is absolutely no difference between the population parameters; as he noted, this situation rarely, if ever, occurs. A loose null hypothesis is a range of values around a sharp null hypothesis such that any difference in the interval is too small for us to conclude that the null hypothesis is false. By adopting this approach, we do not accept the sharp null hypothesis; rather we say that the difference is too small to be meaningful for us.

In summary, when our research yields findings that do not fall in the critical region, we fail to reject the null hypothesis. Therefore, we "retain" it, but we do not officially "accept" it. Failing to find an effect is not the same as verifying that the effect does not exist. Indeed, absence of evidence is not evidence of absence. We recommend that researchers who have strong theoretical reasons to predict an effect, but do not find it in their studies, consider conducting their studies again with stronger manipulations, more reliable measures, and generally greater power. If the effect is out there and is strong enough to warrant interest (that is, it has a nontrivial effect size), you will likely find it. On the other hand, if your data from multiple well-designed studies fail to allow you to reject the null hypothesis, then the effect may be too small to be of value, or, possibly, the null hypothesis may be true—whether or not you can technically conclude that is the case.

A FINAL WORD

Although the principles described in this chapter were developed for the two-sample *z*-test for population means when the population variance is known, they apply to all statistical hypothesis tests using the classical statistical model. The two-sample *z*-test was chosen to illustrate those basic principles in a direct way. In subsequent chapters, we will apply these principles to other test statistics. Therefore, it is important that we master these basic principles before moving forward.

In the next chapter, we will use Student's *t* to analyze the results from Barlett's (2015) two-group experiment described in Chapter 1. As we will see, the shape of the *t*-distribution is affected by the sample size, and the shape and location of the non-central *t*-distribution is affected by the sample size and the effect size. Nevertheless, the basic principles described here apply.

Summary

Starting with our research question, we generate a set of mutually exclusive and exhaustive experimental hypotheses stated in terms of the independent and dependent variables in our research study. With the classical statistical model, we create a parallel set of statistical hypotheses stated in terms of the unknown parameters of the populations from which our random samples are obtained. We obtain estimates of these unknown parameters from our samples. (See the summary flowchart in Figure 6.7.)

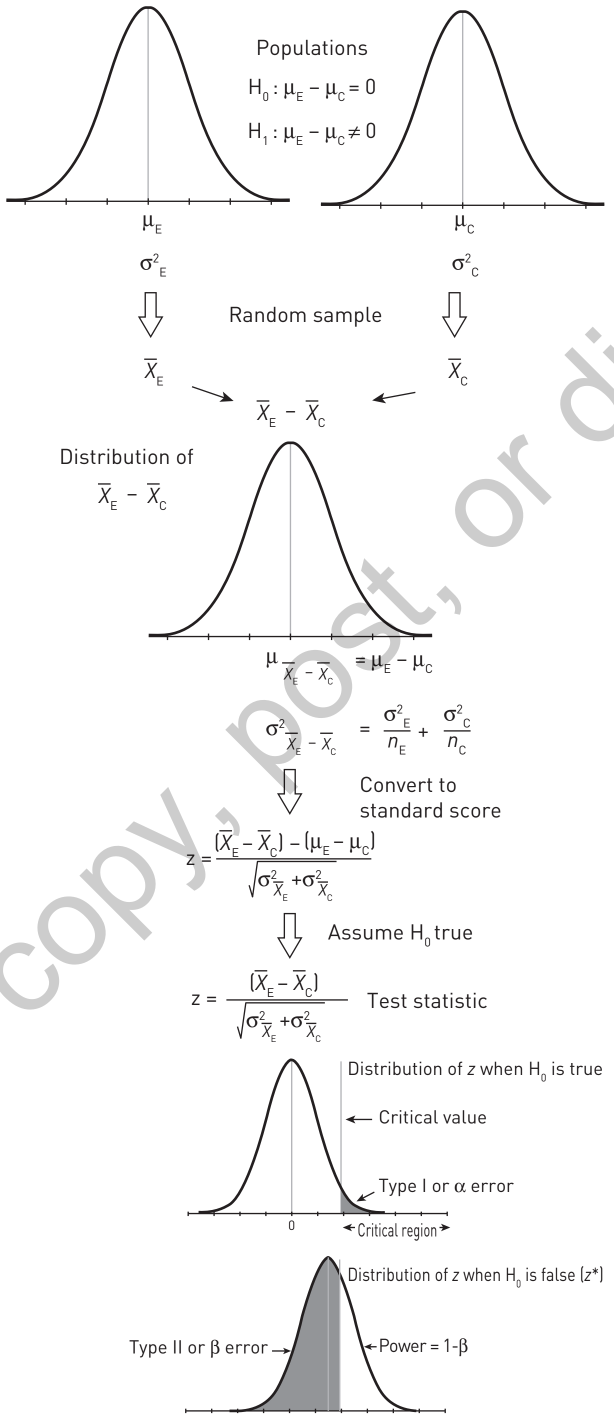
In a two-group experiment in which the statistical hypotheses are about the difference between the means of the populations ($\mu_E - \mu_C$), the data of interest are $\bar{X}_E - \bar{X}_C$. Assuming that the populations are normal distributions with known variances, the distribution of $\bar{X}_E - \bar{X}_C$ is a normal distribution. To find the probability of obtaining our observed value of $\bar{X}_E - \bar{X}_C$, we apply the *z*-score transformation. Then, we create our test statistic by assuming the hypothesis we can test (the null hypothesis) is true. When our null hypothesis is true, our test statistic has a normal distribution with $\mu_z = 0$ and $\sigma_z^2 = 1$, and we can create a critical region (based on our definition of an improbable event) such that, if the value of our test statistic fell in this region, we would reject our null hypothesis. The procedure for creating our test statistic and performing our statistical test is summarized in the flowchart in Figure 6.7.

When our null hypothesis is false, our test statistic is a non-central normal distribution (symbolized as z^*) with $\mu_{z^*} \neq 0$ and $\sigma_{z^*}{}^2 = 1$. The value of μ_{z^*} is a function of the actual value of $\mu_E - \mu_C$, the variances of the populations, and the sample sizes. The greater the value of μ_{z^*}, the greater the power of our test (probability of rejecting a false null hypothesis). We can use power functions to help us decide how large a sample size to use to obtain a high level of power.

When our null hypothesis is true, a result falling in the critical region will lead us to make a type I error or α error (we reject the null hypothesis when it is true). On the other hand, when our null hypothesis is false, a result falling outside the critical region will lead us to make a type II error or β error (we do not reject the null hypothesis when it is false). The convention is to use $\alpha = .05$ as our definition of an improbable event when the null hypothesis is true. We do not know the probability of making a β error, but because power = 1 − β, those factors that increase the power also decrease the probability of making a β error.

The method for hypothesis testing described in this chapter is a hybrid of models described by R. A. Fisher and by J. Neyman and E. S. Pearson. Because these two models are in some ways incompatible, we need to be careful about how we use concepts like *p*-value, α, and α error.

FIGURE 6.7 ■ Summary flowchart


Populations
$H_0: \mu_E - \mu_C = 0$
$H_1: \mu_E - \mu_C \neq 0$
μ_E
σ^2_E
μ_C
σ^2_C
Random sample
$\overline{X}_E$
$\overline{X}_C$
$\overline{X}_E - \overline{X}_C$
Distribution of
$\overline{X}_E - \overline{X}_C$
$\mu_{\overline{X}_E - \overline{X}_C} = \mu_E - \mu_C$
$\sigma^2_{\overline{X}_E - \overline{X}_C} = \frac{\sigma^2_E}{n_E} + \frac{\sigma^2_C}{n_C}$
Convert to standard score
$z = \frac{(\overline{X}_E - \overline{X}_C) - (\mu_E - \mu_C)}{\sqrt{\sigma^2_{\overline{X}_E} + \sigma^2_{\overline{X}_C}}}$
Assume H_0 true
$z = \frac{(\overline{X}_E - \overline{X}_C)}{\sqrt{\sigma^2_{\overline{X}_E} + \sigma^2_{\overline{X}_C}}}$
Test statistic
Distribution of z when H_0 is true
Critical value
Type I or α error
0
Critical region
Distribution of z when H_0 is false (z^*)
Type II or β error
Power = 1-β

Conceptual Exercises

1. Your test statistic is in the critical region, and you reject the null hypothesis. What is the probability that you made a type I or α error? Why?
2. Comment on the appropriateness of the following statement:

   My test statistic is not in the critical region, but the observed power of my test is high; therefore, the data indicate that I should accept the null hypothesis.
3. Fisher argued that the smaller the *p*-value, the stronger the evidence against the null hypothesis. That is, $p < .001$ indicates stronger evidence against the null hypothesis than $p < .01$. Does it follow that $p < .001$ indicates a larger treatment effect than $p < .01$? Why or why not?
4. Are you more likely to make an alpha (or type I) error with a large sample size or with a small sample size? Why or why not?
5. What are the effects of increasing sample size on:

   a. the probability of making a type I (or alpha) error?

   b. the probability of making a type II (or beta) error?

   Why in each case?
6. What does $p < .05$ tell us about the null hypothesis? About the alternative hypotheses?
7. Comment on the following statements and explain what, if anything, is wrong:

   a. We will make a type I error a small proportion of the time—the exact proportion being specified by our alpha level.

   b. Although I did not reject the null hypothesis, the power of my test indicates that I would have rejected the null hypothesis if I had enrolled 5 more participants in each group.

   c. If the sample sizes are equal, then a *p*-value of .001 represents a larger treatment effect than does a *p*-value of .05.
8. In general terms, how does one construct a power function for a statistical test? (Hint: Define the *power* of a test. What part of what distribution contains the power? How is this fact then translated into a power function?)
9. Why does a test statistic for which $p < .\ 001$ not necessarily represent a large treatment effect?
10. Although we usually expect to collect data to support a research hypothesis that more closely matches the alternative hypothesis, the hypothesis we actually test is the null hypothesis. What are the two reasons for why we do so?
11. Respond to the following statement:

    I am not convinced that there really is a difference here because the sample size is so small. We all know that with small sample sizes, there is a lot of variability in the sample means; the central limit theorem tells us that! Therefore, I am not convinced that the significant result is real. If the experimenter had used a larger sample size and gotten a significant result, then I would believe that there is something there.

    In your response, be sure to respond to the assertions about the sample size and the variability of the sample means, as well as the other parts of the statement. What false assumptions is the author making, and why are they false? What is the correct state of affairs? How should we interpret the significance of an experimental result?

Student Study Site

Visit the Student Study Site at **https://study.sagepub.com/friemanstats** for a variety of useful tools including data sets, additional exercises, and web resources.