3

STATISTICAL SIGNIFICANCE TESTING

3.1 ♦ The Logic of Null Hypothesis Significance Testing (NHST)

Chapter 2 reviewed the procedures used to set up a confidence interval (CI) to estimate an unknown population mean (μ) using information from a sample, including the sample mean (M), the number of cases in the sample (N), and the population standard deviation (σ). On the basis of what is known about the magnitude of sampling error, we expect values of the sample mean, M, to be normally distributed around a mean of μ with a standard deviation or standard error of σ_M. The value of σ_M depends on the values of σ and N:

$$\sigma_M = \sigma / \sqrt{N}. \tag{3.1}$$

Because the theoretical distribution of sample means has a normal shape when σ is known, we can use the standard *normal* distribution to evaluate distances from the mean of this sampling distribution in research situations where the value of σ is known. To set up a CI with a 95% level of confidence, we use the values of z that correspond to the middle 95% of the area in a normal distribution ($z = -1.96$ and $z = +1.96$), along with the value of σ_M, to estimate the lower and upper boundaries of the CI. About 95% of the CIs set up using these procedures should contain the actual population mean, μ, across many replications of this procedure.

However, there is another way that the sample mean, M, can be used to make inferences about likely values of an unknown population mean, μ. Values of the sample mean, M, can be used to test hypotheses about a specific value of an unknown population mean through the use of null hypothesis significance test (**NHST**) procedures.

What is the logic behind NHST? At the simplest level, when a researcher conducts a **null hypothesis significance test**, the following steps are involved.

First, the researcher makes a "guess" about the specific value of μ for a population of interest; this guess can be written as a formal **null hypothesis** (H_0). For example, suppose that the variable of interest is the driving speed on Route 95 (in mph) and the population of interest is all the drivers on Route 95. A researcher might state the following null hypothesis:

H_0: $\mu = \mu_{hyp}$, a specific hypothesized value, for example, H_0: $\mu = 65$ mph.

In words, this null hypothesis corresponds to the assumption that the unknown population mean driving speed, μ, is 65 mph. In this example, the value of 65 mph corresponds to the posted speed limit. In a study that tests hypotheses about the value of one population mean, using the mean from a sample drawn from the population, the "effect" that the researcher is trying to detect is the difference between the unknown actual population mean μ and the hypothesized population mean μ_{hyp}.

Next, the researcher selects a random sample from the population of all passing cars on Route 95, measures the driving speed for this sample, and computes a sample mean, M (e.g., the sample mean might be $M = 82.5$ mph).

Then, the researcher compares the observed sample mean, M ($M = 82.5$ mph), with the hypothesized population mean, μ ($\mu = 65$ mph), and asks the following question: If the true population mean driving speed, μ, is really 65 mph, is the obtained sample mean ($M = 82.5$ mph) a likely outcome or an unlikely outcome? A precise standard for the range of outcomes that would be viewed *unlikely* is established by choosing an **alpha** (α) level and deciding whether to use a **one-** or **two-tailed test**, as reviewed later in this chapter; in general, outcomes that would be expected to occur less than 5% of the time when H_0 is true are considered unlikely.

The theoretical unlikeliness of a specific value of M (when H_0 is true) is evaluated by calculating how far M is from the hypothesized value of μ_{hyp} in the number of standard errors. The distance of M from a hypothesized value of μ is called a z ratio; the areas that correspond to values of z are evaluated using the table of the standard normal distribution in situations where the population standard deviation, σ, is known:

$$z = \frac{M - \mu_{hyp}}{\sigma_M}. \tag{3.2}$$

As discussed in Chapters 1 and 2, for a normal distribution, there is a fixed relationship between a distance from the mean given by a z score and the proportion of the area in the distribution that lies beyond the z score. Recall that in Chapter 2, a similar z ratio was used to evaluate the location of an individual X score relative to a distribution of other individual X scores. In Chapter 2, a z score was calculated to describe the distance of an individual X score from a sample or population mean. In this chapter, the z ratio provides information about the location of an individual sample mean, M, relative to a theoretical distribution of many different values of M across a large number of independent samples. Once we convert an observed sample mean, M, into a z score, we can use our knowledge of the fixed relationship between z scores and areas under the normal distribution to evaluate whether the obtained value of M was "close to" $\mu = 65$ and, thus, a likely outcome when $\mu = 65$ or whether the obtained value of M was "far from" $\mu = 65$ and, thus, an unlikely outcome when $\mu = 65$.

When a researcher obtains a sample value of M that is very far from the hypothesized value of μ, this translates into a large value of z. A z ratio provides precise information about the distance of an individual sample mean M from μ in the number of standard errors.

The basic idea behind NHST is as follows. The researcher assumes a value for the unknown population mean, μ (in this example, H_0: $\mu = 65$). The researcher evaluates the obtained value of the sample mean, M (in this case, $M = 82.5$), relative to the distribution

of values of M that would be expected if μ were really equal to 65. The researcher wants to make a decision whether to reject H_0: $\mu = 65$ as implausible, given the value that was obtained for M. If the sample mean, M, is a value that is likely to occur by chance when H_0 is true, the decision is not to reject H_0. If the sample mean, M, is an outcome that is very unlikely to occur by chance when H_0 is true, the researcher may decide to reject H_0.

Subsequent sections review the details involved in the NHST decision process. How can we predict what outcomes of M are likely and unlikely to occur, given an assumed value of μ for a specific null hypothesis along with information about σ and N? How can we evaluate the likelihood of a specific observed value of M relative to the distribution of values of M that would be expected if H_0 were true? What conventional standards are used to decide when an observed value of M is so unlikely to occur when H_0 is true that it is reasonable to reject H_0?

There is a bit of logical sleight of hand involved in NHST, and statisticians disagree about the validity of this logic (for further discussion of these issues, see Cohen, 1994; Greenwald, Gonzalez, Harris, & Guthrie, 1996). The logic that is generally used in practice is as follows: If a researcher obtains a value of M that would be unlikely to occur if H_0 is true, then the researcher rejects H_0 as implausible.

The logic involved in NHST is confusing and controversial for several reasons. Here are some of the issues that make NHST logic potentially problematic.

1. In everyday thinking, people have a strong preference for stating hypotheses that they believe to be correct and then looking for evidence that is consistent with their stated hypotheses. In NHST, researchers often (but not always) hope to find evidence that is inconsistent with the stated null hypothesis.[1] This in effect creates a double negative: Researchers often state a null hypothesis that they believe to be incorrect and then seek to reject that null hypothesis. Double negatives are confusing, and the search for "disconfirmatory" evidence in NHST is inconsistent with most people's preference for "confirmatory" evidence in everyday life.

2. NHST logic assumes that the researcher has a random sample from a well-defined population of interest and that, therefore, the mean, M, and standard deviation, s, based on the sample data should be good estimates of the corresponding population parameters, μ and σ. However, in many real-life research situations, researchers use convenience samples rather than random samples selected from a well-defined actual population. Thus, in many research situations, it is unclear what population, if any, the researcher is in a position to make inferences about. NHST logic works best in research situations (such as industrial quality control studies) where the researcher is able to draw a random sample from the entire population of interest (e.g., a random sample from all the products that come off the assembly line). NHST is more problematic in situations where the study uses a convenience sample; in these situations, the researcher can at best make tentative inferences about some hypothetical broader population that has characteristics similar to those of the sample (as discussed in Chapter 1).

3. As pointed out by Cohen (1994) and many other critics, NHST does not tell researchers what they really want to know. Researchers really want to know, given the value of M in a batch of sample data, how likely it is that H_0, the null hypothesis, is correct. However, the probability estimate obtained using NHST refers to the probability of something quite different. A p value is *not* the conditional probability

that H_0 is correct, given a sample value of M. It is more accurate to interpret the p value as the theoretical probability of obtaining a value of M farther away from the hypothesized value of μ than the value of M obtained in the study, given that the null hypothesis H_0 is correct. When we use a z ratio to look up a p value, as discussed later in this chapter, the p value is the (theoretical) probability of obtaining an observed value of M as large as or larger than the one in the sample data, given that H_0 is correct and given that all the assumptions and procedures involved in NHST are correct. If we obtain a value of M that is very far away from the hypothesized value of the population mean, μ (and that, therefore, would be very unlikely to occur if H_0 is correct), we typically decide to reject H_0 as implausible. Theoretically, setting up the decision rules in this manner yields a known risk of committing a **Type I error** (a Type I error occurs when a researcher rejects H_0 when H_0 is correct).

Given these difficulties and the large number of critiques of NHST that have been published in recent years, why do researchers continue to use NHST? NHST logic provides a way to make yes/no decisions about null hypotheses with a (theoretically) known risk of Type I error. Note, however, that this estimated risk of Type I error is correct only when all the assumptions involved in NHST procedures are satisfied and all the rules for conducting significance tests are followed. In practice, because of violations of assumptions and departures from the ideal NHST procedure, the actual risk of Type I error is often higher than the expected risk given by the choice of the α level for the statistical significance test.

In real-world situations, such as medical research to assess whether a new drug significantly reduces disease symptoms in a sample compared with an untreated control population, researchers and practitioners need to make yes/no decisions about future actions: Is the drug effective, and should it be adopted as a medical treatment? For such situations, the objective yes/no decision standards provided by NHST may be useful, provided that the limitations of NHST are understood (Greenwald et al., 1996). On the other hand, when research focuses more on theory development, it may be more appropriate to think about the results of each study as cumulative evidence, in the manner suggested by Rozeboom (1960): "The primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one ... believes the hypothesis ... being tested" (p. 420).

3.2 ♦ Type I Versus Type II Error

Two different types of errors can occur when statistical significance tests are used to make binary decisions about the null hypothesis. In other words, a researcher who uses NHST typically reports one of two possible decisions about the status of H_0: either "reject H_0" or "do not reject H_0."[2] In actuality, H_0 may be either true or false. When we evaluate whether an NHST decision is correct, there are thus four possible outcomes; these are summarized in Table 3.1. There are two possible correct decisions: A researcher may decide to reject H_0 when H_0 is actually false; a researcher may decide not to reject H_0 when H_0 is true. On the other hand, there are two different possible types of error. A researcher may reject H_0 when it is actually correct; this is called a Type I error. A researcher may fail to reject H_0 when H_0 is actually false; this is called a **Type II error.**

Table 3.1 ♦ Type I Versus Type II Error

	Actual State of the World	
Researcher Decision	H_0 *Is True*	H_0 *Is False*
Reject H_0	Type I error (α)	Correct decision ($1 - \beta$)
Do not reject H_0	Correct decision ($1 - \alpha$)	Type II error (β)

The symbols α and β in Table 3.1 correspond to the theoretical risk of committing a Type I error (α) and a Type II error (β).

When the null hypothesis is actually correct, a researcher may obtain a large difference between the sample mean, M, and the actual population mean, μ (because of sampling error), that leads to an (incorrect) decision to reject H_0; this is called a Type I error. When a researcher commits a Type I error, he or she rejects the null hypothesis when in fact the null hypothesis correctly specifies the value of the population mean.

The theoretical risk of committing a Type I error is related to the choice of the alpha (α) level; *if* a researcher sets the alpha level that is used to look up values for the reject regions at $\alpha = .05$, and *if* all other assumptions of NHST are satisfied, then in theory, the use of NHST leads to a 5% risk of rejecting H_0 when H_0 is actually correct. The theoretical risk of Type I error can be reduced by making the alpha level smaller—for example, setting α at .01 rather than the conventional .05.

On the other hand, when the null hypothesis is incorrect (i.e., the population mean is actually different from the value of μ that is stated in the null hypothesis), the researcher may fail to obtain a statistically significant outcome. Failure to reject H_0 when H_0 is incorrect is called a Type II error. The probability of failing to reject H_0 when H_0 is incorrect—that is, the risk of committing a Type II error—is denoted by β.

Researchers want the risk of both types of error (α and β) to be reasonably low. The theoretical risk of Type I error, α, is established when a researcher selects an alpha level (often $\alpha = .05$) and uses that alpha level to decide what range of values for the test statistic such as a z or ***t* ratio** will be used to reject H_0. If all the assumptions involved in NHST are satisfied and the rules for significance testing are followed, then in theory, the risk of committing a Type I error corresponds to the alpha level chosen by the researcher. However, if the assumptions involved in NHST procedures are violated or the rules for conducting hypothesis tests are not followed, then the true risk of Type I error may be higher than the nominal alpha level, that is, higher than the alpha level named or selected by the investigator as the desired standard for risk of Type I error.

The risk of committing a Type II error depends on several factors, including sample size. Factors that influence the theoretical risk of committing a Type II error are discussed in Section 3.9 on statistical power. Statistical power is the probability of correctly rejecting H_0 when H_0 is false, denoted as $(1 - \beta)$. Therefore, the statistical power $(1 - \beta)$ is the complement of the risk of Type II error (β); factors that increase statistical power also decrease the risk of Type II error.

3.3 ♦ Formal NHST Procedures: The *z* Test for a Null Hypothesis About One Population Mean

Let's consider an application of NHST to a research question about the value of one unknown population mean, μ, in a situation where the population standard deviation, σ, is known.[3] The first step is to identify a variable and population of interest. Suppose that a researcher wants to test a hypothesis about the mean intelligence score for a population of all the residents in a large nursing home. The variable of interest is a score on a widely used intelligence quotient (IQ) test, the Wechsler Adult Intelligence Scale (WAIS). These scores are normally distributed; in the IQ test norms, the scores are scaled to have a mean of $\mu = 100$ points and a population standard deviation of $\sigma = 15$ points for the overall adult population.

3.3.1 ♦ Obtaining a Random Sample From the Population of Interest

Suppose that the entire nursing home population consists of hundreds of residents and the director can afford to test only $N = 36$ people. Therefore, the researcher plans to obtain a random sample of $N = 36$ persons from the population of the entire nursing home and to compute the mean IQ score, M, for this sample. The mean IQ score for the sample, M, will be used to decide whether it is plausible to believe that the mean IQ for the entire nursing home population is equal to 100, the mean value of the IQ for the general adult population.

3.3.2 ♦ Formulating a Null Hypothesis (H_0) for the One-Sample *z* Test

Next the researcher sets up a null hypothesis that specifies a "guessed" value for an unknown population mean. One possible null hypothesis for the nursing home study is based on the norms for IQ scores; for the general adult population, there is a mean IQ of 100 points. The researcher could state the following null hypothesis:

$$H_0: \mu = \mu_{hyp}, \text{ in this example, } H_0: \mu = 100. \tag{3.3}$$

In this example, $H_0: \mu = 100$, where μ is the unknown population mean for the variable of interest (IQ) in the population of interest (all the nursing home residents), and μ_{hyp} is the mean IQ from the WAIS test norms. In words, then, the researcher hypothesizes that the IQ for the nursing home population corresponds to an average ability level compared with the intelligence for the general population.

Note that there are several likely sources of values for μ_{hyp}, the hypothesized population mean. When the variable of interest is a standardized psychological test, the value of μ_{hyp} might be selected based on knowledge of the normative values of scores. A researcher might select the overall population mean as the point of reference for hypothesis tests (as in this example) or use a **cutoff value** that corresponds to a clinical diagnosis as the point of reference for hypothesis tests. For example, on the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), a score of 19 or more corresponds to moderate to severe depression. A researcher might set up the null hypothesis that the mean depression score for a population of nursing home patients equals or exceeds this cutoff value of 19 points. Similarly, for physiological measures, such as systolic blood pressure, a clinical cutoff value (such as 130 mm Hg, which is sometimes used to diagnose borderline hypertension) might be used as the value of μ_{hyp}. A legal standard (such as the posted speed limit on a

highway, as in the earlier example) is another possible source for a specific value of μ_{hyp}. In industrial quality control applications, the hypothesized population mean, μ_{hyp}, often corresponds to some production goal or standard. For example, B. Warner and Rutledge (1999) tested the claim of a cookie manufacturer that the population mean number of chocolate chips in each bag of cookies was equal to or greater than 1,000.

Most applications of NHST presented subsequently in this book involve testing null hypotheses that correspond to an assumption that there is no difference between a pair of means for populations that have received different treatments, or no relationship between a pair of variables, rather than hypotheses about the value of the population mean for scores on just one variable.

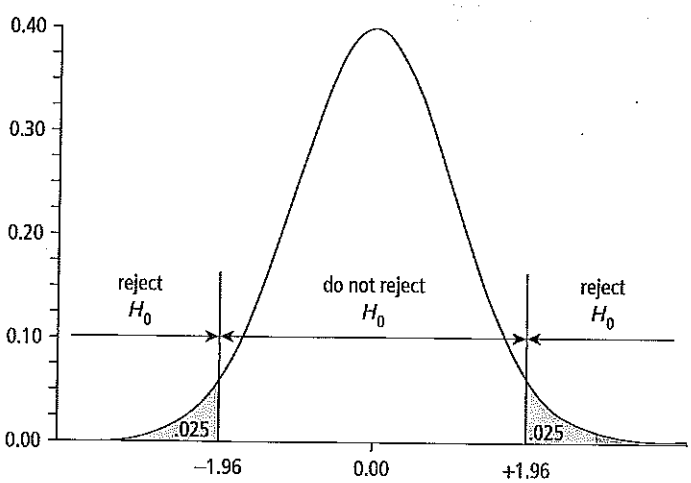
3.3.3 ♦ Formulating an Alternative Hypothesis (H_1)

The **alternative hypothesis** (sometimes called the **research hypothesis**), H_1, can take one of three possible forms. The first form is called a nondirectional or two-tailed alternative hypothesis:

$$H_1: \mu \neq 100. \tag{3.4}$$

Using this alternative hypothesis, the researcher will reject H_0 for values of M that are either much higher or much lower than 100. This is called a **nondirectional** or two-tailed **test;** when we figure out a decision rule (the ranges of values of M and z for which we will reject H_0), the reject region includes outcomes in both the upper and lower tails of the sampling distribution of M, as shown in Figure 3.1. For $\alpha = .05$, two-tailed, we would reject H_0 for obtained values of $z < -1.96$ and $z > +1.96$.

Figure 3.1 ♦ Standard Normal Distribution Curve Showing the Reject Regions for a Significance Test Using $\alpha = .05$, Two-Tailed



NOTES: Reject regions for a z ratio for $\alpha = .05$, two-tailed: Reject H_0 for $z < -1.96$ and for $z > +1.96$; do not reject H_0 for values of z between -1.96 and $z = +1.96$.

For a normal distribution, the proportion of the area that lies in the tail below $z = -1.96$ is .025, and the proportion of the area that lies in the tail above $z = +1.96$ is .025. Therefore, the part of the distribution that is "far" from the center of the distribution, using $\alpha = .05$ as the criterion for far, corresponds to the range of z values less than -1.96 and greater than $+1.96$.

The second and third possible forms of H_1, the alternative hypothesis, are called one-tailed or directional alternative hypotheses. They differ in the direction of the inequality that is stated—that is, whether the true population mean is predicted to be lower than or higher than the specific value of μ stated in the null hypothesis. The second form of H_1 is a one-tailed test that corresponds to a reject region in the lower tail of the distribution of outcomes for M.

$$H_1: \mu < 100. \tag{3.5}$$

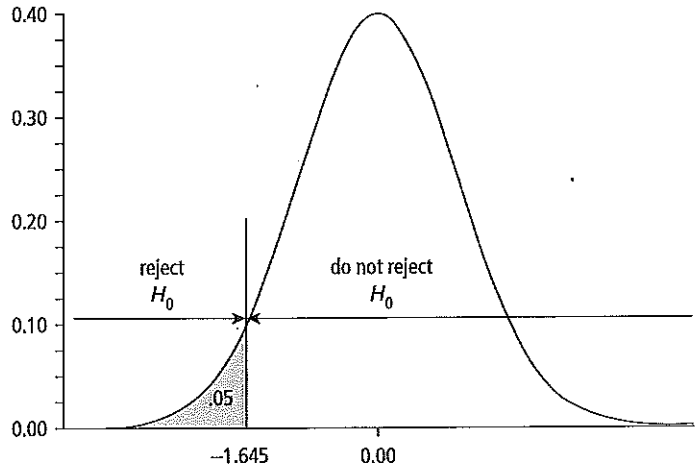
If we use the alternative hypothesis stated in Equation 3.5, we will reject H_0 only for values of M that are much lower than 100. This is called a **directional** or one-tailed **test.** The decision to reject H_0 will be made only for values of M that are in the lower tail of the sampling distribution, as illustrated in Figure 3.2.

The third and last version of the alternative hypothesis is stated in Equation 3.6:

$$H_1: \mu > 100. \tag{3.6}$$

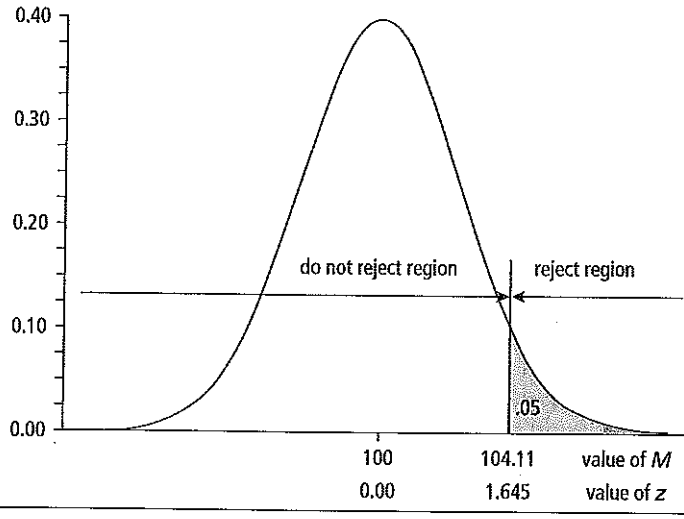
If we use the alternative hypothesis specified in Equation 3.6, we will reject H_0 only for values of M that are much higher than 100. This is also called a directional or one-tailed test; the decision to reject H_0 will be made only for values of M that are in the upper tail of the sampling distribution of values of M, as shown in Figure 3.3; we reject H_0 only for large positive values of z.

Figure 3.2 ♦ Standard Normal Distribution Curve With Reject Region for Significance Test With $\alpha = .05$, One-Tailed, $H_1: \mu < \mu_{hyp}$



NOTES: Reject region for a one-tailed test, $\alpha = .05$, shown as the shaded area in the lower tail of the normal distribution. This reject region is for the directional alternative hypothesis $H_1: \mu < 100$. Reject H_0 if obtained $z < -1.645$. Do not reject H_0 if obtained z value is ≥ -1.645.

Figure 3.3 ♦ Normal Distribution Curve With Reject Region for Significance Test With $\alpha = .05$, One-Tailed, $H_1: \mu > \mu_{hyp}$



NOTES: X-axis shows values of M and corresponding values of z. The reject regions for the one-tailed test, $\alpha = .05$, correspond to the shaded area in the upper tail of the normal distribution. For the directional test (with $H_1: \mu > 100$), we would reject H_0 for values of $M > 104.11$ that correspond to values of $z > +1.645$.

In the following example, we will test the null hypothesis, $H_0: \mu = 100$, against the alternative hypothesis, $H_1: \mu > 100$. The choice of $H_1: \mu > 100$ as the alternative hypothesis means that we will reject H_0 only if the sample mean, M, is substantially greater than 100 IQ points; that is, we will use the reject region shown in Figure 3.3.

3.3.4 ♦ Choosing a Nominal Alpha Level

Next the researcher must choose a nominal alpha level or **level of significance**. The nominal alpha level is a theoretical risk of committing a Type I error, that is, the probability of rejecting the null hypothesis $H_0: \mu = 100$ when the null hypothesis $H_0: \mu = 100$ is actually correct. We call this a *nominal* alpha level because it is named or *nominated* as a standard for making judgments about statistical significance by the researcher. A nominal alpha level is chosen arbitrarily by the researcher. Following the example set by Sir Ronald Fisher, most users of statistics assume that an α value of .05 represents an acceptably small risk of Type I error in most situations. However, in exploratory research, investigators are sometimes willing to use alpha levels (such as $\alpha = .10$) that correspond to a higher risk of Type I error. Sometimes, investigators prefer to use smaller alpha levels; $\alpha = .01$ and $\alpha = .001$ are common choices when researchers want to keep the theoretical risk of Type I error very small.

3.3.5 ♦ Determining the Range of z Scores Used to Reject H_0

Next we need to use the alpha level and the choice of a nondirectional or directional alternative hypothesis to formulate a decision rule: For what range of values of z will we

decide to reject H_0? When we set $\alpha = .05$ as the acceptable level of risk of Type I error and use a one-tailed test with the reject region in the upper tail, our *reject region* is the range of z scores that corresponds to the top 5% of the area in a normal distribution. The z value that corresponds to a .05 proportion of area in the upper tail of a normal distribution can be found in the table in Appendix A, "Proportions of Area Under a Standard Normal Curve." The column that corresponds to "Area C" is examined to locate the table entries that are closest to a proportion of .05 of the area in the upper tail; from the second page of this table in the far right-hand column, we find that an area of .0505 corresponds to $z = +1.64$, and an area of .0495 corresponds to $z = +1.65$. An area of exactly .05 corresponds to a z value of +1.645.[4] In other words, when a variable (such as the value of M across many samples drawn randomly from the same population) has a normal distribution, 5% of the outcomes for the normally distributed variable will have z values $> +1.645$, or, to say this another way, z scores $> +1.645$ correspond to the top 5% of outcomes in a normal distribution.

The graph in Figure 3.3 shows a normal distribution; the tail area that lies beyond the *critical value* of $z = +1.645$ is shaded. A decision rule based on $\alpha = .05$ and a one-tailed or directional version of the alternative hypothesis H_1 that should theoretically give us a 5% risk of Type I error is as follows:

Reject H_0 for obtained values of $z > +1.645$.

Do not reject H_0 for obtained values of $z \leq +1.645$.

Recall that the z value shown in Equation 3.2 provides information about the direction and magnitude of the difference between an obtained sample mean, M, and the hypothesized population mean, μ, stated in the null hypothesis. A large value of z corresponds to a value of M that is "far away" from the hypothesized value of μ.

3.3.6 ♦ Determining the Range of Values of M Used to Reject H_0

It is helpful to understand that the "reject" regions for the outcome of the study can also be stated in terms of values obtained for M, the sample mean. For a critical value of z (such as +1.645) that is obtained from a table of the standard normal distribution, we can figure out the corresponding value of M that would be used to make a decision whether to reject H_0 if we know the hypothesized value of μ and the calculated value of σ_M.

Recall that from Equation 3.2, $z = (M - \mu_{hyp})/\sigma_M$. In this example, the hypothesized population mean $\mu = 100$ and $\sigma_M = \sigma/\sqrt{N} = 15/\sqrt{36} = 2.50$. We can translate the reject regions (given above in ranges of values for the z ratio) into reject regions given in terms of values of M, the sample mean. In this problem, we obtain a z score to evaluate the location of any specific sample mean, M, relative to μ_{hyp} by computing the corresponding z value: $(M - \mu_{hyp})/\sigma_M = (M - 100)/2.5$. We can rearrange this equation to calculate the value of M ($M_{critical}$) that corresponds to a specific critical value of z, $z_{critical}$.

If

$$z_{\text{critical}} = (M_{\text{critical}} - \mu_{\text{hyp}})/\sigma_M,$$

then

$$z_{\text{critical}} \times \sigma_M = (M_{\text{critical}} - \mu_{\text{hyp}}).$$

Rearranging the expression above to isolate the value of M on one side of the equation yields the following equation for M_{critical}, the boundary of the reject region in terms of outcome values for M, the sample mean:

$$M_{\text{critical}} = [z_{\text{critical}} \times \sigma_M] + \mu_{\text{hyp}}. \tag{3.7}$$

Equation 3.7 tells us the value of the sample mean, M (M_{critical}), that corresponds to a specific critical value of z (such as $z = +1.645$). Using the specific numerical values of μ_{hyp} and σ_M in this situation and the critical value $z = +1.645$, we can calculate the boundary for the reject region in terms of values of the sample mean, M:

$$M_{\text{critical}} = [+1.645 \times 2.5] + 100 = 104.11.$$

Figure 3.3 shows the value of M that corresponds to the critical value of z for a directional test with $\alpha = .05$. A critical value of M was calculated given specific numerical values of μ and σ_M, and the critical value of M appears on the X axis of the normal distribution. Given $\sigma = 15$, $N = 36$, and $\sigma_M = 2.50$, and H_0: $\mu = \mu_{\text{hyp}} = 100$, we can state the decision rule in terms of obtained values of z (as at the end of the previous section):

Reject H_0 for any sample mean that corresponds to a value of $z > +1.645$.

But we can also state our decision rule directly in terms of values of the sample mean. In this case, we would reject H_0 for obtained sample values of $M > +104.11$.

It may be helpful to stop and think about the reasoning behind this decision rule. Knowledge about sampling error that was reviewed in Chapter 2 makes it possible for us to predict the amount of variation in the magnitude of M when samples of N cases are randomly drawn from a population with known values of μ and σ. In other words, given specific numerical values of N and σ and an assumed value of μ_{hyp} stated in a null hypothesis, we can predict the distribution of values of M that are expected to occur if H_0 is true. If all our assumptions are correct, if $N = 36$ and $\sigma = 15$, and if $\mu = 100$, then the values of M should be normally distributed around a mean of 100 with a standard deviation or standard error of σ_M where $\sigma_M = 15/\sqrt{36} = 2.5$. This corresponds to the distribution that appears in Figure 3.3. If μ is really 100, then most sample values of M are expected to be "fairly close to" this population mean of 100. We use an alpha level (such as $\alpha = .05$) and an alternative hypothesis (such as H_1: $\mu > 100$) to decide what set of outcomes for the sample mean, M, would be less consistent with the null hypothesis than with the alternative hypothesis. In this example, using the alternative hypothesis H_1: $\mu > 100$, we will

reject H_0 and prefer H_1 only for values of M that are substantially greater than 100. Because we have set the nominal alpha level at $\alpha = .05$ and this alternative hypothesis corresponds to a one-tailed test with a reject region that corresponds to values of M that are greater than 100, we identify the top 5% of possible values of M as the set of outcomes that are least likely to occur when H_0 is correct. Because the top 5% of the area in a normal distribution corresponds to a z score location of +1.645 standard deviations or standard errors above the mean, we decide to use only those values of M that correspond to distances from the mean greater than $z = +1.645$ as evidence to reject H_0. We can convert this z score distance into a corresponding value of the sample mean, M, using Equation 3.7. If H_0 is true, then 95% of the time, we would expect to observe an outcome value for M that is less than or equal to $M = 104.11$. If H_0 is true, we would expect to observe values of $M > 104.11$ only 5% of the time.

The basic idea behind NHST can be summarized as follows. The researcher chooses a nominal alpha level and formulates a null and an alternative hypothesis. The alpha level and hypotheses are used to decide whether to use a one- or two-tailed reject region for the decision about H_0 and how much area to include in one or both tails. For example, when we use a one-tailed test with H_1: $\mu > 100$ and $\alpha = .05$, we reject H_0 for the range of values of M that corresponds to the top 5% of values for M that we would expect to see across many samples if H_0 were true.

Note that the probability that we can assess when we conduct a significance test estimates how likely the obtained value of M is, given the assumption that H_0 is correct. If we obtain a value of M in our sample that is very unlikely to arise just by chance due to sampling error when H_0 is correct, it is reasonable to make the decision to reject H_0. If we obtain a value of M in our sample that is very likely to arise just by chance due to sampling error when H_0 is correct, it is more reasonable to make the decision to not reject H_0. The specific criterion for "very unlikely" is determined by the nominal alpha level (usually, $\alpha = .05$) and the choice of an alternative hypothesis (H_1) that corresponds to a reject region that includes one tail or both tails of the distribution.

Suppose the researcher collected IQ scores for a sample of nursing home patients and obtained a sample mean $M = 114.5$ points. In the situation described in this section, with $N = 36$, $\sigma = 15$, H_0: $\mu = 100$, and H_1: $\mu > 100$, this value of M would lead to a decision to reject H_0. If the population mean were really equal to 100 points, the probability of obtaining a sample mean greater than 104.11 would be, in theory, less than 5%. Thus, based on our understanding of sampling error, a sample mean of $M = 114.5$ falls within a range of values of M that are theoretically very unlikely outcomes if H_0 is correct.

3.3.7 ♦ Reporting an "Exact" *p* Value

Most introductory textbooks present NHST as a yes/no decision rule about H_0. If the obtained value of z exceeds the critical values of z (that correspond to the chosen alpha level and directional or nondirectional alternative hypothesis) or if the obtained value of M exceeds the critical value of M that corresponds to this critical value of z, the researcher decision is to reject H_0. If the obtained value of z does not exceed the critical value(s) of z that correspond to the criteria (such as $\alpha = .05$, one-tailed), the researcher decision is to

not reject H_0. However, there is another way of reporting the outcome for a significance test; an **"exact" *p* value** can be reported, in addition to or instead of a decision whether to reject H_0.

What is an exact *p* value? The exact *p* value is the (theoretical) probability of obtaining a sample mean, *M*, farther away from the hypothesized value of μ specified in the null hypothesis than the value of *M* in the sample in the study, *if* H_0 *is actually correct.* For a sample mean, the exact *p* value for a one-tailed test corresponds to the proportion of outcomes for the sample mean that would be theoretically expected to be greater than the obtained specific value for the sample mean if H_0 were true. We can determine an exact *p* value that corresponds to a sample value of *M*, given that we have the following information: the value of μ specified in the null hypothesis, information on whether the test is one- or two-tailed, and the value of σ_M.

To determine an exact one-tailed *p* value for an obtained sample mean, *M*, of 106 (we continue to use H_0: $\mu = 100$, H_1: $\mu > 100$, and $\sigma_M = 2.5$), we first compute the corresponding value of *z* that tells us how far the sample mean, *M*, is from the hypothesized value of μ (in the number of standard errors):

$$z = (M - \mu)/\sigma_M = (106 - 100)/2.5 = 2.40.$$

The use of a directional or one-tailed alternative hypothesis (H_1: $\mu > 100$) means that we need to look at only one tail of the distribution—in this case, the area in the upper tail, the area that lies above a *z* score of +2.40. From the table of areas that correspond to *z* score distances from the mean in Appendix A, we find that the tail area to the right of $z = 2.40$ is .0082, or .8% of the area. This value, .0082, is therefore the exact one-tailed *p* value that corresponds to a *z* value of +2.40; that is, exactly .0082 of the area in a normal distribution lies above $z = +2.40$. This can be interpreted in the following manner. If H_0 is correct, the (theoretical) likelihood of obtaining a sample mean larger than the one in this sample, $M = 106$, is .0082. Because this is a very unlikely outcome when H_0 is correct, it seems reasonable to doubt that H_0 is correct.

When a program such as SPSS is used to compute a significance test, an exact *p* value is generally reported as part of the results (often this is denoted by **sig**, which is an abbreviation for "statistical significance level"). Where there is a possibility of using a one-tailed or two-tailed test, it is usually possible to select either of these by making a check box selection in one of the SPSS menus.

In practice, users of SPSS and other statistical programs who want to make a binary decision about a null hypothesis (i.e., either "reject H_0" or "do not reject H_0") make this decision by comparing the obtained *p* value that appears on the SPSS printout with a preselected alpha level. If the *p* value reported on the SPSS printout is less than the preselected nominal alpha level (usually $\alpha = .05$), the decision is to reject H_0. If the *p* value reported on the SPSS printout is greater than the preselected alpha level, the decision is to not reject H_0.

Recent changes in publication guidelines in some academic disciplines such as psychology call for the reporting of exact *p* values. The fifth edition of the *Publication Manual of the American Psychological Association* (APA, 2001) summarized these recommendations as follows:

> Two types of probabilities are generally associated with the reporting of significance levels in inferential statistics. One refers to the a priori probability you have selected as an acceptable risk of falsely rejecting a given null hypothesis. This probability, called the "alpha level" (or "significance level"), is the probability of a Type I error in hypothesis testing and is commonly set at .05 or .01. The other kind of probability, the [exact] p value (or significance probability), refers to the a posteriori likelihood of obtaining a result that is as extreme as or more extreme than the observed value you obtained, assuming that the null hypothesis is true. . . . Because most statistical packages now report the [exact] p value (given the null and alternative hypotheses provided) and because this probability can be interpreted according to either mode of thinking, in general it is the exact probability (p value) that should be reported. (pp. 24–25)

When it is inconvenient to report exact p values—for example, in large tables of correlations—it is common practice to highlight the subset of statistical significance tests in the table that have p values below conventional prespecified alpha levels. It is fairly common practice to use one asterisk (*) to indicate $p < .05$, two asterisks (**) for $p < .01$, and three asterisks (***) for $p < .001$. When an exact p value is reported, the researcher should state what alpha level was selected. To report the numerical results presented earlier in this section, a researcher could write either of the following.

Results

Version 1 (statement whether the obtained exact *p* value was less than an a priori alpha level):

The null hypothesis that the mean IQ for the entire population of nursing home residents was equal to the general population mean for all adults (H_0: $\mu = 100$) was tested using a directional alternative hypothesis (H_1: $\mu > 100$) and $\alpha = .05$. The obtained sample mean IQ was $M = 106$ for a random sample of $N = 36$ residents. Because the population standard deviation, σ, was known, a z test was used to test statistical significance. The z value for this sample mean was statistically significant; $z = +2.40$, $p < .05$, one-tailed. Thus, given this sample mean, the null hypothesis that the mean IQ for the entire population of nursing home residents is equal to 100 was rejected.

Version 2 (report of exact *p* value):

For all significance tests in this research report, the significance level that was used was $\alpha = .05$, one-tailed. The sample mean IQ was $M = 106$ for a random sample of $N = 36$ residents. Because the population standard deviation, σ, was known, a z test was used to test statistical significance. The z value for this sample mean was statistically significant, $z = +2.40$, $p = .008$, one-tailed.

Most statistical programs report p values to three decimal places. Sometimes an exact obtained p value has zeros in the first three decimal places and appears as .000 on the printout. Note that p represents a theoretical probability of incorrectly rejecting H_0 and this theoretical risk is never 0, although it becomes smaller as the value of z increases. When SPSS shows a significance value of .000 on the printout, this should be reported as $p < .001$ rather than as $p = .000$.

3.4 ♦ Common Research Practices Inconsistent With Assumptions and Rules for NHST

NHST assumes that we have a random sample of N independent observations from the population of interest and that the scores on the X variable are quantitative and at least approximately interval/ratio level of measurement. It is desirable that the scores on the X variable be approximately normally distributed. If scores on X are at least approximately normally distributed, then the distribution of values of M approaches a normal distribution shape even for rather small values of N and rather small numbers of samples; however, even when scores on the original X variable are nonnormally distributed, the distribution of values of M is approximately normal, provided that the value of N in each sample is reasonably large and a large number of samples are obtained.

The rules for NHST (when the population standard deviation σ is known) involve the following steps. First select a nominal alpha level and state a null and an alternative hypothesis. Next, take a random sample from the population of interest and compute the sample mean, M; evaluate how far this sample mean, M, is from the hypothesized value of the population μ_{hyp} by calculating a z ratio. Then use a decision rule based on the alpha level and the nature of the alternative hypothesis (i.e., a one-tailed or two-tailed reject region) to evaluate the obtained z value to decide whether to reject H_0, or report the exact one- or two-tailed p value that corresponds to the obtained value of z and state the alpha level that should be used to decide whether this obtained p value is small enough to judge the outcome "statistically significant." The risk of Type I error should, theoretically, correspond to the nominal alpha level if only one significance test is conducted and the decision rules are formulated before looking at the numerical results in the sample.

The goal of NHST is to be able to make a reject/do not reject decision about a null hypothesis, with a known risk of committing a Type I error. The risk of Type I error should correspond to the nominal alpha level when the assumptions for NHST are satisfied and the rules are followed. Ideally, the researcher should be able to identify the population of interest, identify the members of that population, and obtain a truly random sample from the population of interest. The researcher should state the null and alternative hypotheses and select an alpha level and formulate a decision rule (the ranges of values of z for which H_0 will be rejected) *before* looking at the value of the sample mean, M, and a corresponding test statistic such as z. The researcher should conduct only one statistical significance test or only a limited number of statistical significance tests.

In practice, researchers often violate one or more of these assumptions and rules. When assumptions are violated and these rules are not followed, the actual risk of Type I error may be higher than the nominal $\alpha = .05$ level that the researcher sets up as the standard.

3.4.1 ♦ Use of Convenience Samples

One common violation of the rules for NHST involves sampling. Ideally, the sample should be selected randomly from a well-defined population. In laboratory research, investigators often work with convenience samples (as discussed in Chapter 1). The consequence of using convenience samples is that the convenience sample may not be representative of any specific real-world population. A convenience sample may not provide adequate information to test hypotheses about population means for any well-defined real-world population.

3.4.2 ♦ Modification of Decision Rules After the Initial Decision

Second, researchers occasionally change their decisions about the alpha level and/or whether to perform a two-tailed/nondirectional test, versus a one-tailed/directional test, after examining the obtained values of M and z. For example, a researcher who obtains a sample mean that corresponds to a z value of +1.88 would find that this z value does not lie in the range of values for the decision to reject H_0 if the criterion for significance that is used is $\alpha = .05$, two-tailed. However, if the researcher subsequently changes the alpha level to $\alpha = .10$, two-tailed, or to $\alpha = .05$, one-tailed (after examining the values of M and z), this would redefine the decision rules in a way that would make the outcome of the study "statistically significant." This is not a legitimate practice; just about any statistical outcome can be judged significant after the fact if the researcher is willing to increase the alpha value to make that judgment. Another way of redefining the decision rules after the fact would be to change H_0; for example, if the researcher cannot reject H_0: $\mu = 100$ and wants to reject H_0, he or she might change the null hypothesis to H_0: $\mu = 90$. It should be clear that reverse engineering the decision rules to reject H_0 based on the obtained value of M for a sample of data is not a legitimate practice, and doing this leads to a much higher risk of committing a Type I error.

3.4.3 ♦ Conducting Large Numbers of Significance Tests

Another common deviation from the rules for NHST occurs in exploratory research. In some exploratory studies, researchers run a large number of tests; for example, a researcher might want to evaluate whether each of 200 Pearson correlations differs significantly from hypothesized correlations of 0. One way of thinking about the nominal alpha level is as a prediction about the number of Type I errors that can be anticipated when large numbers of significance tests are performed. If we set up a decision rule that leads to a Type I error 5% of the time when H_0 is actually correct, it follows that—even if we generated all the scores in our study using a random number generator and, in the population, all the variables in our study had correlations of 0 with each other—we would expect that if we ran 100 correlations and tested each one for statistical significance, we would find approximately 5 of these 100 correlations statistically significant. In other words, we would expect to find about 5 significant correlations that are statistically significant even if we generate the data using a process for which the real population correlations are 0 among all pairs of variables. When a researcher reports a large number of significance tests, the likelihood that at least *one* of the outcomes is an instance of a Type I error increases as the number of tests increases.

The increased risk of Type I error that arises when a large number of statistical significance tests are performed is called an **inflated risk of Type I error.** When a researcher reports significance tests for 100 correlations using $\alpha = .05$ as the nominal criterion for the significance of each individual correlation, the true risk of committing at least one Type I error in the set of 100 statistical significance tests is typically substantially greater than the nominal alpha level of .05.

3.4.4 ♦ Impact of Violations of Assumptions on Risk of Type I Error

It is extremely important for researchers to understand that an exact p value, or a statement that the obtained p value is smaller than some preselected alpha value such

as $\alpha = .05$, involves making a theoretical estimate of the risk of committing a Type I error that is based on a large number of assumptions and conditions. An estimated risk of Type I error can be expected to be accurate only when a large number of assumptions about the procedure are met. *If* the researcher begins with a well-defined actual population; *if* the scores on the variable of interest are quantitative and reasonably close to interval/ratio level of measurement; *if* the researcher decides upon H_0, H_1, and α prior to examining the data; *if* the researcher obtains a random sample of independent observations from the population of interest; and *if* the researcher performs only one statistical significance test, *then* the preselected alpha level theoretically corresponds to the risk of Type I error. When one or more of these assumptions are not satisfied, then the actual risk of Type I error may be considerably higher than the nominal alpha level. Unfortunately, in many real-life research situations, one or more of these conditions are frequently not satisfied.

3.5 ♦ Strategies to Limit Risk of Type I Error

3.5.1 ♦ Use of Random and Representative Samples

For all types of research, both experimental and nonexperimental, samples should be representative of the populations about which the researcher wants to make inferences. When the population of interest can be clearly defined, all members can be identified, and a representative sample can be obtained by using random selection methods (possibly combined with systematic sampling procedures such as stratification), then the application of NHST logic to sample means provides a reasonable way of making inferences about corresponding population means. However, in some research domains, the use of convenience samples is common. As discussed in Chapter 1, some convenience samples may not correspond to any well-defined real-life population. This lack of correspondence between the sample in the study and any well-defined real-world population can make the use of NHST procedures invalid.

3.5.2 ♦ Adherence to the Rules for NHST

It is dishonest to change the decision standards (the alpha level and the choice of a directional vs. nondirectional research hypothesis) after examining the outcome of a study. In the extreme, if a researcher is willing to raise the alpha level high enough, just about any outcome can be judged statistically significant. The risk of Type I error can be limited by adhering to the rules for NHST, for example, deciding on the alpha level before looking at data.

3.5.3 ♦ Limit the Number of Significance Tests

A simple way of limiting the risk of Type I error is to conduct only one significance test or a limited number of significance tests. It is generally easier to limit the number of tests in experimental studies that involve manipulating just one or two variables and measuring just one or two outcomes; the number of possible analyses is limited by the small number of variables.

Researchers often find it more difficult to limit the number of significance tests in nonexperimental, exploratory studies that include measures of large numbers of variables. It is relatively common to report dozens, or even hundreds, of correlations in large-scale exploratory studies. One possible way for researchers to reduce the risk of Type I error in exploratory studies is to decide ahead of time on a limited number of analyses (instead of running a correlation of every variable with every other variable). Another strategy is to make specific predictions about the pattern of outcomes that is expected (e.g., what pairs of variables are expected to have correlations that are large positive, large negative, or close to 0?) and then to assess how closely the set of obtained correlations matches the predicted pattern (Westen & Rosenthal, 2003).

When researchers set out to do exploratory nonexperimental research, however, they often do not have enough theoretical or empirical background to make such detailed predictions about the pattern of outcomes. The next few sections describe other ways of trying to limit the risk of Type I error in studies that include large numbers of analyses.

3.5.4 ♦ Bonferroni-Corrected Per-Comparison Alpha Levels

A commonly used method in limiting the risk of Type I error when multiple significance tests are performed in either experimental or exploratory studies is the Bonferroni correction. Suppose that the researcher wants to conduct $k = 3$ different significance tests—for example, significance tests about the means on three variables such as intelligence, blood pressure, and depression for the nursing home population. The researcher wants the overall "experiment-wise" risk of Type I error for this entire set of $k = 3$ tests to be limited to .05. The overall or **experiment-wise** α is denoted by $\mathbf{EW}_\alpha = .05$. (The term *experiment-wise* is commonly used to describe an entire set of significance tests even when the study is not, strictly speaking, an experiment.) To limit the size of the experiment-wise Type I error risk, the researcher may decide to use a more conservative "corrected" alpha level for each test. The Bonferroni correction is quite simple. The per-comparison alpha level (PC_α) is given as follows:

$$PC_\alpha = EW_\alpha / k, \tag{3.8}$$

where

EW_α is the experiment-wise α, often set at $EW_\alpha = .05$, and

k is the number of significance tests performed in the entire experiment or study.

For example, if the researcher wanted to set up three different z tests (one each to test hypotheses about the population means for intelligence, blood pressure, and depression for the nursing home population) and keep the experiment-wise overall risk of Type I error limited to .05, the researcher could use a **per-comparison alpha** level of $PC_\alpha = EW_\alpha / k = .05/3 = .017$ as the criterion for statistical significance for each of the individual z tests.

The advantage of the Bonferroni correction procedure is its simplicity; it can be used in a wide range of different situations. The disadvantage is that when the number of tests (k) becomes very large, the per-comparison alpha levels become so small that very few

outcomes can be judged statistically significant. Relative to many other methods of trying to control the risk of Type I error, the **Bonferroni procedure** is very conservative. One way to make the Bonferroni procedure somewhat less conservative is to set the experiment-wise alpha to a larger value (such as $EW_\alpha = .10$ instead of .05).

3.5.5 ♦ Replication of Outcome in New Samples

A crucial consideration in both experimental and nonexperimental research is whether a statistically significant finding from a single study can be replicated in later studies. Even in a carefully controlled experiment that includes only one statistical significance test, a decision to reject H_0 may be an instance of Type I error. Because of sampling error, a single study cannot provide conclusive evidence for or against the null hypothesis; in addition, any single study may have methodological flaws. When a statistically significant difference is found between a sample mean and a population mean using the one-sample z test, it is important to replicate this finding across new samples. If the statistically significant finding is actually an instance of Type I error, it is not likely to occur repeatedly across new samples. If successive samples or successive studies replicate the difference, then each replication should gradually increase the researcher's confidence that the outcome is not attributable to Type I error.

3.5.6 ♦ Cross-Validation

Cross-validation is a method that is related to replication. In a cross-validation study, the researcher typically begins with a large sample. The cases in the sample are randomly divided into two separate datasets (a data sampling procedure in SPSS can be used to obtain a random sample, either a specific number of cases or a percentage of cases from the entire dataset). The first dataset may be subjected to extensive exploratory analyses; for example, many one-sample z tests may be run on different variables, or many correlations may be run. After running many exploratory analyses, the researcher chooses a limited number of analyses that are theoretically interesting. The researcher then runs that small number of analyses on the second half of the data, to assess whether significant z tests or correlations can be replicated in this "new" set of data. A cross-validation study is a useful way of trying to reduce the risk of Type I error. If a pair of variables X and Y are significantly correlated with each other when the researcher runs a hundred correlations on the first batch of data but are not significantly correlated in the second batch of data, it is reasonable to conclude that the first significant correlation may have been an instance of Type I error. If the X, Y correlation remains significant when this analysis is performed on a second, new set of data, it is less probable that this correlation is a Type I error.

To summarize, in experiments, the most commonly used method of controlling Type I error is to limit the number of significance tests that are conducted. If a large number of follow-up tests such as comparisons of many pairs of group means are performed, researchers may use **"protected" tests** such as the Bonferroni-corrected PC_α levels to control the risk of Type I error. In nonexperimental or exploratory studies that include measurements of large numbers of variables, researchers often do not limit the number of significance tests. However, they may try to limit the risk of Type I error by using Bonferroni-corrected alpha levels or by running cross-validation analyses. For both

experimental and nonexperimental research, the replication of statistically significant outcomes across new samples and new studies is extremely important; the results of a single study should not be viewed as conclusive. The results of any one study could be due to Type I error or methodological flaws; numerous successful replications of a finding gradually increase researcher confidence that the finding is not just an instance of Type I error.

3.6 ♦ Interpretation of Results

3.6.1 ♦ Interpretation of Null Results

Most authorities agree that when a study yields a nonsignificant result, it is not correct to conclude that we should "accept H_0." There are many possible explanations for a nonsignificant outcome. It is inappropriate to interpret the outcome of an individual study as evidence that H_0 is correct unless these other explanations can be ruled out, and in practice, it is very difficult to rule out many of these possible alternative explanations for a nonsignificant outcome. A significance test may yield nonsignificant results when the null hypothesis is false for numerous reasons. We can only make a case for the possible inference that the null hypothesis might be correct if we can rule out all the following alternative explanations for a nonsignificant outcome, and in practice, it is not possible to rule out all these alternative explanations completely.

1. The effect size that the researcher is trying to detect (e.g., the magnitude of the difference between μ and μ_{hyp}) is very small.
2. The number of cases in the study (N) may be too small to provide adequate statistical power for the significance test. Sample sizes that are too small to have sufficient statistical power are fairly common (Maxwell, 2004).
3. The measure of the outcome variable may be unreliable, invalid, or insensitive to the effects of an intervention.
4. In experiments that involve comparisons of outcomes for groups that receive different dosage levels or types of treatment, the manipulation of the independent variable may be weak, not implemented consistently, or not a valid manipulation of the theoretical construct of interest.
5. The relationship between variables is of a type that the analysis cannot detect (e.g., Pearson's r is not appropriate for detecting curvilinear relationships between variables).
6. A nonsignificant result can arise due to sampling error.

For example, suppose that a developmental psychologist would like to show that the cognitive outcomes for children who are cared for at home (Group 1) do not differ from the outcomes for children who spend at least 20 hours a week in day care (Group 2). If a t test is performed to compare mean cognitive test scores between these two groups and the result is nonsignificant, this result cannot be interpreted as proof that day care has no effect on cognitive outcomes. It is possible that the nonsignificant outcome of the study

was due to a small effect size, small sample sizes, variations in the way day care and home care were delivered, unreliable or invalid outcome measures, failure to include the outcome measures that would reflect differences in outcome, and a number of other limitations of the study.

A researcher can present evidence to try to discount each of these alternative explanations. For example, if a study has an N of 10,000 and used an outcome measure that is generally viewed as appropriate, reliable, and valid, these design factors strengthen the possibility that nonsignificant results might be evidence of a lack of difference in the population. However, the results of one study are not conclusive proof of the null hypothesis. If a nonsignificant difference in cognitive test score outcomes for day care versus home care groups can be replicated across many studies with large samples and good-quality outcome measures, then as evidence accumulates across repeated studies, the degree of belief that there may be no difference between the populations may gradually become stronger.

Usually, researchers try to avoid setting up studies that predict a nonsignificant outcome. Most researchers feel that a statistically significant outcome represents a "success," and a review of publication bias by Hubbard and Armstrong (1992) concluded that many journal editors are reluctant to publish results that are not statistically significant. If both researchers and journal editors tend to regard nonsignificant research outcomes as "failures," then it seems likely that Type I errors may be overrepresented among published studies and Type II errors may be overrepresented among unpublished studies. However, it is worth remembering that a "null" outcome is sometimes the correct answer to a research question.

3.6.2 ♦ Interpretation of Statistically Significant Results

The interpretation of a statistically significant outcome must be made as a carefully qualified statement. A study may yield a statistically significant outcome when there is really no effect in the population for a variety of reasons, such as the following:

1. A statistically significant outcome may arise due to sampling error. That is, even when the null hypothesis H_0: $\mu = \mu_{hyp}$ is correct, a value of the sample mean, M, that is quite different from μ_{hyp} can arise just due to sampling error or chance. By definition, when the nominal alpha level is set at .05, values of M that are far enough away from μ_{hyp} to meet the criterion for the decision to reject H_0 do occur about 5% of the time when the null hypothesis is actually correct. NHST procedures involve a (theoretically) known risk of Type I error; this theoretical level of risk is determined by the selection of an α level before the data are analyzed. However, the actual risk of Type I error corresponds to the nominal alpha level only when all the assumptions for NHST are met and the rules for NHST (such as conducting a limited number of significance tests) are followed.
2. Statistically significant outcomes sometimes occur due to experimenter expectancy effects. There is substantial evidence that when researchers have a preferred outcome for a study or an idea about a likely outcome for the study, they communicate these expectations to research participants and this, in turn, influences

research participants so that they tend to behave in the ways that the researcher expects (Rosenthal, 1966; Rosenthal & Rosnow, 1980). Experimenter expectancy effects occur in research with human participants, and they are even stronger in research with animal subjects. In addition, Rosenthal has reported that errors in data entry and computation tend to be in the direction of the researcher's hypothesis.

3. Statistically significant outcomes sometimes occur because of unrecognized confounds; that is, the treatment co-occurs with some other variable, and it is that other variable that influences the outcome.

If any of these problems are present, then drawing the conclusion that the variables are related in some broader population, based on the results of a study, is incorrect. The accuracy of an obtained p value as an estimate of the risk of Type I error is conditional; the p value reported on the computer printout is only an accurate estimate of the true risk of Type I error if all the assumptions involved in null hypothesis significance testing are satisfied. Unfortunately, in many studies, one or more of these assumptions are violated. When the assumptions involved in NHST are seriously violated, the nominal p values reported by SPSS or other computer programs usually underestimate the true risk of Type I error; sometimes this underestimation of risk is substantial. To evaluate whether a p value provides accurate information about the magnitude of risk of Type I error, researchers need to understand the logic and the assumptions of NHST, recognize how the procedures in their studies may violate those assumptions, and realize how violations of assumptions may make their nominal p values inaccurate estimates of the true risk of Type I error.

3.7 ♦ When Is a *t* Test Used Instead of a *z* Test?

When σ is not known and we have to use $SE_M = s/\sqrt{N}$ to estimate sampling error, we evaluate the distance between a sample mean, M, and the hypothesized population mean, μ, by setting up a t ratio. When we use SE_M to estimate σ_M, the size of the resulting t ratio is evaluated by looking up values in a t distribution with degrees of freedom $(df) = N - 1$, where N is the number of scores in the sample used to estimate M and s. When σ is not known and we use the sample standard deviation, s, to estimate the amount of sampling error, the test statistic for H_0: $\mu = \mu_{hyp}$ has a form similar to the z test that appeared in Equation 3.1; however, the ratio is denoted t (instead of z) as a reminder that a t distribution should be used (instead of the standard normal distribution of z scores) to evaluate tail areas or probabilities.

$$t = \frac{M - \mu_{hyp}}{SE_M}. \tag{3.9}$$

As discussed in Chapter 2, when we use SE_M in place of σ_M, this increases the magnitude of sampling error. We evaluate the size of a t ratio by looking at a t distribution with $N - 1$ df, where N is the number of scores in the sample used to estimate M and s.

Note the consistency in the logic across procedures. Back in Chapter 2, when we wanted to evaluate the location of an individual X score relative to a distribution of other individual

X scores, we set up a z ratio to evaluate the distance of an individual X score from the mean of a distribution of other individual X scores, in number of population standard deviation units ($z = (X - M)/\sigma$). The t ratio in Equation 3.9 provides similar information about the location of an individual sample mean, M, relative to a theoretical distribution of possible outcomes for M that is assumed to have a mean of μ_{hyp} and a standard error of SE_M.

In many research situations, the population standard deviation, σ, is not known. When σ is not known, we have to make several changes in the test procedure. First, we replace σ_M with SE_M, an estimate of sampling error based on the sample standard deviation, s. Second, we use a t distribution with $N - 1$ df (rather than the standard normal distribution) to look up areas that correspond to distances from the mean and to decide whether the sample mean, M, is far away from the hypothesized value, μ_{hyp}. A numerical example of a one-sample t test (using SPSS) appears in Section 3.10.

3.8 ♦ Effect Size

There are several different ways of describing effect size; two of these are discussed here. For the one-sample z test or t test, one way of describing the effect size is to simply look at the magnitude of the obtained difference between M, the sample mean, and μ_{hyp}, the hypothesized population mean, in the units that are used to measure the outcome variable of interest. When the variable is measured in meaningful units, this difference can be interpreted in terms of practical significance. A second way to describe effect size is by using a unit-free effect size such as Cohen's d.

3.8.1 ♦ Evaluation of "Practical" (vs. Statistical) Significance

It is useful to distinguish between "statistical significance" and "**practical**" or "**clinical significance**" (Kirk, 1996). A result that is statistically significant may be too small to have much real-world value. A difference between M and μ_{hyp} can be statistically significant and yet be too small in actual units to be of much practical or clinical significance. Consider research that has been done to compare the mean IQ for twins with the test norm population mean IQ for "singletons"—that is, individual birth children. Suppose that a researcher obtains a sample mean IQ of $M = 98.1$ for $N = 400$ twins (similar to results reported by Record, McKeown, & Edwards, 1970). This sample mean value of $M = 98.1$ can be evaluated relative to the mean IQ for the general population, using H_0: $\mu_{hyp} = 100$, $\sigma = 15$, and $N = 400$ and $\alpha = .05$, two-tailed. Given $\sigma = 15$ and $N = 400$, the value of $\sigma_M = .75$. For this hypothetical example $z = 98.1 - 100/.75 = 2.53$; this obtained value $z = 2.53$ exceeds the critical value of $z = +1.96$, and so this difference would be judged to be statistically significant. However, the actual difference in IQ between twins and singletons was only $98.1 - 100 = -1.9$ points. Is a difference of less than two IQ points large enough to be of any practical or clinical importance? Most researchers do not think this difference is large enough to cause concern. (Note that if N had been smaller, for example, if we recalculated the z test above using $N = 100$, this 1.9–IQ point difference would not be judged statistically significant.)

Statistical significance alone is not a guarantee of practical significance or usefulness (Vacha-Haase, 2001). For example, a meta-analysis of the results of studies on the effect

of short-term coaching on Scholastic Aptitude Test scores (Powers & Rock, 1993) suggested that the average difference in scores between the coached and control groups was of the order of 15 to 25 points. While this difference could be judged statistically significant in a study with a large N, the improvement in scores may be too small to be of much practical value to students. A 15-point improvement would not be likely to improve a student's chance of being accepted by a highly selective university.

On the other hand, when the N values are very small, a real effect of a treatment variable may not be detected even when it is relatively strong. For a z or t ratio to be large when the N values are small, it is necessary to have a very large difference between group means (a big difference in dosage levels) and/or very small variances within groups (good control over extraneous variables) to have a reasonable chance the t ratio will turn out to be large.

There are trade-offs among the design decisions that can affect the size of z or t. If a researcher knows ahead of time that the effect size he or she wants to detect may be quite small, then the researcher may want to make the sample large. The next few sections of this chapter provide more specific guidance for decisions about sample size, based on beliefs about the magnitude of the effect that the researcher wants to be able to detect.

3.8.2 ♦ Formal Effect-Size Index: Cohen's *d*

The problem with $(M - \mu_{hyp})$ as an effect-size index is that the size of the difference depends on the units of measurement, and this difference is not always easily interpretable. We may be able to obtain a more interpretable index of effect size by taking another piece of information into account—that is, the standard deviation of the scores on the variable of interest. For many purposes, such as comparing outcomes across different variables or across different studies, it may be useful to have an index of effect size that is "unit free" or standardized—that is, not dependent on the units in which the original measurements were made. Cohen's d is one possible effect-size index; it describes the difference between two means in terms of number of standard deviations, as shown in the following equation:

$$d = \frac{\mu - \mu_{hyp}}{\sigma} \text{ or } \frac{M - \mu_{hyp}}{\sigma}, \tag{3.10}$$

where μ is the unknown mean X score of the population about which inferences are to be made (based on a random sample drawn from that population); μ_{hyp} is the hypothesized population mean for the variable of interest, X; σ is the population standard deviation of scores on the variable of interest, X; and M is the mean of the X scores in the sample.

In the previous example, the sample mean IQ for twins was $M = 98.1$ IQ points, the hypothesized population mean (based on the mean IQ for the general population) was $\mu_{hyp} = 100$, and the standard deviation for IQ test scores in the population was $\sigma = 15$. The corresponding d effect size is $d = (98.1 - 100)/15 = -.13$; that is, the sample mean for twins was .13 standard deviations below the mean for the general population. When the population standard deviation, σ, is not known, the sample standard deviation, s, can be used to estimate it, giving the following equation for **Cohen's *d*:**

$$d = (M - \mu_{hyp})/s, \tag{3.11}$$

where M is the sample mean; μ_{hyp} is the hypothesized value of the population mean—that is, the specific numerical value in H_0, the null hypothesis; and s is the sample standard deviation of individual X scores.

It is useful to have an effect-size index (such as Cohen's d) that is independent of the size of N, the number of cases in the sample; it is also independent of the units in which the X variable is measured (these units of measurement are removed by dividing the difference between means by σ or s). There are three ways we can use an effect-size index. First, it is useful and informative to report effect-size information along with the results of a statistical significance test. Guidelines for reporting research results in some disciplines such as psychology now call for the inclusion of effect-size information along with statistical significance tests (APA, 2001). Second, effect-size information can be useful when making decisions about the minimum sample size that is needed to have adequate statistical power (as discussed in Section 3.9). Third, effect-size indexes that are unit free and independent of N provide information that can be summarized across studies using meta-analysis.

In later chapters, Cohen's d will also be used to index the magnitude of the difference between the means of two different populations, as illustrated in the following example concerning the mean height for males versus the mean height for females:

$$d = \frac{\mu_1 - \mu_2}{s_{\text{pooled}}}, \tag{3.12}$$

where μ_1 is the mean for Population 1 (e.g., mean height for males), μ_2 is the mean for Population 2 (e.g., mean height for females), and s_{pooled} is the overall standard deviation for the variable of interest, averaged or pooled across the two populations. This is typically estimated by averaging the sample variances as shown below:

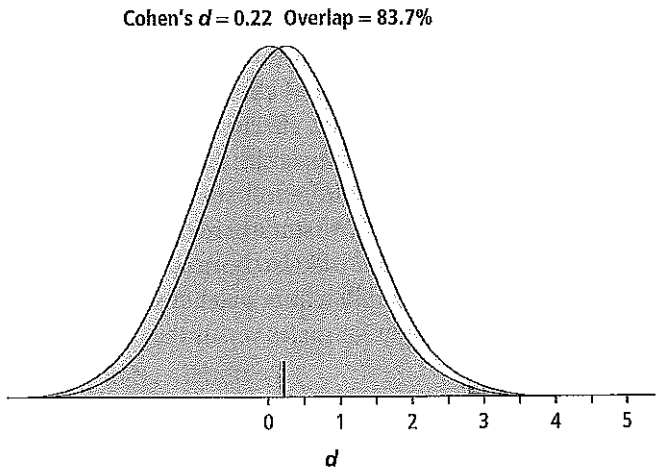
$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}, \tag{3.13}$$

where n_1 is the number of scores in the sample drawn from Population 1, s_1^2 is the sample variance in scores of the sample drawn from Population 1, n_2 is the number of scores in the sample drawn from Population 2, and s_2^2 is the sample variance for the sample drawn from Population 2.

Cohen's d can be illustrated graphically; the distance between the means of two normal distributions is related to the amount of overlap between the distributions. A larger value of d corresponds to a larger and therefore more easily detectable difference between two distributions of scores. For example, Kling, Hyde, Showers, and Buswell (1999) conducted a review of studies that examined differences in self-esteem between women and men; they found that across a large number of studies, the mean size of the Cohen's d effect size was approximately $d = .20$, which corresponds to a rather small effect size. Figure 3.4 illustrates the overlap in the distribution of self-esteem scores for male and female populations, with an effect size of $d = .22$; that is, the mean self-esteem for the male population is about two tenths of a standard deviation higher than the mean self-esteem for the female population. This difference is small; there is substantial overlap in the scores on self-esteem for males and

females. On the other hand, there is a much larger gender difference for height. Data reported on Wikipedia (http://en.wikipedia.org/wiki/Effect_size) suggest that the Cohen's *d* value for gender differences in height is about $d = 2.00$; this is a very large effect size. Figure 3.5 illustrates the amount of overlap in the distributions of female and male heights, given a Cohen's

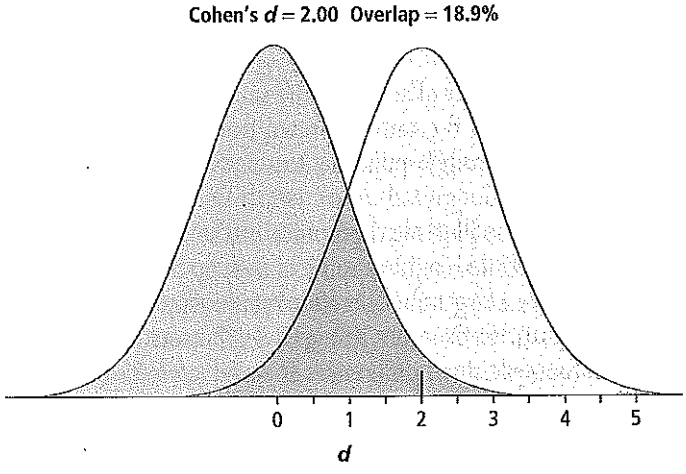
Figure 3.4 ◆ Example of a Small Effect Size: Cohen's *d* = .22



SOURCE: Kling, Hyde, Showers, and Buswell (1999).

NOTES: Across numerous studies, the average difference in self-esteem between male and female samples is estimated to be about .22; the mean self-esteem for males is typically about two tenths of a standard deviation higher than the mean self-esteem for females.

Figure 3.5 ◆ Example of a Large Effect Size: Cohen's *d* = 2.00



SOURCE: http://en.wikipedia.org/wiki/Effect_size

NOTES: From samples of men and women in the United Kingdom, mean height for males = 1,754 mm; mean height for females = 1,620 mm. The standard deviation for height = 67.5 mm. Therefore, Cohen's $d = (M_{male} - M_{female})/s = (1{,}754 - 1{,}620)/67.5 \approx 2.00$.

Table 3.2 ♦ Suggested Verbal Labels for Cohen's *d* Effect-Size Index in Behavioral and Social Science Research

Verbal Label	*Magnitude of* d
Small	$d \leq .20$
Medium	d of the order of .5 (e.g., d between .20 and .79)
Large	$d \geq .80$

NOTES: Population value of Cohen's $d = (\mu - \mu_{hyp})/\sigma$, where μ is the mean for the population from which the study sample was drawn, μ_{hyp} is the hypothesized value of the population mean, and σ is the standard deviation of scores in the population from which the sample was drawn. In other words, d is the distance in number of standard deviation units (σ) between the actual population mean (μ) and the hypothesized population mean stated in the null hypothesis (μ_{hyp}).

d value of 2.00; that is, the population mean for male height is about two standard deviations higher than the population mean for female height.

When the magnitude of a difference indexed by Cohen's d is small, the difference between means may be relatively difficult to detect. When the real magnitude of the difference between population means is relatively small, as in Figure 3.4, a researcher typically needs to have a relatively large sample to obtain a t or z value large enough to reject H_0 (the null hypothesis that the male and female population means are equal). On the other hand, when the magnitude of the difference between population means is large—for example, when Cohen's $d = 2.00$, as in Figure 3.5—a researcher may be able to obtain a t or z value large enough to reject H_0 even when the sample size is relatively small.

Suggested verbal labels for small, medium, and large values of Cohen's d in behavioral and social science research are summarized in Table 3.2.

3.9 ♦ Statistical Power Analysis

Statistical power is defined as the probability of obtaining a value of z or t that is large enough to reject H_0 when H_0 is actually false. In most (although not all) applications of NHST, researchers hope to reject H_0. In many experimental research situations, H_0 corresponds to an assumption that the treatment has no effect on the outcome variable; in many nonexperimental research situations, H_0 corresponds to an assumption that scores on a Y outcome variable are not predictable from scores on an X independent variable. Researchers often hope to demonstrate that a treatment does have an effect on some outcome variable and that they can predict scores on an outcome variable. In other words, they often hope to be able to reject a null hypothesis that states that there is no treatment effect or no relationship between variables.

Refer back to Table 3.1 to see the four possible outcomes when decisions are made whether to reject or not reject a null hypothesis. The outcome of interest, at this point, is the one in the upper right-hand corner of the table: the probability of correctly rejecting H_0 when H_0 is false, which is called *statistical power*. Note that within each column of Table 3.1, the probabilities of the two different possible outcomes sum to 1.00. Statistical power corresponds to $(1 - \beta)$, where β is the risk of committing a Type II error. The

factors that reduce the probability of committing a Type II error (β) also increase statistical power ($1 - \beta$). Researchers want statistical power to be reasonably high; often, statistical power of .80 is suggested as a reasonable goal.

The risk of committing a Type I error (α) is, in theory, established by the choice of a nominal alpha level (assuming that all assumptions for NHST are met and all the rules for NHST are followed). The risk of committing a Type II error (β), on the other hand, depends on several factors, including the nominal alpha level, the magnitude of the true population effect size (such as Cohen's *d*), and the sample size, *N*. Because statistical power ($1 - \beta$) is the complement of the risk of committing a Type II error (β), factors that decrease the risk of Type II error also increase statistical power.

We will begin with some general qualitative statements about how these factors (α, effect size, and *N*) are related to power and then give an empirical example to illustrate how statistical power varies as a function of these factors. In each case, when a statement is made about the effect of changes in one factor (such as *N*, sample size), we assume that the other factors are held constant. Here is a qualitative summary about factors that influence statistical power and the risk of committing a Type II error:

1. Assuming that the population effect size, *d*, and the sample size, *N*, remain constant, statistical power increases (and risk of Type II error decreases) as the value of the nominal alpha level is increased. Usually researchers set the nominal alpha level at .05. In theory, statistical power can be increased by raising the alpha level to $\alpha = .10$ or $\alpha = .20$. However, most researchers are unwilling to accept such high levels of risk of Type I error and, therefore, prefer changing other features of the research situation (rather than increasing α) to improve statistical power.

2. To see how the other two factors (the effect size, Cohen's *d;* the sample size, *N*) influence statistical power, it is useful to reexamine the equations that are used to compute the *t* ratio. Recall that the one-sample *t* test is calculated as follows: $t = (M - \mu_{hyp})/(s/\sqrt{N})$. We can "factor out" the term involving the square root of *N* to show that the size of *t* depends on two things: the magnitude of *d* and the magnitude of the square root of *N*.

   Recall that $t = (M - \mu_{hyp})/(s/\sqrt{N})$; this can be rearranged as follows:

$$\frac{M - \mu_{hyp}}{s} \times \frac{1}{1/\sqrt{N}}.$$

   Recall that $(M - \mu_{hyp})/s$ is Cohen's *d*. And note that $\frac{1}{(1/\sqrt{N})}$ can be simplified to $\sqrt{N}$. Thus, we have

$$t = d \times \sqrt{N}. \tag{3.14}$$

   In other words, the size of the *t* ratio is related to both the magnitude of the population effect size, Cohen's *d*, and the sample size, *N*. (This equation is similar to equations presented by Rosenthal & Rosnow, 1991, which show how the overall independent samples *t* test can be factored into two parts: one term, *d* in this

example, to represent effect size and the other term, square root of N in this example, to represent sample size.) If the sample size, N, is held constant, t increases as d increases. If the effect size, d, is held constant, t increases as $\sqrt{N}$ increases. Therefore, assuming that the sample size, N, remains constant, as the population effect size represented by Cohen's d increases, we expect that the size of the sample t ratio will also tend to increase. The implication of this is that, other factors being equal, we are more likely to obtain a large value of the t ratio (large enough to reject H_0) when the population effect size indexed by d is large.

We can summarize by saying that, other factors being equal, as the magnitude of the population effect size, d, increases, statistical power tends to increase, and the risk of committing a Type II error (β) tends to decrease.

3. It also follows from the previous argument that, if all other factors in the equation for t remain constant, as N increases, the size of the obtained t ratio will tend to be larger. Therefore, other factors being equal, as the size of N increases, statistical power tends to increase (and the risk of committing a Type II error, β, decreases).

The problem with the preceding qualitative statements about the connection between statistical power and the values of α, d, and N is that they do not take sampling error into account in a precise quantitative manner. We can evaluate statistical power more precisely by setting up a graph of the distributions of outcomes of t that are expected if H_0 is true and a separate graph of the distribution of outcomes of t that would be expected for a specific population effect size, d. In the following example, let's continue to consider testing hypotheses about intelligence scores. Suppose that the null hypothesis is

$$H_0: \mu = 100,$$

the sample standard deviation $s = 15$, and the sample size is $N = 10$ (therefore $df = 9$).

$$SE_M = 15/\sqrt{N} = 15/\sqrt{10} = 15/3.162 = 4.74\,.$$

Now let's suppose that the actual population mean is 115. This would make the value of Cohen's $d = [\mu - \mu_{hyp}]/\sigma = [115 - 100]/15 = 1.00$.

From the table of critical values for the t distribution, which appears in Appendix B, the critical values of t for $\alpha = .05$, two-tailed, and $df = 9$ are $t = +2.262$ and $t = -2.262$. Based on Equation 3.7, the critical values of M would therefore be

$$100 - 2.262 \times 4.74 = 89.28,$$

and

$$100 + 2.262 \times 4.74 = 110.72.$$

In other words, we would reject H_0: $\mu = 100$ if we obtain a sample mean, M, that is less than 89.28 or greater than 110.72.

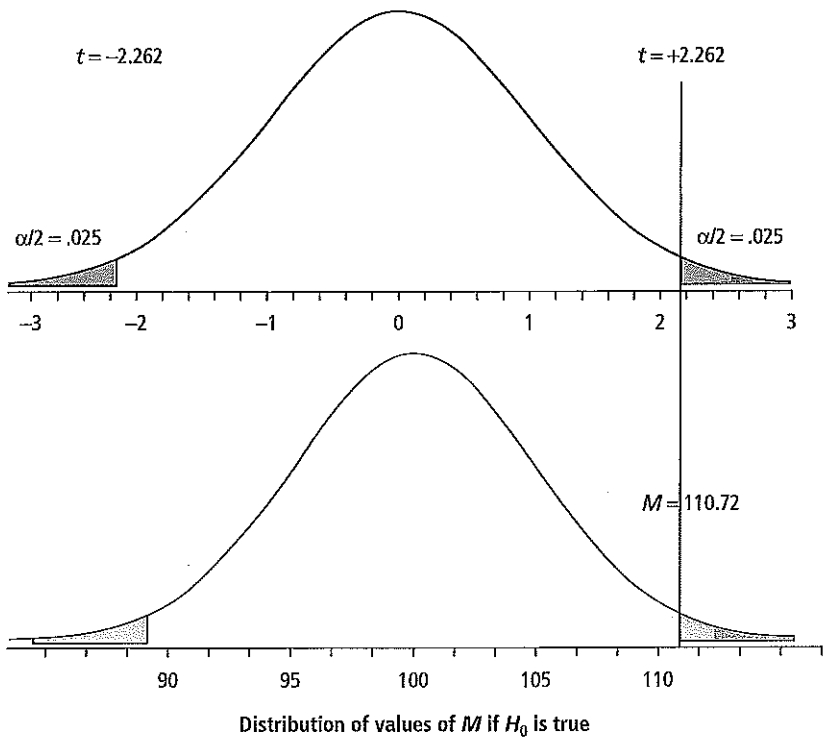
To evaluate statistical power, we need to think about two different possible distributions of outcomes for M, the sample mean. The first is the distribution of outcomes that

would be expected if H_0 were true; the "reject regions" for the statistical significance test are based on this first distribution. The second is the distribution of outcomes for M that we would expect to see if the effect size $d = 1.00$, that is, if the real population mean (115) were one standard deviation above the hypothesized population mean of 100 points.

The upper panel of Figure 3.6 shows the expected distribution of outcome values of t given H_0: $\mu = 100$, H_1: $\mu \neq 100$, $df = 9$, and $\alpha = .05$ (two-tailed). Using the fact that $N = 10$ and $df = 9$, we can find the critical values of t from the table of the t distribution in Appendix B. For $\alpha = .05$ (two-tailed) with 9 df, we would reject H_0 for values of $t > +2.262$ and for values of $t < -2.262$.

The lower panel of Figure 3.6 shows how these critical values of t correspond to values of M. Using Equation 3.7 along with knowledge of H_0 and SE_M, we can convert each critical value of t into a corresponding critical value of M. For example, a t value of +2.262 corresponds to a sample mean, M, of 110.72. The reject regions for H_0 can be given in terms of obtained values of M. We would reject H_0: $\mu = 100$ for values of $M > 110.72$ and for values of $M < 89.28$.

Figure 3.6 ♦ Diagram Showing Reject Regions for the Following Null Hypothesis: H_0: $\mu = 100$, $SE_M = 3$, $\alpha = .05$, Two-Tailed



NOTES: Reject regions (in terms of z values) are as follows: Reject H_0 for $z > +1.96$ and for $z < -1.96$. Reject regions (in terms of values of M) are as follows: Reject H_0 for $M < 89.28$ and for $M > 110.72$.

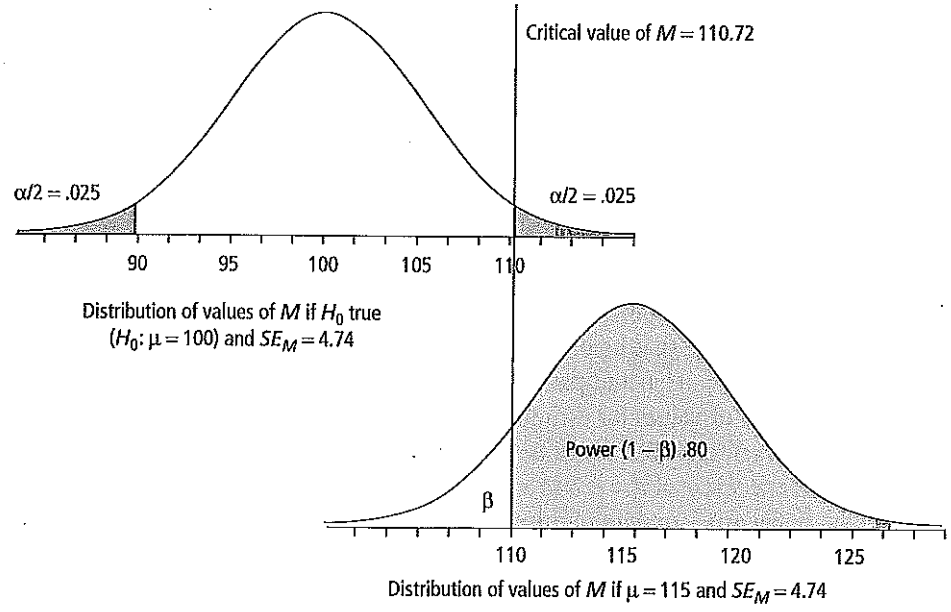
The preceding discussion shows how the distribution of outcomes for M that is (theoretically) expected when H_0 is assumed to be true is used to figure out the reject regions for H_0 (in terms of values of t or M).

The next step is to ask what values of M would be expected to occur if H_0 is false (e.g., one of many ways that H_0 can be false is if μ is actually equal to 115). An actual population mean of $\mu = 115$ corresponds to a Cohen's d effect size of +1 (i.e., the actual population mean $\mu = 115$ is one standard deviation higher than the value of $\mu_{hyp} = 100$ given in the null hypothesis).

The lower panel of Figure 3.7 illustrates the theoretical sampling distribution of M if the population mean is really equal to 115. We would expect most values of M to be fairly close to 115 if the real population mean is 115, and we can use SE_M to predict the amount of sampling error that is expected to arise for values of M across many samples.

The final step involves asking this question: Based on the distribution of outcomes for M that would be expected if μ is really equal to 115 (as shown in the bottom panel of Figure 3.7), how often would we expect to obtain values of the sample mean, M, that are larger than the critical value of $M = 110.72$ (as shown in the upper panel of Figure 3.7)? Note that values of M below the lower critical value of $M = 89.28$ would occur so rarely when μ really is equal to 115 that we can ignore this set of possible outcomes.

Figure 3.7 ♦ Statistical Power: Probability of Obtaining a Sample Value of M That Exceeds the Critical Value of M (for H_0: $\mu = 100$ and $\alpha = .05$, Two-Tailed)



NOTES: The upper distribution shows how values of M are expected to be distributed if H_0 is true and $\mu = 100$. The shaded regions in the upper and lower tails of the upper distribution correspond to the reject regions for this test. The lower distribution shows how values of M would actually be distributed if the population mean, μ, is actually 115; based on this distribution, we see that if μ is really 115, then about .80 or 80% of the outcomes for M would be expected to exceed the critical value of M (110.72).

To figure out the probability of obtaining sample means, M, greater than 110.72 when $\mu = 115$, we find the t ratio that tells us the distance between the "real" population mean, $\mu = 115$, and the critical value of $M = 110.72$. This t ratio is $t = (M - \mu)/SE_M = (110.72 - 115)/4.74 = -.90$. The likelihood that we will obtain a sample value for M that is large enough to be judged statistically significant given the decision rule developed previously (i.e., reject H_0 for $M > 110.72$) can now be evaluated by finding the proportion of the area in a t distribution with 9 df that lies to the right of $t = -.90$. Tables of the t distribution, such as the one in Appendix B, do not provide this information; however, it is easy to find Java applets on the Web that calculate exact tail areas for any specific value of t and df. Using one such applet, the proportion of the area to the left of $M = 110.72$ and $z = -.90$ (for the distribution centered at $\mu = 115$) was found to be .20 and the proportion of area to the right of $M = 110.72$ and $z = -.90$ was found to be .80. The shaded region on the right-hand side of the distribution in the lower part of Figure 3.7 corresponds to statistical power in this specific situation; that is, *if* we test the null hypothesis H_0: $\mu = 100$, using $\alpha = .05$, two-tailed, and a t test with $df = 9$, and *if* the real value of Cohen's $d = +1.00$ (i.e., the real population mean is equal to $\mu = 115$), *then* there is an 80% chance that we will obtain a value of M (and therefore a value of t) that is large enough to reject H_0.

We can use this logic to figure out what sample size, N, is required to achieve a specific desired level of statistical power (usually the desired level of power is at least 80%) when we are planning a study. Tables have been published that map out the minimum N needed to achieve various levels of statistical power as a function of the alpha level and the population effect size Cohen's d. An example of a statistical power table appears in Table 3.3.

Table 3.3 can be used to look up the statistical power that corresponds to the situation in the previous example. The previous example involved a single sample t test with an effect size $d = +1.00$, a sample size $N = 10$, and $\alpha = .05$, two-tailed. Table 3.3 provides estimates of statistical power for a one-sample t test with $\alpha = .05$, two-tailed. If you look up the value of d across the top of the table and find the column for $d = +1.00$, as well as look up the value of N in the rows of the table and find the row that corresponds to $N = 10$, the table entry that corresponds to the estimated power is .80. This agrees with the power estimate that was based on an examination of the distributions of the values of M shown in Figures 3.6 and 3.7.

The power table can be used to look up the minimum N value required to achieve a desired level of power. For example, suppose that a researcher believes that the magnitude of difference that he or she is trying to detect using a one-sample t test corresponds to Cohen's $d = +.50$ and plans to use $\alpha = .05$, two-tailed. The researcher can read down the column of values for estimated power under the column headed $d = +.50$. Based on the values in Table 3.3, the value of N required to have a statistical power of about .80 to detect an effect size of $d = +.5$ in a one-sample t test with $\alpha = .05$, two-tailed, is about $N = 35$.

The true strength of the effect that we are trying to detect—for example, the degree to which the actual population mean, μ, differs from the hypothesized value, μ_{hyp}, as indexed by the population value of Cohen's d—is usually not known. The sample size needed for adequate statistical power can be approximated only by making an educated guess about the true magnitude of the effect, as indexed by d. If that guess about the population effect size, d, is wrong, then the estimate of power based on that guess will also be wrong. Information from past studies can often be used to make at least approximate estimates of population effect size.

Table 3.3 ♦ Power Tables for the One-Sample *t* Test Using $\alpha = .05$, Two-Tailed

	Cohen's d																		
N	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*	*.50*	*.55*	*.60*	*.65*	*.70*	*.75*	*.80*	*.90*	*1.00*	*1.10*	*1.20*	*1.30*	*1.40*	*1.50*
3	.05	.05	.05	.05	.05	.05	.06	.06	.06	.06	.06	.06	.06	.07	.08	.08	.09	.10	.11	.13
4	.05	.05	.06	.06	.06	.07	.07	.08	.08	.09	.10	.10	.11	.14	.17	.21	.25	.31	.37	.44
5	.06	.06	.06	.07	.08	.09	.10	.11	.12	.13	.15	.17	.19	.25	.31	.39	.47	.55	.63	.70
6	.06	.07	.07	.08	.09	.11	.12	.14	.17	.19	.22	.25	.29	.37	.46	.55	.64	.72	.79	.84
7	.06	.07	.08	.10	.11	.13	.16	.18	.21	.25	.29	.33	.38	.48	.58	.67	.75	.82	.87	.91
8	.07	.08	.09	.11	.13	.16	.19	.23	.27	.31	.36	.41	.46	.57	.67	.76	.83	.88	.92	.95
9	.07	.09	.10	.13	.15	.19	.22	.27	.31	.37	.42	.48	.54	.65	.75	.83	.88	.93	.95	.97
10	.08	.09	.12	.14	.18	.21	.26	.31	.36	.42	.48	.54	.60	.71	.80	.87	.92	.95	.97	.98
11	.08	.10	.13	.16	.20	.24	.29	.35	.41	.47	.54	.60	.66	.77	.85	.91	.94	.97	.98	.99
12	.09	.11	.14	.17	.22	.27	.33	.39	.45	.52	.59	.65	.71	.81	.88	.93	.96	.98	.99	.99
13	.09	.12	.15	.19	.24	.30	.36	.42	.49	.56	.63	.70	.75	.85	.91	.95	.97	.99	.99	
14	.10	.13	.16	.21	.26	.32	.39	.46	.53	.61	.67	.74	.79	.88	.93	.96	.98	.99		
15	.10	.13	.17	.22	.28	.35	.42	.49	.57	.64	.71	.77	.82	.90	.95	.97	.99	.99		
16	.11	.14	.19	.24	.30	.37	.45	.53	.60	.68	.74	.80	.85	.92	.96	.98	.99			
17	.11	.15	.20	.26	.32	.40	.48	.56	.64	.71	.77	.83	.87	.93	.97	.99	.99			
18	.12	.16	.21	.27	.34	.42	.50	.59	.67	.74	.80	.85	.89	.95	.98	.99				
19	.12	.17	.22	.29	.36	.45	.53	.62	.69	.76	.82	.87	.91	.96	.98	.99				
20	.13	.17	.23	.30	.38	.47	.56	.64	.72	.79	.84	.89	.92	.97	.99	.99				
21	.13	.18	.24	.32	.40	.49	.58	.67	.74	.81	.86	.90	.94	.97	.99					
22	.14	.19	.26	.33	.42	.51	.60	.69	.76	.83	.88	.92	.95	.98	.99					
23	.14	.20	.27	.35	.44	.53	.63	.71	.78	.85	.89	.93	.95	.98	.99					
24	.15	.21	.28	.36	.46	.55	.65	.73	.80	.86	.91	.94	.96	.99						
30	.18	.25	.35	.45	.56	.66	.75	.83	.89	.93	.96	.98	.99							
40	.23	.33	.45	.58	.69	.79	.87	.92	.96	.98	.99									
50	.28	.41	.54	.68	.79	.88	.93	.97	.98	.99										
60	.33	.47	.63	.76	.86	.93	.97	.99	.99											
70	.37	.54	.70	.82	.91	.96	.98	.99												
80	.42	.60	.75	.87	.94	.98	.99													
90	.46	.65	.80	.91	.96	.99														
100	.51	.70	.84	.93	.98	.99														

SOURCE: Reprinted with permission from Dr. Victor Bissonnette.

When is statistical power analysis needed? It is a good idea whenever a researcher plans to conduct a study to think about whether the expected effect size, alpha level, and sample size are likely to be adequate to obtain reasonable statistical power. People who write proposals to compete for research funds from government grant agencies are generally required to include a rationale for decisions about planned sample size that takes statistical power into account. SPSS has an add-on program that can be used to estimate statistical power for more complex designs.

In summary, a researcher can do several things to try to reduce the magnitude of β (i.e., reduce the risk of Type II error and increase statistical power). The researcher can set a higher alpha level (but this trade-off, which involves increasing the risk of Type I error, is usually not considered acceptable). The researcher can increase the size of N, the sample size. Another way to reduce β and increase power that is possible in some but not all research situations is to increase the size of the difference that the researcher is trying to detect—that is, increase the effect size that corresponds to Cohen's d. One way of increasing the effect size is by increasing the difference $\mu - \mu_{hyp}$, that is, the difference between the actual population mean, μ, and the hypothesized value of the population mean, μ_{hyp}, given in the null hypothesis. Other factors being equal, as this difference $\mu - \mu_{hyp}$ increases, the likelihood of obtaining values of M and z large enough to reject H_0 also increases. Another way of increasing the effect size given by Cohen's d is through reduction of the within-group standard deviation, σ or s. Typically, s can be made smaller by selecting a homogeneous sample of participants and standardizing measurement procedures.

3.10 ♦ Numerical Results for a One-Sample *t* Test Obtained From SPSS

A one-sample t test (to test a null hypothesis about the value of the mean of one population) can be performed using SPSS. The data for this empirical example are in the SPSS file named wais.sav; this file contains hypothetical IQ scores for a sample of $N = 36$ randomly sampled residents from a nursing home population. In the following example, we will assume that the researcher does not know the value of σ, the population standard deviation, and must therefore use s, the sample standard deviation. Using the data from this sample, the researcher can do the following things:

1. Set up a CI for μ based on the sample values of M and s (using procedures reviewed in Chapter 2).
2. Test a specific null hypothesis about the value of μ, such as H_0: $\mu = 100$ (using a preselected alpha level such as .05 and a specific alternative hypothesis such as H_1: $\mu \neq 100$). When σ is known, a z test may be used (as described in Section 3.1). In the following example, we will assume that σ is not known and that we use the sample standard deviation s to set up a one-sample t test.
3. Calculate a sample estimate of Cohen's d as an index of effect size.

The menu selections that are needed to run a one-sample t test appear in Figure 3.8; from the top-level menu bar, choose <Analyze>, and then from the pull-down menus,

click on <Compare Means> and <One-Sample T Test>. These menu selections open up the SPSS dialog window for the One-Sample T Test, which appears in Figure 3.9. Initially, the names of all the variables in the dataset appear in the left-hand window; the variable for the one-sample t test is selected from this list and moved into the right-hand window (here, the variable is WAIS IQ score for each person in the sample). The hypothesized value of μ that is specified in the null hypothesis should be typed into the window that has the heading Test Value. In this example, H_0: $\mu = 100$, and therefore the test value is 100.

The p value that is reported by SPSS for the one-sample t test is a two-tailed or nondirectional p value. The Options button opens up a new dialog window where the user can specify the width of the CI; the default choice is a 95% CI (see Figure 3.10).

The results for the one-sample t test appear in Figure 3.11. The first panel reports the sample descriptive statistics; that is, the mean IQ for this sample was $M = 103.83$. This sample mean, M, is almost four points higher than the value of μ predicted in the null hypothesis, H_0: $\mu = 100$. The obtained t ratio was $t(35) = 2.46$, $p = .019$, two-tailed.

If the population mean, μ, really were equal to 100, we would expect that the means for 95% of all samples would lie rather close to 100. Our definition of "rather close" depends on the selected alpha level ($\alpha = .05$, two-tailed, in this example) and on the magnitude of the sampling error, which is given by the sample value of $SE_M = 1.556$. For a t distribution

Figure 3.8 ♦ SPSS Menu Selections for a One-Sample t Test

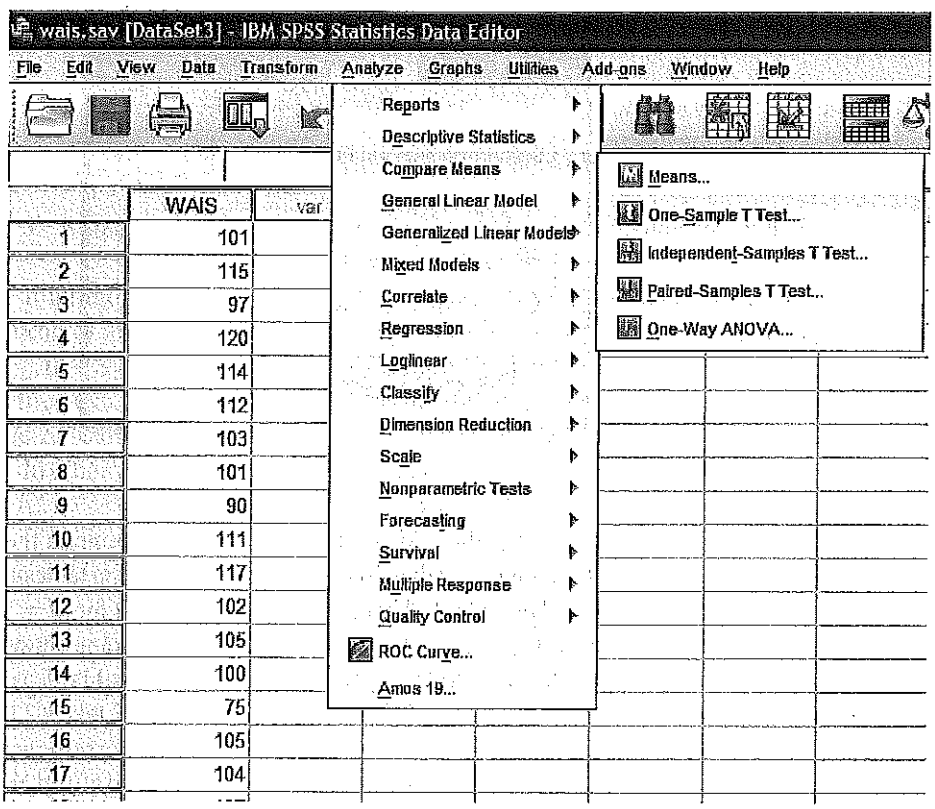


Figure 3.9 ♦ SPSS Dialog Window for the One-Sample *t* Test Procedure

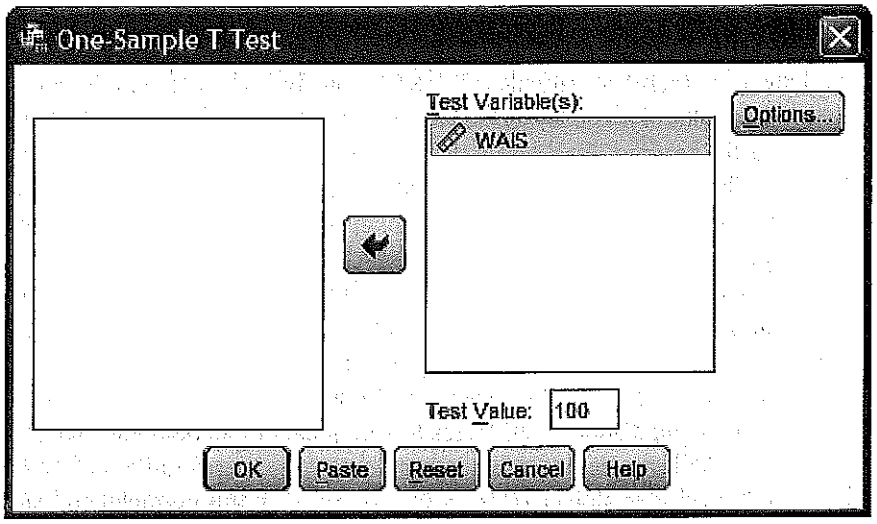
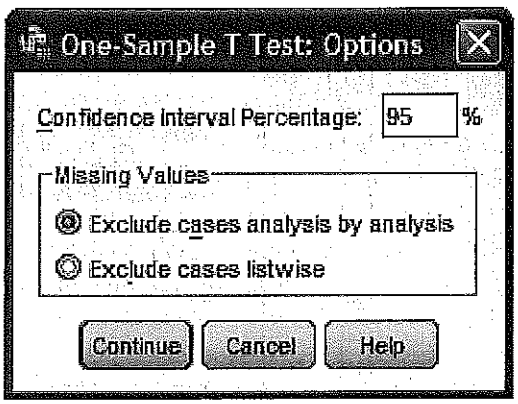


Figure 3.10 ♦ Options for the One-Sample *t* Test: Width of the Confidence Interval for the Sample Mean *M*



with $df = N - 1 = 36 - 1 = 35$, the critical values of t that correspond to the top 2.5% and the bottom 2.5% of the area of the t distribution can be found in the table of critical values for the t distribution in Appendix B. This table shows values of t for $df = 30$ and $df = 40$. We can figure out the value of t for $df = 35$ by taking the midpoint between the critical values of t that appear for $df = 30$ ($t_{critical} = 2.042$) and for $df = 40$ ($t_{critical} = 2.021$) in the column for $\alpha = .05$, two-tailed. The midpoint of these values is $(2.042 + 2.021)/2 = 2.031$. Thus, for a t distribution with 35 df, the critical values of t that correspond to the bottom 2.5% and top 2.5% of the area are as follows: The bottom 2.5% of the outcomes correspond to $t < -2.031$, and the top 2.5% of the outcomes correspond to $t > +2.031$.

Figure 3.11 ♦ SPSS Output for One-Sample *t* Test on Nursing Home Sample IQ Data

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
WAIS	36	103.83	9.337	1.556

One-Sample Test

	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
WAIS	2.463	35	.019	3.833	.67	6.99

Given all this information, we can predict what the distribution of outcomes of sample values of M would look like across hundreds or thousands of different samples when we use the sample value s (because we do not know the population value of σ). The distribution of values of M (if H_0 is true) is expected to have the following characteristics:

1. A mean of 100
2. A shape described by t with $df = 35$ (in other words, we use values from a t distribution with 35 df to figure out how distances from the mean are related to tail areas or proportions)
3. A standard deviation or standard error, SE_M, that depends on s and N

If all the conditions stated so far are correct, then we would expect that 95% of the outcomes for the sample mean, M, would lie between $\mu - 2.031 \times SE_M$ and $\mu + 2.031 \times SE_M$, that is, between $[100 - 2.031 \times 1.556]$ and $[100 + 2.031 \times 1.556]$—in this example, between 96.84 and 103.16. (Conversely, 2.5% of the values of M would be less than 96.84, and 2.5% of the values of M would be greater than 103.16.)

In other words, our criterion for "values of M that would be unlikely to occur if H_0 were true," in this case, is values of $M < 96.84$ and values of $M > 103.16$.

If H_0 is correct ($\mu = 100$) and if $SE_M = 1.556$, it would be quite unusual to see sample means less than 96.84 or greater than 103.16 for samples of size $N = 36$. When we use $\alpha = .05$, two-tailed, as our criterion for statistical significance, "quite unusual" corresponds to the most extreme 5% of outcomes (the bottom 2.5% and the top 2.5% of the distribution of likely values for M).

The obtained sample mean $M = 103.83$ is large enough in this example for it to fall within the top 2.5% of the distribution of predicted outcomes for M (for the null hypothesis H_0: $\mu = 100$). The obtained t ratio, $t = 2.463$, tells us the distance (in number of standard errors) of the sample mean, M, from the hypothesized value $\mu = 100$. The exact p value given on the

SPSS printout ($p = .019$, two-tailed) corresponds to the theoretical probability of obtaining a sample mean that is more than 2.463 standard errors above or below the hypothesized population mean, if H_0 is true and all our other assumptions are correct. If μ really were equal to 100 for the entire nursing home population, the probability of obtaining a sample mean this far away from μ just due to sampling error is $p = .019$—that is, less than 2%. In other words, this outcome $M = 103.83$ is an outcome that is very unlikely to occur if H_0 is true. We can, therefore, reject H_0: $\mu = 100$ (because the outcome of the study, $M = 103.83$, would be an unlikely outcome if this null hypothesis were correct).

A CI was set up for the value of the difference between the unknown population mean, μ, and the hypothesized value $\mu = 100$; this 95% CI has a lower boundary of .67 and an upper boundary of 6.99. In other words, given the size of the sample mean IQ, the difference between the unknown population mean IQ, μ, for the entire population of nursing home residents and the hypothesized IQ $\mu_{hyp} = 100$ probably lies between .67 and 6.99 points. As discussed in Chapter 2, in the long run, 95% of the CIs set up using the value of μ, the critical values of t, and the sample value of SE_M should actually contain the true population mean, μ. It seems reasonable in this case to conclude that the true magnitude of the difference between the mean IQ for the nursing home population and the value of 100 stated in the null hypothesis is probably between .67 points and 6.99 points; the nursing home population IQ is probably a few points higher than the value of 100 points given in the null hypothesis.

Finally, we can also report the outcome for this study using Cohen's d to index the effect size. In this example, we have $M = 103.83$, $\mu_{hyp} = 100$, and $s = 9.337$; $d = (103.83 - 100)/9.337 = .41$. Using the verbal labels for effect size in Table 3.2, this would be judged a moderate effect size.

3.11 ◆ Guidelines for Reporting Results

An APA Task Force report (Wilkinson & Task Force on Statistical Inference, 1999) concluded that reports of significance tests alone are not sufficient; additional information about the outcomes of studies (CIs and effect-size information) should be reported. The fifth edition of the APA publication manual (2001) says,

> When reporting inferential statistics (e.g., t tests, F tests, and chi-square), include information about the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme as or more extreme than the one obtained, and the direction of the effect. Be sure to include sufficient descriptive statistics (e.g., per-cell sample size, means, correlations, standard deviations) so that the nature of the effect being reported can be understood by the reader and for future meta-analysis. This information is important, even if no significant effect is reported. When point estimates are provided, always include an associated measure of variability (precision) specifying its nature (e.g., the standard error). (p. 22)

The APA guidelines also state that the reporting of CIs is strongly recommended. Note that certain types of information discussed in this chapter are usually not included; for example, it is uncommon to see statements of a formal null hypothesis in a journal article.

3.12 ◆ Summary

3.12.1 ◆ Logical Problems With NHST

The reasons why NHST procedures are problematic are complex; some of the potential problems with NHST are mentioned here. To begin with, the basic logic involved in using the proposition that "the results in the sample would be unlikely to occur if H_0 is true" to make the decision to reject H_0 is controversial; Sir Ronald Fisher argued in favor of making a decision to reject H_0 in this situation, but other major figures such as Karl Pearson argued against this reasoning.

Even though NHST procedures involve looking for disconfirmatory evidence, they do not involve the kind of theory testing that Karl Popper advocated when he said that theories need to be potentially falsifiable (see Meehl, 1978). Usually researchers want to reject H_0, but that is not usually equivalent to falsifying a theory. In fact, the decision to reject H_0 is often interpreted as (indirect) support for a theory. The researcher's theory more often corresponds to the prediction made in the alternate or research hypothesis than to the prediction made in the null hypothesis. Some theorists argue that in a sense, H_0 is virtually always false (see Krueger, 2001, for a discussion of this point). If that is the case, then results of NHST will always turn out to be statistically significant if sample sizes are made large enough.

In practice, many of the assumptions for NHST are frequently violated; for example, samples are often not randomly selected from any real population, and researchers often report large numbers of significance tests. The desire to obtain statistically significant results can tempt researchers to engage in "data fishing"; researchers may "massage" their data (e.g., by running many different analyses or by deleting extreme scores) until they manage to obtain statistically significant results. When any of these violations of assumptions are present, researchers should explain that their reported p values do not accurately represent the true risk of incorrectly rejecting H_0. Another name for these researcher behaviors is "data torturing" (Mills, 1993).

Misunderstandings about the meaning of a statistically significant result are fairly common. A few people mistakenly think that p is the probability that the null hypothesis is true, that $1 - p$ is the probability that the alternative hypothesis is true, or that $1 - p$ represents some sort of probability that the results of a study are replicable.[5] None of these interpretations are correct. These misunderstandings occur partly because statistical significance tests do not tell us what we really want to know. As Cohen (1994) said,

> What we want to know is "Given these data, what is the probability that H_0 is true?" But as most of us know, what [NHST] tells us is, "Given that H_0 is true, what is the probability of these (or more extreme) data?" These are not the same, as has been pointed out many times over the years. (p. 997)

In addition to difficulties and disputes about the logic of statistical significance testing, there are additional reasons why the results of a single study should not be interpreted as conclusive evidence that the null hypothesis is either true or false. A study can be flawed in many ways that make the results uninformative, and even when a study is well designed and carefully conducted, statistically significant outcomes sometimes arise just by

chance. Therefore, the results of a single study should never be treated as conclusive evidence. To have enough evidence to be confident that we know how variables are related, it is necessary to have many replications of a result based on methodologically rigorous studies.

Despite potential logical and practical problems with NHST, the APA Task Force did not recommend that this approach to evaluate data should be entirely abandoned. NHST can help researchers to evaluate whether chance or sampling error is a likely explanation for an observed outcome of a study, and at a minimum, researchers should be able to rule out chance as an explanation for their results before they begin to suggest other interpretations. Instead of abandoning NHST, which can provide useful information about the expected magnitude of sampling error when it is judiciously applied and cautiously and appropriately interpreted, the APA Task Force advocated more complete reporting of other information about the outcomes of studies, including CIs and effect-size information.

3.12.2 ◆ Other Applications of the *t* Ratio

The most general form of the *t* ratio is as follows:

$$t = \frac{\text{Sample statistic} - \text{Hypothesized population parameter}}{SE_{\text{sample statistic}}}. \tag{3.15}$$

In other words, in its most general form, a *t* test provides information about the magnitude of the difference between a sample statistic (such as M) and the corresponding value of the population parameter that is given in the null hypothesis (such as μ_{hyp}) in number of standard errors. The first application of this test that is generally covered in introductory statistics involves using a *t* ratio to evaluate whether the mean of a single sample (M) is close to, or far from, a hypothesized value of the population mean, μ_{hyp}. However, this test can be applied to several other common sample statistics.

For example, a very common type of experiment involves random assignment of participants to two groups, administration of different treatments to the two groups, and comparisons of the sample means on the outcome variable for Group 1 versus Group 2 (M_1 vs. M_2) to assess whether the difference between means is statistically significant. The population parameter of interest is $(\mu_1 - \mu_2)$, and the null hypothesis (that the means of the treatment and control populations are the same) is H_0: $(\mu_1 - \mu_2) = 0$. The sample statistic of interest is $M_1 - M_2$. The formula for the independent samples *t* ratio for testing a hypothesis about the equality of the means of two separate populations is as follows (note that this is just a specific case of the general form of *t* shown in Equation 3.15):

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{SE_{M_1 - M_2}}. \tag{3.16}$$

Because the difference between the means of populations that receive different treatments in an experiment $(\mu_1 - \mu_2)$ is usually hypothesized to be 0, this is usually reduced to

$$t = \frac{(M_1 - M_2)}{SE_{M_1 - M_2}}. \tag{3.17}$$

Again, the basic idea is simple. The researcher assesses whether the sample statistic outcome $(M_1 - M_2)$ is close to or far from the corresponding hypothesized population parameter (usually 0) by setting up a t ratio. The t ratio is the difference $M_1 - M_2$ in terms of number of standard errors. (Note again the parallel to the evaluation of the location of a single score using a z value, $z = (X - M)/s$. When we look at a z score for a single X value, we want to know how far X is from M in number of standard deviations; when we set up a t ratio, we want to know how far $M_1 - M_2$ is from $\mu_1 - \mu_2$ in number of standard errors.) A large-difference $M_1 - M_2$ (and a correspondingly large t ratio) is interpreted as evidence that the results of the study are inconsistent with the null hypothesis. As in the preceding section, the critical value for t is obtained from a table of the appropriate t distribution. For the independent samples t test, the applicable $df = n_1 + n_2 - 2$, where n_1 and n_2 are the numbers of participants in Groups 1 and 2, respectively.

The t ratio for testing a hypothesis about a Pearson correlation between two variables is yet another variation on the general form of the t ratio shown in Equation 2.14:

$$t = \frac{r - \rho_{hyp}}{SE_r}. \tag{3.18}$$

(*Note:* ρ_{hyp} is the unknown population correlation, which we try to estimate by looking at the value of r in our sample. The test formula in Equation 3.18 works only for $\rho_{hyp} = 0$, not for other hypothesized values of the population correlation. More information about hypothesis tests for correlations appears in Chapters 8 and 9.)

The t test for assessing whether the (raw score) regression slope in an equation of the form $Y' = b_0 + b_1 X$ differs from 0 also has the same basic form:

$$t = \frac{b_1 - b_{hyp}}{SE_{b_1}}, \tag{3.19}$$

where b_1 is the sample estimate of the regression slope, SE_{b_1} is the standard error of the slope estimate, and b_{hyp} is the corresponding hypothesized population slope, which is usually hypothesized to be 0.

3.12.3 ◆ What Does It Mean to Say "$p < .05$"?

At this point, let's return to the fundamental question, What does it mean to say "$p < .05$"? An obtained p value represents a (theoretical) risk of Type I error; researchers want this risk of error to be low, and usually that means they want p to be less than .05. A very brief answer to the question, What does it mean to say "$p < .05$"? is as follows: *A* p *value is a theoretical estimate of the probability (or risk) of committing a Type I error.* A Type I error occurs when the null hypothesis, H_0, is true and the researcher decides to reject H_0. However, researchers need to be aware that the "exact" p values that are given on computer printouts may seriously underestimate the true risk of Type I error when the assumptions for NHST are not met and the rules for carrying out statistical significance tests are not followed.

Notes

1. Although the use of NHST involves setting up a decision where researchers tend to look for "disconfirmatory" evidence, NHST should not be confused with the kind of falsification that Karl Popper advocated as a preferred scientific method. In many applications of NHST, the prediction made by the researcher's theory corresponds to H_1 rather than to H_0. Therefore, a decision to reject H_0 is often interpreted as support for a theory (rather than falsification of the researcher's theory). The logic involved in NHST is problematic even when it is well understood, and it is often misunderstood. See Kline (2004) for further discussion of the logical problems involved in NHST.

2. There is a difference between the conclusion "Do not reject H_0" and the statement "Accept H_0." The latter statement ("Accept H_0") is generally considered an inappropriate interpretation; that is, it is too strong a conclusion. Most textbooks explicitly say that researchers should not report their conclusions as "Accept H_0."

3. Beginning in Chapter 5, the null hypothesis generally corresponds to a statement that scores on a predictor and outcome variable are not related. For example, in experiments, H_0 generally corresponds to a hypothesis that the mean outcome scores are the same across groups that receive different treatments or, in other words, the hypothesis that the manipulated treatment variable has no effect on outcomes. In nonexperimental studies, H_0 generally corresponds to the null hypothesis that scores on an outcome variable Y cannot be predicted from scores on an independent variable X. In most experiments, researchers hope to obtain evidence that a treatment does have an effect on outcomes, and in most nonexperimental research, researchers hope to obtain evidence that they can predict scores on outcome variables. Thus, in later applications of NHST, researchers usually hope to reject H_0 (because H_0 corresponds to "no treatment effect" or "no relationship between scores on the predictor and outcome variables").

4. This value was obtained by linear interpolation. The tabled values of z and their corresponding tail areas were as follows:

Value of z	*Value of* p, *Tail Area to the Right of* z
1.64	.045
1.65	.055

The desired exact tail area of .05 is halfway between the two table entries for the tail area; that is, .05 is (.045 + .055)/2. We can therefore find the corresponding exact value of z by finding the midpoint between the two values of z that appear in the table. By linear interpolation, a z score of $(1.64 + 1.65)/2 = 1.645$ corresponds to a tail area of .05.

5. It is not correct to interpret $1 - p$ as "the probability of replicating a significant outcome in future studies" because the probability of obtaining statistically significant outcomes (statistical power) in future studies depends on other factors such as N (sample size). Killeen (2005) suggested that $(1 - p)$ might be *related* to the probability of future replication, and for a time, the editorial board of the journal *Psychological Science* called for the inclusion of p_{rep}, that is, a calculated probability of replication of results. Subsequent critiques such as Maraun and Gabriel (2010) have questioned Killeen's reasoning; the use of Killeen's p_{rep} is not recommended.

Comprehension Questions

1. What research decision influences the magnitude of risk of a Type I error?
2. What factors influence the magnitude of risk of a Type II error?
3. How are the risk of a Type II error and the statistical power related?
4. Other factors being equal, which type of significance test requires a value of t that is larger (in absolute value) to reject H_0—a directional or a nondirectional test?
5. In your own words, what does it mean to say "$p < .05$"?
6. Describe at least two potential problems with NHST.
7. What is a null hypothesis? An alternative hypothesis?
8. What is an alpha level? What determines the value of α?
9. What is an "exact" p value?
10. What is the difference between a directional and a nondirectional significance test?
11. Why do reported or "nominal" p values often seriously underestimate the true risk of a Type I error?
12. What is statistical power? What information is needed to decide what sample size is required to obtain some desired level of power (such as 80%)?
13. What recommendations did the APA Task Force (Wilkinson & Task Force on Statistical Inference, 1999) make about reporting statistical results? Are significance tests alone sufficient?
14. What conclusions can be drawn from a study with a null result?
15. What conclusions can be drawn from a study with a "statistically significant" result?
16. Briefly discuss: What information do you look at to evaluate whether an effect obtained in an experiment is large enough to have "practical" or "clinical" significance?
17. When a researcher reports a p value, "p" stands for "probability" or risk.
    a. What probability does this p refer to?
    b. Do we typically want p to be large or small?
    c. What is the conventional standard for an "acceptably small" p value?
18. Suppose a researcher writes in a journal article that "the obtained p was $p = .032$; thus, there is only a 3.2% chance that the null hypothesis is correct." Is this a correct or incorrect statement?

19. A p value can be interpreted as a (conditional) risk that a decision to reject H_0 is a Type I error, but the p values reported in research papers are valid indications of the true risk of Type I error *only if* the data meet the assumptions for the test and the researcher has followed the rules that govern the use of significance tests. Identify *one* of the most common researcher behaviors that make the actual risk of Type I error much higher than the "nominal" risk of Type I error that is set by choosing an alpha level.

20. Use Table 3.3: Suppose you are planning to do a study where you will use a one-sample t test. On the basis of past research, you think that the effect size you are trying to detect may be approximately equal to Cohen's $d = .30$. You plan to use $\alpha = .05$, nondirectional (two-tailed). (a) If you want to have power of .80, what minimum N do you need in the sample? (b) If you can afford to have only $N = 20$ participants in the sample, approximately what is the expected level of statistical power?