CHAPTER 7

Making Sense of Statistical Significance

Effect Size and Statistical Power

Chapter Outline

Statistical significance is extremely important in behavioral and social science research, but sophisticated researchers and readers of research understand that there is more to the story of a research result than $p < .05$ or *ns* (not significant). This chapter helps you become sophisticated about statistical significance. Gaining this sophistication means learning about two closely interrelated issues: effect size and statistical power.

TIP FOR SUCCESS

This chapter builds directly on Chapters 5 and 6. We do not recommend embarking on this chapter until you have a good understanding of the key material in those chapters, especially hypothesis testing and the distribution of means.

Effect Size

Consider again the example from Chapter 6 of giving special instructions to fifth-graders taking a standard achievement test. In the hypothesis-testing process for this example (the *Z* test), we compared two populations:

Population 1: Fifth-graders receiving special instructions.
Population 2: Fifth-graders in general (who do not get the special instructions).

The research hypothesis was that Population 1 would have a higher mean score on the test than Population 2. Population 2 (that is, how fifth-graders perform on this test when given in the usual way) is known to have a mean of 200. In the example, we said the researchers found that their sample of 64 fifth-graders who were given the special instructions had a mean score on the test of 220. Following the hypothesis-testing procedure, we rejected the null hypothesis that the two populations are the same. This was because it was extremely unlikely that we would get a sample with a score as high as 220 from a population like Population 2 (see Figure 7–1, which is the same as Figure 6–7 from the last chapter). Thus, we could conclude the result is "statistically significant." In this example, our best estimate of the mean of Population 1 is the sample's mean, which is 220. Thus, we

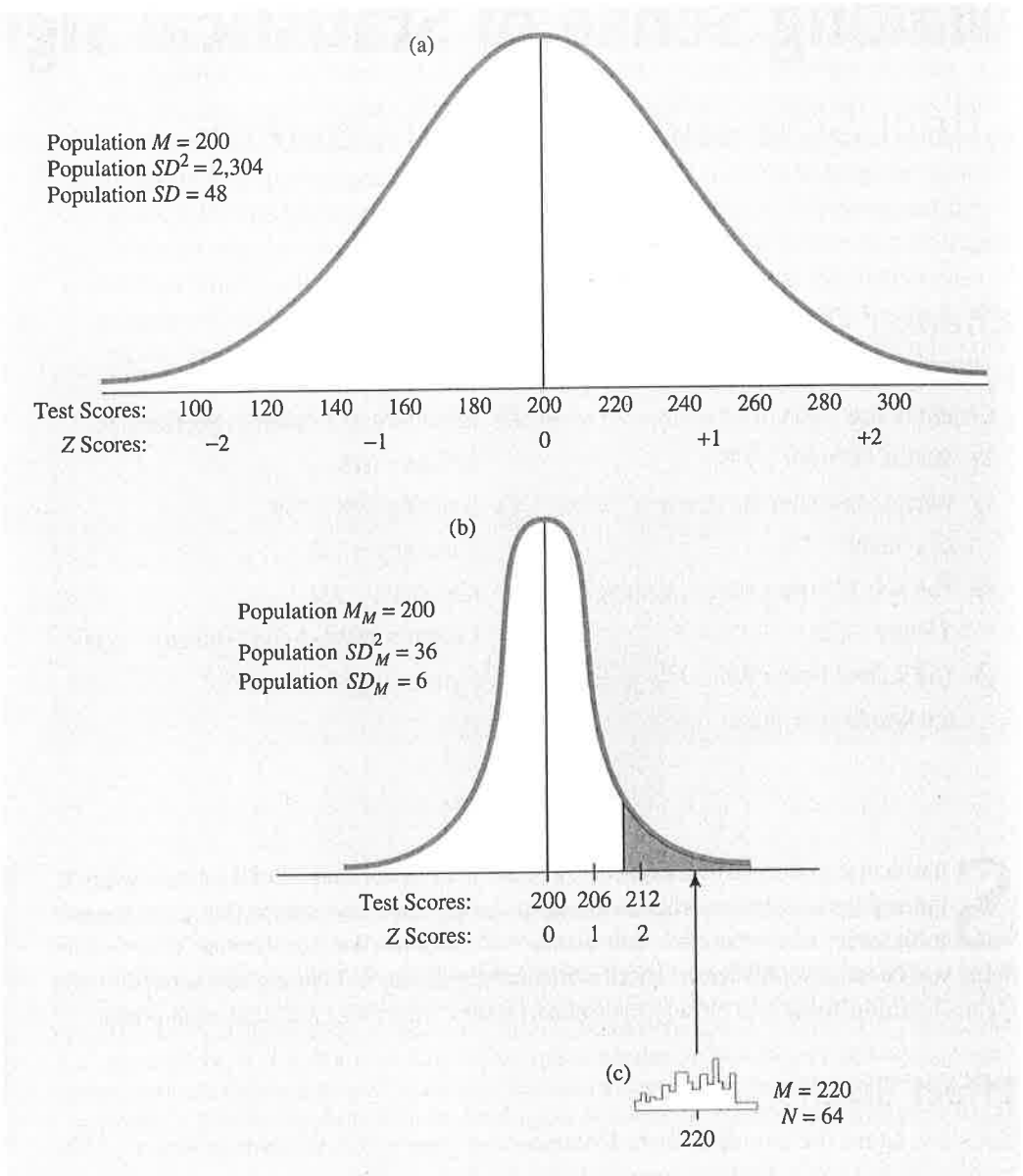

Figure 7–1 For the fictional study of fifth-graders' performance on a standard school achievement test, (a) the distribution of the population of individuals, (b) the distribution of means (the comparison distribution), and (c) the sample's distribution. The shaded area in the distribution of means is the rejection region—the area in which the null hypothesis will be rejected if the study sample mean turns out to be in that area. (See discussion in Chapter 6.)

can estimate that giving the special instructions has an average effect of increasing a fifth-grader's score by 20 points.

Now look again at Figure 7–1. Suppose the sample's score had been only 210. This would have had a Z score of 1.67 [$(210 - 200)/6 = 1.67$]. This is more extreme than the cutoff in this example, which was 1.64, so the result would still have been significant. However, in this situation we would estimate that the average effect of the special instructions was only 10 points.

Notice that both results are significant, but in one example the effect is twice as big as in the other example. The point is that knowing statistical significance does not give you much information about the *size* of the effect. Significance tells us that the results of the experiment should convince us that there *is* an effect (that it is not "due to chance"). But significance does not tell us how *big* this nonchance effect is.

Put another way, **effect size** is a measure of the difference between populations. You can think of effect size as how much something changes after a specific intervention. Effect size indicates the extent to which two populations do *not* overlap—that is, how much they are separated due to the experimental procedure. In the fifth-grader example, Population 2 (the known population) had a mean of 200; based on our original sample's mean of 220, we estimated that Population 1 (those getting the special instructions) would have a mean of 220. The left curve in Figure 7–2 (page 206) is the distribution (*of individual scores*) for Population 2; the right curve is the distribution for Population 1. Now look at Figure 7–3 (page 206). Again, the left curve is for Population 2 and is the same as in Figure 7–2. However, this time the right curve for Population 1 is estimated based on a sample (the sample getting the special instructions) with a mean of 210. Here you can see that the effect size is smaller and that the two populations overlap even more. The amount that two populations do not overlap is called the effect size because it is the extent to which the experimental procedure has an *effect* of separating the two populations.

We often very much want to know not only whether a result is significant, but how big the effect is. As we mentioned in the last chapter (and discuss in more detail later), an effect could well be statistically significant but not of much practical significance. (For example, suppose an increase of only 10 points on the test is not considered important.) Also, as you will see later in the chapter, effect size plays an important role in two other important statistical topics: meta-analysis and power.

Figuring Effect Size

You just learned that effect size is a measure of the difference between two population means. In Figure 7–2, the effect size is shown as the difference between the Population 1 mean and the Population 2 mean, which is 20 (that is, $220 - 200 = 20$). This effect size of 20 is called a *raw score effect size,* because the effect size is given in terms of the raw score on the measure (which, in this case, is an achievement test score, from a low of 0 to a high of, say, 300). But what if you want to compare this effect size with the result of a similar study that used a different achievement test? This similar study used a test with possible scores from 0 to 100, and the researchers reported an estimated Population 2 mean of 80, a Population 1 mean of 85, and a population standard deviation of 10? The raw score effect size in this study is 5 (that is, $85 - 80 = 5$). How do we compare this raw score effect size of 5 with the raw score effect size of 20 in our original study? The solution to this problem is to use a *standardized effect size*—that is, to divide the raw score effect size for each study by each study's population standard deviation.

In the original example of giving special instructions to fifth-graders taking a standard achievement test, the population standard deviation (of individuals) was 48. Thus,

effect size Standardized measure of difference (lack of overlap) between populations. Effect size increases with greater differences between means.

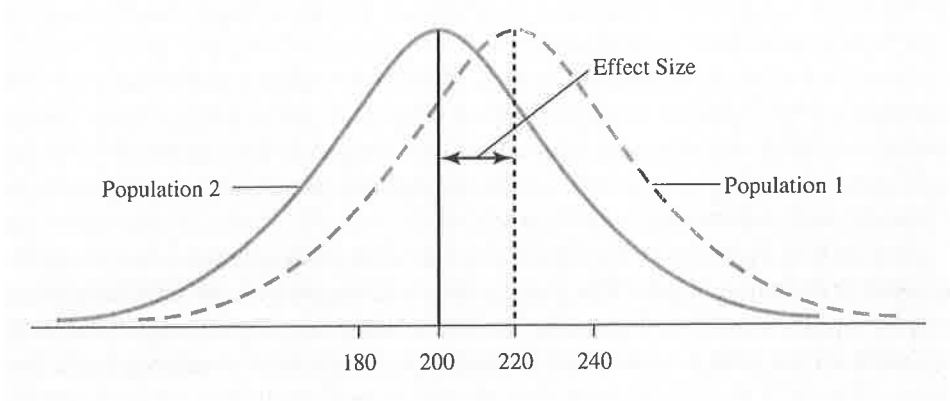

Figure 7-2 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of individuals for Population 1, those given the special instructions (right curve), and for Population 2, those not given special instructions (left curve). Population 1's mean is estimated based on the sample mean of 220, as originally described in Chapter 6; its standard deviation of 48 is assumed to be the same as Population 2's, which is known.

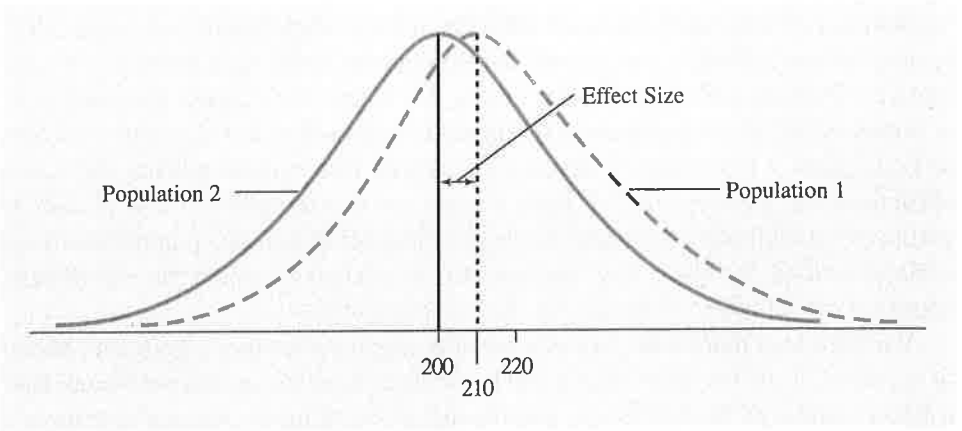

Figure 7-3 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of individuals for Population 1, those given the special instructions (right curve), and for Population 2, those not given special instructions (left curve). Population 1's mean is estimated based on a sample with a mean of 210; its standard deviation of 48 is assumed to be the same as population 2's, which is known.

TIP FOR SUCCESS

Notice that what you are doing here is basically the same as figuring *Z* scores. And, as when figuring *Z* scores, you can compare apples to oranges in the sense that you can compare results on different measures with different means and standard deviations.

a raw score effect size of 20 gives a standardized effect size of 20/48, which is .42. That is, the effect of giving the special instructions was to increase test scores by .42 of a standard deviation. The raw score effect size of 5 in the similar study (which had a population standard deviation of 10) is a standardized effect size of $5/10 = .50$. Thus, in this similar study, the effect was to increase the test scores by .50 (half) of a standard deviation. So, in this case the effect size in our original example is slightly smaller than the effect size in the similar study. Usually, when behavioral and social scientists refer to an effect size in a situation like we are considering, they mean a standardized effect size.

Here is the rule for calculating standardized effect size: Divide the predicted difference between the population means by the population standard deviation.[1] Stated as a formula,

[1]This procedure gives a measure of effect size called "Cohen's *d*." It is the preferred method for the kind of hypothesis testing you have learned so far (the *Z* test). (In later chapters, you learn some additional measures of effect size that are appropriate to particular hypothesis-testing situtations.)

$$\text{Effect Size} = \frac{\text{Population 1 } M - \text{Population 2 } M}{\text{Population } SD} \quad (7\text{-}1)$$

The standardized effect size is the difference between the two population means divided by the population's standard deviation.

In this formula, Population 1 M is the mean for the population that receives the experimental manipulation, Population 2 M is the mean of the known population (the basis for the comparison distribution), and Population SD is the standard deviation of the population of individuals. Notice that when figuring effect size you don't use the standard deviation of the distribution of means. Instead, you use the standard deviation of the population of individuals. Also notice that you are concerned with only one population's SD. This is because in hypothesis testing you usually assume that both populations have the same standard deviation. (We say more about this in later chapters.)

Consider again the fifth-grader example shown in Figure 7–1. The best estimate of the mean of Population 1 is the sample mean, which was 220. (In hypothesis-testing situations, you don't know the mean of Population 1, so you use an *estimated mean;* thus, you are actually figuring an *estimated effect size.*) The mean of Population 2 was 200, and the population standard deviation was 48. The difference between the two population means is 20 and the standard deviation of the populations of individuals is 48. Thus, the effect size is 20/48, or .42. In terms of the formula,

$$\text{Effect Size} = \frac{\text{Population 1 } M - \text{Population 2 } M}{\text{Population } SD} = \frac{220 - 200}{48} = \frac{20}{48} = .42$$

For the example in which the sample mean was 210, we would estimate Population 1's mean to be 210. Thus,

$$\text{Effect Size} = \frac{\text{Population 1 } M - \text{Population 2 } M}{\text{Population } SD} = \frac{210 - 200}{48} = \frac{10}{48} = .21$$

In both of these examples, the effect size is positive. If the effect size is negative, it just means that the mean of Population 1 is lower than the mean of Population 2.

Effect Size Conventions

What should you consider to be a "big" effect, and what is a "small" effect? Jacob Cohen (1988, 1992), a researcher who developed the effect size measure among other major contributions to statistical methods, has helped solve this problem (see Box 7–2 on page 215). Cohen came up with some **effect size conventions** based on the effects found in many actual studies. Specifically, Cohen recommended that, for the kind of situation we are considering in this chapter, we should think of a small effect size as about .20. With an effect size of .20, the populations of individuals have an overlap of about 85%. This small effect size of .20 is, for example, the average difference in height between 15- and 16-year-old girls (see Figure 7–4a on page 208), which is about a half-inch difference with a standard deviation of about 2.1 inches. Cohen considered a medium effect size to be .50, which means an overlap of about 67%. This is about the average difference in heights between 14- and 18-year-old girls (see Figure 7–4b). Finally, Cohen defined a large effect size as .80. This is only about a 53% overlap. It is about the average difference in height between 13- and 18-year-old girls (see Figure 7–4c). These three effect size conventions are summarized in Table 7–1. (Note that these effect size conventions apply in the same way to both positive and negative effect sizes. So, −.20 is a small effect size, −.50 is a medium effect size, and −.80 is a large effect size.)

Consider another example. As noted earlier in the book, many IQ tests have a standard deviation of 16 points. An experimental procedure with a small effect size

effect size conventions Standard rules about what to consider a small, medium, and large effect size, based on what is typical in behavioral and social science research; also known as Cohen's conventions.

Table 7-1 Summary of Cohen's Effect Size Conventions for Mean Differences

Verbal Description	Effect Size
Small	.20
Medium	.50
Large	.80

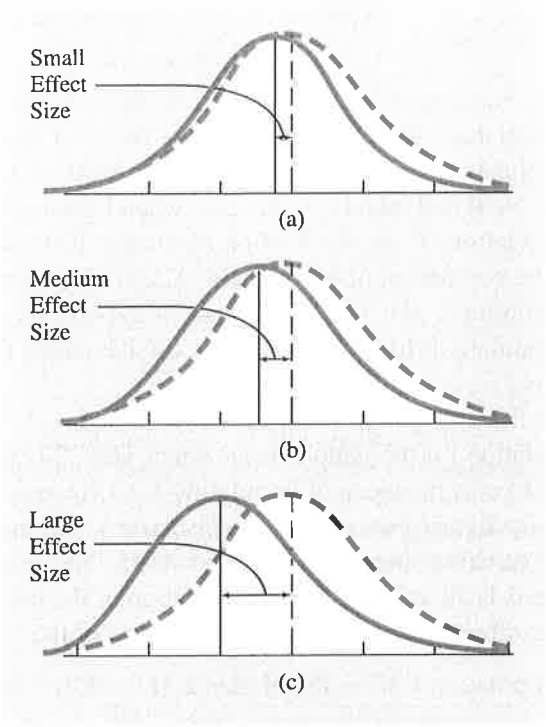

Figure 7-4 Comparisons of pairs of population distributions of individuals showing Cohen's conventions for effect size: (a) small effect size (.20), (b) medium effect size (.50), (c) large effect size (.80).

would be an increase of 3.2 IQ points. (A difference of 3.2 IQ points between the mean of the population who goes through the experimental procedure and the mean of the population that does not, divided by the population standard deviation of 16, gives an effect size of .20.) An experimental procedure with a medium effect size would increase IQ by 8 points. An experimental procedure with a large effect size would increase IQ by 12.8 points.

A More General Importance of Effect Size

Effect size, as we have seen, is the difference between means divided by the population standard deviation. This division by the population standard deviation standardizes the difference between means, in the same way that a *Z* score gives a standard for comparison to other scores, even scores on different scales. Especially by using the standard deviation of the population of individuals, we bypass the variation from study to study of different sample sizes, making comparison even easier and effect size even more of a standard.

Knowing the effect size of a study lets you compare results with effect sizes found in other studies, even when the other studies have different population standard deviations. Equally important, knowing the effect size lets you compare studies using different measures, even if those measures have different means and variances.

Also, within a particular study, our general knowledge of what is a small or a large effect size helps you evaluate the overall importance of a result. For example, a result may be statistically significant but not very large. Or a result that is not statistically significant (perhaps due to a small sample) may have just as large an effect size as another study (perhaps one with a larger sample) where the result was significant. Knowing the effect sizes of the studies helps us make better sense of

such results. We examine both of these important implications of effect size in later sections of this chapter.

An important development in using statistics in the behavioral and social sciences (and also in medicine and other fields) in the last few decades is a procedure called **meta-analysis.** This procedure combines results from different studies, even results using different methods of measurement. When combining results, the crucial thing is the *effect sizes.* As an example, a sociologist might be interested in the effects of cross-race friendships on prejudice, a topic on which there has been a large number of studies. Using meta-analysis, the sociologist could combine the results of these studies. This would provide an overall average effect size. It would also tell how the average effect sizes differ for studies done in different countries or about prejudice toward different ethnic groups. (For an example of such a study, see Davies, Tropp, Aron, Pettigrew, & Wright, 2010.) An educational researcher might be interested in the effects of different educational methods on students' educational achievement. Walberg and Lai (1999) carried out a large meta-analysis on this topic and provided effect size estimates for 275 educational methods and conditions. The effect sizes for selected general educational methods are shown in Table 7–2. As you can see in the table, many of the methods are associated with medium effect sizes and several have large (or very large) effect sizes. For another example of meta-analysis, see Box 7–1 (page 211).

Reviews of the collection of studies on a particular topic that use meta-analysis are an alternative to the traditional "narrative" literature review article. Such traditional reviews describe and evaluate each study and then attempt to draw some overall conclusion.

meta-analysis Statistical method for combining effect sizes from different studies.

Table 7–2 Effect Sizes of Selected General Educational Methods

Elements of Instruction	
Cues	1.25
Reinforcement	1.17
Corrective feedback	.94
Engagement	.88
Mastery Learning	.73
Computer-Assisted Instruction	
For early elementary students	1.05
For handicapped students	.66
Teaching	
Direct instruction	.71
Comprehension instruction	.55
Teaching Techniques	
Homework with teacher comments	.83
Graded homework	.78
Frequent testing	.49
Pretests	.48
Adjunct questions	.40
Goal setting	.40
Assigned homework	.28
Explanatory Graphics	.75

Source: Adapted from Walberg, H. J., & Lai, J.-S. (1999). Meta-analytic effects for policy. In G. J. Cizek (Ed.), *Handbook of educational policy* (pp. 419–453). San Diego, CA: Academic Press.

How are you doing?

1. What does effect size add to just knowing whether a result is significant?
2. Why do researchers usually use a *standardized* effect size?
3. Write the formula for effect size in the situation we have been considering.
4. On a standard test, the population is known to have a mean of 500 and a standard deviation of 100. Those receiving an experimental treatment have a mean of 540. What is the effect size?
5. What are the effect size conventions?
6. (a) What is meta-analysis? (b) What is the role of effect size in a meta-analysis?

Answers

1. A significant result can be just barely big enough to be significant or much bigger than necessary to be significant. Thus, knowing effect size tells you how big the effect is.
2. A standardized effect size makes the results of studies using different measures comparable.
3. Effect Size = (Population 1 M − Population 2 M)/Population SD.
4. Effect Size = (Population 1 M − Population 2 M)/Population SD = (540 − 500)/100 = .40.
5. Effect size conventions: small = .20, medium = .50, large = .80.
6. (a) Meta-analysis is a systematic procedure for combining results of different studies. (b) Meta-analyses usually come up with an average effect size across studies and also sometimes compare average effect sizes for different subgroups of studies.

TIP FOR SUCCESS

If you are at all unsure about Type I and Type II errors, take some time now to review the "Decision Errors" section in Chapter 5. As a brief reminder, you make a Type I error if the hypothesis-testing procedure leads you to decide that a study supports the research hypothesis when in reality the research hypothesis is false. You make a Type II error if the hypothesis-testing procedure leads you to decide that the results of a study are inconclusive, when in reality the research hypothesis is true. Remember that these errors do not come about due to errors in figuring or poor decision making; they occur because in the hypothesis-testing process you are making probabilistic decisions about populations based on information in samples.

Statistical Power

Power is the ability to achieve your goals. A goal of a researcher conducting a study is to get a significant result—but only *if* the research hypothesis really is true. The **statistical power** of a research study is the probability that the study will produce a statistically significant result if the research hypothesis is true. Power is *not* simply the probability that a study will produce a statistically significant result. The power of a study is the probability that it will produce a statistically significant result *if the research hypothesis is true.* If the research hypothesis is false, you do not want to get significant results. (That would be a Type I error, as you learned in Chapter 5.) Remember, however, even if the research hypothesis is true, an experiment will not automatically give a significant result. The particular sample that happens to be selected from the population may not turn out to be extreme enough to reject the null hypothesis.

Statistical power is important for several reasons. As you will learn later in the chapter, figuring power when planning a study helps you determine how many participants you need. Also, understanding power is extremely important when you read a research article, particularly for making sense of results that are not significant or results that are statistically significant but not of practical importance.

Consider once again our example of the effects of giving special instructions to fifth-graders taking a standard achievement test. Recall that we compared two populations:

Population 1: Fifth-graders receiving special instructions.
Population 2: Fifth-graders in general (who do not receive special instructions).

Also recall that the research hypothesis was that Population 1 would score higher than Population 2 on the achievement test.

statistical power Probability that the study will give a significant result if the research hypothesis is true.

BOX 7–1 Effect Sizes for Relaxation and Meditation: A Restful Meta-Analysis

In the 1970s and 1980s, the results of research on meditation and relaxation were the subject of considerable controversy. Eppley, Abrams, and Shear (1989) decided to look at the issue systematically by conducting a meta-analysis of the effects of various relaxation techniques on trait anxiety (that is, ongoing anxiety as opposed to a temporary state). Eppley and colleagues chose trait anxiety for their meta-analysis because it is related to many other mental health issues, yet in itself is fairly consistent from test to test.

Following the usual procedure, the researchers searched the scientific literature for studies—not only research journals but also books and doctoral dissertations. Finding all the relevant research studies is one of the most difficult parts of meta-analysis.

To find the "bottom line," the researchers compared effect sizes for each of the four widely studied methods of meditation and relaxation. The result was that the average effect size for the 35 Transcendental Meditation (TM) studies was .70 (meaning an average difference of .70 standard deviation in anxiety scores between those who practiced this meditation procedure and those in the control groups). This effect size was significantly larger than the average effect size of .28 for the 44 studies on all other types of meditation, the average effect size of .38 for the 30 studies on "progressive relaxation" (a widely used method at the time by clinical psychologists), and the average effect size of .40 for the 37 studies on other forms of relaxation.

Looking at different populations of research participants, they discovered that people screened to be highly anxious contributed more to the effect size, and prison populations and younger participants seemed to gain more from TM. There was no significant impact on effect size of the skill of the instructors, expectations of the participants, or whether participants had volunteered or been randomly assigned to conditions.

The researchers thought that one clue to TM's high performance might be that techniques involving concentration produced a significantly smaller effect, whereas TM makes a point of teaching an "effortless, spontaneous" method.

Whatever the reasons, Eppley et al. (1989) concluded that there are "grounds for optimism that at least some current treatment procedures can effectively reduce trait anxiety" (p. 973). So if you are prone to worry about matters like statistics exams, consider these results. (For an overview of several meta-analyses of such meditation effects, see Walton et al., 2002.)

The curve in Figure 7–5 (page 212) shows the distribution of means for Population 2. (Be careful: When discussing effect size, we showed figures, such as Figures 7–2 and 7–3, for populations of individuals; now we are back to focusing on distributions of means.) This curve is the comparison distribution, the distribution of means that you would expect for both populations if the null hypothesis were true. The mean of this distribution of means is 200 and its standard deviation is 6. In Chapter 6, we found that using the 5% significance level, one-tailed, you need a Z score for the mean of your sample of at least 1.64 to reject the null hypothesis. Using the formula for converting Z scores to raw scores, this comes out to a raw score of 209.84; that is, $(1.64)(6) + 200 = 209.84$. Therefore, we have shaded the tail of this distribution above a raw score of 209.84 (a Z score of 1.64 on this distribution). This is the area where you would reject the null hypothesis if, as a result of your study, the mean of your sample was in this area.

Imagine that the researchers predict that giving students the special instructions will increase students' scores on the achievement test to 208. (This is an increase of 8 points from the mean of 200 when no special instructions are given.) If this prediction is correct, the research hypothesis is true and the mean of Population 1 (the population of students who receive the special instructions) is indeed greater than the mean of Population 2. The distribution of means for Population 1 for this *hypothetical predicted situation* is shown in the top part of Figure 7–6 (page 213). Notice that the distribution has a mean of 208.

Now take a look at the curve shown in the bottom part of Figure 7–6. This curve is exactly the same as the one shown in Figure 7–5; it is the comparison distribution, the distribution of means for Population 2. Notice that the distribution of means

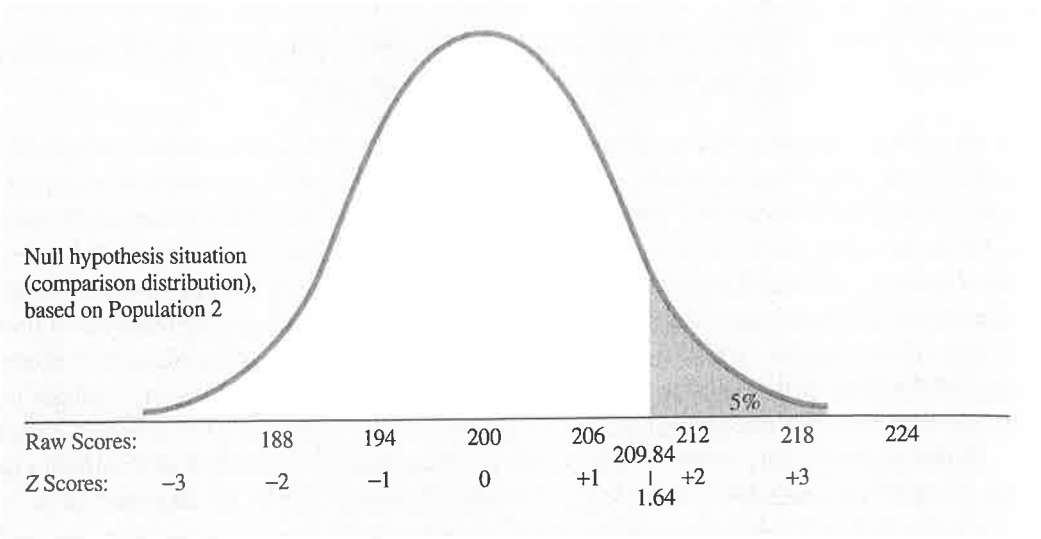

Figure 7–5 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 2 (the comparison distribution), those not given special instructions. Significance cutoff score (209.84) shown for $p < .05$, one-tailed.

for Population 1 (the top curve) is set off to the right of the distribution of means for Population 2 (the bottom curve). This is because the researchers predict the mean of Population 1 to be higher (a mean of 208) than the mean of Population 2 (which we know is 200). (If Population 1's mean is predicted to be *lower* than Population 2's mean, then Population 1 would be set off to the *left*.) If the null hypothesis is true, the true distribution for Population 1 is the same as the distribution based on Population 2. Thus, the Population 1 distribution would be lined up directly above the Population 2 distribution and would not be set off to the right (or the left).

Recall that the cutoff score for rejecting the null hypothesis in this example is 209.84. Thus, the shaded rejection area for Population 2's distribution of means (shown in the bottom curve in Figure 7–6) starts at 209.84. We can also create a rejection area for the distribution of means for Population 1. This rejection area will also start at 209.84 (see the shaded area in the top curve in Figure 7–6). Remember that, in this example, Population 1's distribution of means represents the possible sample means that we would get if we randomly selected 64 fifth-graders from a population of fifth-graders with a mean of 208 (and a standard deviation of 48).

Now, suppose the researchers carry out the study. They give the special instructions to a randomly selected group of 64 fifth-graders and find their mean score on the achievement test. And suppose this sample's mean turns out to be in the shaded area of the distribution (that is, a mean of 209.84 or higher). If that happens, the researchers will reject the null hypothesis. What Figure 7–6 shows us is that most of the means from Populations 1's distribution of means (assuming that its mean is 208) will not be large enough to reject the null hypothesis. Less than half of the upper distribution is shaded. Put another way, if the research hypothesis is true, as the researcher predicts, the sample we study is a random sample from this Population 1 distribution of means. However, there is less than a 50–50 chance that the mean of a random sample from this distribution will be in the shaded rejection area.

Recall that the statistical power of a study is the probability that the study will produce a statistically significant result, if the research hypothesis is true. Since we are assuming the research hypothesis is true in this example, the shaded region in the upper distribution represents the power of the study. It turns out that the power for this situation

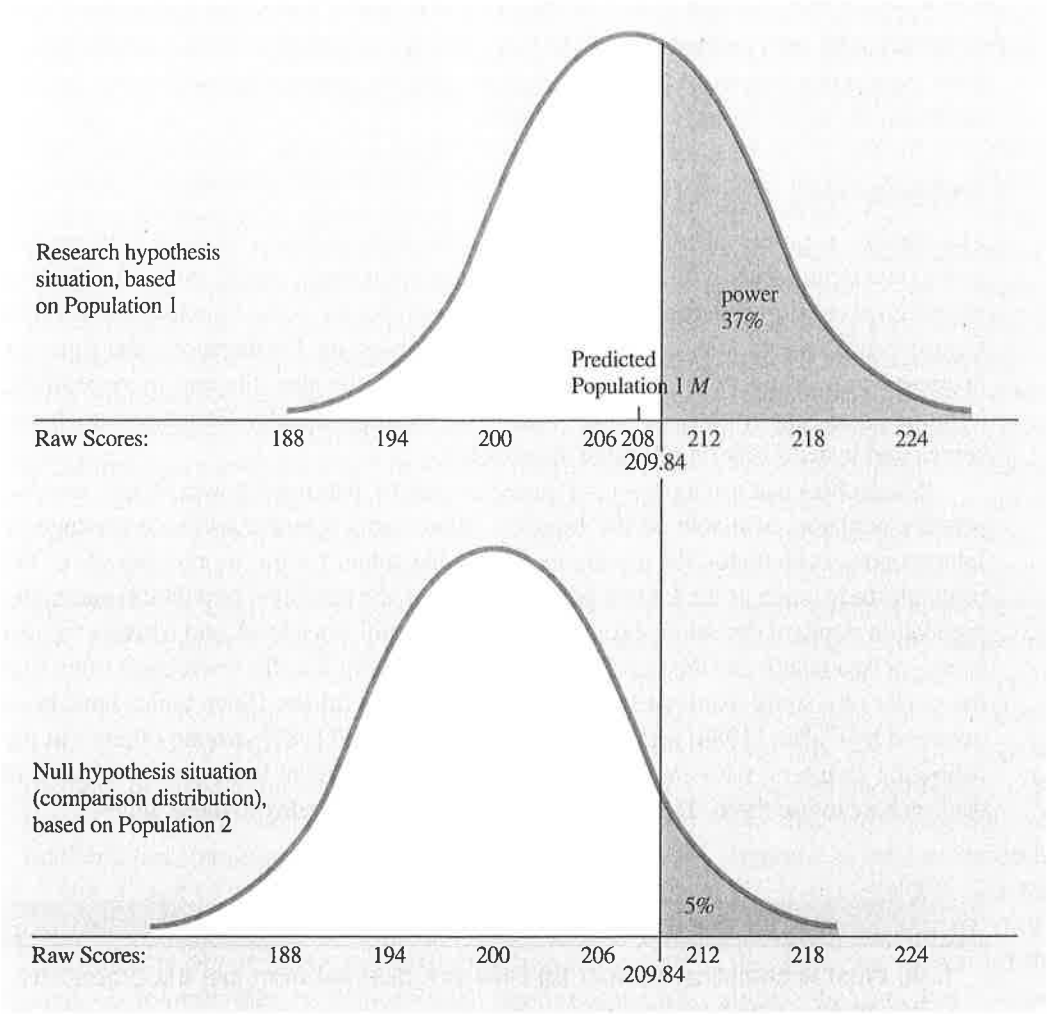

Figure 7–6 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). Significance cutoff score (209.84) shown for $p < .05$, one-tailed. Shaded sections of both distributions are the area in which the null hypothesis will be rejected. Power = 37%.

(shown in Figure 7–6) is only 37%. Therefore, assuming the researcher's prediction is correct, the researcher has only a 37% chance that the sample of 64 fifth-graders will have a mean high enough to make the result of the study statistically significant.

Suppose that the particular sample of 64 fifth-graders studied had a mean of 203.5. Since you would need a mean of at least 209.84 to reject the null hypothesis, the result of this study would not be statistically significant. It would not be significant, even though the research hypothesis really is true. This is how you would come to make a Type II error.

It is entirely possible that the researchers might select a sample from Population 1 with a mean far enough to the right (that is, with a high enough mean test score) to be in the shaded rejection area. However, given the way we have set up this particular example, there is a less-than-even chance that the study will turn out significant, *even though we know the research hypothesis is true.* (Of course, once again, the researcher would not know this.) When a study like the one in this example has only a small chance of being significant even if the research hypothesis is true, we say the study has *low power.*

Suppose, on the other hand, the situation was one in which the upper curve was expected to be way to the right of the lower curve, so that almost any sample taken from the upper curve would be in the shaded rejection area in the lower curve. In that situation, the study would have high power.

Determining Statistical Power

The statistical power of a study can be figured. In a situation like the fifth-grader testing example (when you have a known population and a single sample), figuring power involves figuring out the area of the shaded portion of the upper distribution in Figure 7–6. However, the figuring is somewhat laborious. Furthermore, the figuring becomes quite complex once we consider, starting in the next chapter, more realistic hypothesis-testing situations. Thus, researchers do not usually figure power themselves and instead rely on alternate approaches.

Researchers can use a power software package to determine power. There are also power calculators available on the Internet. When using a power software package or Internet power calculator, the researcher puts in the values for the various aspects of the research study (such as the known population mean, the predicted population mean, the population standard deviation, the sample size, the significance level, and whether the test is one- or two-tailed) and the figuring is done automatically. Finally, researchers often find the power of a study using special charts called **power tables.** (Such tables have been prepared by Cohen [1988] and by Kraemer & Thiemann [1987], among others.) In the following chapters, with each method you learn, we provide basic power tables and discuss how to use them. Table A–5 in the Appendix is an index to these tables.

power table Table for a hypothesis-testing procedure showing the statistical power of a study for various effect sizes and sample sizes.

How are you doing?

1. (a) What is statistical power? (b) How is it different from just the probability of getting a significant result?
2. Give two reasons why statistical power is important.
3. What is the probability of getting a significant result if the research hypothesis is false?
4. (a) Name three approaches that researchers typically use to determine power. (b) Why do researchers use these approaches, as opposed to figuring power by hand themselves?

Answers

1. (a) Statistical power is the probability of getting a significant result if the research hypothesis is true. (b) It is the probability *if the research hypothesis is true.*
2. Statistical power is important because (1) it can help you determine how many participants are needed for a study you are planning, and (2) understanding power can help you make sense of results that are not significant or results that are statistically significant but not of practical importance.
3. The probability of getting a significant result if the research hypothesis is false is the significance level (that is, the probability of making a Type I error).
4. (a) Three approaches that researchers typically use to determine power are (1) power software packages, (2) Internet power calculators, and (3) power tables. (b) Researchers use these approaches because in common hypothesis-testing situations, figuring power by hand is very complicated.

BOX 7–2 Jacob Cohen, the Ultimate New Yorker: Funny, Pushy, Brilliant, and Kind

New Yorkers can be proud of Jacob Cohen, who single-handedly introduced to behavioral and social scientists some of our most important statistical tools, including the main topics of this chapter (power analysis and effect size) as well as many of the sophisticated uses of regression analysis and much more. Never worried about being popular—although he was—he almost single-handedly forced the recent debate over significance testing, which he liked to joke was entrenched like a "secular religion." About the asterisk that accompanies a significant result, he said the religion must be "of Judeo-Christian derivation, as it employs as its most powerful icon a six-pointed cross" (1990, p. 1307).

Cohen entered graduate school at New York University (NYU) in clinical psychology in 1947 and 3 years later had a master's and a doctorate. He then worked in rather lowly roles for the U.S. Veterans Administration, doing research on various practical topics, until he returned to NYU in 1959. There he became a very famous faculty member because of his creative, off-beat ideas about statistics. Amazingly, he made his contributions having no mathematics training beyond high school algebra.

But a lack of formal training may have been Jacob Cohen's advantage, because he emphasized looking at data and thinking about them, not just applying a standard analysis. In particular, he demonstrated that the standard methods were not working very well, especially for "soft" fields of psychology such as clinical, personality, and social psychology, because researchers in these fields had no hope of finding what they were looking for due to a combination of typically small effect sizes of such research and researchers' use of small sample sizes. Entire issues of journals were filled with articles that only had a 50–50 chance of finding what their authors were looking for.

Cohen's ideas were hailed as a great breakthrough, especially regarding power and effect size. Yet, the all-too-common practice of carrying out studies with inadequate power that he railed against as hindering scientific progress stubbornly continued. But even after 20 years of this, he was patient, writing that he understood that these things take time. Cohen's patience must have been part of why behavioral and social scientists from around the world found him a "joy to work with" (Murphy, 1998). Those around him daily at NYU knew him best; one said Cohen was "warm and engaging . . . renowned for his many jokes, often ribald" (Shrout, 2001, p. 166).

But patient or not, Cohen did not let up on researchers. He wanted them to think more deeply about the standard methods. Starting in the 1990s he really began to force the issue of the mindless use of significance testing. But he still used humor to tease behavioral and social scientists for their failure to see the problems inherent in the arbitrary yes–no decision feature of null hypothesis testing. For example, he liked to remind everyone that significance testing came out of Sir Ronald Fisher's work in agriculture (see Box 10–1), in which the decisions were yes–no matters, such as whether a crop needed manure. He pointed out that behavioral and social scientists "do not deal in manure, at least not knowingly" (Cohen, 1990, p. 1307)! He really disliked the fact that Fisher-style decision making is used to determine the fate of not only doctoral dissertations, research funds, publications, and promotions, "but whether to have a baby just now" (p. 1307). And getting more serious, Cohen charged that significance testing's "arbitrary unreasonable tyranny has led to data fudging of varying degrees of subtlety, from grossly altering data to dropping cases where there 'must have been' errors" (p. 1307).

Cohen was active in many social causes, especially desegregation in the schools and fighting discrimination in police departments. He cared passionately about everything he did. He was deeply loved. And he suffered from major depression, becoming incapacitated by it four times in his life.

Got troubles? Got no more math than high school algebra? It doesn't have to stop you from contributing to science.

What Determines the Power of a Study?

It is very important that you understand what power is about. It is especially important to understand the factors that affect the power of a study and how to use power when planning a study and when making sense of a study you read.

The statistical power of a study depends on two main factors: (1) how big an effect (the effect size) the research hypothesis predicts and (2) how many participants

are in the study (the sample size). Power is also affected by the significance level chosen, whether a one-tailed or two-tailed test is used, and the kind of hypothesis-testing procedure used.

Effect Size

Figure 7–6 shows the situation in our special test-instructions example in which the researchers had reason to predict that fifth-graders who got the special instructions (Population 1, the top curve) would have a mean score *8 points higher* than fifth-graders in general (Population 2, the bottom curve). Figure 7–7 shows the same study for a situation in which the researchers would have reason to expect that Population 1 (those who got the special instructions) would have a mean score *16 points higher* than Population 2 (fifth-graders in general). Compare Figure 7–7 to Figure 7–6. You are more likely to get a significant result in the situation shown in Figure 7–7. This is because there is more overlap of the top curve with the shaded area on the comparison

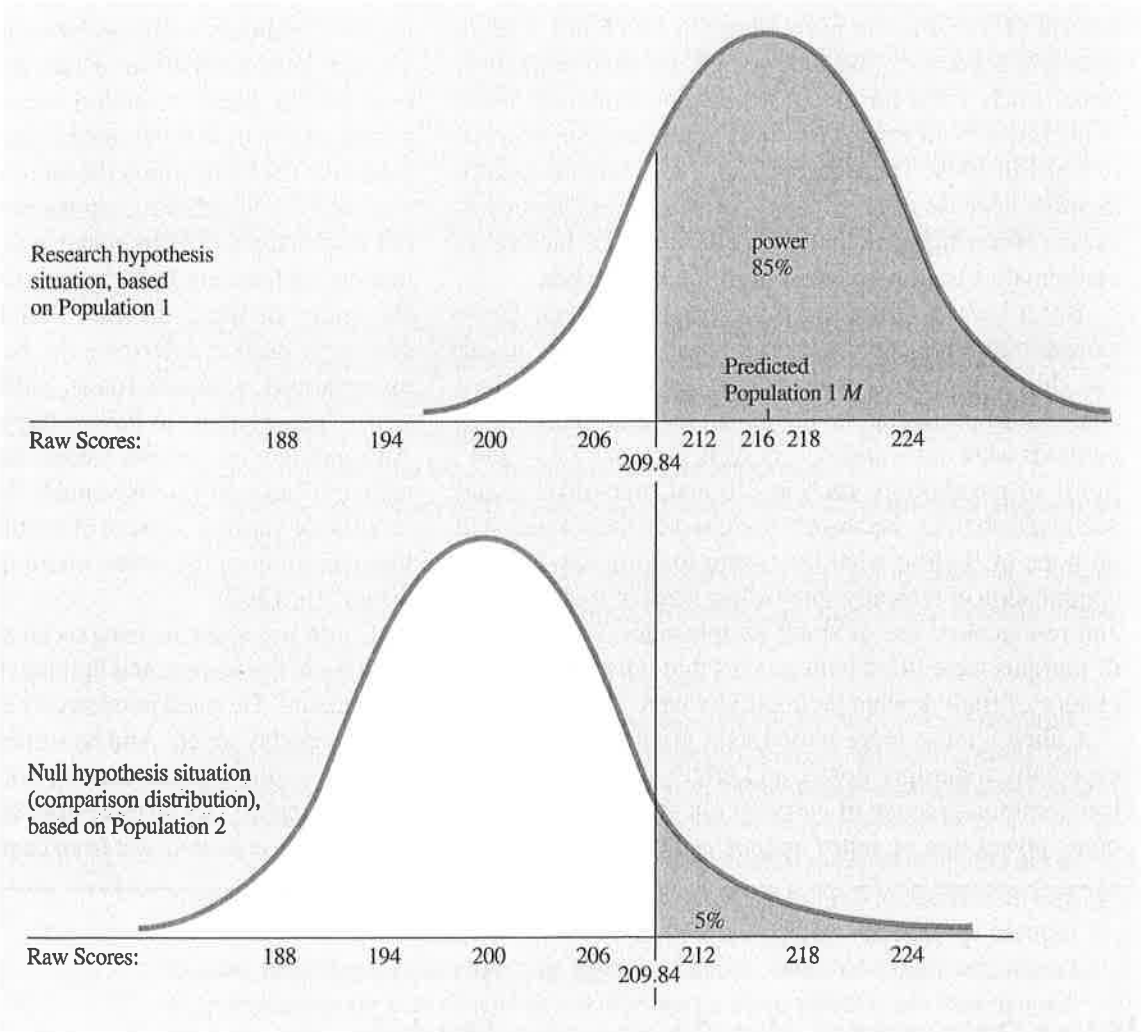

Figure 7–7 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 216 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). Significance cutoff score (209.84) shown for $p < .05$, one-tailed. Power = 85%. Compare with Figure 7–6, in which the predicted population mean was 208 and power was 37%.

distribution. Recall that the probability of getting a significant result (the power) for the situation in Figure 7–6, in which there was a basis for the researchers to predict only a mean of 208, is only 37%. However, for the situation in Figure 7–7, in which there was a basis for the researchers to predict a mean of 216, the power is 85%. In any study, the bigger the difference that your theory or previous research says you should expect between the means of the two populations, the more power there is in the study. That is, if in fact there is a big mean difference in the population, you have more chance of getting a significant result in the study. So if you predict a bigger mean difference, the power you figure based on that prediction will be greater. (Thus, if you figure power based on a prediction that is unrealistically big, you are just fooling yourself about the power of the study.)

The difference in the means between populations we saw earlier is part of what goes into effect size. Thus, the bigger the effect size is, the greater the power is. The effect size for the situation in Figure 7–6, in which the researchers predicted Population 1 to have a mean of 208, is .17. That is, Effect Size = (Population 1 M − Population 2 M)/Population SD = $(208 - 200)/48 = 8/48 = .17$.

The effect size for the situation in Figure 7–7, in which the researchers predicted Population 1 to have a mean of 216, is .33. That is, Effect Size = (Population 1 M − Population 2 M)/Population SD = $(216 - 200)/48 = 16/48 = .33$.

Effect size, however, is also affected by the population standard deviation. The smaller the standard deviation is, the bigger the effect size is. In terms of the effect size formula, this is because if you divide by a smaller number, the result is bigger. In terms of the actual distributions, this is because if two distributions that are separated are narrower, they *overlap less*. Figure 7–8 shows two distributions of means based on the same example. However, this time we have changed the example so that the population standard deviation is exactly half of what it was in Figure 7–6. In this version, the predicted mean is the original 208. However, both distributions of means are much narrower. Therefore, there is much less overlap between the upper curve and the lower curve (the comparison distribution). The result is that the power is 85%, much higher than the power of 37% in the original situation. The idea here is that the smaller the population standard deviation becomes, the greater the power is.

Overall, these examples illustrate the general principle that the less overlap between the two distributions, the more likely it is that a study will give a significant result. Two distributions might have little overlap overall either because there is a large difference between their means (as in Figure 7–7) or because they have such a small standard deviation that even with a small mean difference they do not overlap much (as in Figure 7–8 on page 218). This principle is summarized more generally in Figure 7–9 (page 219).

Sample Size

The other major influence on power, besides effect size, is the number of people in the sample studied, the sample size. Basically, the more people there are in the study, the more power there is.

Sample size affects power because the larger the sample size is, the smaller the standard deviation of the distribution of means becomes. If these distributions have a smaller standard deviation, they are narrower. And if they are narrower, there is less overlap between them. Figure 7–10 (page 220) shows the situation for our fifth-grader example if the study included 100 fifth-graders instead of the 64 in the original example, with a predicted mean of 208 and a population standard deviation of 48.

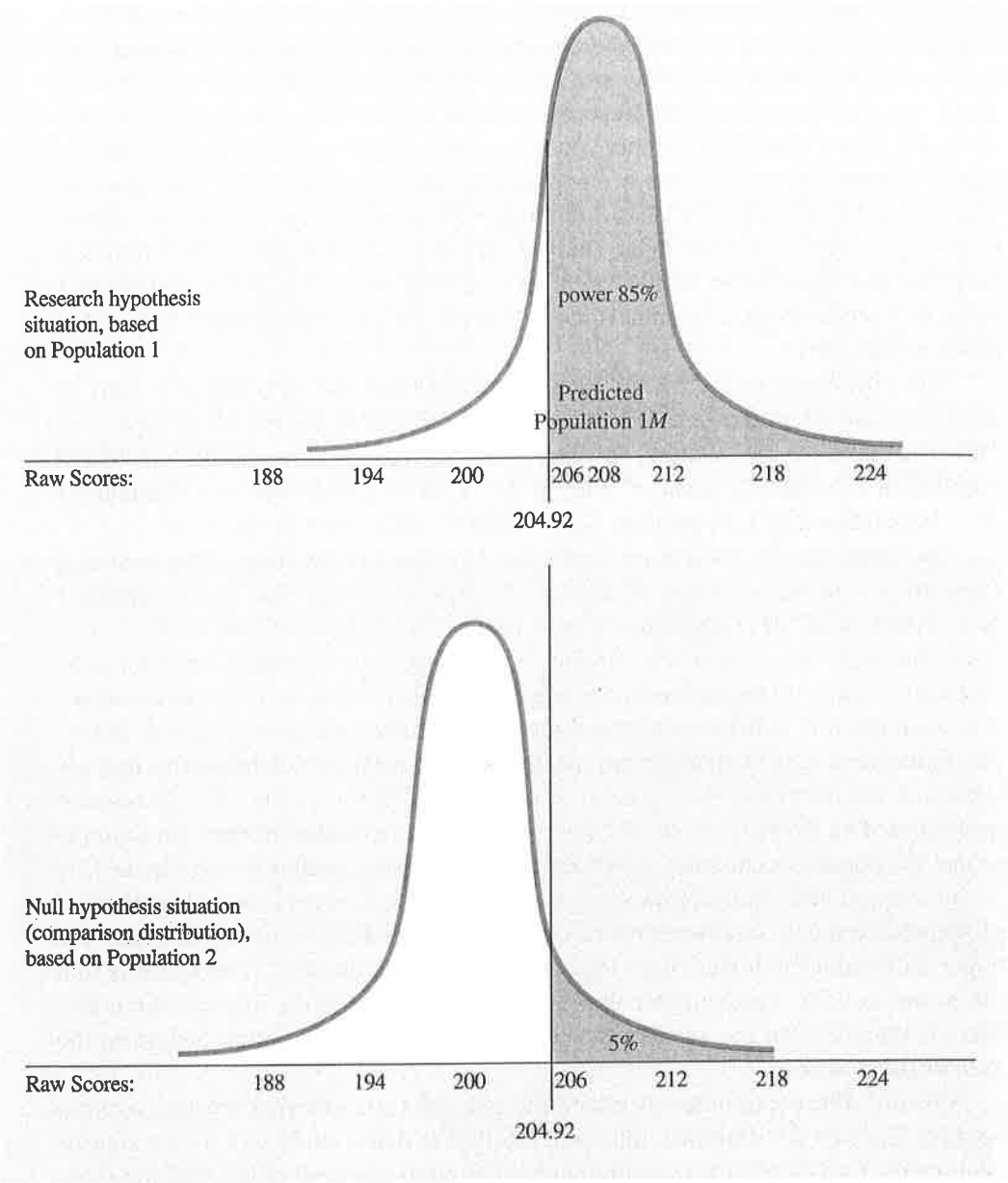

Figure 7–8 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). Significance cutoff score (204.92) shown for $p < .05$, one-tailed. In this example, the population standard deviation is half as large as that shown for this example in previous figures. Power = 85%. Compare with Figure 7–6, which had the original population standard deviation and a power of 37%.

The power now is 51%. (It was 37% with 64 fifth-graders.) With 500 participants in the study, the power is 99% (see Figure 7–11 on page 221).

Don't get mixed up. The distributions of means can be narrow (and thus have less overlap and more power) for two very different reasons. One reason is that the populations of individuals may have small standard deviations. This reason has to do with effect size. The other reason is that the sample size is large. This reason is completely separate. Sample size has nothing to do with effect size. Both effect size and sample size influence power. However, as we will see shortly, these two different

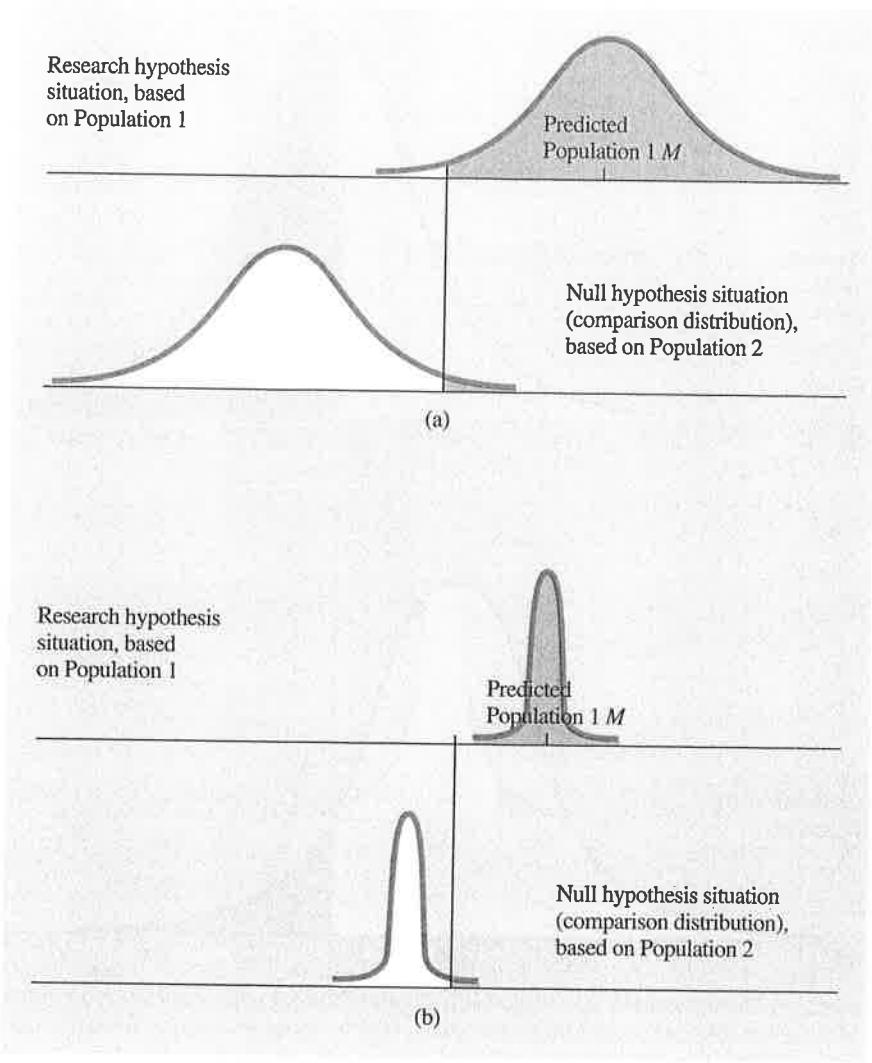

Figure 7-9 The predicted and comparison distributions of means might have little overlap (and thus the study would have high power) because either (a) the two means are very different or (b) the population standard deviation is very small.

influences on power lead to completely different kinds of practical steps for increasing power when planning a study.

Figuring Needed Sample Size for a Given Level of Power

When planning a study, the main reason researchers consider power is to help decide how many people to include in the study. Sample size has an important influence on power. Thus, a researcher wants to be sure to have enough people in the study for the study to have fairly high power. (Too often, researchers carry out studies in which the power is so low that it is unlikely they will get a significant result even if the research hypothesis is true.)

Suppose the researchers in our fifth-grader example were planning the study and wanted to figure out how many students to include in the sample. Let us presume that based on previous research for a situation like theirs, the researchers predicted a mean difference of 8 and there is a known population standard deviation of 48. In this case, it turns out that the researchers would need 222 fifth-graders to have 80% power. In practice, researchers use power software packages, Internet power calculators, or special

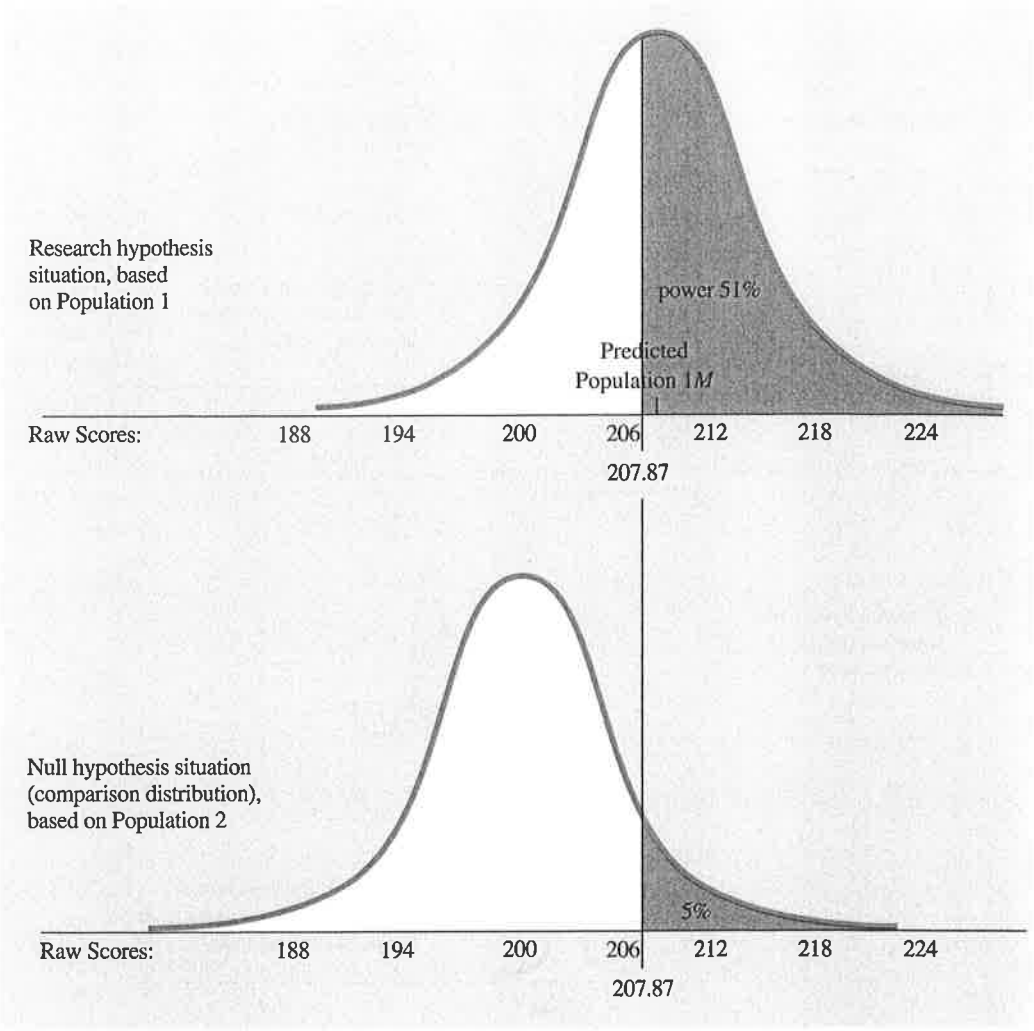

Figure 7–10 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). In this example, the sample size is 100, compared to 64 in the original example. Significance cutoff score (207.87) shown for $p < .05$, one-tailed. Power = 51%. Compare with Figure 7–6, which had the original sample size of 64 fifth-graders and a power of 37%.

power tables that tell you how many participants you would need in a study to have a high level of power, given a predicted effect size. We provide simplified versions of power tables for each of the main hypothesis-testing procedures you learn in upcoming chapters.

Other Influences on Power

Three other factors, besides effect size and sample size, affect power:

1. *Significance level.* Less extreme significance levels (such as $p < .10$ or $p < .20$) mean more power. More extreme significance levels (.01 or .001) mean less power. Less extreme significance levels result in more power because the shaded rejection area on the lower curve is bigger. Thus, more of the area in the upper curve is shaded. More extreme significance levels result in less power because the shaded rejection region on the lower curve is smaller. Suppose in our original version of the fifth-grader example we had instead used the .01 significance level. The power would have dropped from 37% to only 16% (see Figure 7–12 on

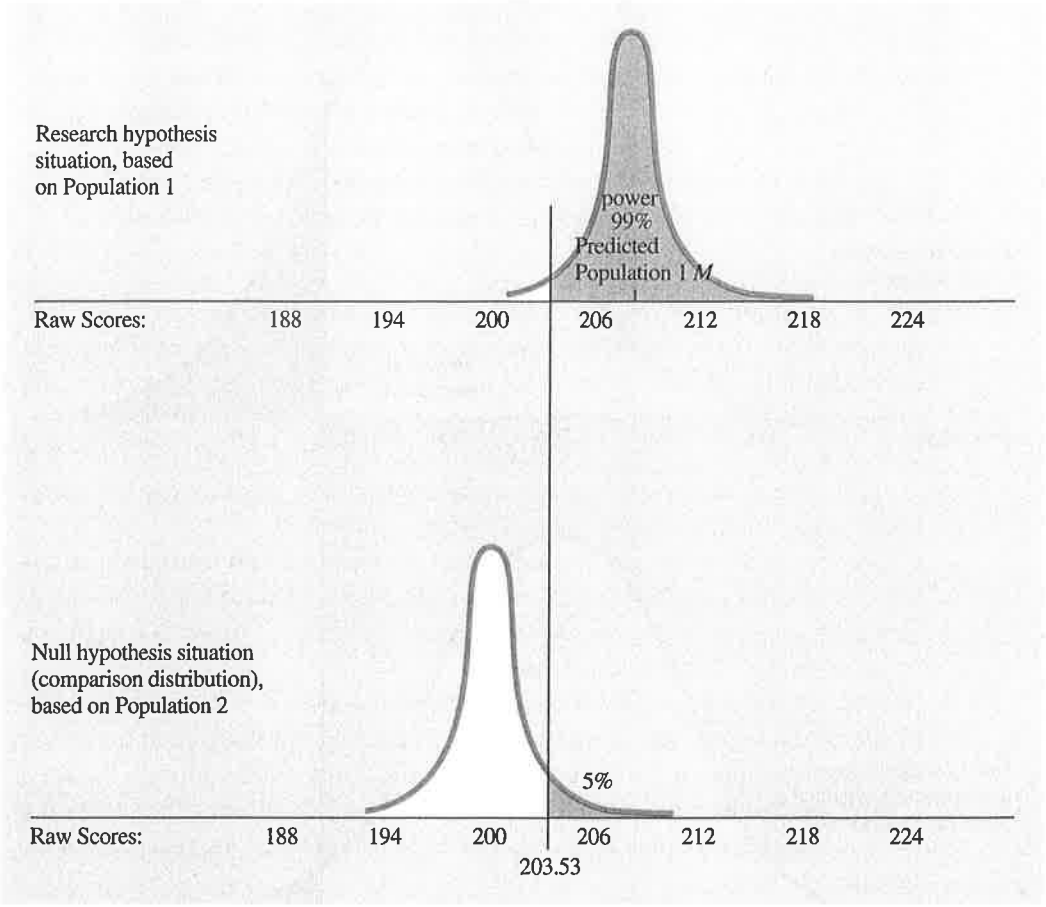

Figure 7-11 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). In this example, the sample size is 500, compared to 64 in the original example. Significance cutoff score (203.53) shown for $p < .05$, one-tailed. Power = 99%. Compare with Figure 7–6, which had the original sample size of 64 fifth-graders and power was 37%, and Figure 7-10, which had a sample size of 100 and a power of 51%.

page 222). It is important to bear in mind that using a less extreme significance level (such as $p < .10$ or $p < .20$) increases the chance of making a Type I error. Also, using an extreme significance level (such as $p < .01$ or $p < .001$) increases the chance of making a Type II error.

2. *One- versus two-tailed tests.* Using a two-tailed test makes it harder to get significance on any one tail. Thus, keeping everything else the same, power is less with a two-tailed test than with a one-tailed test. Suppose in our fifth-grader testing example we had used a two-tailed test instead of a one-tailed test (but still using the 5% level overall). As shown in Figure 7–13 (page 224), power would be only 26% (compared to 37% in the original one-tailed version shown in Figure 7–6).
3. *Type of hypothesis-testing procedure.* Sometimes the researcher has a choice of more than one hypothesis-testing procedure to use for a particular study. We have not considered any such situations so far in this book but we will do so in Chapter 11.

Summary of Influences on Power

Table 7–3 (page 222) summarizes the effects of various factors on the power of a study.

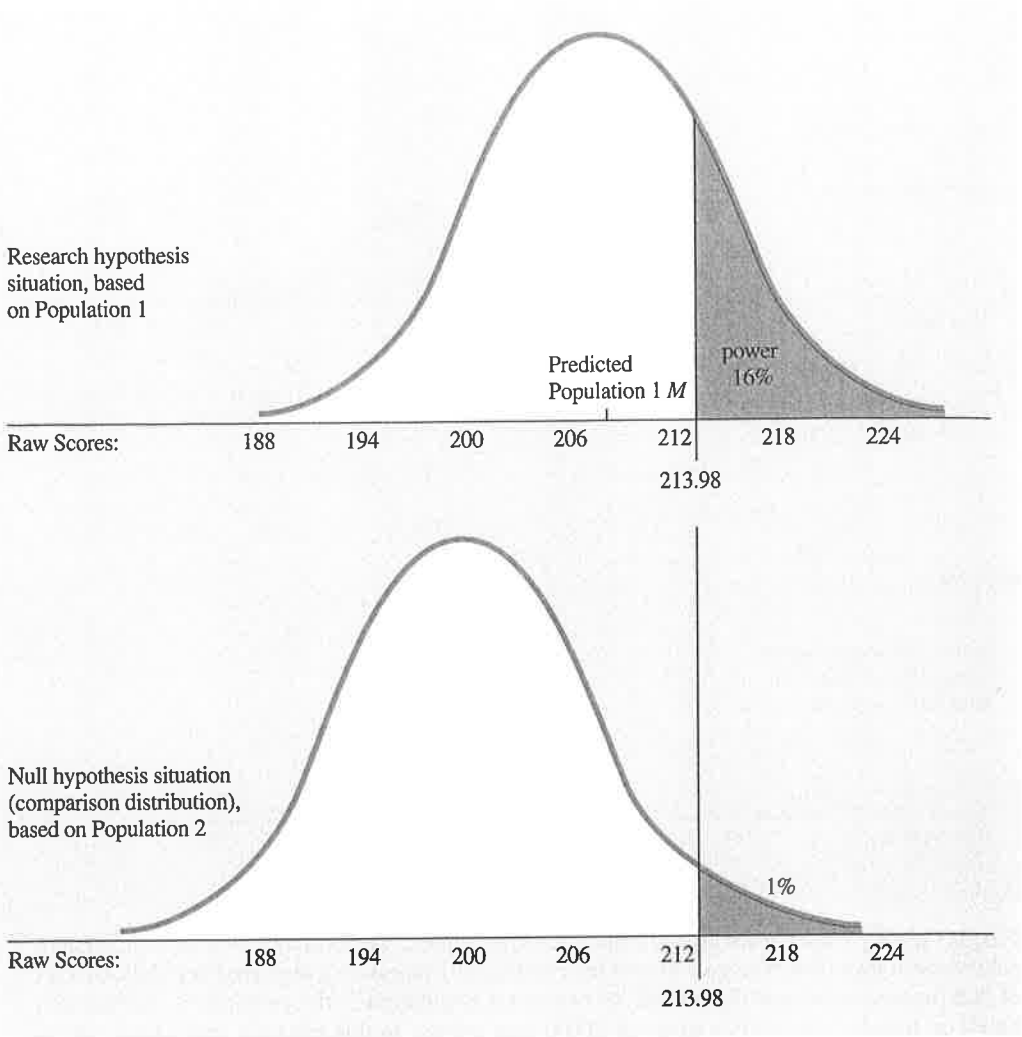

Figure 7-12 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). Significance cutoff score (213.98) now shown for $p < .01$, one-tailed. Power = 16%. Compare with Figure 7–6, which used a significance level of $p < .05$ and a power of 37%.

Table 7-3 Influences on Power

Feature of the Study	Increases Power	Decreases Power
Effect size	Large	Small
Effect size combines the following two features:		
Predicted difference between population means	Large differences	Small differences
Population standard deviation	Small Population *SD*	Large Population *SD*
Sample size (*N*)	Large *N*	Small *N*
Significance level	Lenient (such as .10)	Stringent (such as .01)
One-tailed versus two-tailed test	One-tailed	Two-tailed
Type of hypothesis-testing procedure used	Varies	Varies

How are you doing?

1. (a) What are the two factors that determine effect size? For each factor, (b) and (c), explain how and why it affects power.
2. (a) How and (b) why does sample size affect power?
3. (a) How and (b) why does the significance level used affect power?
4. (a) How and (b) why does using a one-tailed versus a two-tailed test affect power?

Answers

1. (a) The two factors that determine effect size are (1) the difference between the population means and (2) the population standard deviation.
(b) The more difference there is between the population means, the larger the effect size, and the more power. This is because it drives the distribution of means farther apart and thus they have less overlap. Therefore, the area in the predicted distribution that is more extreme than the cutoff in the known distribution is greater.
(c) The smaller the population standard deviation is, the larger the effect size becomes, and the greater the power. This is because it makes the distribution of means narrower and thus have less overlap. Thus, the area in the predicted distribution that is more extreme than the cutoff in the known distribution is greater.
2. (a) The larger the sample size is, the more power there is. (b) This is because a large sample size makes the distribution of means narrower (because the standard deviation of the distribution of means is the square root of the result of dividing the population variance by the sample size) and thus have less overlap; so the area in the predicted distribution more extreme than the cutoff in the known distribution is greater.
3. The more liberal the significance level is (for example, $p < .10$ vs. $p < .05$), the more power there is. (b) This is because it makes the cutoff in the known distribution less extreme; so the corresponding area that is more extreme than this cutoff in the predicted distribution of means is larger.
4. A study with a one-tailed test has more power (for a result in the predicted direction) than a two-tailed test. (b) This is because with a one-tailed test, the cutoff in the predicted direction in the known distribution is less extreme; so the corresponding area that is more extreme than this cutoff in the predicted distribution of means is larger. There is an added cutoff in the opposite side with a two-tailed test, but this is so far out on the distribution that it has little effect on power.

The Role of Power When Planning a Study

Determining power is very important when planning a study. If you do a study in which the power is low, even if the research hypothesis is true, the study will probably not give statistically significant results. Thus, the time and expense of carrying out the study, as it is currently planned, would probably not be worthwhile. So when the power of a planned study is found to be low, researchers look for practical ways to increase the power to an acceptable level.

What is an acceptable level of power? A widely used rule is that a study should have 80% power to be worth doing (see Cohen, 1988). Power of 80% means that there is an 80% chance that the study will produce a statistically significant result if the

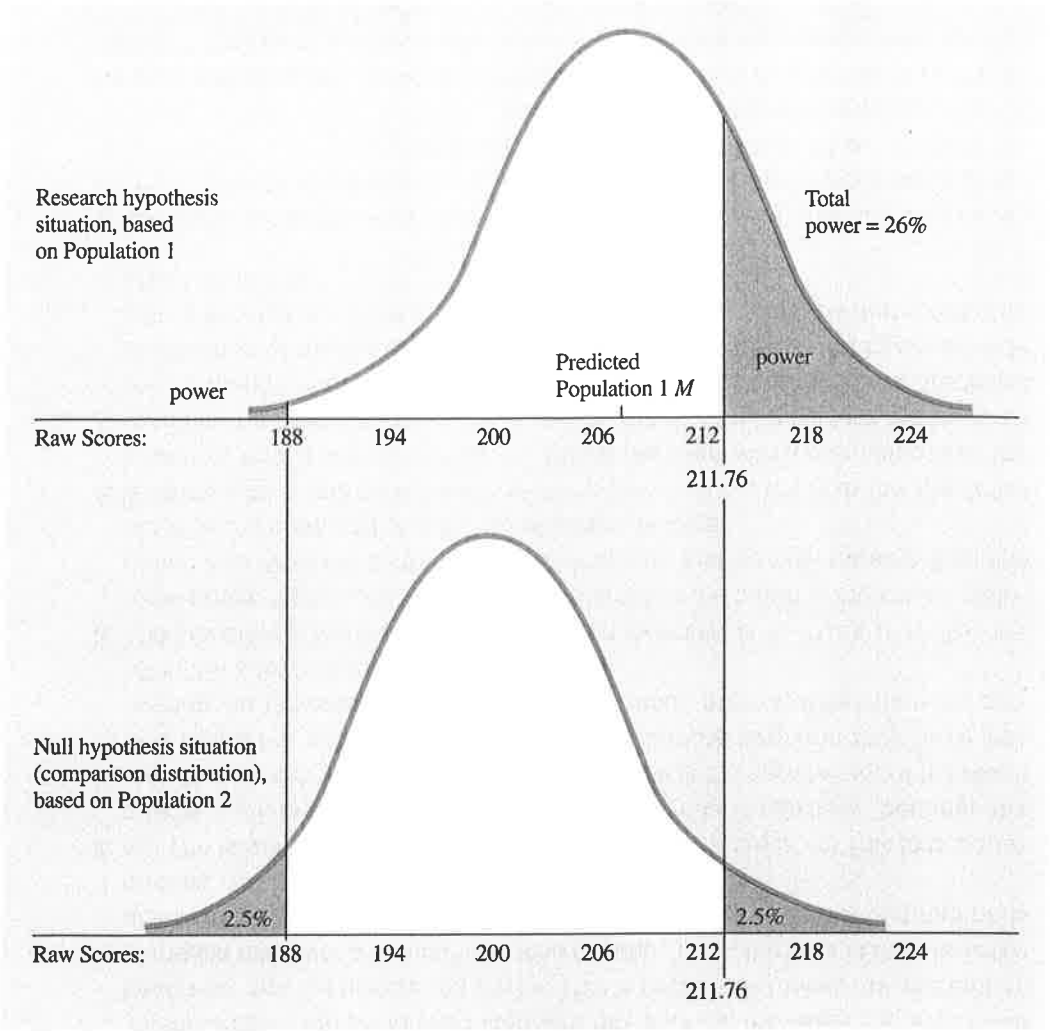

Figure 7-13 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). Significance cutoff scores (188.24 and 211.76) now shown for $p < .05$, two-tailed. Power = 26%. Compare with Figure 7–6, which used a significance level of $p < .05$, one-tailed, and a power of 37%.

research hypothesis is true. Obviously, the more power the better. However, the costs of greater power, such as studying more people, often make even 80% power beyond your reach.

How can you increase the power of a planned study? In principle, you can do so by changing any of the factors summarized in Table 7–3. Let's consider each.

1. **Increase the effect size by increasing the predicted difference between population means.** You can't just arbitrarily predict a bigger difference. There has to be a sound basis for your prediction. Thus, to increase the predicted difference, your method in carrying out the study must make it reasonable to expect a bigger difference. Consider again our example of the experiment about the impact of special instructions on fifth-graders' test scores. One way to increase the expected mean difference would be to make the instructions more elaborate, spending more time explaining them, perhaps allowing time for practice, and so forth. In some studies, you may be able to increase the expected mean difference

by using a more intense experimental procedure. A disadvantage of this approach of increasing the impact of the experimental procedure is that you may have to use an experimental procedure that is not like the one to which you want the results of your study to apply. It can also sometimes be difficult or costly.

2. **Increase effect size by decreasing the population standard deviation.** You can decrease the population standard deviation in a planned study in at least two ways. One way is to use a population that has less variation within it than the one originally planned. With the fifth-grader testing example, you might only use fifth-graders in a particular suburban school system. The disadvantage is that your results will then apply only to the more limited population.

   Another way to decrease the population standard deviation is to use conditions of testing that are more standardized and measures that are more precise. For example, testing in a controlled laboratory setting usually makes for smaller overall variation among scores in results (meaning a smaller standard deviation). Similarly, using measures and tests with very clear wording also reduces variation. When practical, this is an excellent way to increase power, but often the study is already as rigorous as it can be.
3. **Increase the sample size.** The most straightforward way to increase power in a study is to study more people. Of course, if you are studying billionaires who have made their fortune by founding an Internet company, there is a limit to how many are available. Also, using a larger sample size often adds to the time and cost of conducting the research. In most research situations, though, increasing sample size is the main way to change a planned study to raise its power.
4. **Use a less extreme level of significance (such as $p < .10$ or $p < .20$).** Ordinarily, the level of significance you use should be the least extreme that reasonably protects against Type I error. Normally, this will be $p < .05$. In general, we don't recommend using a less extreme significance level to increase power because this increases the chances of making a Type I error.
5. **Use a one-tailed test.** Whether you use a one- or two-tailed test depends on the logic of the hypothesis being studied. As with significance level, it is rare that you have much of a choice about this factor.
6. **Use a more sensitive hypothesis-testing procedure.** This is fine if alternatives are available. We consider some options of this kind in Chapter 11. Usually, however, the researcher begins with the most sensitive method available, so little more can be done.

Table 7–4 (page 226) summarizes some practical ways to increase the power of a planned experiment.

The Role of Power When Interpreting the Results of a Study

Understanding statistical power and what affects it is very important in drawing conclusions from the results of research.

Role of Power When a Result Is Statistically Significant: Statistical Significance versus Practical Significance

You have learned that a study with a larger effect size is more likely to come out statistically significant. It also is possible for a study with a very small effect size to come out significant. This is likely to happen when a study has high power due to other factors, especially a large sample size. Consider a sample of 10,000 adults who

Table 7-4 Summary of Practical Ways of Increasing the Power of a Planned Experiment

Feature of the Study	Practical Way of Raising Power	Disadvantages
Predicted difference between population means	Increase the intensity of experimental procedure.	May not be practical or may distort study's meaning.
Population standard deviation	Use a less diverse population.	May not be available; decreases generalizability.
	Use standardized, controlled circumstances of testing or more precise measurement.	Not always practical.
Sample size (N)	Use a larger sample size.	Not always practical; can be costly.
Significance level	Use a more lenient level of significance (such as .10).	Raises the probability of Type I error.
One-tailed versus two-tailed test	Use a one-tailed test.	May not be appropriate for the logic of the study.
Type of hypothesis-testing procedure	Use a more sensitive procedure.	None may be available or appropriate.

complete a new Internet-based counseling program designed to increase their level of happiness. At the end of the program, their mean happiness score is 100.6, compared to the mean happiness score of 100 (Population $SD = 10$) for adults in general. This result would be significant at the $p < .001$ level. So the researchers could be confident that the new program increases people's level of happiness. But the effect size is a tiny .06. This means that the new program increases happiness by only a very small amount. Such a small increase is not likely to make a noticeable difference in people's lives and thus the researchers might conclude that the effect of the program is statistically significant but has little practical significance.

In the fields of clinical psychology and medicine, researchers and clinicians often distinguish between a result being statistically significant and *clinically significant.* The latter phrase means that the result is big enough to make a difference that matters in treating people. Chambless and Hollon (1998) stated the issue quite simply: "If a treatment is to be useful to practitioners it is not enough for treatment effects to be statistically significant: they also need to be large enough to be clinically meaningful" (p. 11).

The message here is that when judging a study's results, there are two questions. First, is the result statistically significant? If it is, you can consider there to be a *real effect.* The next question then is whether the effect size is large enough for the result to be *useful or interesting.* This second question is especially important if the study has practical implications. (Sometimes, in a study testing purely theoretical issues, it may be enough just to be confident that there is an effect at all in a particular direction.)

If the sample was small, you can assume that a statistically significant result is probably also practically significant. On the other hand, if the sample size is very large, you must consider the effect size directly, because it is quite possible that the effect size is too small to be useful. As Bakeman (2006) succinctly noted: "... statistical significance should not overly impress us. After all, even the most miniscule effect can achieve statistical significance if the sample size is large enough" (pp. 136–137).

What we just said may seem a bit of a paradox. Most people assume that the more people there are in the study, the more important its results will be. In a sense,

just the reverse is true. All other things being equal, if a study with only a few participants manages to be significant, that significance must be due to a large effect size. A study with a large number of people that is statistically significant may or may not have a large effect size.

Notice that it is not usually a good idea to compare the significance level of two studies to see which has the more important result. For example, a study with a small number of participants that is significant at the .05 level might well have a large effect size. At the same time, a study with a large number of participants that is significant at the .001 level might well have a small effect size.

The most important lesson from all this is that the word *significant* in statistically significant has a very special meaning. It means that you can be pretty confident that there is some real effect. But it does *not* tell you much about whether that real effect is significant in a practical sense, that it is important or noteworthy.

Role of Power When a Result Is Not Statistically Significant

We saw in Chapter 5 that a result that is not statistically significant is inconclusive. Often, however, we really would like to conclude that there is little or no difference between the populations. Can we ever do that?

Consider the relationship of power to a nonsignificant result. Suppose you carried out a study that had low power and did not get a significant result. In this situation, the result is entirely inconclusive. Not getting a significant result may have come about because the research hypothesis was false or because the study had too little power (for example, because it had too few participants).

On the other hand, suppose you carried out a study that had high power and you did not get a significant result. In this situation, it seems unlikely that the research hypothesis is true. In this situation (where there is high power), a nonsignificant result is a fairly strong argument against the research hypothesis. This does not mean that all versions of the research hypothesis are false. For example, it is possible that the research hypothesis is true and the populations are only very slightly different (and you figured power based on predicting a large difference).

In sum, a nonsignificant result from a study with low power is truly inconclusive. However, a nonsignificant result from a study with high power does suggest either that the research hypothesis is false or that there is less of an effect than was predicted when figuring power.

Summary of the Role of Significance and Sample Size in Interpreting Research Results

Table 7–5 summarizes the role of significance and sample size in interpreting research results.

Table 7-5 Role of Significance and Sample Size in Interpreting Research Results

Result Statistically Significant	Sample Size	Conclusion
Yes	Small	Important result
Yes	Large	Might or might not have practical importance
No	Small	Inconclusive
No	Large	Research hypothesis probably false

How are you doing?

1. (a) What are the two basic ways of increasing the effect size of a planned study? For each, (b) and (c), how can it be done, and what are the disadvantages?
2. What is usually the easiest way to increase the power of a planned study?
3. What are the disadvantages of increasing the power of a planned study by using (a) a more lenient significance level or (b) a one-tailed test rather than a two-tailed test?
4. Why is statistical significance not the same as practical importance?
5. You are comparing two studies in which one is significant at $p < .01$ and the other is significant at $p < .05$. (a) What can you conclude about the two studies? (b) What can you *not* conclude about the two studies?
6. When a result is significant, what can you conclude about effect size if the study had (a) a very large sample size or (b) a very small sample size?
7. When a result is not significant, what can you conclude about the truth of the research hypothesis if the study had (a) a very large sample size or (b) a very small sample size?

Answers

1. (a) The two basic ways of increasing the effect size of a planned study are (1) increase the predicted difference between the population means, and (2) reduce the population standard deviation.
(b) You can increase the predicted difference between the population means by increasing the intensity of the experimental procedure. The disadvantages are that it might change the meaning of the procedure you really want to study and it might not be practical.
(c) You can decrease the population standard deviation by using a less diverse population. This has the disadvantage of not permitting you to apply your results to a more general population. Another way to decrease the population standard deviation is to use more standardized procedures or more accurate measurement. However, this may not be practical.
2. The easiest way to increase the power of a planned study is to increase the sample size.
3. (a) Increasing the power of a planned study by using a more lenient significance level increases the probability of a Type I error. (b) Using a one-tailed test rather than a two-tailed test may not be appropriate to the logic of the study; and if the result comes out opposite to predictions, in principle, it would have to be considered nonsignificant.
4. A statistically significant result means that you can be confident the effect did not occur by chance; it does not, however, mean that it is a large or substantial effect.
5. (a) We can be more confident that the first study's result is not due to chance.
(b) We *cannot* conclude which one has the bigger effect size.
6. (a) Given a very large sample size, the effect size could be small or large.
(b) Given a very small sample size, the effect size is probably large.
7. (a) The research hypothesis is probably not true (or has a much smaller effect size than predicted). (b) You can conclude very little about the truth of the research hypothesis.

Effect Size and Power in Research Articles

It is common for articles to mention effect size. For example, Morehouse and Tobler (2000) studied the effectiveness of an intervention program for "high-risk, multiproblem, inner-city, primarily African-American and Latino youth." The authors reported "Youth who received 5–30 hours of intervention ([the high-dosage group], $n = 101$) were compared with those who received 1–4 hours (the low-dosage group, $n = 31$).... The difference between the groups in terms of reduction in [alcohol and drug] use was highly significant. A between-groups effect size of .68 was achieved for the high-dosage group when compared with the low-dosage group." The meaning of the .68 effect size is that the group getting 5 to 30 hours of intervention was .68 standard deviations higher in terms of reduction of their drug and alcohol use than the group getting only 1 to 4 hours of the intervention. This is a medium to large effect size. Effect sizes are also almost always reported in meta-analyses, in which results from different articles are being combined and compared (for an example, see Box 7–1 earlier in the chapter).

As was the case with decision errors, you usually think about power when planning research and evaluating the results of a research study. (Power, for example, is often a major topic in grant proposals requesting funding for research and in thesis proposals.) As for research articles, power is sometimes mentioned in the final section of an article where the author discusses the meaning of the results or in discussions of results of other studies. In either situation, the emphasis tends to be on the meaning of nonsignificant results. Also, when power is discussed, it may be explained in some detail. This is because it has been only recently that most behavioral and social scientists have begun to be knowledgeable about power.

For example, Denenberg (1999), in discussing the basis for his own study, makes the following comments about a relevant previous study by Mody, Studdert-Kennedy, and Brady (1997) that had not found significant results.

> [T]hey were confronted with the serious problem of having to accept the null hypothesis.... We can view this issue in terms of statistical power.... A minimal statistical power of .80 [80%] is required before one can consider the argument that the lack of significance may be interpreted as evidence that Ho [the null hypothesis] is true. To conduct a power analysis, it is necessary to specify an expected mean difference, the alpha [significance] level, and whether a one-tailed or two-tailed test will be used. Given a power requirement of .8, one can then determine the *N* necessary. Once these conditions are satisfied, if the experiment fails to find a significant difference, then one can make the following kind of a statement: "We have designed an experiment with a .8 probability of finding a significant difference, if such exists in the population. Because we failed to find a significant effect, we think it quite unlikely that one exists. Even if it does exist, its contribution would appear to be minimal...."
>
> Mody et al. never discussed power, even though they interpreted negative findings as evidence for the validity of the null hypothesis in all of their experiments.... Because the participants were split in this experiment, the *n*s [sample sizes] were reduced to 10 per group. Under such conditions one would not expect to find a significant difference, unless the experimental variable was very powerful. In other words it is more difficult to reject the null hypothesis when working with small *n*s [sample sizes]. The only meaningful conclusion that can be drawn from this study is that no meaningful interpretation can be made of the lack of findings.... (pp. 380–381)*

*Excerpt from A critique of Mody, Studdert-Kennedy, and Brady's "Speech perception deficits in poor readers: Auditory processing or phonological coding?" Victor H. Denenberg. *Journal of Learning Disabilities.* Austin: Sep/Oct 1999. Vol. 32, Iss. 5; p. 379. Copyright © 1999, Hammill Institute on Disabilities. Reprinted by permission of Sage Publications.

Here is another example. Huey and Polo (2008) conducted a review of research on psychological treatments for a variety of emotional and behavioral problems (such as anxiety, depression, and substance use) among ethnic minority youth. In discussing their results, they noted the following: "[A] concern is whether sample sizes have been sufficient to test key hypotheses. The absence of difference does not necessarily indicate group equivalence, and may suggest that studies lack adequate statistical power" (p. 295). They went on to state that "larger samples are needed to better answer key questions of theoretical interest to minority mental health researchers. Although there are other methods for maximizing statistical power (e.g., using more sensitive measures, adjusting alpha [significance] level), increasing sample size is perhaps the most practical approach" (p. 295).

Learning Aids

Summary

1. Effect size is a measure of the difference between population means. In the hypothesis-testing situations you learned in this chapter, you can think of effect size as how much something changes after a specific intervention. The effect size is figured by dividing the difference between population means by the population standard deviation. Cohen's effect size conventions consider a small effect to be .20, a medium effect to be .50, and a large effect to be .80. Effect size is important in its own right in interpreting results of studies. It is also used to compare and combine results of studies, as in meta-analysis, and to compare different results within a study.
2. The statistical power of a study is the probability that it will produce a statistically significant result *if the research hypothesis is true.* Researchers usually figure the power of a study using power software packages, Internet power calculators, or special tables.
3. The larger the effect size is, the greater the power is. This is because the greater the difference between means or the smaller the population standard deviation is (the two ingredients in effect size), the less overlap there is between the known and predicted populations' distributions of means. Thus, the area in the predicted distribution that is more extreme than the cutoff in the known distribution is greater.
4. The larger the sample size is, the greater the power is. This is because the larger the sample is, the smaller is the variance of the distribution of means. So, for a given effect size, there is less overlap between distributions of means.
5. Power is also affected by significance level (the more extreme, such as $p < .01$, the lower the power), by whether a one- or two-tailed test is used (with less power for a two-tailed test), and by the type of hypothesis-testing procedure used (in the occasional situation where there is a choice of procedure).
6. Statistically significant results from a study with high power (such as one with a large sample size) may not have practical importance. Results that are not statistically significant from a study with low power (such as one with a small sample size) leave open the possibility that statistically significant results might show up if power were increased.
7. Research articles commonly report effect size, and effect sizes are almost always reported in meta-analyses. Research articles sometimes include discussions of power, especially when evaluating nonsignificant results.

Key Terms

effect size (p. 205)
effect size conventions (p. 207)
meta-analysis (p. 209)
statistical power (p. 210)
power tables (p. 214)

Example Worked-Out Problem

In a known population with a normal distribution, Population $M = 40$ and Population $SD = 10$. A sample given an experimental treatment has a mean of 37. What is the effect size? Is this approximately small, medium, or large?

Answer

Effect size = (Population 1 M − Population 2 M)/Population $SD = (37 - 40)/10 = -3/10 = -.30$. Approximately small.

Outline for Writing Essays on Effect Size and Power for Studies Involving a Single Sample of More than One Individual and a Known Population

1. Explain the idea of effect size as the degree of overlap between distributions. Note that this overlap is a function of mean difference and population standard deviation (and describe precisely how it is figured and why it is figured that way). If required by the question, discuss the effect size conventions.
2. Explain the idea of power as the probability of getting significant results if the research hypothesis is true. Be sure to mention that the usual minimum acceptable level of power for a research study is 80%. Explain the role played by power when you are interpreting the results of a study (both when a study is and is not significant), taking into account significance levels and sample size in relation to the likely effect size.
3. Explain the relationship between effect size and power.

Practice Problems

These problems involve figuring. Most real-life statistics problems are done on a computer with special statistical software. Even if you have such software, do these problems by hand to ingrain the method in your mind.

Set I (for answers, see p. 453)

1. In a completed study, there is a known population with a normal distribution, Population $M = 25$, and Population $SD = 12$. What is the estimated effect size if a sample given an experimental procedure has a mean of (a) 19, (b) 22, (c) 25, (d) 30, and (e) 35? For each part, also indicate whether the effect is approximately small, medium, or large.
2. In a planned study, there is a known population with a normal distribution, Population $M = 50$, and Population $SD = 5$. What is the predicted effect size if the researchers predict that those given an experimental treatment have a mean of (a) 50, (b) 52, (c) 54, (d) 56, and (e) 47? For each part, also indicate whether the predicted effect is approximately small, medium, or large.

3. Here is information about several possible versions of a planned study, each involving a single sample. Figure the predicted effect size for each study:

	Population 2		Predicted
Study	*M*	*SD*	*Population 1 M*
(a)	90	4	91
(b)	90	4	92
(c)	90	4	94
(d)	90	4	86

4. You read a study in which the result is significant ($p < .05$). You then look at the size of the sample. If the sample is very large (rather than very small), how should this affect your interpretation of (a) the probability that the null hypothesis is actually true and (b) the practical importance of the result? (c) Explain your answers to a person who understands hypothesis testing but has never learned about effect size or power.
5. Aron et al. (1997) placed strangers in pairs and asked them to talk together following a series of instructions designed to help them become close. At the end of 45 minutes, individuals privately answered some questions about how close they now felt to their partners. (The researchers combined the answers into a "closeness composite.") One key question was whether closeness would be affected by either (a) matching strangers based on their attitude agreement or (b) leading participants to believe that they had been put together with someone who would like them. The result for both agreement and expecting to be liked was that "there was no significant differences on the closeness composite" (p. 367). The researchers went on to argue that the results suggested that there was little true effect of these variables on closeness (note that the symbol *d* in the text below means effect size):

   > There was about 90% power in this study of achieving significant effects ... for the two manipulated variables if in fact there were a large effect of this kind (d [effect size] = .8). Indeed, the power is about 90% for finding at least a near significant ($p < .10$) medium-sized effect (d [effect size] = .5). Thus, it seems unlikely that we would have obtained the present results if in fact there is more than a small effect.... (p. 367)

   Explain this result to a person who understands hypothesis testing but has never learned about effect size or power.
6. How does each of the following affect the power of a planned study?
   (a) A larger predicted difference between the means of the populations
   (b) A larger population standard deviation
   (c) A larger sample size
   (d) Using a more extreme significance level (e.g., $p < .01$ instead of $p < .05$)
   (e) Using a two-tailed test instead of a one-tailed test
7. List two situations in which it is useful to consider power, indicating what the use is for each.

Set II

8. In a completed study, there is a known population with a normal distribution, Population $M = 122$, and Population $SD = 8$. What is the estimated effect size if a sample given an experimental procedure has a mean of (a) 100, (b) 110, (c) 120,

(d) 130, and (e) 140? For each part, also indicate whether the effect is approximately small, medium, or large.

9. In a planned study, there is a known population with a normal distribution, Population $M = 0$, and Population $SD = 10$. What is the predicted effect size if the researchers predict that those given an experimental treatment have a mean of (a) −8, (b) −5, (c) −2, (d) 0, and (e) 10? For each part, also indicate whether the predicted effect is approximately small, medium, or large.
10. Here is information about several possible versions of a planned study, each involving a single sample. Figure the predicted effect size for each study:

	Population 2		Predicted
Study	*M*	*SD*	*Population 1 M*
(a)	90	2	91
(b)	90	1	91
(c)	90	2	92
(d)	90	2	94
(e)	90	2	86

11. What is meant by effect size? (Write your answer for a layperson.)
12. In the "Effect Size and Power in Research Articles" section earlier in the chapter, you read about a review study conducted by Huey and Polo (2008) that examined psychological treatments for clinical problems among ethnic minority youth. As part of their review, the researchers identified 25 studies that compared the effect of a psychotherapy treatment versus a control treatment on youths' clinical problems. They conducted a meta-analysis of the 25 studies and reported the results as follows (note that the symbol *d* in the text below means effect size):

    > [T]he mean effect size was $d = .44$. Because coefficients of .20 or lower represent "small" effects, coefficients around .50 "medium" effects, and coefficients of .80 or higher "large effects," the overall *d* reported here falls somewhat below the standard for a "medium" effect (Cohen, 1988). (p. 282)

    Explain the purpose and results of this meta-analysis to someone who is familiar with effect size but has never heard of meta-analysis.
13. What is meant by the statistical power of an experiment? (Write your answer for a layperson.)
14. You read a study that just barely fails to be significant at the .05 level. That is, the result is not statistically significant. You then look at the size of the sample. If the sample is very large (rather than very small), how should this affect your judgment of (a) the probability that the null hypothesis is actually true and (b) the probability that the null hypothesis is actually false? (c) Explain your answers to a person who understands hypothesis testing but has never learned about power.
15. Caspi et al. (1997) analyzed results from a large-scale longitudinal study of a sample of children born around 1972 in Dunedin, New Zealand. As one part of their study, the researchers compared the 94 in their sample who were, at age 21, alcohol dependent (clearly alcoholic) versus the 863 who were not alcohol dependent. Caspi et al. compared these two groups in terms of personality test scores from when they were 18 years old. After noting that all results were significant, they reported the following results (note that the symbol *d* in the text below means effect size):

    > Young adults who were alcohol dependent at age 21 scored lower at age 18 on Traditionalism ($d = .49$), Harm Avoidance ($d = .44$), Control ($d = .64$), and

Social Closeness (d = .40), and higher on Aggression (d = .86), Alienation (d = .66), and Stress Reaction (d = .50).

Explain these results, including why it was especially important for the researchers in this study to give effect sizes, to a person who understands hypothesis testing but has never learned about effect size or power.

16. Tsang, Colley, and Lynd (2009) conducted a review to examine the statistical power of studies that had compared patients' experiences of serious adverse events (such as a life-threatening medical event) during randomized controlled trials of medical treatments. They identified six studies that reported the results of statistical analyses to test whether the number of adverse effects experienced by patients receiving one medical treatment differed from the number experienced by those receiving a different treatment. Tsang et al. summarized their results as follows: "Three of the six studies included in this analysis reported non-statistically significant differences in serious adverse event rates, and concluded that there was no difference in risk despite [having power] of less than 0.37 to detect the reported differences" (p. 610). They also noted: "A high probability of type II error may lead to erroneous clinical inference resulting in harm. The statistical power for nonsignificant tests should be considered in the interpretation of results" (p. 609). Explain the results of this review to a person who understands hypothesis testing and decision errors but has never learned about effect size or power.
17. You are planning a study that you determine from a power table as having quite low power. Name six things that you might do to increase power.