Hypothesis Tests with Means of Samples

Chapter Outline

In actual practice, behavioral and social science research almost always involve a sample of many individuals. In this chapter, we build on what you have learned so far and consider hypothesis testing involving a sample of more than one individual. For example, a team of educational researchers is interested in the effects of different kinds of instructions on timed school achievement tests. These educational researchers have a theory that makes the following prediction: People will do better on a test if they are given instructions to answer each question with the first response that comes to mind. To test this theory, the researchers give a standard school achievement test to 64 randomly selected fifth-grade schoolchildren. They give the test in the usual way, except that they add to the instructions a statement that children are to answer each question with the first response that comes to mind. From previous testing, the researchers know the population mean and standard deviation of the test when it is given in the usual way (without any special instructions). In this chapter, you will learn how to test hypotheses in situations like this example: situations in which the population has a *known mean and standard deviation* and in which a sample has

From Chapter 6 of *Statistics for the Behavioral & Social Sciences: A Brief Course*, Fifth Edition. Arthur Aron, Elliot J. Coups, Elaine N. Aron. Copyright © 2011 by Pearson Education, Inc. Published by Pearson Prentice Hall. All rights reserved.

more than one individual. Mainly, this requires examining in some detail a new kind of distribution, called a "distribution of means." (We return to the educational researchers' special instructions example later in the chapter.)

The Distribution of Means

Hypothesis testing in the usual research situation, where you are studying a sample of many individuals, is exactly the same as you have already learned —with an important exception. When you have more than one person in your sample, there is a special problem with Step ❷, determining the characteristics of the comparison distribution. In each of our examples so far, the comparison distribution has been a distribution of *individual scores* (such as the population of ages when individual babies start walking). A distribution of individual scores has been the correct comparison distribution because we have used examples with a sample of *one individual*. That is, there has been a match between the type of sample score we have been dealing with (a score from *one individual*) and the comparison distribution (a distribution of *individual scores*).

Now, consider the situation when you have a sample of, say, 64 individuals. You now have a *group of 64 scores*. The mean is a very useful representative value of a group of scores. Thus, the score you care about when there is more than one individual in your sample is the *mean of the group of scores*. In this example, you would focus on the mean of the 64 individuals. If you were to compare the mean of this sample of 64 individuals to a distribution of a population of individual scores, this would be a mismatch—like comparing apples to oranges. Instead, when you are interested in the mean of a sample of 64 scores you need a comparison distribution that is a distribution of means of samples of 64 scores. We call such a comparison distribution a **distribution of means.** The scores in a distribution of means are *means*, not scores of individuals.

A distribution of means is a distribution of the means of each of lots and lots of samples of the same size, with each sample randomly taken from the same population of individuals. (Statisticians also call this distribution of means a *sampling distribution of the mean*. In this text, however, we use the term *distribution of means* to make it clear that we are talking about populations of *means*, not samples or a distribution of samples.)

The distribution of means is the correct comparison distribution when there is more than one person in a sample. Thus, in most research situations, determining the characteristics of a distribution of means is necessary for Step ❷ of the hypothesis-testing procedure (determining the characteristics of the comparison distribution).

Building a Distribution of Means

To help you understand the idea of a distribution of means, we consider how you could build up such a distribution from an ordinary population distribution of individual scores. Suppose our population of individual scores was the grade levels of the 90,000 elementary and junior high school children in a particular region. Suppose further (to keep the example simple) that there are exactly 10,000 children at each grade level, from first through ninth grade. This population distribution would be rectangular, with a mean of 5, a variance of 6.67, and a standard deviation of 2.58 (see Figure 1).

Next, suppose that you wrote each child's grade level on a table tennis ball and put all 90,000 balls into a giant tub. The tub would have 10,000 balls with a 1 on them, 10,000 with a 2 on them, and so forth. Stir up the balls in the tub, and then take two of them out. You have taken a random sample of two balls. Suppose one ball has

distribution of means Distribution of means of samples of a given size from a particular population (also called a sampling distribution of the mean); comparison distribution when testing hypotheses involving a single sample of more than one individual.

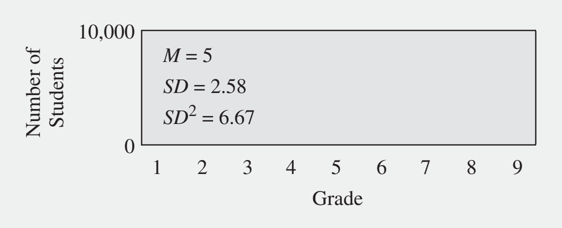


Figure 1 Distribution of grade levels among 90,000 schoolchildren (fictional data).

a 2 on it and the other has a 9 on it. The mean grade level of this sample of two children's grade levels is 5.5, the average of 2 and 9. Now you put the balls back, mix up all the balls, and select two balls again. Maybe this time you get two 4s, making the mean of your second sample 4. Then you try again; this time you get a 2 and a 7, making your mean 4.5. So far you have three means: 5.5, 4, and 4.5.

Each of these three numbers is a mean of a sample of grade levels of two schoolchildren. And these three means can be thought of as a small distribution in its own right. The mean of this little distribution of means is 4.67 (the sum of 5.5, 4, and 4.5 divided by 3). The variance of this distribution of means is .39 (the variance of 5.5, 4, and 4.5). The standard deviation of this distribution of means is .62 (the square root of .39). A histogram of this distribution of three means is shown in Figure 2.

Suppose you continued selecting samples of two balls and taking the mean of the numbers on each pair of balls. The histogram of means would continue to grow. Figure 3 shows examples of distributions of means with just 50 means, up to a distribution of means with 1,000 means (with each mean being of a sample of two randomly drawn balls). (We actually made the histograms shown in Figure 3 using a computer to make the random selections instead of using 90,000 table tennis balls and a giant tub.)

As you can imagine, the method we just described is not a practical way of determining the characteristics of a distribution of means. Fortunately, you can figure out the characteristics of a distribution of means directly, using some simple rules, without taking even one sample. The only information you need is (1) the characteristics of the distribution of the population of individuals and (2) the number of scores in each sample. (Don't worry for now about how you could know the characteristics of the population of individuals.) The laborious method of building up a distribution of means in the way we have just considered and the concise method you will learn shortly give the same result. We have had you think of the process in terms of the painstaking method only because it helps you understand the idea of a distribution of means.

TIP FOR SUCCESS

Before moving on, be sure you fully understand the idea of a distribution of means (and why it is the correct comparison distribution when a sample has more than one individual). You may need to go through this chapter a couple of times to achieve full understanding of this crucial concept.

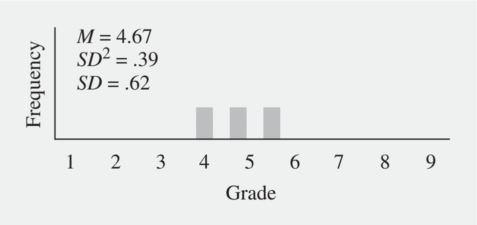


Figure 2 Distribution of the means of three randomly taken samples of two schoolchildren's grade levels, each from a population of grade levels of 90,000 schoolchildren (fictional data).

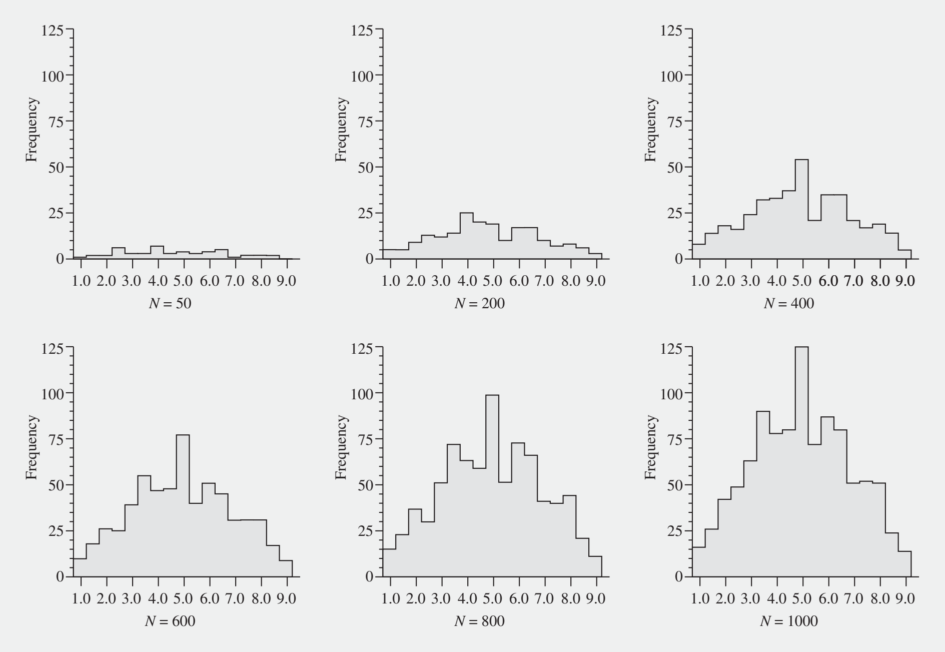


Figure 3 Histograms of means of two grade levels randomly selected from a large group of students with equal numbers of grades 1 through 9. Histograms are shown for 50 such means, 200 such means, and so forth, up to 1000 such means. Notice that the histograms become increasingly like a normal curve as the number of means increases.

Determining the Characteristics of a Distribution of Means

Recall that Step ❷ of hypothesis testing involves determining the characteristics of the comparison distribution. The three characteristics of the comparison distribution that you need are as follows:

1. Its mean
2. Its spread (which you can measure using the variance and standard deviation)
3. Its shape

Notice three things about the distribution of means we built up in our example, as shown in Figure 3:

1. The mean of the distribution of means is about the same as the mean of the original population of individuals (both are 5).
2. The spread of the distribution of means is less than the spread of the distribution of the population of individuals.
3. The shape of the distribution of means is approximately normal.

The first two observations, regarding the mean and the spread, are true for all distributions of means. The third, regarding the shape, is true for most distributions of means. These three observations, in fact, illustrate three basic rules that you can use to find the mean, the spread (that is, the variance and standard deviation), and the shape of any distribution of means without having to write on plastic balls and take endless samples.

Now, let's look at the three rules more closely. The first is for the **mean of a distribution of means.**

mean of a distribution of means The mean of a distribution of means of samples of a given size from a particular population; it comes out to be the same as the mean of the population of individuals.

Rule 1: The mean of a distribution of means is the same as the mean of the population of individuals. Stated as a formula,

$$\text{Population } M_M = \text{Population } M \qquad (1)$$

The mean of the distribution of means is equal to the mean of the population of individuals.

Population M_M is the mean of the distribution of means. It uses the word *Population* because the distribution of means is also a kind of population. Population *M* is the mean of the population of individuals.

Each sample is based on randomly selected individuals from the population of individuals. Thus, the mean of a sample will sometimes be higher and sometimes lower than the mean of the whole population of individuals. However, because the selection process is random and because we are taking a very large number of samples, eventually the high means and the low means perfectly balance each other out.

In Figure 3, as the number of sample means in the distributions of means increases, the mean of the distribution of means becomes more similar to the mean of the population of individuals, which in this example was 5. It can be proven mathematically that, if you took an infinite number of samples, the mean of the distribution of means of these samples would come out to be exactly the same as the mean of the distribution of individuals.

The second rule is about spread. This second rule has two parts: the first part, Rule 2a, is for the **variance of a distribution of means.**

Rule 2a: The variance of a distribution of means is the variance of the population of individuals divided by the number of individuals in each sample. A distribution of means will be less spread out than the population of individuals from which the samples are taken. If you are taking a sample of two scores, it is unlikely that *both* scores will be extreme. Furthermore, for a particular random sample to have an extreme mean, the two extreme scores would both have to be extreme in the same direction (both very high or both very low). Thus, having more than a single score in each sample has a moderating effect on the mean of such samples. In any one sample, the extremes tend to be balanced out by a middle score or by an extreme in the opposite direction. This makes each sample mean tend toward the middle and away from extreme values. With fewer extreme means, there is less spread, and the variance of the means is less than the variance of the population of individuals.

Consider again our example. There were plenty of 1s and 9s in the population, making a fair amount of spread. That is, about a ninth of the time, if you were taking samples of single scores, you would get a 1 and about a ninth of the time you would get a 9. If you are taking samples of two at a time, you would get a sample with a mean of 1 (that is, in which *both* balls were 1s) or a mean of 9 (both balls 9s) much less often. Getting two balls that average out to a middle value such as 5 is much more likely. (This is because several combinations could give this result—1 and 9, 2 and 8, 3 and 7, 4 and 6, or two 5s.)

The more individuals in each sample, the less spread out will be the means of the samples. This is because the more scores in each sample, the rarer it will be for extremes in any particular sample not to be balanced out by middle scores or extremes in the other direction. In terms of the plastic balls in our example, we rarely got a mean of 1 when taking samples of two balls at a time. If we were taking three balls at a time, getting a sample with a mean of 1 (all three balls would have to be 1s) is even less likely. Getting middle values for the means becomes more likely.

Using samples of two balls at a time, the variance of the distribution of means came out to about 3.34. This is half of the variance of the population of individuals, which was 6.67. If we had built up a distribution of means using samples of three

variance of a distribution of means Variance of the population divided by the number of scores in each sample.

balls each, the variance of the distribution of means would have been 2.22. This is one-third of the variance of the population of individuals. Had we randomly selected five balls for each sample, the variance of the distribution of means would have been one-fifth of the variance of the population of individuals.

These examples follow a general rule—our Rule 2a for the distribution of means: The variance of a distribution of means is the variance of the population of individuals divided by the number of individuals in each of the samples. This rule holds in all situations and can be proven mathematically.

Here is Rule 2a stated as a formula:

The variance of a distribution of means is the variance of the population of individuals divided by the number of individuals in each sample.

$$\text{Population } SD_M^2 = \frac{\text{Population } SD^2}{N} \tag{2}$$

Population SD_M^2 is the variance of the distribution of means, Population SD^2 is the variance of the population of individuals, and N is the number of individuals in each sample.

In our example, the variance of the population of individual children's grade levels was 6.67, and there were two children's grade levels in each sample. Thus,

$$\text{Population } SD_M^2 = \frac{\text{Population } SD^2}{N} = \frac{6.67}{2} = 3.34.$$

To use a different example, suppose a population had a variance of 400 and you wanted to know the variance of a distribution of means of 25 individuals each:

$$\text{Population } SD_M^2 = \frac{\text{Population } SD^2}{N} = \frac{400}{25} = 16.$$

The second rule also tells us about the **standard deviation of a distribution of means.**

Rule 2b: The standard deviation of a distribution of means is the square root of the variance of the distribution of means. Stated as a formula,

The standard deviation of a distribution of means is the square root of the variance of the distribution of means.

$$\text{Population } SD_M = \sqrt{\text{Population } SD_M^2} \tag{3}$$

Population SD_M is the standard deviation of the distribution of means.

The standard deviation of the distribution of means also has a special name of its own, the **standard error (*SE*)** (or *standard error of the mean*, abbreviated *SEM*). It has this name because it tells you how much the means of samples are typically "in error" as estimates of the mean of the population of individuals. That is, it tells you how much the various means in the distribution of means deviate from the mean of the population. (We have more to say about this idea in the "Advanced Topic" section on confidence intervals later in the chapter.)

Finally, the third rule for finding the characteristics of a distribution of means focuses on its shape.

Rule 3: The shape of a distribution of means is approximately normal if either (a) each sample is of 30 or more individuals or (b) the distribution of the population of individuals is normal. Whatever the shape of the distribution of the population of individuals, the distribution of means tends to be unimodal and symmetrical. In the grade-level example, the population distribution was rectangular. (It had an equal number at each grade level.) However, the **shape of the distribution of means** (Figure 3) was roughly that of a bell—unimodal and symmetrical. Had we taken many more than 1,000 samples, the shape would have been even more clearly unimodal and symmetrical.

standard deviation of a distribution of means (Population SD_M) Square root of the variance of the distribution of means; same as *standard error (SE).*

standard error (*SE*) Same as *standard deviation of a distribution of means*; also called *standard error of the mean* (*SEM*).

shape of a distribution of means Contour of a histogram of a distribution of means, such as whether it follows a normal curve or is skewed; in general, a distribution of means will tend to be unimodal and symmetrical and is often normal.

A distribution of means tends to be unimodal because of the same basic process of extremes balancing each other out that we noted in the discussion of the variance: middle scores for means are more likely, and extreme means are less likely. A distribution of means tends to be symmetrical because a lack of symmetry (skew) is caused by extremes. With fewer extremes, there is less asymmetry. In our grade-level example, the distribution of means we built up also came out so clearly symmetrical because the population distribution of individual grade levels was symmetrical. Had the population distribution of individuals been skewed to one side, the distribution of means would have still been skewed, but not as much.

The more individuals in each sample, the closer the distribution of means will be to a normal curve. Although the distribution of means will rarely be an exactly normal curve, with samples of 30 or more individuals (even with a non-normal population of individuals), the approximation of the distribution of means to a normal curve is very close and the percentages in the normal curve table will be extremely accurate.[1, 2] (That is, samples that are larger than 30 make for even slightly better approximations, but for most practical research purposes, the approximation with 30 is quite good enough.) Finally, whenever the population distribution of individuals is normal, the distribution of means will be normal, regardless of the number of individuals in each sample.

Summary of Rules and Formulas for Determining the Characteristics of a Distribution of Means

Rule 1: The mean of a distribution of means is the same as the mean of the population of individuals:

$$\text{Population } M_M = \text{Population } M$$

Rule 2a: The variance of a distribution of means is the variance of the population of individuals divided by the number of individuals in each sample:

$$\text{Population } SD_M^2 = \frac{\text{Population } SD^2}{N}$$

Rule 2b: The standard deviation of a distribution of means is the square root of the variance of the distribution of means:

$$\text{Population } SD_M = \sqrt{\text{Population } SD_M^2}$$

[1] We have ignored the fact that a normal curve is a smooth theoretical distribution. However, in most real-life distributions, scores are only at specific numbers, such as a child being in a particular grade and not in a fraction of a grade. So, one difference between our example distribution of means and a normal curve is that the normal curve is smooth. However, in behavioral and social science research, even when our measurements are at specific numbers, we usually treat the situation as if the underlying thing being measured is continuous.

[2] Think of any distribution of individual scores in nature as a situation in which each score is actually an average of a random set of influences on that individual. Consider the distribution of weights of pebbles. Each pebble's weight is a kind of average of all the different forces that went into making the pebble have a particular weight. Statisticians refer to this general principle as the Central Limit Theorem.

Rule 3: The shape of a distribution of means is approximately normal if either (a) each sample is of 30 or more individuals or (b) the distribution of the population of individuals is normal.

Figure 4 shows these three rules graphically.

Example of Determining the Characteristics of a Distribution of Means

Consider the population of scores of students who have taken the Graduate Record Examinations (GREs): Suppose the distribution is approximately normal with a mean of 500 and a standard deviation of 100. What will be the characteristics of the distribution of means of samples of 50 students?

Rule 1: The mean of a distribution of means is the same as the mean of the population of individuals. The mean of the population is 500. Thus, the mean of the distribution of means will also be 500. That is, Population M_M = Population $M = 500$.

Rule 2a: The variance of a distribution of means is the variance of the population of individuals divided by the number of individuals in each sample. The standard deviation of the population of individuals is 100; thus, the variance of the population of individuals is 100^2, which is 10,000. The variance of the distribution of means is therefore 10,000 divided by 50 (the size of the sample). This comes out to 200. That is, Population SD_M^2 = Population $SD^2/N = 10{,}000/50 = 200$.

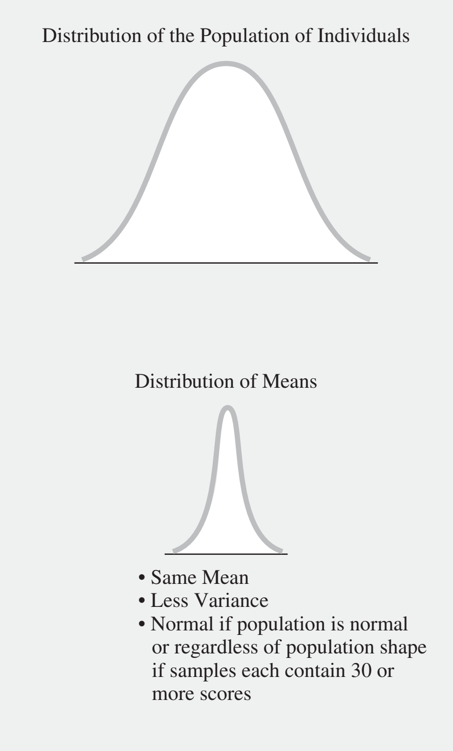

Distribution of the Population of Individuals
Distribution of Means
• Same Mean
• Less Variance
• Normal if population is normal or regardless of population shape if samples each contain 30 or more scores

Figure 4 Comparing the distribution of the population of individuals (upper curve) and the distribution of means (lower curve).

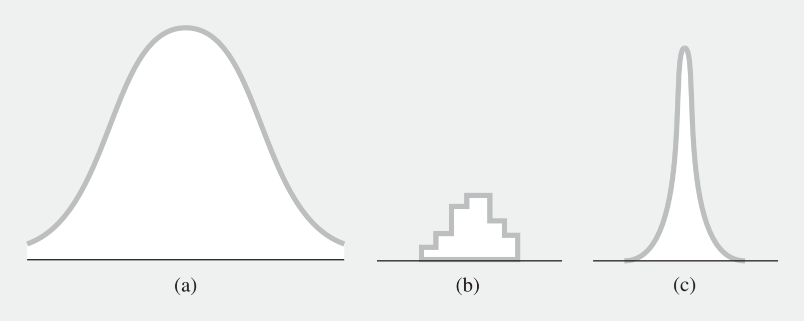


Figure 5 Three kinds of distributions: (a) the distribution of a population of individuals, (b) the distribution of a particular sample taken from that population, and (c) the distribution of means of samples taken from that population.

Rule 2b: The standard deviation of a distribution of means is the square root of the variance of the distribution of means. The standard deviation of the distribution of means is the square root of 200, which is 14.14. That is, Population $SD_M = \sqrt{\text{Population } SD^2_M} = \sqrt{200} = 14.14$.

Rule 3: The shape of a distribution of means is approximately normal if either (a) each sample is of 30 or more individuals or (b) the distribution of the population of individuals is normal. Our situation meets both of these conditions—the sample of 50 students is more than 30, and the population of individuals follows a normal distribution. Thus, the distribution of means will follow a normal curve. (It would have been enough even if only one of the two conditions had been met.)

Review of the Three Kinds of Distributions

We have considered three different kinds of distributions: (1) the distribution of a population of individuals, (2) the distribution of a particular sample of individuals taken from that population, and (3) the distribution of means. Figure 5 shows these three kinds of distributions graphically and Table 1 describes them.

Table 1 Comparison of Three Types of Distributions

	Population's Distribution	Particular Sample's Distribution	Distribution of Means
Content	Scores of all individuals in the population	Scores of the individuals in a single sample	Means of samples randomly taken from the population
Shape	Could be any shape; often normal	Could be any shape	Approximately normal if samples have ≥30 individuals in each or if population is normal
Mean	Population M	$M = (\Sigma X)/N$; figured from scores in the sample	Population M_M $=$ Population M
Variance	Population SD^2	$SD^2 = [\Sigma(X - M)^2]/N$	Population SD^2_M $=$ Population SD^2/N
Standard Deviation	Population SD	$SD = \sqrt{SD^2}$	Population SD_M $= \sqrt{\text{Population } SD^2_M}$

TIP FOR SUCCESS

Be sure you fully understand the different types of distributions shown in Table 1 before you move on. To check your understanding, cover up portions of the table and then try to recall the information that is covered up.

How are you doing?

1. What is a distribution of means?
2. Explain how you could create a distribution of means by taking a large number of samples of four individuals each.
3. (a) Why is the mean of the distribution of means the same as the mean of the population of individuals? (b) Why is the variance of a distribution of means smaller than the variance of the distribution of the population of individuals?
4. Write the formula for the variance of the distribution of means, and define each of the symbols.
5. (a) What is the standard error (*SE*)? (b) Why does it have this name?
6. A population of individuals has a normal distribution, a mean of 60, and a standard deviation of 10. What are the characteristics of a distribution of means from this population for samples of four each?

Answers

1. A distribution of means is a distribution of the means of a very large number of samples, each of the same size, randomly taken from the population of individuals.
2. Take a random sample of four scores from the population and figure its mean. Do this a very large number of times. Make a distribution of all of the means.
3. (a) With randomly taken samples, some will have higher means and some lower means than the population of individuals; in the long run these have to balance out. (b) You are less likely to get a sample of several scores that has an extreme mean than you are to get a single extreme score. This is because in any random sample it is highly unlikely to get several extremes in the same direction; extreme scores tend to be balanced out by middle scores or extremes in the opposite direction. Thus, with fewer extreme scores and more middle scores, there is less variance.
4. The formula for the variance of the distribution of means is: Population SD^2_M = Population SD^2/N. Population SD^2_M is the variance of the distribution of means; Population SD^2 is the variance of the population of individuals; N is the number of individuals in your sample.
5. (a) The standard error is the standard deviation of the distribution of means. (b) It has this name because it tells you about how much means of samples typically (standardly) differ from the population mean, and thus tells you the typical amount that the means of samples are in error as estimates of the population mean.
6. The characteristics of a distribution of means are calculated as follows: Population M_M = Population $M = 60$. Population SD^2_M = Population SD^2/N $= 10^2/4 = 25$; Population $SD_M = 5$. Shape = normal.

Hypothesis Testing with a Distribution of Means: The *Z* Test

Now we are ready to turn to hypothesis testing when there is more than one individual in the study's sample. The hypothesis-testing procedure you will learn is called a **Z test** (or a *Z test for a single sample*), because it is the *Z* score that is checked against the normal curve.

The Distribution of Means as the Comparison Distribution in Hypothesis Testing

Z **test** Hypothesis-testing procedure in which there is a single sample and the population variance is known.

In the usual research situation in the behavioral and social sciences, a researcher studies a sample of more than one person. In this situation, the distribution of means is the comparison distribution. It is the distribution whose characteristics need to be determined in Step ❷ of the hypothesis-testing process. The distribution of means is the distribution to which you compare your sample's mean to see how likely it is that you could have selected a sample with a mean that extreme *if the null hypothesis were true*.

Figuring the *Z* Score of a Sample's Mean on the Distribution of Means

There can be some confusion in figuring the location of your sample on the comparison distribution in hypothesis testing with a sample of more than one. In this situation, you are finding a *Z* score of your sample's mean on a distribution of means. (Before, you were finding the *Z* score of a single individual on a distribution of a population of single individuals.) The method of changing the sample's mean to a *Z* score is the same as the usual way of changing a raw score to a *Z* score. However, you have to be careful not to get mixed up because more than one mean is involved. It is important to remember that you are treating the sample mean like a single score. Recall that the ordinary formula for changing a raw score to a *Z* score is $Z = (X - M)/SD$. In the present situation, you are actually using the following conceptually identical formula:

$$Z = \frac{(M - \text{Population } M_M)}{\text{Population } SD_M} \quad (4)$$

The *Z* score for the sample's mean on the distribution of means is the sample's mean minus the mean of the distribution of means, divided by the standard deviation of the distribution of means.

For example, suppose your sample's mean is 18 and the distribution of means has a mean of 10 and a standard deviation of 4. The *Z* score of this sample mean is +2. Using the formula,

$$Z = \frac{(M - \text{Population } M_M)}{\text{Population } SD_M} = \frac{(18 - 10)}{4} = \frac{8}{4} = 2.$$

This is shown in Figure 6.

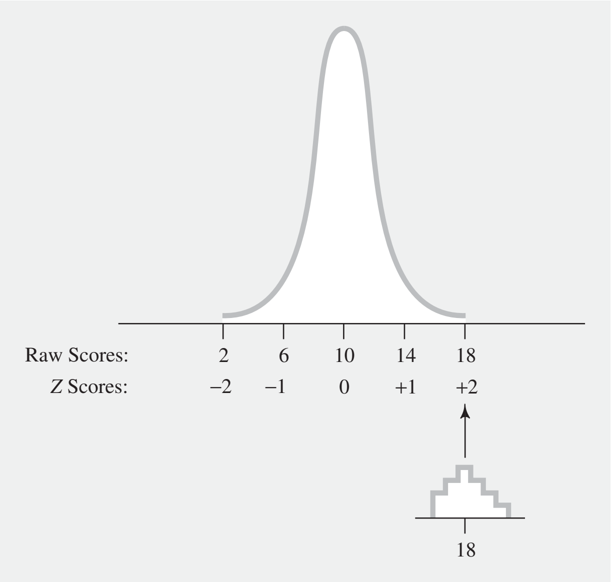


Figure 6 *Z* score for the mean of a particular sample on the distribution of means.

Example. Let's return to the example at the start of the chapter in which a team of educational researchers is interested in the effects of instructions on timed school achievement tests. The researchers give a standard school achievement test to 64 randomly selected fifth-grade schoolchildren. They give the test in the usual way, except that they add to the instructions a statement that children are to answer each question with the first response that comes to mind. When given in the usual way, the test is known to have a mean of 200, a standard deviation of 48, and an approximately normal distribution. This distribution is shown in Figure 7a.

Now let's carry out the *Z* test by following the five steps of hypothesis testing.

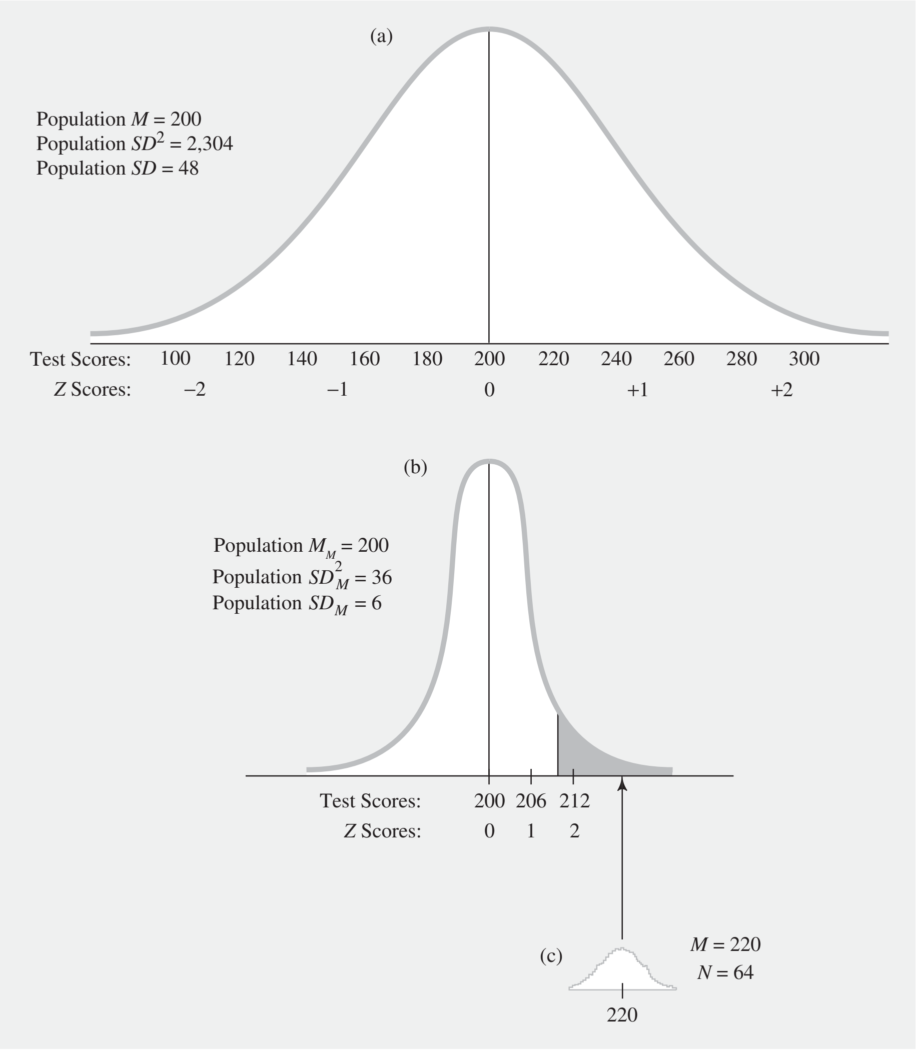


Figure 7 For the fictional study of performance on a standard school achievement test: (a) the distribution of the population of individuals, (b) the distribution of means (the comparison distribution), and (c) the sample's distribution. The shaded area in the distribution of means is the rejection region—the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

TIP FOR SUCCESS

Population 2 is the population for the comparison distribution, which is the distribution that represents the population situation if the null hypothesis is true.

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** The two populations are these:

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

The research hypothesis is that the population of fifth-graders who take the test with the special instructions will have a higher mean score on the test than the population of fifth-graders who take the test in the normal way. The null hypothesis is that the mean of Population 1's scores will not be higher than the mean of Population 2's scores. Note that these are directional hypotheses. The researchers want to know if their special instructions will increase test scores; results in the opposite direction would not be relevant to the theory the researchers are testing.

❷ **Determine the characteristics of the comparison distribution.** The result of the study will be a mean of a sample of 64 individuals (of fifth-graders in this case). Thus, the comparison distribution has to be the distribution of means of samples of 64 individuals each. This comparison distribution of means will have a mean of 200 (the same as the population mean). That is, Population $M_M = 200$. Its variance will be the population variance divided by the number of individuals in the sample. The population variance, Population SD^2, is 2,304 (the population standard deviation of 48 squared); the sample size is 64. Thus, the variance of the distribution of means, Population SD_M^2, will be 36 (that is, 2,304/64). The standard deviation of the distribution of means, Population SD_M, is 6 (the square root of 36). Finally, because there are more than 30 individuals in the sample, the shape of the distribution of means will be approximately normal. Figure 7b shows this distribution of means.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Let's assume the researchers decide to use the standard 5% significance level. As we noted in Step ❶, the researchers are making a directional prediction. Hence, the researchers will reject the null hypothesis if the result is in the top 5% of the comparison distribution. The comparison distribution (the distribution of means) is a normal curve. Thus, the top 5% can be found from the normal curve table. It starts at a Z of $+1.64$. This top 5% is shown as the shaded area in Figure 7b.

❹ **Determine your sample's score on the comparison distribution.** Let's suppose the result of the (fictional) study is that the 64 fifth-graders given the special instructions had a mean of 220. (This sample's distribution is shown in Figure 7c.) A mean of 220 is 3.33 standard deviations above the mean of the distribution of means:

$$Z = \frac{(M - \text{Population } M_M)}{\text{Population } SD_M} = \frac{(220 - 200)}{6} = \frac{20}{6} = 3.33.$$

❺ **Decide whether to reject the null hypothesis.** We set the minimum Z score to reject the null hypothesis at $+1.64$. The Z score of the sample's mean is $+3.33$. Thus, the educational researchers can reject the null hypothesis and conclude that the research hypothesis is supported. To put this another way, the result of the Z test is statistically significant at the $p < .05$ level. You can see this in Figure 7b. Note how extreme the sample's mean is on the distribution of means (the distribution that would apply if the null hypothesis were true). The final conclusion is that, among fifth-graders like those studied, the special instructions do improve test scores.

How are you doing?

1. How is hypothesis testing with a sample of more than one person different from hypothesis testing with a sample of a single person?
2. How do you find the Z score for the sample's mean on the distribution of means?
3. A researcher predicts that showing a certain film will change people's attitudes toward alcohol. The researcher then randomly selects 36 people, shows them the film, and gives them an attitude questionnaire. The mean score on the attitude test for these 36 people is 70. The score for people in general on this test is 75, with a standard deviation of 12. Using the five steps of hypothesis testing and the 5% significance level, carry out a Z test to see if showing the film changed people's attitudes toward alcohol.

Answers

1. In hypothesis testing with a sample of more than one person, the comparison distribution is a distribution of means.
2. You use the usual formula for changing a raw score to a Z score, using the mean and standard deviation of the distribution of means. The formula is $Z = (M - \text{Population } M_M)/\text{Population } SD_M$.
3. ❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** The two populations are these:

   Population 1: People shown the film.
   Population 2: People in general (who are not shown the film).

   The research hypothesis is that the mean attitude of the population shown the film is different from the mean attitude of the population of people in general; the null hypothesis is that the populations have the same mean attitude score.

   ❷ **Determine the characteristics of the comparison distribution.** Population $M_M = \text{Population } M = 75$. Population $SD^2_M = \text{Population } SD^2/N = 12^2/36 = 4$. Population $SD_M = 2$. Shape is normal.

   ❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Two-tailed cutoffs, 5% significance level, are $+1.96$ and -1.96.

   ❹ **Determine your sample's score on the comparison distribution.** $Z = (M - \text{Population } M_M)/\text{Population } SD_M = (70 - 75)/2 = -2.50$.

   ❺ **Decide whether to reject the null hypothesis.** The Z score of the sample's mean is -2.50, which is more extreme than -1.96; reject the null hypothesis. Seeing the film does change attitudes toward alcohol.

Hypothesis Tests About Means of Samples (*Z* Tests) in Research Articles

As we have noted several times, research in which there is a known population mean and standard deviation is rare in behavioral and social science research. Thus, you will not often see a Z test in a research article. We have asked you to learn about this situation mainly as another crucial building block for understanding hypothesis testing in more common research situations. Still, Z tests do show up now and then.

BOX 1 More about Polls: Sampling Errors and Errors in Thinking about Samples

Consider the statement, “From a telephone poll of 1,000 American adults taken on June 4 and 5. Sampling error ±3%.” First, you might wonder how such small numbers like 1,000 (but rarely much less) can be used to predict the opinion of the entire U.S. population. Second, after working through the material in this chapter on the standard deviation of the distribution of means, you may wonder what the term *sampling error* means when a sample is not randomly sampled but rather selected by the complicated probability method used for polls.

Regarding sample size, you know from this chapter that large sample sizes, like 1,000, greatly reduce the standard deviation of the distribution of means. That is, the curve becomes very high and narrow, gathered all around the population mean. The mean of any sample of that size is very close to being the population mean.

When a sample is only a small part of a very large population, the sample’s absolute size is the only determiner of accuracy. This absolute size determines the impact of the random errors of measurement and selection. What remains important is reducing bias or systematic error, which can be done only by careful planning.

As for the term *sampling error*, it is worked out according to past experience with the sampling procedures used. It is expressed in tables for different sample sizes (usually below 1,000, because that is where error increases dramatically).

So the number of people polled is not very important, but what matters very much are the methods of sampling and estimating error, which will not be reported in the detail necessary to judge whether the results are reliable. If the sampling and error-estimating approach is not revealed at all, be cautious. For more information about how polls are conducted, go to http://media.gallup.com/PDF/FAQ/HowArePolls.pdf (note that sampling error is referred to as “margin of error” on the Web site).

Here is an example. As part of a larger study, Wiseman (1997) gave a loneliness test to a group of college students in Israel. As a first step in examining the results, Wiseman checked that the average score on the loneliness test was not significantly different from a known population distribution based on a large U.S. study of university students that had been conducted earlier by other researchers (Russell, Peplau, & Cutrona, 1980). Wiseman reported:

> . . . [T]he mean loneliness scores of the current Israeli sample were similar to those of Russell et al.’s (1980) university sample for both males (Israeli: $M = 38.74$, $SD = 9.30$; Russell: $M = 37.06$, $SD = 10.91$; $z = 1.09$, *NS*) and females (Israeli: $M = 36.39$, $SD = 8.87$; Russell: $M = 36.06$, $SD = 10.11$; $z = .25$, *NS*). (p. 291)

In this example, the researcher gives the standard deviation for both the sample studied (the Israeli group) and the population (from the Russell et al. study). However, in the steps of figuring each Z (the sample’s score on the distribution of means), the researcher would have used the standard deviation only of the population. Notice also that the researcher took the nonsignificance of the difference as support for the sample means being “similar” to the population means. However, the researcher was very careful not to claim that these results showed there was “no difference.”

Of the topics we have covered in this chapter, the one you are most likely to see discussed in a research article is the standard deviation of the distribution of means, used to describe the amount of variation that might be expected among means of samples of a given size from this population. In this context, it is usually called the *standard error*, abbreviated *SE* (or *SEM*, for *standard error of the mean*). Standard errors are often shown in research articles as the lines that go above (and sometimes also below) the tops of the bars in a bar graph—these lines are called *error bars*. For example, Maier, Elliot, and Lichtenfeld (2008) conducted a study to examine whether the perception of the color red can adversely affect intellectual performance. The researchers explained the theory behind their hypothesis as follows: "In achievement situations, success and failure are at stake, and individuals can be motivated to approach success or avoid failure. We posit that in such situations, red is linked to the psychological danger of failure. . . . In many cultures, teachers use red to mark students' mistakes and errors, and over time this repeated pairing is presumed to create a learned association between red and failure." To test their hypothesis, Maier and colleagues asked 20 German high school students to take a 20-item numeric IQ test and randomly assigned them to one of two experimental conditions. In one condition, a large red rectangle appeared on the cover page of the test. In the other condition, the rectangle was a gray color. The results are shown in Figure 8, which includes error bars. You may be interested to know that the results supported the researchers' hypothesis: Students who viewed the red rectangle solved fewer of the IQ items than students who viewed the gray rectangle.

Be careful to read the fine print when you see lines above the tops of bars in a bar graph. Sometimes the bars are not for standard error bars, but instead are standard deviations or "confidence intervals" (see the "Advanced Topic" section below)! In Figure 8, the note under the figure states that the figure shows the mean and standard error of the number of correctly solved items. In some cases you would only know by reading the text of the article what the bars represent.

Advanced Topic: Estimation and Confidence Intervals

Hypothesis testing is our main focus. However, there is another kind of statistical question related to the distribution of means that is also important in the behavioral and social sciences: estimating the population mean based on the scores in a sample. Traditionally, this has been very important in survey research. In recent years it is also becoming important in experimental research and can even serve as an alternative approach to hypothesis testing.

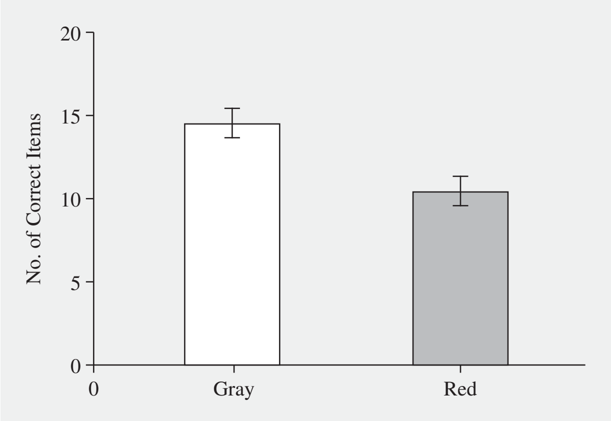


Figure 8 The effect of color on IQ test (numeric subtest) performance in Experiment 1.
NOTE: Mean and standard error of the number of correctly solved items by color on the cover of the test.
Source: Maier et al., Mediation of the Negative Effect of Red on Intellectual Performance. From *Personality and Social Psychology Bulletin,* Vol. 34, No. 11, 1530–1540 (2008). Copyright © 2008, Society for Personality and Social Psychology, Inc. Reprinted by permission of Sage Publications.

Estimating the Population Mean When It Is Unknown

When the population mean is unknown, the best estimate of the population mean is the sample mean. In the study of fifth-graders who were given the special instructions, the mean score for the sample of 64 fifth-graders was 220. Thus, 220 is the best estimate of the mean for the unknown population of fifth-graders who would ever receive the special instructions.

How accurate is the sample mean as an estimate of the population mean? A way to get at this question is to ask, "How much do means of samples from a population vary?" Fortunately we have already thought about this question when considering the distribution of means. The variation in means of samples from a population is the variation in the distribution of means. The standard deviation of this distribution of means, the standard error of the mean, is thus a measure of how much the means of samples vary from the overall population mean. (As we noted earlier, just because researchers are often interested in using a mean of a sample to estimate the population mean, this variation in the distribution of means is thought of as "error" and we give the name "standard error of the mean" to the standard deviation of a distribution of means.)

In our example, the accuracy of our estimate of 220 for the mean of the population of fifth-graders who get the special instructions would be the standard deviation of the distribution of means (also called the standard error), which we figured earlier to be 6.

Range of Possible Means Likely to Include the Population Mean

You can also estimate the *range* of possible means that are likely to include the population mean. Consider our estimate of 220 with a standard error of 6. Now follow this closely: Suppose you took a mean from our distribution of means; it is 34% likely you would get a mean between 220 (the mean of the distribution of means) and 226 (one standard error above 220). This is because the distribution of means is a normal curve. Thus, the standard error is 1 standard deviation on that curve, and 34% of a normal curve is between the mean and 1 standard deviation above the mean. From this reasoning, we could also figure that another 34% should be between 220 and 214 (1 standard error below 220). Putting this together, we have a region from 214 to 226 that we are 68% confident should include the population mean (see Figure 9a).

This is an example of a **confidence interval (CI).** We would call it the "68% confidence interval." The upper and lower ends of a confidence interval are called **confidence limits.** In this example, the confidence limits for the 68% confidence interval are 214 and 226 (see Figure 9a).

Let's review the logic. Based on our knowledge of a sample's mean, we are trying to estimate the mean of the population from which that sample came. Our best estimate of the population mean has to be our sample mean. What we don't know is how good an estimate it is. If sample means from that population could vary a lot, then we cannot be very confident that our estimate is close to the true population mean. But if the sample means are likely all to be very close to the true population mean, we can assume our estimate is pretty close. To get a sense of how accurate our estimate is, we can use our knowledge of the normal curve to estimate the *range* of possible means that are likely to include the population mean. This estimate of the range of means is called a confidence interval.

confidence interval (CI) Roughly speaking, the region of scores (that is, the scores between an upper and lower value) that is likely to include the true population mean; more precisely, the range of possible population means from which it is not highly unlikely that you could have obtained your sample mean.

confidence limit Upper or lower value of a confidence interval.

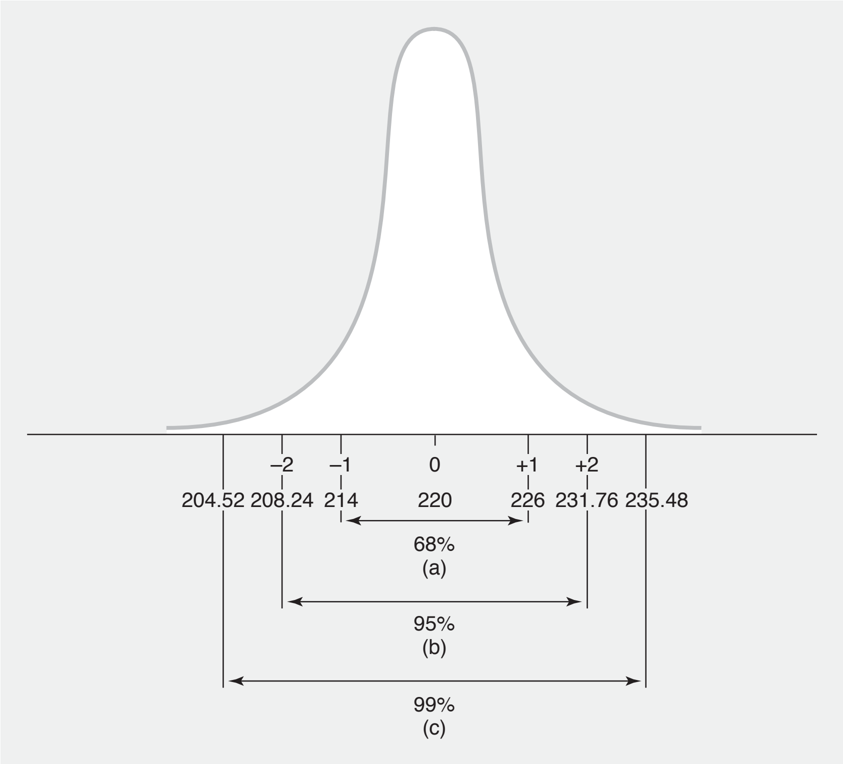


Figure 9 A distribution of means and the (a) 68%, (b) 95%, and (c) 99% confidence intervals for fifth-graders taking a test with special instructions (fictional data).

The 95% and 99% Confidence Intervals

Normally, you would want to be more than 68% confident about your estimates. Thus, when figuring confidence intervals, behavioral and social scientists use 95% or even 99% confidence intervals. These are figured based on the distribution of means for the area that includes the middle 95% or middle 99%. For the **95% confidence interval,** you want the area in a normal curve on each side between the mean and the *Z* score that includes 47.5% (47.5% plus 47.5% adds up to 95%). The normal curve table shows this to be 1.96. Thus, in terms of *Z* scores, the 95% confidence interval is from -1.96 to $+1.96$ on the distribution of means. Changing these *Z* scores to raw scores for the school achievement test example gives an interval of 208.24 to 231.76 (see Figure b). That is, for the lower confidence limit, $(-1.96)(6) + 220 = 208.24$; for the upper confidence limit, $(1.96)(6) + 220 = 231.76$. In sum, based on the sample of 64 fifth-graders who get the special instructions, you can be 95% confident that the true population mean for such fifth-graders is between 208.24 and 231.76 (see Figure 9b).

For a **99% confidence interval,** you use the *Z* scores for the middle 99% of the normal curve (the part that includes 49.5% above and below the mean). This comes out to $+2.58$ and -2.58. Changing these *Z* scores to raw scores, the 99% confidence interval is from 204.52 to 235.48 (see Figure 9c).

Notice in Figure 9 that the greater the confidence is, the broader is the confidence interval. In our example, you could be 68% confident that the true population mean is between 214 and 226; but you could be 95% confident that it is between 208.24 and 231.76 and 99% confident it is between 204.52 and 235.48. This is a general

95% confidence interval Confidence interval in which, roughly speaking, there is a 95% chance that the population mean falls within this interval.

99% confidence interval Confidence interval in which, roughly speaking, there is a 99% chance that the population mean falls within this interval.

principle. It makes sense that you need a wider range of possibility to be more sure you are right.

Steps for Figuring the 95% and 99% Confidence Intervals

Here are three steps for figuring a confidence interval. These steps assume that the distribution of means is approximately a normal distribution.

❶ **Estimate the population mean and figure the standard deviation of the distribution of means.** The best estimate of the population mean is the sample mean. Next, find the variance of the distribution of means in the usual way: Population SD_M^2 = Population SD^2/N. Take the square root of the variance of the distribution of means to find the standard deviation of the distribution of means: Population $SD_M = \sqrt{\text{Population } SD_M^2}$.

❷ **Find the *Z* scores that go with the confidence interval you want.** For the 95% confidence interval, the *Z* scores are $+1.96$ and -1.96. For the 99% confidence interval, the *Z* scores are $+2.58$ and -2.58.

❸ **To find the confidence interval, change these *Z* scores to raw scores.** To find the lower limit, multiply -1.96 (for the 95% confidence interval) or -2.58 (for the 99% confidence interval) by the standard deviation of the distribution of means (Population SD_M) and add this to the population mean. To find the upper limit, multiply $+1.96$ (for the 95% confidence interval) or $+2.58$ (for the 99% confidence interval) by the standard deviation of the distribution of means (Population SD_M) and add this to the population mean.

Example. Let's return again to the study of 64 fifth-graders who were given special instructions on a test. Recall that in that example, the test scores of fifth-graders in general (who were not given special instructions) were normally distributed, with a mean of 200 and a standard deviation of 48. The mean of the sample of 64 fifth-graders who were given the special instructions was 220. We figured the 99% confidence limit earlier, but now we are going to do the figuring using the three steps:

❶ **Estimate the population mean and figure the standard deviation of the distribution of means.** The best estimate of the population mean (of fifth-graders who take the test with the special instructions) is the sample mean of 220.

Population SD_M^2 = Population $SD^2/N = 48^2/64 = 36$. Thus, Population $SD_M = \sqrt{\text{Population } SD_M^2} = \sqrt{36} = 6$.

❷ **Find the *Z* scores that go with the confidence interval you want.** For the 99% confidence interval, the *Z* scores are $+2.58$ and -2.58.

❸ **To find the confidence interval, change these *Z* scores to raw scores.** Lower confidence limit $= (-2.58)(6) + 220 = -15.48 + 220 = 204.52$; upper confidence limit $= (+2.58)(6) + 220 = 15.48 + 220 = 235.48$. The 99% confidence interval is from 204.52 to 235.48. Thus, based on the sample of 64 fifth-graders, you can be 99% confident that an interval from 204.52 to 235.48 includes the true population mean of fifth-graders who take the test with the special instructions.

Confidence Intervals and Hypothesis Testing

You can use confidence intervals to do hypothesis testing! If the confidence interval does not include the mean of the null hypothesis distribution, then the result is significant. For example, in the fifth-grader study, the 99% confidence interval for those who got the special instructions was from 204.52 to 235.48. However, the population that did not get the special instructions had a mean of 200. This population mean is outside the range of the confidence interval. Thus, if you are 99% confident that the true range is 204.52 to 235.48 and the population mean for those who didn't get the special instructions is not in this range, you are 99% confident that that population is not the same as the one from which your sample came.

Another way to understand this is in terms of the idea that the confidence limits are the points at which a more extreme true population would not include your sample mean 99% of the time (or 95% of the time for the 95% confidence interval). The population mean for those not getting the special instructions was 200. If this were the true mean also for the group that got the special instructions, 99% of the time it would not produce a sample mean as high as the one we got.

Most behavioral and social science research uses ordinary hypothesis testing. However, sometimes you will see the confidence-interval method used instead. And sometimes you will see both.

How are you doing?

1. (a) What is the best estimate of a population mean? (b) Why?
2. (a) What number is used to indicate the accuracy of an estimate of the population mean? (b) Why?
3. What is a 95% confidence interval?
4. A researcher predicts that showing a certain film will change people's attitudes toward alcohol. The researcher then randomly selects 36 people, shows them the film, and gives them an attitude questionnaire. The mean score on the attitude test for these 36 people is 70. The score on this test for people in the general population (who do not see the film) is 75, with a standard deviation of 12. (a) Find the best estimate of the mean of people in general who see the film and (b) its 95% confidence interval. (c) Compare this result to the conclusion you drew when you used this example in the "How are you doing?" section for hypothesis testing with a distribution of means.

Answers

1. (a) The best estimate of a population mean is the sample mean. (b) It is more likely to have come from a population with the same mean than from any other population.
2. (a) The standard deviation of the distribution of means (or standard error) is used to indicate the accuracy of an estimate of the population mean. (b) The standard deviation of the distribution of means (standard error) is roughly the average amount that means vary from the mean of the distribution of means.
3. A 95% confidence interval is the range of values that you are 95% confident includes the population mean, estimated based on the scores in a sample.
4. (a) The best estimate is the sample mean: 70.

   (b) Population SD^2_M = Population $SD^2/N = 12^2/36 = 4$. Thus, the standard deviation of the distribution of means, Population $SD_M = \sqrt{\text{Population } SD^2_M} = \sqrt{4} = 2$. The lower confidence limit $= (-1.96)(2) + 70 = -3.92 + 70 = 66.08$; the upper confidence limit $= (1.96)(2) + 70 = 73.92$. The 95% confidence interval is from 66.08 to 73.92.

   (c) The 95% confidence interval does not include the mean of the general population (which was 75). Thus, you can reject the null hypothesis that the two populations are the same. This is the same conclusion as when using this example for hypothesis testing.

Advanced Topic: Confidence Intervals in Research Articles

Confidence intervals (usually abbreviated as CIs) are becoming increasingly common in research articles in some fields, such as medical research. For example, Morey and colleagues (2009) conducted a study to test a diet and exercise intervention for overweight individuals age 65 years and older who had been diagnosed with cancer. Participants were randomly assigned to be in the intervention group (a 12-month program of mailed materials and telephone counseling) or a control group (no intervention until the study was done). Morey et al. reported the following results: “Participants in the intervention group reported a mean weight loss of 2.06 kg (95% CI, 1.69–2.43 kg), which was more than twice that reported by the control group (0.92 kg; 95% CI, 0.51–1.33 kg)” (p. 1888). This means that we can be 95% confident that the true amount of weight lost on average by participants in the intervention group was between 1.69 kg and 2.43 kg, and for control group participants it was between 0.51 kg and 1.33 kg. As another example, a researcher might explain that the average amount of overtime hours worked in a particular industry is 3.7 with a 95% confidence interval of 2.5 to 4.9. This would tell you that the true amount of overtime hours is probably somewhere between 2.5 and 4.9.

Learning Aids

1. When studying a sample of more than one individual, the comparison distribution in the hypothesis-testing process is a distribution of means. It can be thought of as what would result from (a) taking a very large number of samples, each of the same number of scores taken randomly from the population of individuals, and then (b) making a distribution of the means of these samples.
2. The distribution of means has the same mean as the corresponding population of individuals. However, it has a smaller variance because the means of samples are less likely to be extreme than individual scores. (In any one sample, extreme scores are likely to be balanced by middle scores or extreme scores in the other direction.) Specifically, the variance of the distribution of means is the variance of the population of individuals divided by the number of individuals in each sample. Its standard deviation is the square root of its variance. The shape of the distribution of means approximates a normal curve if either (a) the samples are each of 30 or more individuals or (b) the population of individuals follows a normal curve.
3. Hypothesis tests with a single sample of more than one individual and a known population are called *Z* tests and are done in the same as the hypothesis tests where the studies were of a single individual compared to a population of individual scores. The main exception is that, in a hypothesis test with a single sample of more than one individual and a known population, the comparison distribution is a distribution of means.
4. The kind of hypothesis test described in this chapter (the *Z* test) is rarely used in research practice; you have learned it as a stepping stone. The standard deviation of the distribution of means (the standard error, *SE*), is often used to describe the expected variability of means, particularly in bar graphs in which the standard error may be shown as the length of a line above (and sometimes below) the top of each bar.

5. ADVANCED TOPIC: The sample mean is the best estimate for the population mean when the population mean is unknown. The accuracy of the estimate is the standard deviation of the distribution of means (also known as the standard error, *SE*), which tells you roughly the amount by which means vary. Based on the distribution of means, you can figure the range of possible means that are likely to include the population mean. If we assume the distribution of means follows a normal curve, the 95% confidence interval includes the range from 1.96 standard deviations below the sample mean (the lower confidence limit) to 1.96 standard deviations above the sample mean (the upper confidence limit). The 99% confidence interval includes the range from 2.58 standard deviations below the sample mean (the lower confidence limit) to 2.58 standard deviations above the sample mean (the upper confidence limit). Confidence intervals are sometimes reported in research articles, usually with the abbreviation CI.

Key Terms

distribution of means
mean of a distribution of means (Population M_M)
variance of a distribution of means (Population SD^2_M)
standard deviation of a distribution of means (Population SD_M)
standard error (*SE*)
shape of the distribution of means
Z test
confidence interval (CI)
confidence limits
95% confidence interval
99% confidence interval

Figure the Standard Deviation of the Distribution of Means

Find the standard deviation of the distribution of means for a population with Population $SD = 13$ and a sample size of 20.

Answer

Using Rules 2a and 2b for the characteristics of a distribution of means: **The variance of a distribution of means is the variance of the population of individuals divided by the number of individuals in each sample; the standard deviation of a distribution of means is the square root of the variance of the distribution of means.** The variance of the population is 169 (that is, 13 squared is 169); dividing this by 20 gives a variance of the distribution of means of 8.45. The square root of this, 2.91, is the standard deviation of the distribution of means.

Using the formulas, Population SD^2_M = Population $SD^2/N = 13^2/20 = 8.45$.

$$\text{Population } SD_M = \sqrt{\text{Population } SD^2_M} = \sqrt{8.45} = 2.91.$$

Hypothesis Testing with a Sample of More than One: The *Z* Test

A sample of 75 given an experimental treatment had a mean of 16 on a particular measure. The general population of individuals has a mean of 15 on this measure and a standard deviation of 5. Carry out a *Z* test using the full five steps of hypothesis testing with a two-tailed test at the .05 significance level and make a drawing of the distributions involved.

Answer

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations. The two populations are:**

Population 1: Those given the experimental treatment.
Population 2: People in the general population (who are not given the experimental treatment).

The research hypothesis is that the population given the experimental treatment will have a different mean on the particular measure from the mean of people in the general population (those not given the experimental treatment). The null hypothesis is that the populations have the same mean score on this measure.

❷ **Determine the characteristics of the comparison distribution.** Population M_M = Population $M = 15$. Population SD^2_M = Population $SD^2/N = 5^2/75 = .33$; Population $SD_M = \sqrt{\text{Population } SD^2_M} = \sqrt{.33} = .57$; shape is normal (sample size is greater than 30).

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Two-tailed cutoffs, .05 significance level, are $+1.96$ and -1.96.

❹ **Determine your sample's score on the comparison distribution.** Using the formula $Z = (M - \text{Population } M_M)/\text{Population } SD_M$, $Z = (16 - 15)/.57 = 1/.57 = 1.75$.

❺ **Decide whether to reject the null hypothesis.** The sample's Z score of 1.75 is *not* more extreme than the cutoffs of $+1.96$ and -1.96; do not reject the null hypothesis. The results are inconclusive. The distributions involved are shown in Figure 10.

Outline for Writing Essays for Hypothesis-Testing Problems Involving a Single Sample of More Than One and a Known Population (The *Z* Test)

1. Describe the core logic of hypothesis testing in this situation. Be sure to explain the meaning of the research hypothesis and the null hypothesis in this situation where we focus on the mean of a sample and compare it to a known population mean. Explain the concept of support being provided for the research hypothesis when the study results allow the null hypothesis to be rejected.
2. Explain the concept of the comparison distribution. Be sure to mention that with a sample of more than one, the comparison distribution is a distribution of means because the information from the study is a mean of a sample. Mention that the distribution of means has the same mean as the population mean because there is no reason for random samples in the long run to have a different mean; the distribution of means has a smaller variance (the variance of the population divided by the number in each sample) because it is harder to get extreme means than extreme individual cases by chance, and the larger the samples are, the rarer it is to get extreme means.
3. Describe the logic and process for determining (using the normal curve) the cutoff sample score(s) on the comparison distribution at which the null hypothesis should be rejected.

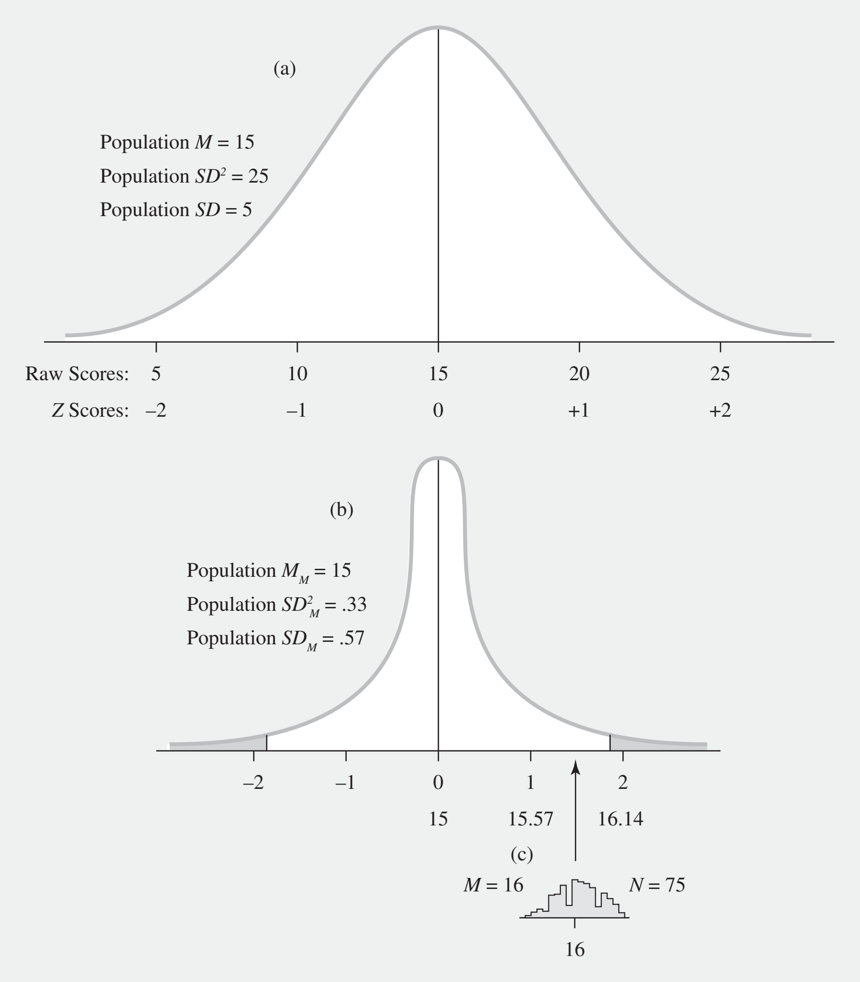


Figure 10 Answer to the hypothesis-testing problem in Example Worked-Out Problems: (a) the distribution of the population of individuals, (b) the distribution of means (the comparison distribution), and (c) the sample's distribution.

4. Describe why and how you figure the Z score of the sample mean on the comparison distribution.
5. Explain how and why the scores from Steps ❸ and ❹ of the hypothesis-testing process are compared. Explain the meaning of the result of this comparison with regard to the specific research and null hypotheses being tested.

Advanced Topic: Finding Confidence Intervals

Find the 99% confidence interval for the sample mean in the study just described.

Answer

❶ **Estimate the population mean and figure the standard deviation of the distribution of means.** The best estimate of the population mean in the preceding problem is the sample mean of 16. The standard deviation of the distribution of means, Population SD_M, in the problem above was .57.

❷ **Find the *Z* scores that go with the confidence interval you want.** For the 99% confidence interval, the *Z* scores are $+2.58$ and -2.58.

❸ **To find the confidence interval, change these *Z* scores to raw scores.** Lower confidence limit $= (-2.58)(.57) + 16 = -1.47 + 16 = 14.53$; upper confidence limit $= (2.58)(.57) + 16 = 1.47 + 16 = 17.47$. The 99% confidence interval is from 14.53 to 17.47.

Advanced Topic: Outline for Writing Essays for Finding Confidence Intervals

1. Explain that a confidence interval is an estimate (based on your sample's mean and the standard deviation of the distribution of means) of the range of values that is likely to include the true population mean for the group studied (Population 1). Be sure to mention that the 95% (or 99%) confidence interval is the range of values you are 95% (or 99%) confident include the true population mean.
2. Explain that the first step in figuring a confidence interval is to estimate the population mean (for which the best estimate is the sample mean) and figure the standard deviation of the distribution of means.
3. Mention that you next find the *Z* scores that go with the confidence interval that you want.
4. Describe how to change the *Z* scores to raw scores to find the confidence interval.

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.

All data are fictional unless an actual citation is given.

Set I (for answers, see the end of this chapter)

1. Why is the standard deviation of the distribution of means generally smaller than the standard deviation of the distribution of the population of individuals?
2. For a population that has a standard deviation of 10, figure the standard deviation of the distribution of means for samples of size (a) 2, (b) 3, (c) 4, and (d) 9.
3. For a population that has a standard deviation of 20, figure the standard deviation of the distribution of means for samples of size (a) 2, (b) 3, (c) 4, and (d) 9.
4. ADVANCED TOPIC: Figure the 95% confidence interval (that is, the lower and upper confidence limits) for each part of problem 2. Assume that in each case the researcher's sample has a mean of 100 and that the population of individuals is known to follow a normal curve.
5. ADVANCED TOPIC: Figure the 99% confidence interval (that is, the lower and upper confidence limits) for each part of problem 3. Assume that in each case the researcher's sample has a mean of 10 and that the population of individuals is known to follow a normal curve.
6. For each of the following samples that were given an experimental treatment, test whether these samples represent populations that are different from the general population: (a) a sample of 10 with a mean of 44, (b) a sample of 1 with a mean of 48. The general population of individuals has a mean of 40, a standard deviation of 6, and follows a normal curve. For each sample, carry out a *Z* test using the five steps of hypothesis testing with a two-tailed test at the .05 significance level, and

make a drawing of the distributions involved (c) ADVANCED TOPIC: Figure the 95% confidence interval for parts (a) and (b).

7. For each of the following samples that were given an experimental treatment, test whether they represent populations that score significantly higher than the general population: (a) a sample of 100 with a mean of 82 and (b) a sample of 10 with a mean of 84. The general population of individuals has a mean of 81, a standard deviation of 8, and follows a normal curve. For each sample, carry out a *Z* test using the five steps of hypothesis testing with a one-tailed test at the .01 significance level, and make a drawing of the distributions involved. (c) ADVANCED TOPIC: Figure the 99% confidence interval for parts (a) and (b).
8. Twenty-five women between the ages of 70 and 80 were randomly selected from the general population of women their age to take part in a special program to decrease reaction time (speed). After the course, the women had an average reaction time of 1.5 seconds. Assume that the mean reaction time for the general population of women of this age group is 1.8, with a standard deviation of .5 seconds. (Also assume that the population is approximately normal.) What should you conclude about the effectiveness of the course? (a) Carry out a *Z* test using the five steps of hypothesis testing (use the .01 significance level). (b) Make a drawing of the distributions involved. (c) Explain your answer to someone who is familiar with the general logic of hypothesis testing, the normal curve, *Z* scores, and probability, but not with the idea of a distribution of means. (d) ADVANCED TOPIC: Figure the 99% confidence interval and explain your answer to someone who is familiar with the general logic of hypothesis testing, the normal curve, *Z* scores, probability, and the idea of a distribution of means, but has not heard of confidence intervals.
9. A large number of people were shown a particular film of an automobile collision between a moving car and a stopped car. Each person then filled out a questionnaire about how likely it was that the driver of the moving car was at fault, on a scale from 0 = *not at fault* to 10 = *completely at fault*. The distribution of ratings under ordinary conditions follows a normal curve with Population $M = 5.5$ and Population $SD = .8$. Sixteen randomly selected individuals are tested in a condition in which the wording of the question is changed so the question asks, “How likely is it that the driver of the car who crashed into the other was at fault?” (The difference is that in this changed condition, instead of describing the event in a neutral way, the question uses the phrase “crashed into.”) Using the changed instructions, these 16 research participants gave a mean at-fault rating of 5.9. Did the changed instructions significantly increase the rating of being at fault? (a) Carry out a *Z* test using the five steps of hypothesis testing (use the .05 significance level). (b) Make a drawing of the distributions involved. (c) Explain your answer to someone who has never taken statistics. (d) ADVANCED TOPIC: Figure the 95% confidence interval.
10. Lee, Byatt, and Rhodes (2000) tested a theory of the role of distinctiveness in face perception. In their study, participants indicated whether they recognized each of 48 faces of male celebrities when they were shown rapidly on a computer screen. A third of the faces were shown in caricature form, in which facial features were electronically modified so that distinctive features were exaggerated; a third were shown in veridical form, in which the faces were not modified at all; and a third were shown in anticaricature form, in which facial features were modified to be slightly more like the average of the celebrities’ faces. The average percentage correct across the participants is shown in Figure 11. Explain the meaning of the error bars in this figure to a person who understands mean, standard deviation, and variance, but nothing else about statistics.

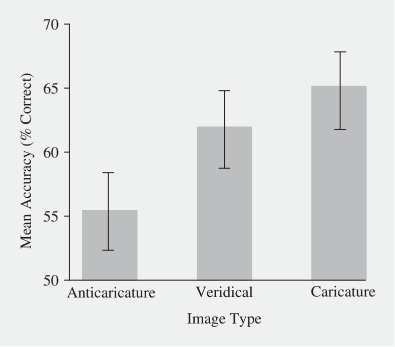


Figure 11 Identification accuracy as a function of image type. Standard error bars are shown. *Source:* Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science, 11,* 379–385. Copyright © 2002, 2000 American Psychological Society. Reprinted by permission of Blackwell Publishing.

11. ADVANCED TOPIC: Anderson, Carey, and Taveras (2000) studied the rate of HIV testing among adults in the United States and reported one of their findings as follows: "Responses from the NHIS [National Health Interview Survey] indicate that by 1995, 39.7% of adults (95% CI = 38.8%, 40.5%) had been tested at least once. . . ." (p. 1090). Explain what "(95% CI = 38.8%, 40.5%)" means to a person who understands hypothesis testing with the mean of a sample of more than one but has never heard of confidence intervals.

Set II

12. Under what conditions is it reasonable to assume that a distribution of means will follow a normal curve?
13. Indicate the mean and the standard deviation of the distribution of means for each of the situations shown to the right.
14. Figure the standard deviation of the distribution of means for a population with a standard deviation of 20 and sample sizes of (a) 10, (b) 11, (c) 100, and (d) 101.
15. For each of the studies shown below, the samples were given an experimental treatment and the researchers compared their results to the general population. (Assume all populations are normally distributed.) For each, carry out a *Z* test using the five steps of hypothesis testing for a two-tailed test, and make a drawing of the distributions involved. ADVANCED TOPIC: Figure the 95% confidence interval for each study.

	Population		**Sample Size**
Situation	*M*	SD^2	*N*
(a)	100	40	10
(b)	100	30	10
(c)	100	20	10
(d)	100	10	10
(e)	50	10	10
(f)	100	40	20
(g)	100	10	20

	Population		**Sample Size**	**Sample Mean**	**Significance Level**
Study	*M*	*SD*	*N*		
(a)	36	8	16	38	.05
(b)	36	6	16	38	.05
(c)	36	4	16	38	.05
(d)	36	4	16	38	.01
(e)	34	4	16	38	.01

16. For each of the following studies, the samples were given an experimental treatment and the researchers compared their results to the general population. (Assume all populations are normally distributed.) For each, carry out a Z test using the five steps of hypothesis testing for a two-tailed test at the .01 level, and make a drawing of the distributions involved. ADVANCED TOPIC: Figure the 99% confidence interval for each study.

	Population		**Sample Size**	**Sample Mean**
Study	*M*	*SD*	*N*	
(a)	10	2	50	12
(b)	10	2	100	12
(c)	12	4	50	12
(d)	14	4	100	12

17. ADVANCED TOPIC: Figure the 95% confidence interval (that is, the lower and upper confidence limits) for each part of problem 13. Assume that in each case the researcher's sample has a mean of 80 and the population of individuals is known to follow a normal curve.
18. ADVANCED TOPIC: Figure the 99% confidence interval (that is, the lower and upper confidence limits) for each part of problem 14. Assume that in each case the researcher's sample has a mean of 50 and that the population of individuals is known to follow a normal curve.
19. A researcher is interested in whether people are able to identify emotions correctly in other people when they are extremely tired. It is known that, using a particular method of measurement, the accuracy ratings of people in the general population (who are not extremely tired) are normally distributed with a mean of 82 and a variance of 20. In the present study, however, the researcher arranges to test 50 people who had no sleep the previous night. The mean accuracy for these 50 individuals was 78. Using a two-tailed test and the .05 significance level, what should the researcher conclude? (a) Carry out a Z test using the five steps of hypothesis testing. (b) Make a drawing of the distributions involved. (c) Explain your answer to someone who knows about hypothesis testing with a sample of a single individual but who knows nothing about hypothesis testing with a sample of more than one individual. (d) ADVANCED TOPIC: Figure the 95% confidence interval and explain your answer to someone who is familiar with the general logic of hypothesis testing, the normal curve, Z scores, probability, and the idea of a distribution of means, but who has not heard of confidence intervals.
20. A researcher is interested in the conditions that affect the number of dreams per month that people report in which they are alone. We will assume that based on extensive previous research, it is known that in the general population the number of such dreams per month follows a normal curve, with Population $M = 5$ and Population $SD = 4$. The researcher wants to test the prediction that the number of such dreams will be greater among people who have recently experienced a traumatic event. Thus, the researcher studies 36 individuals who have recently experienced a traumatic event, having them keep a record of their dreams for a month. Their mean number of alone dreams is 8. Should you conclude that people who have recently had a traumatic experience have a significantly different number of dreams in which they are alone? (a) Carry out a Z test using the five

steps of hypothesis testing (use the .05 significance level). (b) Make a drawing of the distributions involved. (c) Explain your answer to a person who has never had a course in statistics. (d) ADVANCED TOPIC: Figure the 95% confidence interval.

21. A government-sponsored telephone counseling service for adolescents tested whether the length of calls would be affected by a special telephone system that had a better sound quality. Over the past several years, the lengths of telephone calls (in minutes) were normally distributed with Population $M = 18$ and Population $SD = 8$. They arranged to have the special phone system loaned to them for one day. On that day, the mean length of the 46 calls they received was 21 minutes. Test whether the length of calls has changed using the .05 significance level. (a) Carry out a Z test using the five steps of hypothesis testing. (b) Make a drawing of the distributions involved. (c) Explain your answer to someone who knows about hypothesis testing with a sample of a single individual but who knows nothing about hypothesis testing with samples of more than one individual. (d) ADVANCED TOPIC: Figure the 95% confidence interval.
22. Perna, Antoni, Baum, Gordon, and Schneiderman (2003) tested whether a stress management intervention could reduce injury and illness among college athletes. In their study, 34 college athletes were randomly assigned to be in one of two groups: (1) a stress management intervention group: This group received a cognitive-behavioral stress management (CBSM) intervention during preseason training; (2) a control group: This group did not receive the intervention. At the end of the season, for each athlete, the researchers recorded the number of health center visits (including visits to the athletic training center) and the number of days of illness or injury during the season. The results are shown in Figure 12. In the figure caption, the researchers note that the figure shows the "Mean $(+SE)$." This tells you that the line above the top of each bar represents the standard error. Explain what this means, using one of the error bars as an example, to a person who understands mean and standard deviation, but knows nothing else about statistics.
23. ADVANCED TOPIC: Stankiewicz, Legge, Mansfield, and Schlicht (2006) examined how limitations in human perception and memory (and other factors) affect people's ability to find their way in indoor spaces. In one of their experiments, eight

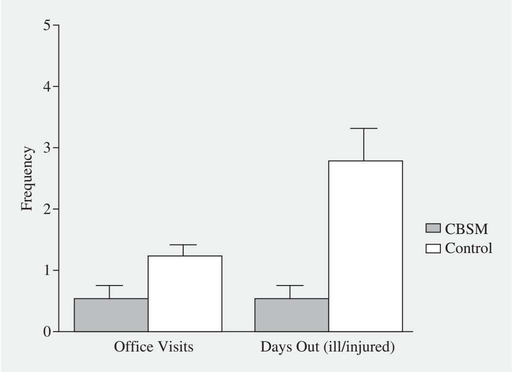


Figure 12 Mean (+*SE*) number of accumulated days injured or ill and athletic training room and health center office visits for cognitive behavioral stress management (CBSM) ($n = 18$) and control group ($n = 16$) from study entry to season's end. *Source:* Perna, F. M., Antoni, M. H., Baum, A., Gordon, P., & Schneiderman, N. (2003). Cognitive behavioral stress management effects on injury and illness among competitive athletes: A randomized clinical trial. *Annals of Behavioral Medicine, 25,* 66–73. Copyright © 2003 by Lawrence Erlbaum Associates, Inc. Reprinted by permission.

students used a computer keyboard to move through a virtual indoor space of corridors and hallways shown on a computer monitor. The researchers calculated how efficiently students moved through the space, with efficiency ranging from 0 (extremely inefficient) to 1 (extremely efficient). The researchers compared the efficiency of moving through the space when students had a limited view of the space with when they had a clear (or unlimited) view of the space. Their results, shown in Figure 13, include error bars. The figure caption notes that "Error bars represent 1 standard error of the mean." Explain what this means, using one of the error bars as an example, to a person who understands mean and standard deviation, but knows nothing else about statistics.

24. Cut up 90 small slips of paper, and write each number from 1 to 9 on 10 slips each. Put the slips in a large bowl and mix them up. (a) Take out a slip, write down the number on it, and put it back. Do this 20 times. Make a histogram, and figure the mean and the variance of the result. You should get an approximately rectangular distribution. (b) Take two slips out, figure out their mean, write it down, and put the slips back.[3] Repeat this process 20 times. Make a histogram; then figure the mean and the variance of this distribution of means. The variance should be about half of the variance of this distribution of means. (c) Repeat the process again, this time taking three slips at a time. Again, make a histogram and figure the mean and the variance of the distribution of means. The distribution of means of three slips each should have a variance of about a third of the distribution of samples of one slip each. Also note that as the sample size increases, your distributions get closer to normal. (Had you begun with a normally distributed distribution of slips, your distributions of means would have been fairly close to normal regardless of the number of slips in each sample.)

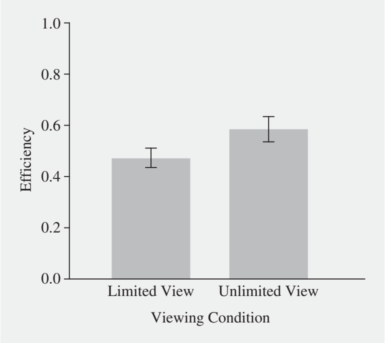


Figure 13 The mean navigation efficiency when navigating in the unlimited and limited viewing condition in Experiment 2. In the limited-view condition, visual information was available as far as the next intersection (further details were obscured by "fog"). In the unlimited-view condition, visual information was available to the end of the corridor. Error bars represent 1 standard error of the mean. *Source:* Stankiewicz, B. J., Legge, G. E., Mansfield, J. S., & Schlicht, E. J. (2006). Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 688–704. Copyright © 2006 by the American Psychological Association. Reproduced with permission. The use of APA information does not imply endorsement by APA.

[3] Technically, when taking the samples of two slips, this should be done by taking one, writing it down, putting it back, then taking the next, writing it down, and putting it back. You would consider these two scores as one sample for which you figure a mean. The same applies for samples of three slips. This is called sampling with replacement. However, with 90 slips in the bowl, taking two or three slips at a time and putting them back will be a close enough approximation for this exercise and will save you some time.

Answers to Set I Practice Problems

1. There is less variation among means of samples of more than one score than there are among individual scores. This is because the likelihood of two extreme scores in the same direction randomly ending up in the same sample is less than the probability of each of those extreme scores being chosen individually.
2. (a) Population $SD^2 = 10^2 = 100$; Population $SD^2_M =$ Population $SD^2/N = 100/2 = 50$; Population $SD_M = \sqrt{\text{Population } SD^2_M} = \sqrt{50} = 7.07$; (b) 5.77; (c) 5; (d) 3.33.
3. (a) Population $SD^2 = 20^2 = 400$; Population $SD^2_M =$ Population $SD^2/N = 400/2 = 200$; Population $SD_M = \sqrt{\text{Population } SD^2_M} = \sqrt{200} = 14.14$; (b) 11.55; (c) 10; (d) 6.67.
4. (a) The best estimate of the population mean is the sample mean of 100. From question 2a, the standard deviation of the distribution of means (Population SD_M) is 7.07. For the 95% confidence limits, the Z scores you want are -1.96 and $+1.96$. Lower limit $= (-1.96)(7.07) + 100 = 86.14$. Upper limit $= (1.96)(7.07) + 100 = 113.86$; (b) 88.69, 111.31; (c) 90.2, 109.8; (d) 93.47, 106.53.
5. (a) The best estimate of the population mean is the sample mean of 10. From question 3a, the standard deviation of the distribution of means (Population SD_M) is 14.14. For the 99% confidence limits, the Z scores you want are -2.57 and $+2.57$. Lower limit $= (-2.57)(14.14) + 10 = -26.34$; upper limit $= (2.57)(14.14) + 10 = 46.34$; (b) -19.68, 39.68; (c) -15.70, 35.70; (d) -7.14, 27.14.
6. (a)

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.** There are two populations of interest:

Population 1: People given the experimental treatment.
Population 2: People in general (who do not get the experimental treatment).

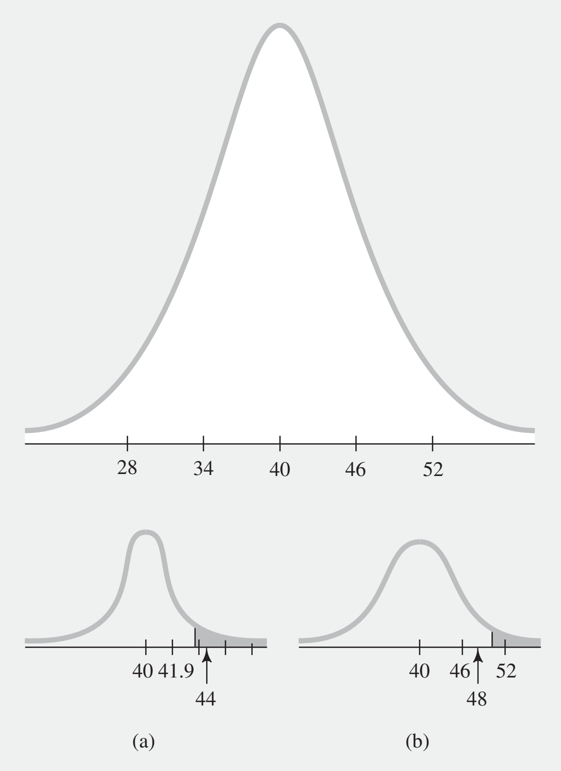
The research hypothesis is that the population given the experimental treatment (Population 1) has a different mean score than people in general (Population 2). The null hypothesis is that Population 1's mean is the same as Population 2's.

❷ **Determine the characteristics of the comparison distribution.** Comparison distribution is a distribution of means of samples of 10 taken from the distribution of Population 2. Population $M_M =$ Population $M = 40$; Population $SD^2_M =$ Population $SD^2/N = 6^2/10 = 3.6$; Population $SD_M = \sqrt{\text{Population } SD^2_M} = \sqrt{3.6} = 1.90$. Because the population is normal, the distribution of means is normal.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** For a two-tailed test at the .05 level, the cutoffs are -1.96 and 1.96.

❹ **Determine your sample's score on the comparison distribution.** $Z = (44 - 40)/1.90 = 2.11$.

❺ **Decide whether to reject the null hypothesis.** 2.11 is more extreme than 1.96. Thus, you can reject the null hypothesis. The research hypothesis is supported; those who receive the experimental treatment score differently from the general population. The distributions involved are shown below.



(b) Hypothesis-testing steps similar to (a) above. Population $SD_M = 6$; $Z = (48 - 40)/6 = 1.33$; do not reject null hypothesis; study is inconclusive as to whether those who receive the experimental treatment are different from those in the general population.

(c) For part (a), 95% confidence interval: Lower limit $= (-1.96)(1.9) + 44 = 40.28$; Upper limit $= (1.96)(1.9) + 44 = 47.72$ for part (b), 95% confidence interval: 36.24 to 59.76.

7. Hypothesis-testing steps and drawing similar to 6 above. (a) Population $SD_M = .8$; $Z = (82 - 81)/.8 = 1.25$; do not reject null hypothesis. (b) Population $SD_M = 2.53$; $Z = (84 - 81)/2.53 = 1.19$; do not reject null hypothesis. (c) For part (a), 99% confidence interval: 79.94 to 84.06; for part (b), 99% confidence interval: 77.50 to 90.50.
8. (a) and (b) Hypothesis-testing steps and drawing similar to 6 above. Population $SD_M = .1$; $Z = (1.5 - 1.8)/.1 = -3$; reject the null hypothesis.

(c) This is a standard hypothesis-testing problem, with one exception. You can't compare directly the reaction times for the group of 25 women tested to a distribution of reaction times for individual women. The probability of a group of scores having an extreme mean just by chance is much less than the probability of any one individual having an extreme score just by chance. (When taking a group of scores at random, any extreme individual scores are likely to be balanced out by less extreme or oppositely extreme scores.) Thus, you need to compare the mean of the group of 25 reaction times to a distribution of what would happen if you were to take many random groups of 25 reaction time scores and find the mean of each group of 25 scores.

Such a distribution of many means of samples has the same mean as the original distribution of individual scores (there is no reason for it to be otherwise). However, it is a narrower curve. This is because the chances of extremes are less. In fact, its variance will be exactly the variance of the original distribution of individuals divided by the number of scores in each sample. In this example, this makes a distribution of means with a mean of 1.8 and a standard deviation of .1 (that is, the square root of the result of $.5^2$ divided by 25). This will be a normal distribution because a distribution of many means from a normally distributed population is also normal.

The cutoff for significance, using the .01 level and a one-tailed test, is -2.33. The mean reaction time of the group of 25 women who received the special program, 1.5, was 3 standard deviations below the mean of the distribution of means, making it clearly more extreme than the cutoff. Thus, you can reject the null hypothesis and conclude that the results support the hypothesis that elderly women who take part in the special program have lower reaction times.

(d) 99% confidence interval: 1.24 to 1.76. The confidence interval is an estimate based on your sample's mean and the standard deviation of the distribution of means. What it estimates is the range of values that is likely to include the true population mean for the group studied. (The group studied is Population 1. In this example, the group studied is women who receive the special reaction-time program.) A 99% confidence interval is the range of values you are 99% confident include the true population mean. The lower end of this interval (in this example, 1.24) is the mean of the lowest distribution of means that would have a 99% chance of including this sample mean; its upper end (in this example, 1.76) is the mean of the highest distribution of means that would have a 99% chance of including this sample mean.

To figure the confidence interval, you first consider that the best single estimate of the mean of Population 1 is the sample's mean (in this case, 1.5). You then assume that the standard deviation of the distribution of means for this population is the same as for the known population (which we figured earlier to be .1). Based on this information, if the true population mean was 1.5, 99% of the time, sample means would fall between a Z score of -2.57 (the point on the normal curve that includes 49.5% of the scores below the mean) and $+2.57$. In our example, these Z scores correspond to raw scores of 1.24 and 1.76.

It turns out that the values figured in this way are the limits (the upper and lower end) of the confidence interval. Why? Suppose the true population mean was 1.24. In this case, there would be a .5% chance of getting a mean as large as or larger than 1.5. (That is, with a mean of 1.24 and a standard deviation of .1, 1.5 is exactly 2.57 standard deviations above the mean. This is the point that corresponds to the cutoff for the top .5% of this curve.) Similarly, if the true population mean was 1.76, there would only be a .5% chance of getting a mean lower than 1.5.

9. (a) and (b) Hypothesis-testing steps and drawing similar to 6 above. Population $SD_M = .2$; $Z = (5.9 - 5.5)/.2 = 2$; reject the null hypothesis. (c) Similar to 8c above, plus an explanation of material from on hypothesis testing, normal curve, means, and standard deviations. (d) 95% confidence interval: 5.51 to 6.29.
10. The error bars are the lines that go above and below the top of each bar. The error bars show, for each particular group, the standard deviation of the distribution of means for people like those in this group. (Then explain a distribution of means as in 8c above.)
11. Similar to confidence interval part of 8d above.

***Z* test**

❶ **Restate the question as a research hypothesis and a null hypothesis about the populations.**

❷ **Determine the characteristics of the comparison distribution.**
(a) Population M_M = Population M;
(b) Population SD^2_M = Population SD^2/N;
Population SD_M =
$\sqrt{\text{Population } SD^2_M}$;
(c) approximately normal if population normal or $N > 30$.

❸ **Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.** Use normal curve table.

❹ **Determine your sample's score on the comparison distribution.** $Z = (M - \text{Population } M_M)/$ Population SD_M

❺ **Decide whether to reject the null hypothesis.** Compare scores from Steps ❸ and ❹.

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

We do not recommend embarking on this chapter until you have a good understanding of the concepts of hypothesis testing and the distribution of means.

Effect Size

Consider an example of giving special instructions to fifth-graders taking a standard achievement test. In the hypothesis-testing process for this example (the Z test), we compare two populations:

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

From Chapter 7 of *Statistics for the Behavioral & Social Sciences: A Brief Course*, Fifth Edition. Arthur Aron, Elliot J. Coups, Elaine N. Aron. Copyright © 2011 by Pearson Education, Inc. Published by Pearson Prentice Hall. All rights reserved.

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, 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 reject the null hypothesis that the two populations are the same. This was because it is extremely unlikely that we would get a sample with a score as high as 220 from a population like Population 2 (see Figure 1). 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

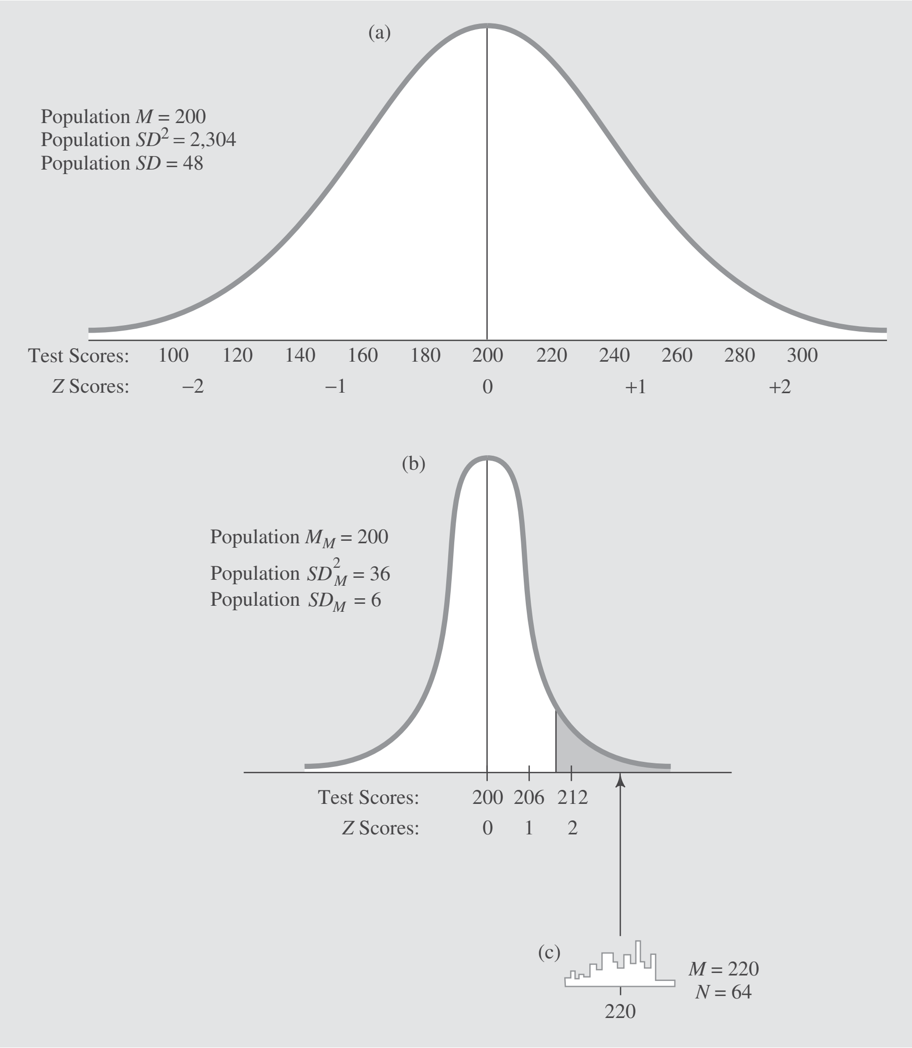


Figure 1 For the fictional study of fifth-graders' performance on a standard school achievement test, (a) the distribution of the population of individuals, (b) the distribution of means (the comparison distribution), and (c) the sample's distribution. The shaded area in the distribution of means is 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.

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 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 2 is the distribution (*of individual scores*) for Population 2; the right curve is the distribution for Population 1. Now look at Figure 3. Again, the left curve is for Population 2 and is the same as in Figure 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. 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 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.

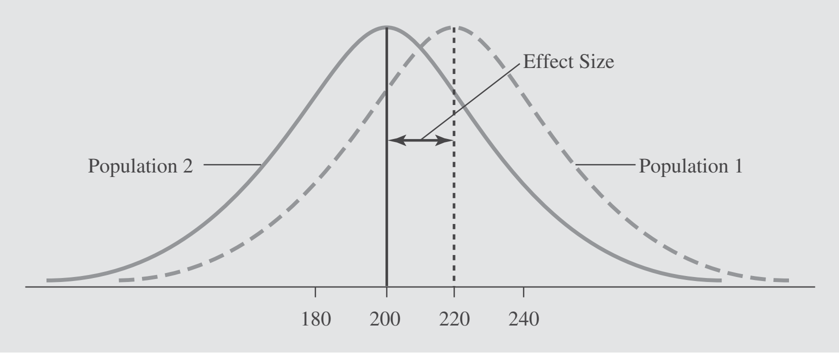


Figure 2 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of individuals for Population 1, those given the special instructions (right curve), and for Population 2, those not given special instructions (left curve). Population 1's mean is estimated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

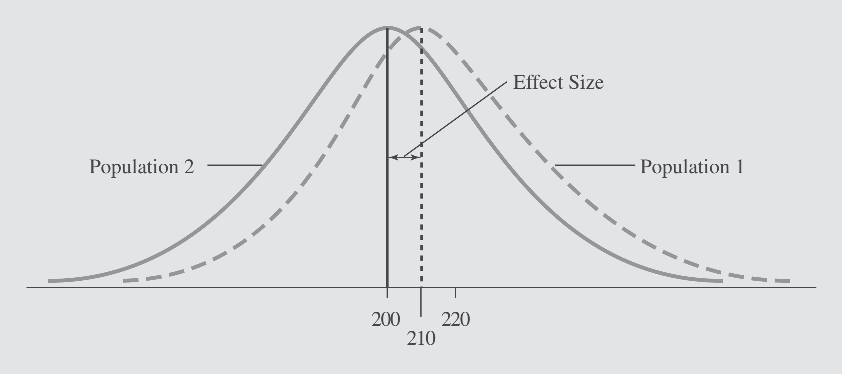


Figure 3 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of individuals for Population 1, those given the special instructions (right curve), and for Population 2, those not given special instructions (left curve). Population 1's mean is estimated 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.

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,

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.

[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).

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

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.

Consider again the fifth-grader example shown in Figure 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 2). 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 4a), 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 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 4c). These three effect size conventions are summarized in Table 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. Many IQ tests have a standard deviation of 16 points. An experimental procedure with a small effect size would be an increase of

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

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

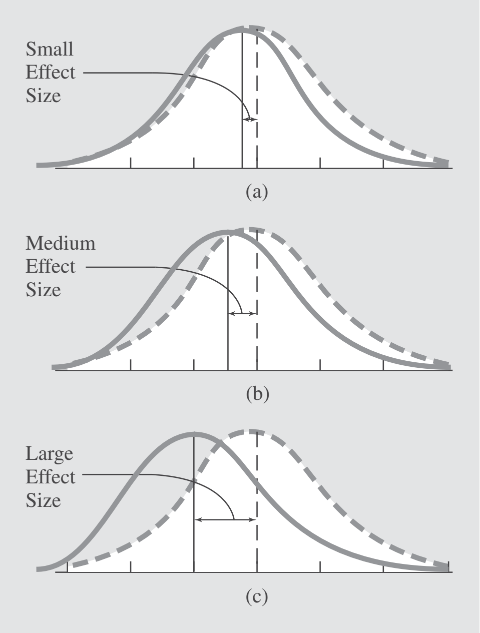


Figure 4 Comparisons of pairs of population distributions of individuals showing Cohen's conventions for effect size: (a) small effect size (.20), (b) medium effect size (.50), (c) large effect size (.80).

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 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 1.

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 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.

Statistical Power

TIP FOR SUCCESS

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.

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.) 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).

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

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

BOX 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 5 shows the distribution of means for Population 2. (Be careful: When discussing effect size, we showed figures, such as Figures 2 and 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. 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 6. Notice that the distribution has a mean of 208.

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

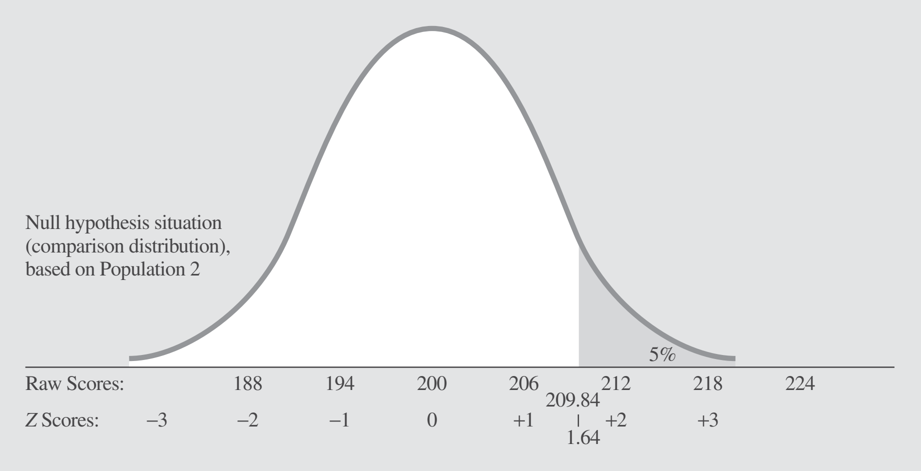


Figure 5 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 2 (the comparison distribution), those not given special instructions. Significance cutoff score (209.84) shown for $p < .05$, one-tailed.

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 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 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 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

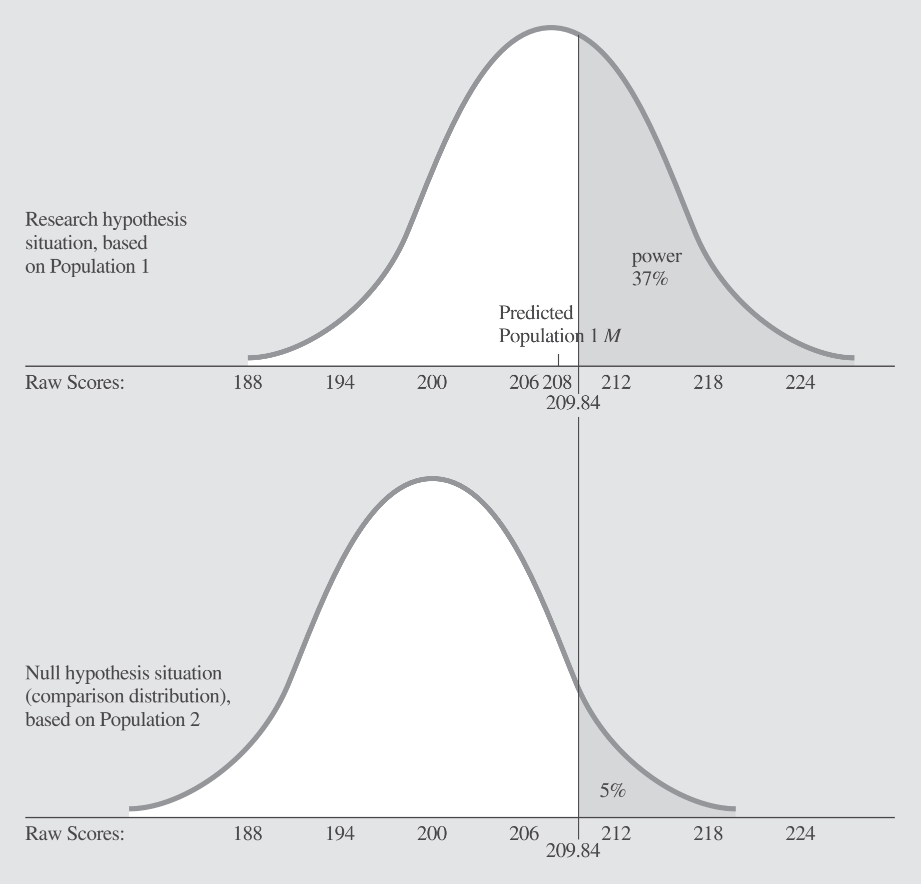


Figure 6 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population 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 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 6. However, the figuring is somewhat laborious. Furthermore, the figuring becomes quite complex once we consider 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.

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 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, 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 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 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 to Figure 6. You are more likely to get a significant result in the situation shown in Figure 7. This is because there is more overlap of the top curve with the shaded area on the comparison

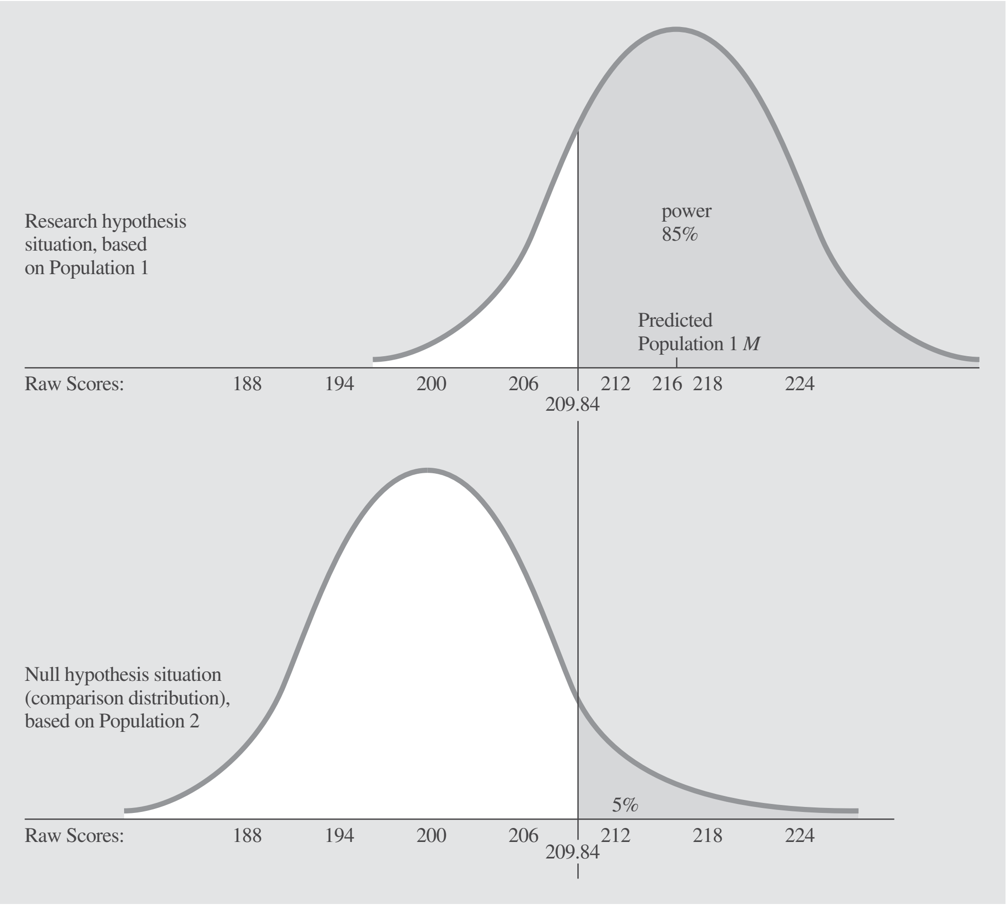


Figure 7 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 216 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population mean of 200 (lower curve). Significance cutoff score (209.84) shown for $p < .05$, one-tailed. Power = 85%. Compare with Figure 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 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, 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 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, 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 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 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) or because they have such a small standard deviation that even with a small mean difference they do not overlap much (as in Figure 8). This principle is summarized more generally in Figure 9.

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 10 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.

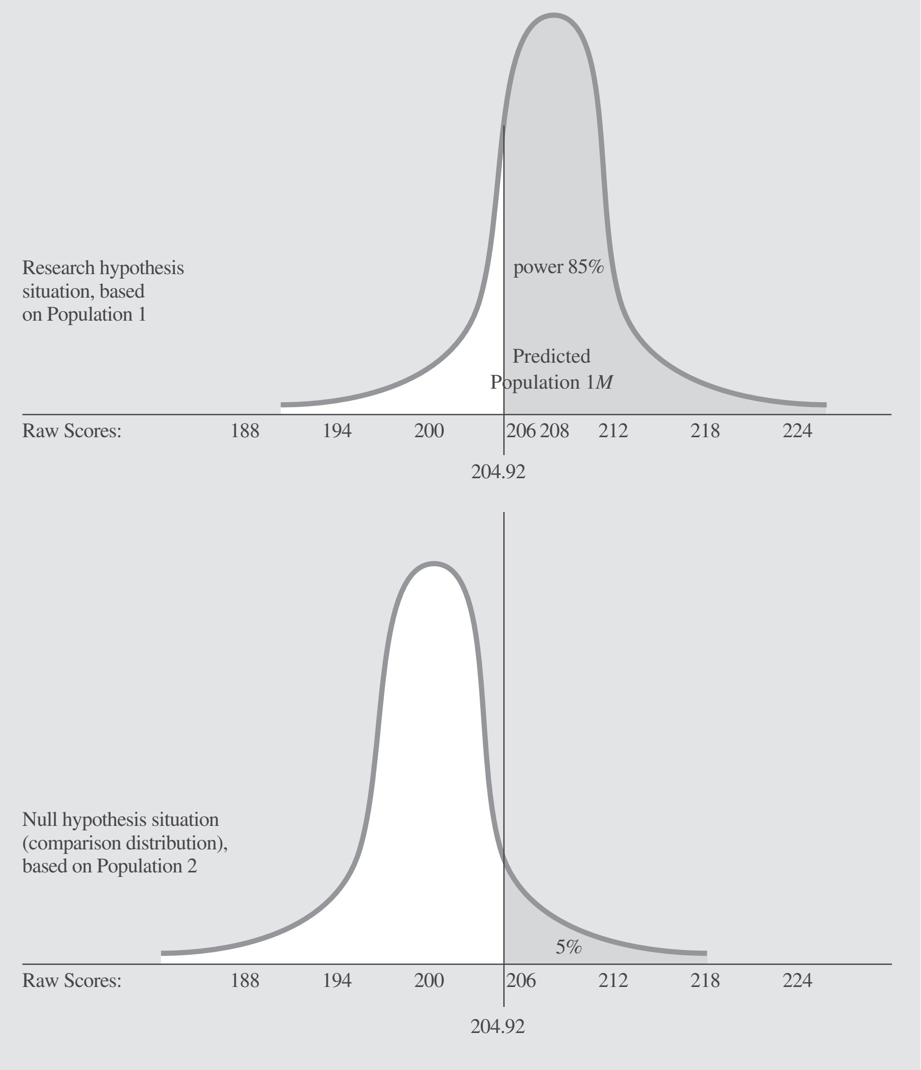


Figure 8 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population 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 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 11).

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

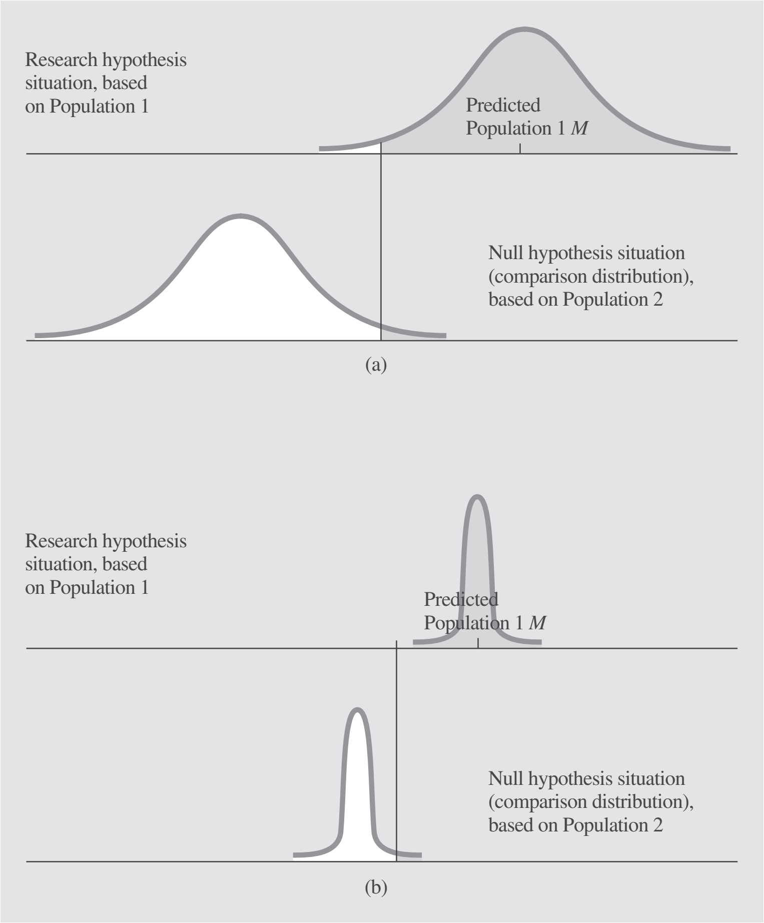


Figure 9 The predicted and comparison distributions of means might have little overlap (and thus the study would have high power) because either (a) the two means are very different or (b) the population standard deviation is very small.

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

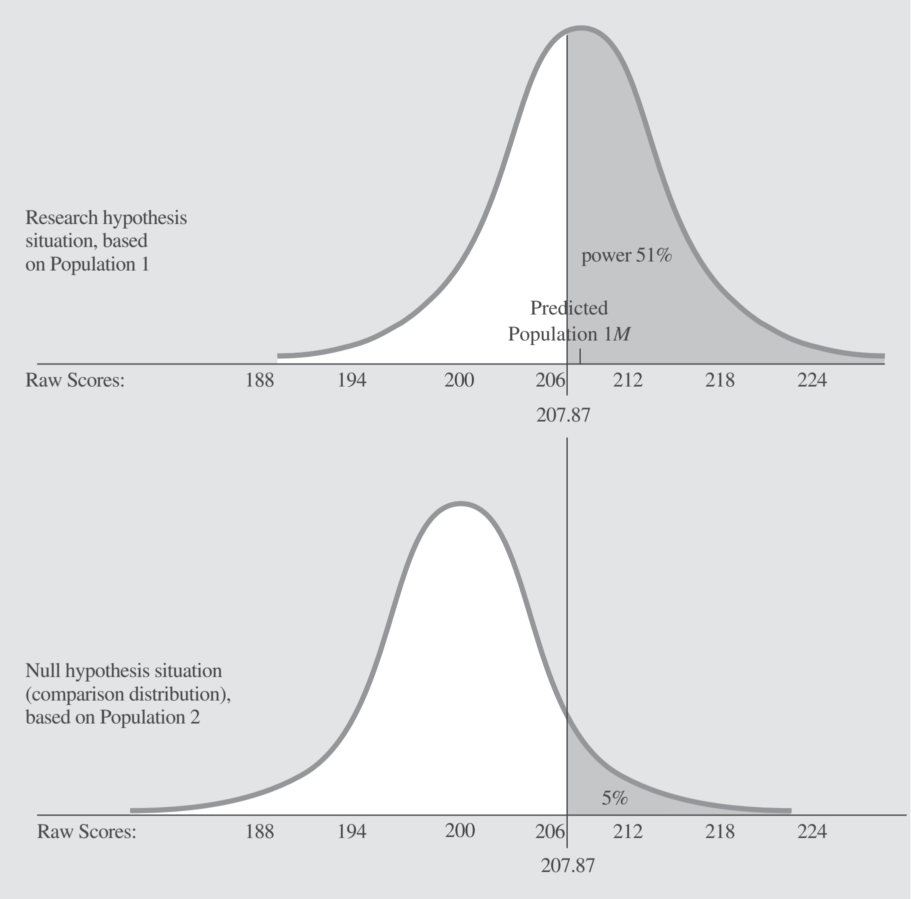


Figure 10 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population 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 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.

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 12).

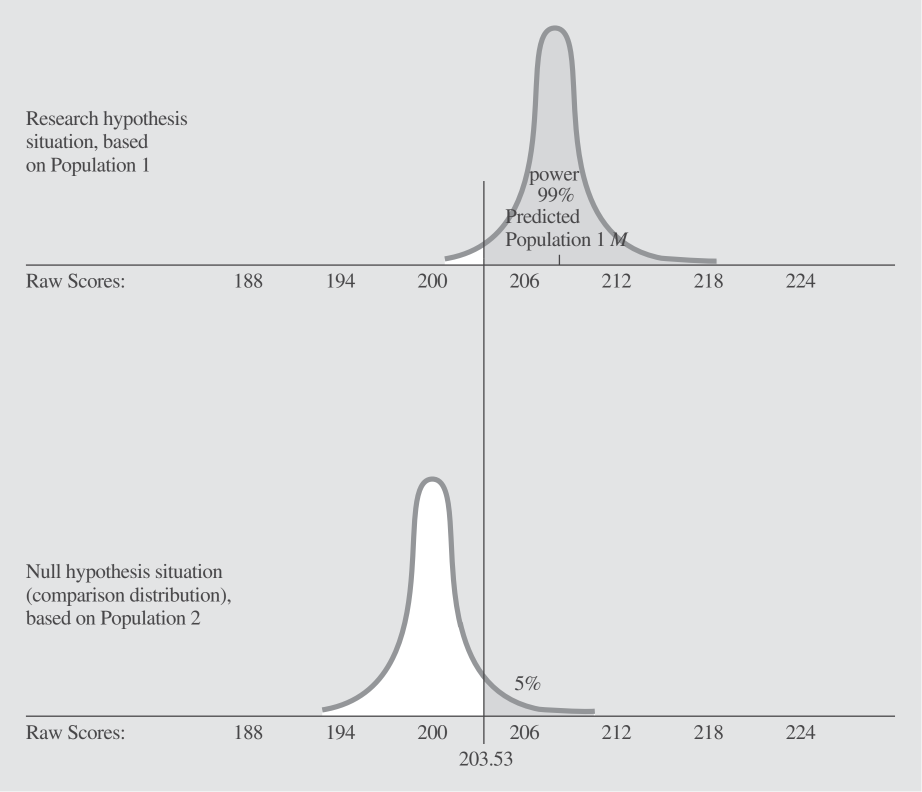


Figure 11 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population 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 6, which had the original sample size of 64 fifth-graders and power was 37%, and Figure 10, which had a sample size of 100 and a power of 51%.

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 13, power would be only 26% (compared to 37% in the original one-tailed version shown in Figure 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.

Summary of Influences on Power

Table 3 summarizes the effects of various factors on the power of a study.

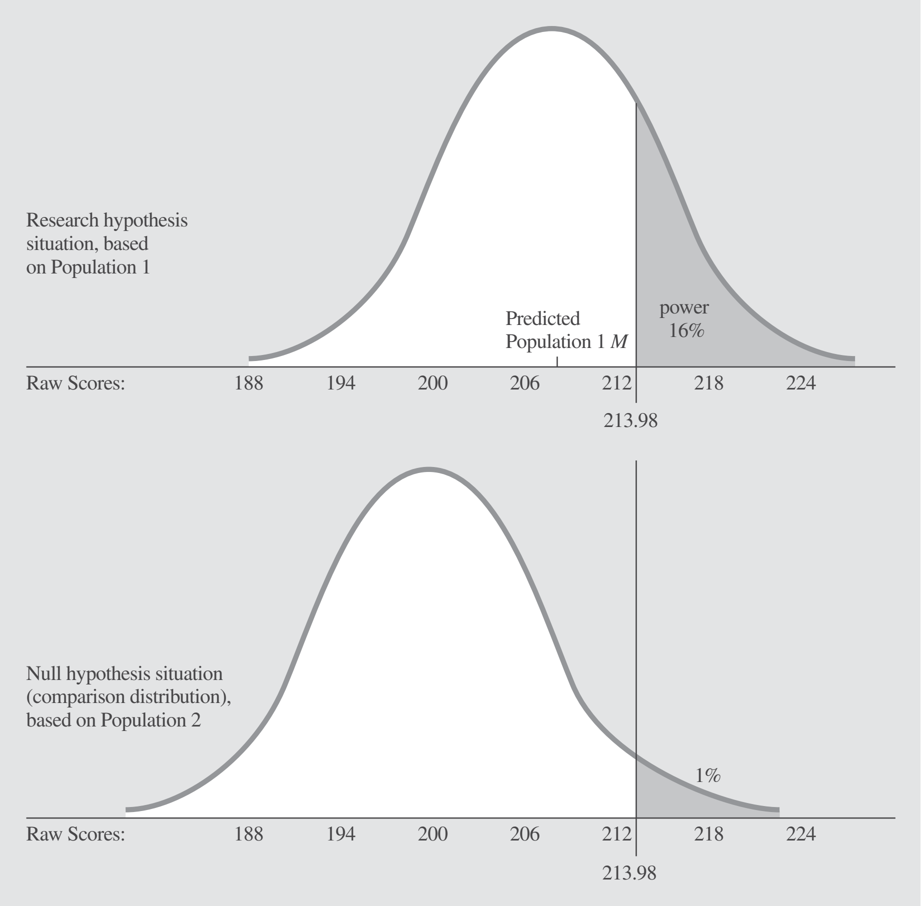


Figure 12 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population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

Table 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 (*M*)	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

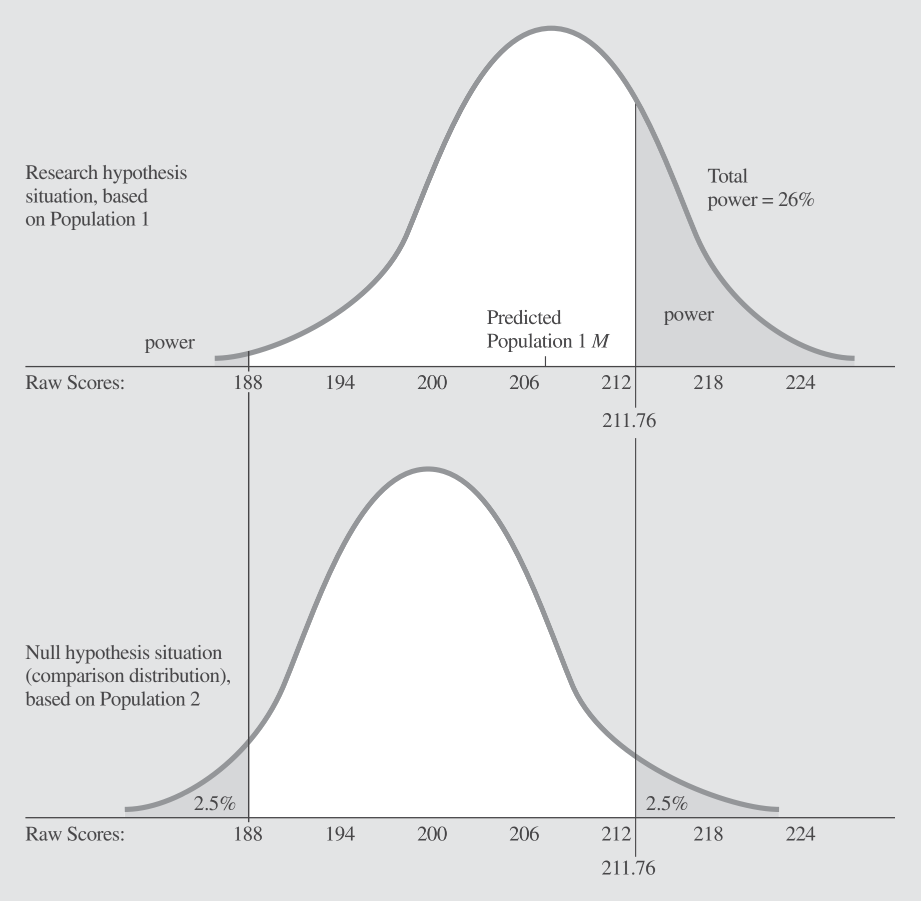


Figure 13 For the fictional study of fifth-graders' performance on a standard school achievement test, distribution of means for Population 1 based on a predicted population mean of 208 (upper curve), and distribution of means for Population 2 (the comparison distribution) based on a known population 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 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 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. Usually, however, the researcher begins with the most sensitive method available, so little more can be done.

Table 4 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 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

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

Table 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 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
effect size conventions
meta-analysis
statistical power
power tables

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.

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 the end of this chapter)

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 ($d\,[\text{effect size}] = .8$). Indeed, the power is about 90% for finding at least a near significant ($p < .10$) medium-sized effect ($d\,[\text{effect size}] = .5$). 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.

1. (a) Effect size = (Population 1 M − Population 2 M)/ Population $SD = (19 - 25)/12 = -.50$, medium; (b) −.25, small; (c) 0, no effect; (d) .42, medium; (e) .83, large.
2. (a) Predicted effect size = Population 1 M − Population 2 M/ Population $SD = (50 - 50)/5 = 0$, no effect; (b) .40, medium; (c) .80, large; (d) 1.20, large; (e) −.60, medium.
3. (a) Predicted effect size = Population 1 M − Population 2 M/Population $SD = (91 - 90)/4 = .25$; (b) .50; (c) 1.00; (d) −1.00.
4. (a) Not affected. (b) Possibly of small importance. (c) Regarding situation (a), the significance tells you the probability of getting your results if the null hypothesis is true; sample size is already taken into account in figuring the significance. Regarding situation (b), it is possible to get a significant result with a large sample even when the actual practical effect is slight—such as when the mean of your sample (and this, your best estimate of the mean of the population that gets the experimental treatment) is only slightly higher than the mean of the known population. This is possible because significance is based on the difference between the mean of your sample and the known population mean with this difference then divided by the standard deviation of the distribution of means. If the sample size is very large, then the standard deviation of the distribution of means is very small. (This is because it is figured by taking the square root of the result of dividing the population variance by the sample size.) Thus, even a small difference between the means when divided by a very small denominator can give a large overall result, making the study significant.
5. Power is the chance of rejecting the null hypothesis if the research hypothesis is true. In other words, the power of a study represents the likelihood that you will get a statistically significant result in your study, if in fact the research hypothesis is true. Ideally, a study should have power of 80% or more. If a study has low power and does not get a statistically significant result, the result of the study is entirely inconclusive. This is because it is not clear whether the nonsignificant result is due to the low power of the study or because the research hypothesis is in fact false.

   Effect size can be thought of as the degree to which distributions do not overlap. The larger the effect size, the larger the power. As noted in the quotation from the research article, the study had a high level of power (about 90%) for detecting both large and medium-sized effects. Given this high level of power, the researchers were able to conclude that the most likely reason for the nonsignificant study results is that the research hypothesis is in fact false. As the researchers noted, with such a high level of power, it is very unlikely that the results of the study would be nonsignificant if there were in fact a medium-sized or large effect in the population. Since smaller effect sizes are associated with lower power, the researchers were careful not to rule out the possibility that there is in fact a small effect in the population (which may not have been detected in the study due to the lower power for identifying a small effect size).
6. (a) Increases power; (b) decreases power; (c) increases power; (d) decreases power; (e) decreases power.

7. One situation is that when planning an experiment, figuring power gives you the chance to make changes of various kinds (or even abandon the project) if power is too low. (Or if power is higher than reasonably needed, you would then be able to make changes to make the study less costly, for example, by reducing the number of participants.) Another situation is figuring power after a study is done that had nonsignificant results. If you figure that power was high in the study, this means you can be pretty confident that the null hypothesis really is true in the population, in the sense that the true difference in the population is really smaller than the effect size you used to figure power. But if you figure the power of the study was low, this tells you that the result really is ambiguous and that it is still reasonable to think that future research testing this hypothesis might have a chance of being significant. A third possibility is figuring power after a study is done that got a significant result and the researchers do not give the effect size. If the study had high power (as when it used a large sample), this tells you that the effect size could have been small and thus the effect not very important for practical application. But if the study seems to have had low power (as from having a small sample), this tells you that the effect size must have been large for them to get a significant result.