Chapter Outline

In Chapter 6, we introduced the basic logic of hypothesis testing. We used as examples studies in which the sample was a single individual. As we noted, however, in actual practice, psychology research usually involves samples of many individuals. In this chapter, we build on what you have learned so far and consider hypothesis testing with a sample of more than one individual. Mainly this requires examining in some detail what we call a distribution of means.

THE DISTRIBUTION OF MEANS

Hypothesis testing in the usual research situation, where we are studying a sample of many individuals, is exactly the same as you learned in Chapter 6—with an important exception. When you have more than one person in your sample, there is a special problem with Step 2, determining the characteristics of the comparison distribution. The problem is that the score you care about in your sample is the mean of the group of scores. The comparison distributions we have been considering so far have been distributions of populations of individual scores (for example, the ages when individual babies start walking or the population of individual scores on a happiness questionnaire). Comparing the mean of a sample of, say, 50 individuals to a distribution of individual scores is a mismatch—like comparing apples and oranges. Instead, when you are interested in the mean of a sample of 50, you need a comparison distribution that is a distribution of means of samples of 50 scores. Such a comparison distribution we will call a **distribution of means.**

Put more formally, a distribution of means is a distribution of the means of each of a very large number of samples of the same size, with each sample randomly drawn from the same population of individuals. (Statisticians also

distribution of means

call this distribution of means a "sampling distribution of the mean." In this book, however, we use the term *distribution of means* to make it clear that we are discussing populations, not samples or distributions of samples.)

The distribution of means is the proper comparison distribution when there is more than one person in a sample. Thus, in most research situations, determining its characteristics is necessary for Step 2 of the hypothesis-testing procedure.

CONSTRUCTING A DISTRIBUTION OF MEANS

The idea of a distribution of means can be understood by considering how one could build up such a distribution from an ordinary distribution of individuals. Suppose our population was of 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 7-1).

Next, suppose that you wrote each child's grade level on a table tennis ball and put all 90,000 plastic balls into a giant tub. The tub would contain 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 a 2 on it and the other has a 9 on it. In that case, the mean grade level of your 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.

These three numbers (each a mean of a sample of grade levels of two school children) can be thought of as a small distribution in its own right. The mean of this little distribution of three numbers is 4.67 (the sum of 5.5, 4, and 4.5, divided by 3). The variance of this distribution is .39 (the variance of 5.5, 4, and 4.5). The standard deviation is .62 (the square root of .39). A histogram of this distribution of three means is shown in Figure 7-2.

If you continued the process, the histogram of means would continue to grow. An example after 10 random samples of two balls each is shown in Figure 7-3a. Figure 7-3b shows the histogram of the distribution of means after

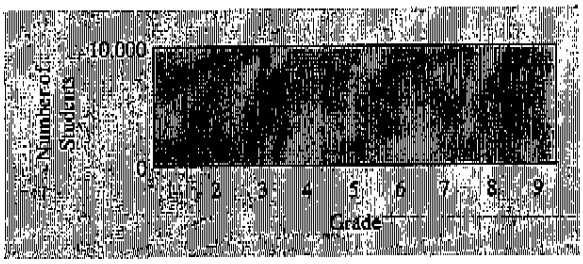

FIGURE 7-1
Distribution of grade levels among 90,000 school children (fictional data).

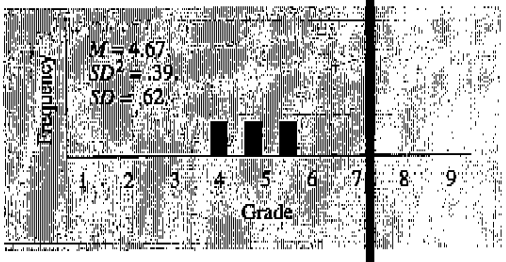

FIGURE 7-2
Distribution of the means of three randomly drawn samples of the grade levels of two school children each from a population of the grade levels of 90,000 school children (fictional data).

20 random samples of two each. After 100 random samples, the histogram of the distribution of the means might look like Figure 7-3c; after 1,000, like Figure 7-3d. (We actually made the histograms in Figure 7-3 using a computer to make the random selections, instead of using 90,000 table tennis balls and a giant tub.)

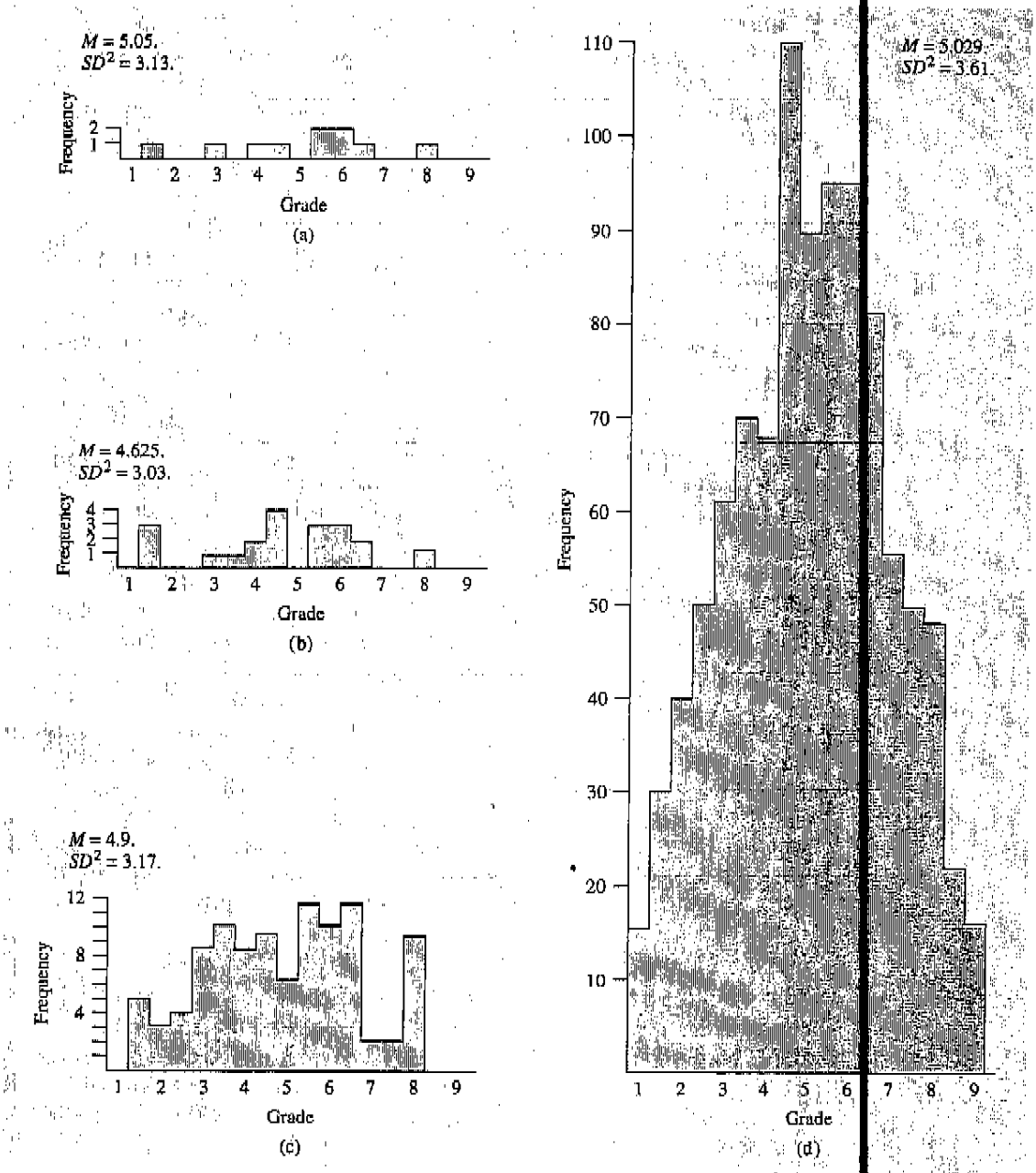

FIGURE 7-3
Distributions of the means of randomly selected samples of two balls each from a population of 90,000 balls, consisting of 10,000 bearing each of the numbers from 1 through 9. Numbers of sample means in each distribution shown are (a) 10 sample means, (b) 20 sample means, (c) 100 sample means, and (d) 1,000 sample means. (Actual sampling simulated by computer.)

In practice, researchers almost never have the opportunity to take many different samples from a population. It takes a lot of work to come up with a single sample and study the people in that sample. Fortunately, however, you can determine the characteristics of a distribution of means directly, using some simple rules, without taking a single sample. The only information you need is (a) the characteristics of the population distribution of individuals and (b) 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.

CHARACTERISTICS OF A DISTRIBUTION OF MEANS

Notice three things about the distribution of means we built in our example (as shown in Figure 7-3):

1. The mean of the distribution of means comes out to be about the same as the mean of the original population of individual grade levels from which the samples were taken (5 in both cases).
2. The spread of the distribution of means comes out to be less than the spread of the distribution of the population of individuals from which the samples were taken.
3. The shape of the distribution of means came out to be approximately normal (or at least unimodal and symmetrical).

It turns out that the first two of these three are true for all distributions of means and the third is true for most distributions of means.

These three links of the distribution of means to the population of individuals are the foundation for a set of simple rules that allow you to determine the mean, variance, and shape of a distribution of means without having to write on plastic balls and take endless samples. These three rules, to which we will turn shortly, are based on the *central limit theorem,* a fundamental principle in mathematical statistics that we mentioned in Chapter 5.

Now, let's examine the three rules.

Rule 1: Determining the Mean of a Distribution of Means

mean of a distribution of means

The first rule is that the **mean of a distribution of means** is the same as the mean of the population of individuals from which the samples are taken. Each sample is based on randomly selected scores from the population of individuals. Thus, sometimes the mean of a sample will be higher than the mean of the whole population of individuals and sometimes the mean of a sample will be lower than the mean of the whole population of individuals. However, there is no reason for the means of these samples to tend overall to be consistently higher or lower than the mean of the population of individuals. If enough samples are taken, the high means and the low means balance each other out.

You can see in Figure 7-3 that with a large number of samples the mean of the distribution of means becomes very similar to the mean of the population

of individuals, which in this case was 5. If we had shown an example with 10,000 means of samples, it would be even closer to 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.

Rule 2: Determining the Variance of a Distribution of Means

Figure 7-3 also shows that a distribution of means will be less spread out than the population of individuals from which the samples are taken. The reason for this is as follows: Any one score, even an extreme score, has some chance of being selected in a random sample. The chance is less of two extreme scores being selected in the same random sample. Further, to create an extreme sample mean, two extreme scores would have to be extreme in the same direction. So increasing numbers has a moderating effect. In any one sample, the deviants tend to be balanced out by middle scores or by extremes in the opposite direction. This makes each sample mean tend toward the middle and away from extreme values. With fewer extreme means, the variance of the means is less.

Consider 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 being 9s) much less often. The chances of 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—a 1 and a 9, a 2 and an 8, a 3 and a 7, a 4 and a 6, and two 5s).

The more scores in each sample, the less spread out the distribution of means of those samples. This is because with several scores in each sample, it is even rarer for extreme scores 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 saw that it was fairly unlikely to get 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 even more likely.

Using samples of two balls at a time in our example, the variance of the distribution of means will come out to about 3.33. This is half of the variance of our population of individual balls, which was 6.67. If we had built up a distribution of means using samples of three balls each, the variance of the distribution of means would have been 2.22, which is one third of the variance of our 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 second rule regarding the distribution of means: The **variance of a distribution of means** is the variance of the distribution of the population of individuals divided by the number of scores in each of the samples being selected. This rule holds in all situations and can be proven mathematically.

variance of a distribution of means

Here is the rule for figuring the variance of the distribution of means stated as a formula:

$$\sigma_M^2 = \frac{\sigma^2}{N} \tag{7-1}$$

In this formula, σ_M^2 is the variance of the distribution of means, σ^2 is the variance of the population of individuals, and N is the number of scores in each sample.

In our example, the variance of the population of individual grade levels was 6.67, and there were two school children in each sample. The variance of the distribution of means is figured as follows:

$$\sigma_M^2 = \frac{\sigma^2}{N} = \frac{6.67}{2} = 3.34$$

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

$$\sigma_M^2 = \frac{\sigma^2}{N} = \frac{400}{25} = 16$$

standard deviation of a distribution of means

The **standard deviation of a distribution of means** is the square root of the variance of the distribution of means. Stated as a formula,

$$\sigma_M = \sqrt{\sigma_M^2} = \sqrt{\frac{\sigma^2}{N}} \tag{7-2}$$

In this formula, σ_M is the standard deviation of the distribution of means.

Sometimes this formula is algebraically manipulated to emphasize the relation between the standard deviation of the population of individuals and the standard deviation of the distribution of means:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{7-3}$$

standard error of the mean

Because of its importance in hypothesis testing, the standard deviation of the distribution of means is sometimes called by a special name of its own, the **standard error of the mean,** or the *standard error,* for short. This name represents the degree to which particular means of samples are typically "in error" as estimates of the mean of the population of individuals. That is, the standard error of the mean tells you how much the particular means in the distribution of means deviate from the mean of the population. We will have more to say about this in our discussion of confidence intervals at the end of the chapter.

Rule 3: The Shape of a Distribution of Means

Regardless of the shape of the original distribution of individuals, the distribution of means will tend to be unimodal and symmetrical. In the grade-level example, the population distribution of students at individual grade levels was rectangular. (It had an equal number of scores at each value.) However,

the **shape of the distribution of means** was roughly that of a bell—unimodal and symmetrical. Had we taken many more than 1,000 samples in our examples in Figure 7-3, the shape would have come out even more clearly unimodal and symmetrical.

shape of the distribution of means

A distribution of means tends to be unimodal due to 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. The distribution tends to be symmetrical because lack of symmetry (skew) is caused mainly by extremes. With fewer extremes, there is less asymmetry. In our grade-level example, the distribution of means we created 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 still have been skewed, but not as much.

The more scores in each sample, the closer the distribution of means will be to a normal curve. Thus, our third rule is that with samples of 30 or more scores, even with a nonnormal 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] Also, whenever the population distribution of individuals is normal, a distribution of means will be normal, regardless of the number of scores in each sample.

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

Here are the three rules:

1. The mean of a distribution of means is the same as the mean of the distribution of the population of individuals.
2. The variance of a distribution of means is the variance of the distribution of the population of individuals divided by the number of scores in each sample ($\sigma_M^2 = \sigma^2/N$). Its standard deviation is the square root of its variance ($\sigma_M = \sqrt{\sigma_M^2}$).
3. The shape of a distribution of means is at least approximately normal if either (a) each sample is of 30 or more scores, or (b) the distribution of the population of individuals is normal. Otherwise, it will still tend to be unimodal and roughly symmetrical.

These principles are shown graphically in Figure 7-4.

[1]We have ignored the fact that a normal curve is a smooth theoretical distribution. In most real-life examples, scores are at specific intervals. So one difference between a normal curve and our example's distribution of table tennis balls' means is that a normal curve is smooth. However, in psychology research, we usually assume that even though our measurements are at specific intervals, the underlying thing being measured is continuous.

[2]We have already considered this principle of a distribution of means tending toward a normal curve in Chapter 5. Though we had not yet discussed the distribution of means, we still used this principle to explain why the distribution of so many things in nature follows a normal curve. In that chapter, we explained it as the various influences balancing each other out, to make an averaged influence come out with most of the scores near the center and a few at each extreme. Now we have made the same point using the terminology of a distribution of means. Think of any distribution of individual scores in nature as representing a situation in which each score is actually an average of a random set of influences on that individual score. Consider the distribution of weights of pebbles. Each pebble's weight represents a kind of average of all the different forces that went into making the pebble have a particular weight.

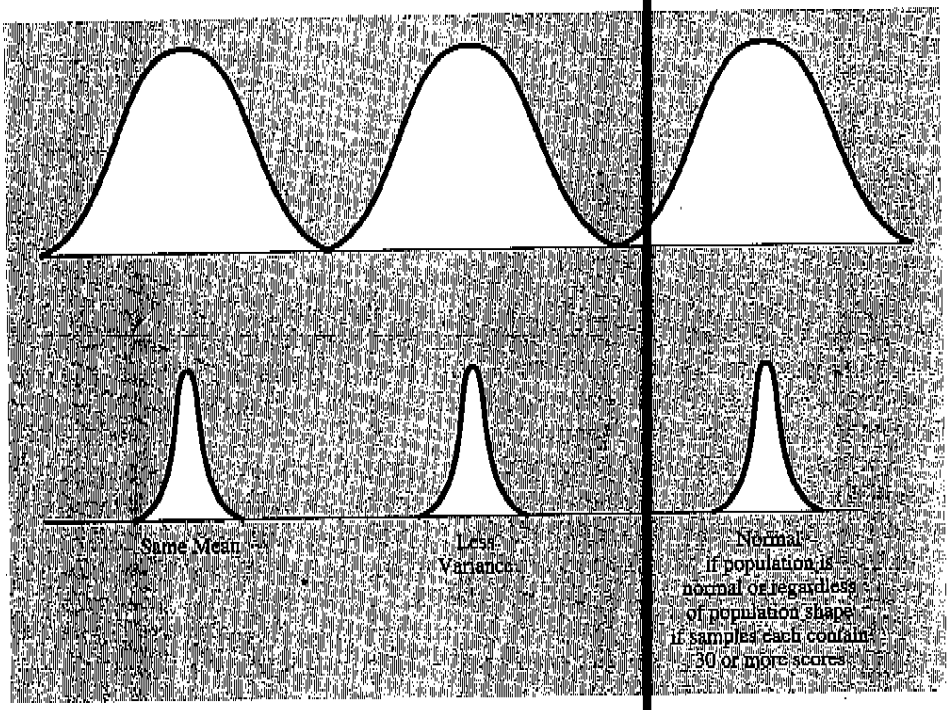

FIGURE 7-4
Illustration of the principles of the relation of the distribution of means (lower curves) to the distribution of the population of individuals (upper curves).

Example of Determining the Characteristics of a Distribution of Means

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

1. Because the mean of the population is 500, the mean of the distribution of means will also be 500.

2. The variance of the distribution of means is the variance of the population of individuals divided by the number of individuals in each sample. Because the standard deviation of the population of individuals is 100, the variance of the population of individuals is 10,000. The variance of the distribution of means is 10,000 divided by 50, which is 200. In terms of the formula,

$$s_M^2 = \frac{s^2}{N} = \frac{10{,}000}{50} = 200$$

The standard deviation of the distribution of means is the square root of the variance of the distribution of means: $\sqrt{200} = 14.14$.

3. The shape of the distribution of means will be normal. Both of our requirements are met: The population distribution of individuals is normal,

and the number of individuals in each sample is 30 or more. (It would have been enough even if only one of these requirements had been met.)

Another Example of Determining the Characteristics of a Distribution of Means

The Adjective Check List (Gough & Heilbrun, 1983) is a widely used personality test. The test consists of a list of adjectives—*able, active, athletic,* and so forth—each of which is checked off by test takers if it applies to them. One of the subtests of the Adjective Check List focuses on aggression (adjectives such as *aggressive, argumentative,* and *assertive*). The test has been given to large numbers of people in the past, and it is known that scores on the Aggression scale have a skewed distribution with a mean of 51 and a variance of 93 (rounded off). What will be the characteristics of a distribution of sample means from this population of individuals if the samples each contain 10 individuals?

1. The mean of the distribution of means will be 51, the same as the mean of the population of individuals.
2. The variance of the distribution of means will be 93, the population variance, divided by 10, the number of scores in each sample. The result is 9.3. Using the formula,

$$\sigma_M^2 = \frac{\sigma^2}{N} = \frac{93}{10} = 9.3$$

The standard deviation of the distribution of means is the square root of 9.3, or 3.05.

3. The distribution of means will not be normal because the population distribution of individuals is not normal and the number in each sample is only 10. However, like any distribution of means, it will tend to be unimodal and closer to symmetrical than the population distribution of individuals.

Review of Three Kinds of Distributions

We have considered three different 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. Figure 7-5 illustrates these three distributions and Table 7-1 compares them.

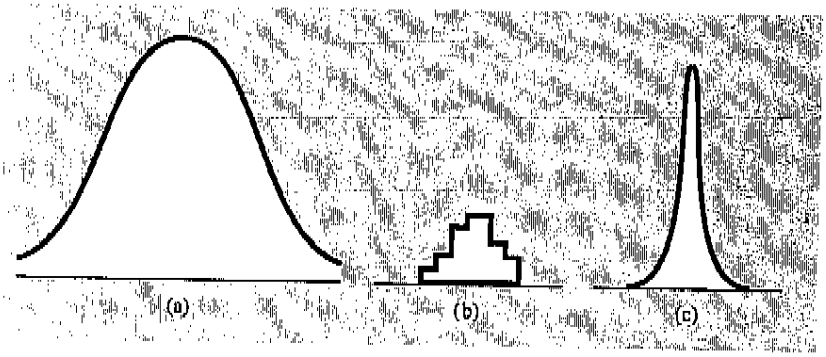

FIGURE 7-5
Three kinds of distributions: (a) the distribution of a population of individuals, (b) the distribution of a particular sample taken from that population, and (c) the distribution of means of all possible samples of a particular size taken from that distribution.

TABLE 7-1
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	Normal if population is normal, approximately normal if samples contain ≥ 30 scores each
Mean	μ	$M = \Sigma X/N$, calculated from scores of those in the sample	$\mu_M = \mu$
Variance	σ^2	$SD^2 = \Sigma(X - M)^2/N$, calculated from scores of those in the sample	$\sigma_M^2 = \sigma^2/N$
Standard deviation	σ	$SD = \sqrt{SD^2}$	$\sigma_M = \sqrt{\sigma_M^2} = \sigma/\sqrt{N}$

HYPOTHESIS TESTING INVOLVING A DISTRIBUTION OF MEANS

Now we are ready to turn to hypothesis testing when there is more than one individual in the study's sample.

The Distribution of Means as the Comparison Distribution in Hypothesis Testing

In this new situation, the distribution of means provides the crucial connection between the sample and the null hypothesis. Suppose we are studying a sample of more than one person (the usual situation in research). In this situation, the distribution of means is the *comparison distribution*—the distribution whose characteristics are determined in Step 2 of the hypothesis-testing process. The distribution of means is the distribution to which the sample mean can be compared to see how likely it is that such a sample mean could have been selected if the null hypothesis were true.

Finding the *Z* Score of a Sample Mean on a Distribution of Means

There can be some confusion in determining 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 no different from the usual way of changing a raw score to a Z score. However, you have to be careful not to get

Box 7-1
More About Polls: Sampling Errors and Errors in Thinking About Samples

If you think back to Box 5-3 on surveys and the Gallup poll, you will recall that we left an important question unanswered about the sort of fine print you find near the results of a poll, saying something like "From a telephone survey of 1,000 American adults taken on June 4–5. Sampling error ±3%." We said that 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. public.

Let's begin with the question of sample size. You know from this chapter that when sample sizes are large, like 1,000, the standard deviation of the distribution of means is greatly reduced. That is, the distribution of sample means becomes very high and narrow, gathered all around the population mean. Thus, the mean of any sample of that size is very close to being the population mean. To put it another way, the variance of the distribution of means—which reflects how much any sample's mean tends to differ from the population's mean—is the variance of the population divided by the sample size. The size of the population (of individuals) itself or the relation of the sample's size to the population's plays no part in this formula.

Still, you might persist in an intuitive feeling that the number required to represent all of the huge U.S. public might need to be larger than just 1,000. However, if you think about it, 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. A sample's size relative to the population does sometimes matter—if the population is so small that "removing" or interviewing some would increase the odds of the remaining ones being interviewed. But when the population is in the millions, removing a thousand or two will have almost no impact on the odds of others being interviewed. A survey of 1,000 out of 1 million voters or out of 10 million or 100 million will have essentially the same chance error. What is important is reducing bias or systematic error, which can be done only by careful planning.

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. In other words, the ordinary formula (from Chapter 2) for converting a raw score to a Z score is $Z = (X - M)/SD$. In the present situation, you are actually using the following formula:

$$Z = \frac{(M - \mu_M)}{\sigma_M} \tag{7-4}$$

For example, suppose your sample 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,

$$Z = \frac{(M - \mu_M)}{\sigma_M} = \frac{18 - 10}{4} = \frac{8}{4} = 2$$

This is illustrated in Figure 7-6.

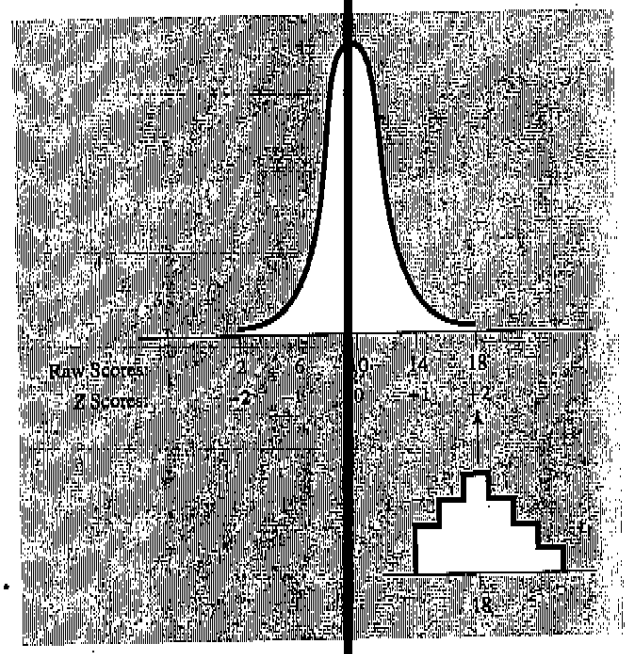

FIGURE 7-6
Z score for the mean of a particular sample on the distribution of means.

Example of Hypothesis Testing With a Sample of More Than One Individual

Recall from Chapters 1 and 2 the fictional experiment involving the reading of ambiguous sentences. In that chapter, we simply looked at the distribution of reading times when the sentences were presented without a context. Now we will suppose that the researchers want to test a theory about the importance of context. Thus, they conduct a study examining reading times when there is some context for the ambiguous sentences, making the meanings clearer. The goal is to see if reading time will be faster under these conditions. Of course, it is also possible that providing a context slows down reading by making the situation more complicated.

We will also assume that the researchers have conducted many previous studies with these ambiguous sentences presented without a context. From this research, we will presume that the researchers are confident that in the general population reading times for ambiguous sentences without any context are roughly normally distributed with a mean of 2.75 seconds and a variance of .02 seconds ($\sigma = .14$ seconds). This population distribution is shown in Figure 7-7a.

In this study, 40 individuals are tested using ambiguous sentences with a context. Their mean reading time is 2.71 seconds. (In this example, we know the population variance before doing the study. In this kind of situation, the variance of the sample is not used in any way in the hypothesis-testing process.) The sample's distribution is shown in Figure 7-7c.[3]

[3]Actually, this study would be much better if the researchers also had another group of research participants who were randomly assigned to be tested for reading speed of ambiguous sentences without a context. Relying on the information from previous studies is a bit hazardous, because the testing circumstances from one study to another may not be identical. However, we have taken liberties with this example to help us introduce the hypothesis-testing process one step at a time. In this example and the others in this chapter, we use situations in which a

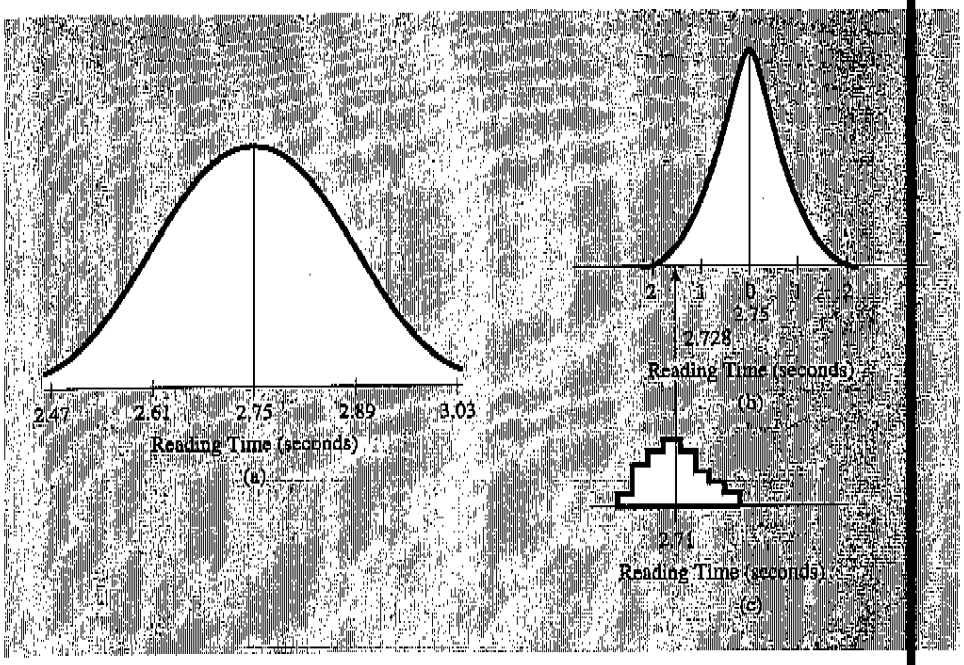

FIGURE 7-7
For the fictional experiment from Chapter 1 involving reading ambiguous sentences, (a) the distribution of the population of individuals, (b) the distribution of means, and (c) the sample's distribution.

What should the researchers conclude? Let's follow the steps of hypothesis testing.

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

Population 1: Research participants who read ambiguous sentences with a context

Population 2: Research participants who read ambiguous sentences without a context

The research hypothesis is that there is a difference in reading time between the two populations, that reading time with a context will be different than reading time without a context; $\mu_1 \neq \mu_2$. The null hypothesis is that there is no difference in reading time between the two populations; $\mu_1 = \mu_2$. Note that these hypotheses are nondirectional. The researchers expect reading time to be faster with a context. However, they cannot rule out the possibility that the context will slow reading time, and such a result would be quite interesting.

2. Determine the characteristics of the comparison distribution. If the null hypothesis is true, the population of individuals our sample comes from is no different than Population 2, for which we know the mean and variance. What we need to figure out now is the characteristics of a distribution of means of samples of 40 scores each, taken from this population of individuals.

---

single sample is contrasted with a "known" population. Starting in Chapter 9, we extend the hypothesis-testing procedure to more realistic research situations, those involving more than one group of participants and those involving populations whose characteristics are not known.

Thus, you follow the rules for determining the characteristics of a distribution of means: (a) Its mean is the same as the population mean, in this case, 2.75 seconds; and (b) its variance is the population variance divided by the number of scores in each sample. Using the formula,

$$\sigma_M^2 = \frac{\sigma^2}{N} = \frac{.02}{40} = .0005$$

The standard deviation is the square root of this, .022. Finally, (c) the shape of the distribution will be close to a normal curve because the samples have more than 30 scores each. This distribution of means is illustrated in Figure 7-7b.

3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected. Assume that the researchers decided on the 5% significance level. As we noted in Step 1, they have made a nondirectional hypothesis. Thus, a two-tailed test is required. We just determined that our comparison distribution is normal, so we can now consult the normal curve table to find the Z score needed for the top and bottom 2 1/2%. The table shows that to reject the null hypothesis at the 5% level, we need a Z score either at or above +1.96 or at or below –1.96. These two 2 1/2% regions in which the null hypothesis would be rejected are shown as small shaded areas (they are very hard to see) in the two tails of the distribution of means illustrated in Figure 7-7b.

4. Determine your sample's score on the comparison distribution. The mean of the sample is 2.71 (see Figure 7-7c). From Step 2 we know that the comparison distribution (our distribution of means) has a mean of 2.75 and a standard deviation of .022. Using the formula,

$$Z = \frac{(M - \mu_M)}{\sigma_M} = \frac{2.71 - 2.75}{.022} = \frac{-.04}{.022} = -1.82$$

5. Compare the scores from Steps 3 and 4 to decide whether to reject the null hypothesis. The Z score needed to reject the null hypothesis is ±1.96. Our obtained Z score is only –1.82. Therefore, we cannot reject the null hypothesis. The experiment is inconclusive. This can be seen graphically in Figure 7-7b, which shows that the location of our sample mean on the distribution of means is not so extreme as to be clearly unlikely to be taken from this distribution.

This result, however, is nearly extreme enough to reject the null hypothesis. Thus, the researchers might note that the result was "near significant" or "approached significance," perhaps adding "$p < .10$." (The cutoff for significance at the .10 level, two-tailed, is ±1.64.) But with a borderline result like this, the best advice is to repeat the experiment, perhaps with more participants. (Chapter 8 includes a discussion of the effects of increasing the number of participants on the probability that your experiment will produce a significant result.)

Another Example of Hypothesis Testing With a Sample of More Than One Individual

Here is another fictional example. Two educational psychologists are studying the effects of instructions on timed scholastic achievement tests. They have a theory which predicts that when test takers are told to answer each

question with the first response they think of, they will do better. To examine this theory, the researchers arrange to have 64 randomly selected fifth-grade schoolchildren take a standard school achievement test. The test is given in the usual way, with one exception. In their study, the test instructions include an additional sentence advising the test takers to answer each question with the first response that comes to mind. When given in the usual way (that is, without the extra sentence in the instructions), the test is known to have a mean of 200, a standard deviation of 48, and an approximately normal distribution, which is shown in Figure 7-8a.

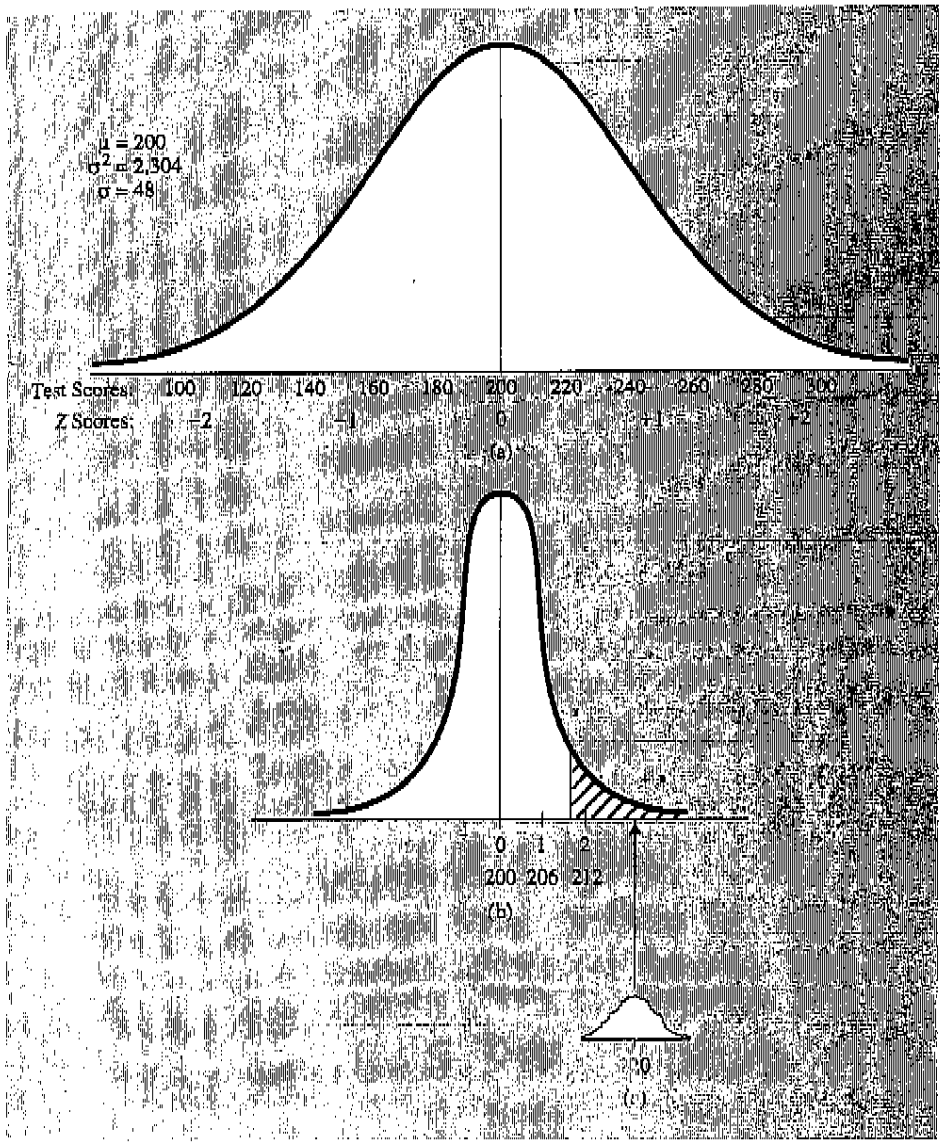

FIGURE 7-8
For the fictional study of performance on a standard school achievement test, (a) the distribution of the population of individuals, (b) the distribution of means (the comparison distribution), and (c) the sample's distribution

What kind of result should lead the educational psychologists to conclude that the procedure makes a difference?

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

Population 1: Fifth graders who get the special instructions
Population 2: Fifth graders 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 higher scores than the population of children who take the test in the normal way; $\mu_1 > \mu_2$. The null hypothesis is that Population 1's scores will not be higher than Population 2's; $\mu_1 \leq \mu_2$. (Note that these are directional hypotheses.)

2. Determine the characteristics of the comparison distribution. Our study gives us a mean of a sample of 64 scores (of fifth graders in this case). The comparison distribution has to be the distribution of means of samples of 64 scores each. This distribution will have a mean of 200 (the same as the population mean). Its variance will be the population variance divided by the number of individuals in the sample. The population variance is 2,304 (the population standard deviation of 48 squared) and the sample size is 64. Thus, the variance of the distribution of means will be 2,304/64, or 36. The standard deviation of the distribution of means is the square root of 36, or 6. Finally, since there are more than 30 individuals in the sample, the shape of the distribution of means will be approximately normal. Figure 7-8b shows this distribution of means.

3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected. Again, assume the researchers decide on the usual 5% significance level. The researchers in this study have a clear directional prediction and are not really interested in any effect in the opposite direction. (If the special instructions do not improve test scores, they would not be used in the future. Any possible results showing a negative effect are irrelevant.) Hence, the researchers will reject the null hypothesis if the result is in the top 5% of the comparison distribution. The comparison distribution (our 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 7-8b.

4. Determine your sample's score on the comparison distribution. The 64 fifth graders tested using the special instructions had a mean score of 220. (This sample's distribution is shown in Figure 7-8c.) A mean of 220 is 3.33 standard deviations above the mean of the distribution of means:

$$Z = \frac{(M - \mu_M)}{\sigma_M} = \frac{220 - 200}{6} = \frac{20}{6} = 3.33$$

5. Compare the scores from Steps 3 and 4 to decide whether to reject the null hypothesis. We set the minimum Z score needed to reject the null hypothesis to +1.64. The Z score of the sample's mean is +3.33. Thus, the educational psychologists can reject the null hypothesis and conclude that the research hypothesis is supported. To put this another way, the result is statistically significant at the $p < .05$ level. This can be seen in Figure 7-8b by noting how extreme the sample 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.

ESTIMATION AND CONFIDENCE INTERVALS

Hypothesis testing is our main focus in this book. However, there is another kind of statistical question related to the distribution of means that is sometimes important in psychology. This other kind of question is estimating an unknown population mean based on the scores in a sample. This is important, for example, in survey research. As we will see, it can also be important as an alternative approach to hypothesis testing.

Point Estimates and Interval Estimates

The best estimate of the population mean is the sample mean. In the study of fifth graders who got the special instructions, the mean score for the sample of 64 individuals tested was 220. Thus, 220 is the best estimate of the mean for the unknown population of fifth graders who might ever receive the special instructions. In this case, we are estimating the specific value of the population mean. Whenever we estimate a specific value of a population parameter, this is called a **point estimate.**

point estimate

You can also estimate a range of possible means that are likely to include the population mean. For example, you might estimate that a range of 200 to 240 includes the true population mean for fifth graders who get the special instructions.[4] This is called an **interval estimate.**

interval estimate

General Principle and Terminology of Confidence Intervals

The wider the interval estimate, the more confident you can be that it will include the true population mean. In our fifth-grader example, you might be quite sure that the range of 100 to 340 includes the true population mean. But you would be sticking your neck out if you estimated that the range of 219 to 221 includes the true population mean.

In general, you want an interval that is wide enough to ensure that it includes the population mean. This is called a **confidence interval** (sometimes abbreviated *CI*). If you want to be 95% sure, you want a **95% confidence interval.** The 95% confidence interval in the fifth-grader example is from 208.24 to 231.76. That is, based on the sample studied, you can be 95% sure that an interval from 208.24 to 231.76 includes the true population mean. (You will learn shortly how to figure this out yourself.) The upper and lower end of the confidence interval are called **confidence limits.** In this example, the confidence limits are 208.24 and 231.76.

confidence interval

95% confidence interval

confidence limits

If you want to be even more sure than 95%, you need a wider interval. In our example, the **99% confidence interval** has confidence limits of 204.58 and 235.42.

99% confidence interval

[4]In the mathematical logic of inferential statistics, we have to think of a population mean as something fixed. Confidence intervals can vary, but the population mean is fixed. Thus, we can say that we are 95% sure that our confidence interval includes the population mean. We should not say that the population mean has a 95% chance of being in the confidence interval.

Finding Confidence Limits

Confidence limits are based on the distribution of means. What you want to know is the points at which the middle 95% of means begin and end on this distribution. Thus, you need to find the cutoff points for the bottom 2.5% and the top 2.5%. This leaves a total of 95% in the middle. (For the 99% confidence interval, you would need to figure the scores that go with the top and bottom .5%, leaving 99% in between.)

Let's start with the lower limit. As usual, it is easiest to think in terms of Z scores. The Z score for the bottom 2.5% on a normal curve is -1.96. (You would find this from the normal curve table.) The example has a mean of 220 and a standard deviation of the distribution of means of 6. Thus, on this distribution of means, a Z score of -1.96 is 208.24. (That is, we converted the Z score of -1.96 to the raw score of 208.24 using the usual procedure for converting a Z score to a raw score.)

Figuring the upper limit works the same way. The Z score for the top 2.5% is $+1.96$. This comes out to 231.76 on our distribution of means.

Steps for Figuring Confidence Intervals

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

1. Determine the characteristics of the distribution of means, using usual computation. However, note that we are interested in the distribution of means for the population that represents the sample we are studying (what we have called Population 1), not the distribution of means for the population to which we are comparing it (Population 2). The mean of the distribution of means is thus estimated as the mean of our sample. As for the variance, fortunately, we usually assume that the variance of the two populations will be the same. Consequently, we can use the known variance from the given population (Population 2) as the basis for computing the variance of the distribution of means for the population we are interested in (Population 1). (The variance of the distribution of means is based only on the variance of the population and the sample size. Thus, the variance of the distribution of means will be the same for both populations.)

2. Use the normal curve table to find the Z scores that go with the upper and lower percentage you want. For a 95% confidence interval, this is the Z score that goes with the top and bottom 2.5%. For a 99% confidence interval, this is the Z score for the top and bottom .5%.

3. Convert these Z scores to raw scores on your distribution of means. These are the upper and lower confidence limits.

Another Example of Computing the Confidence Interval

As another example, let's compute the confidence interval for the ambiguous sentence example for those reading the sentences with a context. In that example, the 40 individuals tested in this way had a mean reading time of 2.71 seconds, and we knew from past research that the population of individuals reading ambiguous sentences without any special context had a variance of .02 seconds. With this information we are prepared to compute our confidence interval.

1. Determine the characteristics of the distribution of means. Its mean will be 2.71 seconds. We assume that the population of individuals tested reading ambiguous sentences with a context will have the same shape and variance as the population reading without a context ($\sigma^2 = .02$). Thus, the distribution of means will be normal and have a variance equal to .02/40, or .0005. The standard deviation is the square root of this, .022. (Note that this comes out to be the same standard deviation of the distribution of means we computed earlier, when doing hypothesis testing and focusing on the distribution of means for the population who read the sentences without a context.)

2. Use the normal curve table to find the Z scores that go with the upper and lower percentage you want. Suppose we want the usual 95% confidence interval, the Z score that goes with the top and bottom 2.5%. As we saw earlier, this comes out to ± 1.96.

3. Convert these Z scores to raw scores on your distribution of means. With a mean of 2.71 and a standard deviation of .022, a Z score of -1.96 is equal to a raw score of $2.71 - (.022 \times 1.96)$, which comes out to 2.667 Similarly, a Z score of $+1.96$ is equal to a raw score of $2.71 + (.022 \times 1.96)$, which comes out to 2.753. Thus, the 95% confidence limits are 2.667 to 2.753. This means that based on the results of the study, we are 95% confident that the true mean reading time for ambiguous sentences presented with a context is between 2.667 and 2.753 seconds.

The Subtle Logic of Confidence Intervals

The logic of confidence intervals is a little more subtle than might appear at first glance. This subtlety has to do with confidence intervals being estimates based only on information about a sample. That is, just as with hypothesis testing, confidence intervals involve statistical inference about a population based on information from a sample.

To make the fine points of the logic involved clear, it helps to imagine that we somehow know about the true population's mean. For example, suppose you somehow know that the population of fifth graders getting special instructions (Population 1 in the examples above) has a mean of 210. (Our focus now is on the population about which we are making estimates based on the sample. Don't confuse this population with Population 2, the one we knew about from the beginning in which the fifth graders did not get special instructions.) For such a population, a 95% confidence interval would be around its mean of 210. Using our computation procedure, you can compute the probability is 95% that any sample mean will be between 198.24 and 221.76 (see Figure 7-9, interval a).

If, in fact, the true population mean were 210, it would not have been surprising for the educational researchers to come up with a mean of 220 when studying a sample of 64 students. Such a sample mean would be well within the 95% limits on this distribution of means.

However, in the usual research situation we do not know the mean of the population we are studying. In reality, the educational researchers would have no way of knowing that the true population mean of fifth graders getting the special instructions is 210. All they know is the mean of their particular sample. Still, they can use their sample's mean as an estimate of the population mean. Based on that estimate they can compute a 95% confidence interval, which we have determined would go from 208.24 and 231.76. Sure

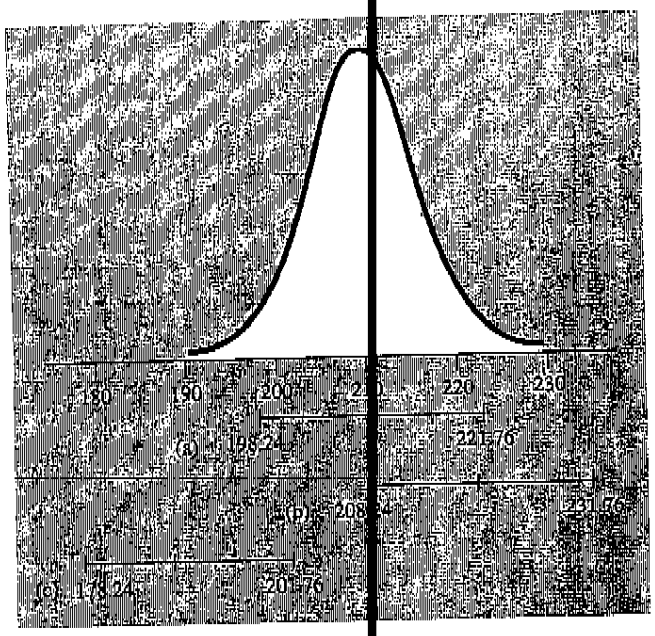

FIGURE 7-9
Examples of 95% confidence intervals displayed against distributions of means based on a (a) known population mean of 210, (b) sample mean of 220, and (c) sample mean of 190.

enough, this confidence interval would contain the true population mean, so the confidence in our interval would be justified. (See Figure 7-9, interval b.)

However, suppose the educational researchers had done the study and found that their sample had a mean of 190. We are still assuming the true mean for this population (unknown to the researchers) is 210, and that 95% of the time samples of 64 from this population should fall between 198.24 and 221.76. Thus, getting a sample with a mean of 190 is quite unlikely; certainly less likely than 5%, but possible. In fact, we expect that 5% of the time our samples will have means outside the 95% interval.

Suppose the educational psychologists go ahead and compute their confidence interval, using their sample's mean of 190 as their estimate of the population mean? Following the usual computation rules, they will get a 95% confidence interval of 178.24 to 201.76. Thus, they would compute a confidence interval that does not include the true population mean. (See Figure 7-9, interval c.)

In sum, when the estimated mean is within the 95% limits of the true population mean, your confidence interval will include the true mean. Fortunately, 95% of the time, the estimated mean is within the 95% limits of the true population mean. Thus, 5% of the time, a confidence interval based on the estimated mean will not include the true mean.

In other words, 95% of the time when you compute a confidence interval, it will include the true mean; 5% of the time it won't. This is why we say we are 95% confident that our interval includes the true mean. However, we can never know for sure that we are in the 95% or the 5% situation. There is always a 5% chance the true mean is not included in this region at all.

Confidence Intervals and Hypothesis Testing

In addition to their value in estimating the population mean, 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. This is because we are 95% confident that our interval includes the true population mean. If this 95% range does not include the Population 2 mean, then there is less than a 5% chance that this sample could have come from Population 2.

In the example of the special instructions for taking the achievement test, the 95% confidence interval of 208.24 to 231.76 does not include the mean of 200 for the population of fifth graders that takes the test without the special instructions. This is consistent with our conclusion earlier in the chapter that the result was significant using the .05 level. In the ambiguous sentences example, the 95% confidence interval for those reading the sentences with a context was 2.667 to 2.753 seconds. This interval does include the mean reading time (2.75) for the population reading the sentences without a context. Thus, just as we concluded when using the hypothesis-testing procedure, the result is not significant at the .05 level.

CONTROVERSIES AND LIMITATIONS: CONFIDENCE INTERVALS OR SIGNIFICANCE TESTS?

You may recall from Chapter 6 that currently there is a lively debate among psychologists about significance testing. Among the major issues in that debate is a proposal that psychologists use confidence intervals instead of significance tests.

Those who favor replacing significance tests with confidence intervals (e.g., Cohen, 1994; Hunter, 1997; Schmidt, 1996) cite several major advantages. First, as we noted above, confidence intervals contain all the key information in a significance test,[5] but also give additional information—the estimation of the range of values that we can be quite confident include the true population mean. A second advantage is that they focus attention on estimation instead of hypothesis testing. Some researchers argue that the goal of science is to provide numeric estimates of effects, not just decisions as to whether an effect is different from zero. That is, with estimation (point and interval estimates), you have a clear idea of how big the effect is and how accurate you are about that estimate. With hypothesis testing you know whether the effect is likely to be in the predicted direction, but not how big this effect is in that direction.

Confidence intervals are particularly valuable when the results are not significant (Frick, 1995), because knowing the confidence interval gives you an idea of just how far from no effect the true mean is likely to be found. If the entire confidence interval is near to no effect, we can feel confident that even if there is some true effect, it is probably small. For example, suppose a group of people is tested after receiving a procedure that claims to affect IQ. The mean for the group is 102 and the confidence interval is 99 to 105. This would be a nonsignificant finding because it includes 100, the mean IQ of the

[5]Some proponents of confidence intervals over significance testing argue that we should ignore the link with hypothesis testing. This is the most radical anti-significance-test position. That is, these psychologists argue that our entire focus should be on estimation, and significance testing of any kind should be irrelevant. In Chapter 8, we will discuss the rationale for their position, along with the counter arguments.

population who do not receive the special procedure. At the same time, since the confidence interval includes numbers other than 100, it is certainly possible that there is a real effect. Nevertheless, the key point is that if there is a real effect it is likely to be very small, since we are 95% confident it won't be more than a 1-point decrease or a 5-point increase. On the other hand, suppose the confidence interval for this same study was 89 to 115. This result would also be nonsignificant (because it includes 100). However, it would tell us that the study is really very inconclusive: It is possible that there is little or no effect (that the population mean of those who receive the procedure is around 100), but it is also possible that there is a substantial effect (that the true population mean for those who receive the procedure involves a decrease of as much as 11 IQ points or an increase of as much as 15 IQ points).

A third advantage claimed by proponents of confidence intervals over significance testing is that researchers are less likely to misuse them. As we noted in Chapter 6, a prevalent error in the use of significance tests is to conclude that a nonsignificant result means there is no effect. With confidence intervals it is harder to fall into this kind of error. The confidence interval with a nonsignificant result will include the mean expected for no effect. However, it will also include other possible values. Thus, we are reminded that the true population mean may very well be other than the no effect mean.

In spite of these apparent advantages, it is extremely rare to find confidence intervals in most types of psychology research articles. In part, this is probably due to tradition and to most psychologists having been trained with and more used to significance tests. Confidence intervals also require more description in a research article. For example, consider a larger table of results. It is easy to put in a star for each number to show its significance, and it is easy to read such a table. With confidence intervals, instead of a star, you would need two extra numbers (the upper and lower confidence limit) for each result.

Other psychologists (e.g., Abelson, 1997; Harris, 1997) note two reasons for not entirely abandoning significance testing in favor of confidence intervals. First, for some advanced statistical procedures, it is possible to do significance testing but not to compute confidence intervals. Second, just as it is possible to make mistakes with significance tests, it is also possible to make other kinds of mistakes with confidence intervals—especially since most research psychologists are less experienced in using them.

Finally, the issue of confidence intervals versus significance is rooted in a larger controversy of estimation versus hypothesis testing, a controversy we will discuss in Chapter 8. However, to preview that discussion, we can point out now that confidence intervals often make the most sense in applied research situations, while significance testing often makes the most sense in more theoretically oriented research.

Whatever the outcome of this controversy about confidence intervals, it is important to understand them since you will run into them occasionally when reading research literature, and it is possible you will see them more often in the future. On the other hand, they now appear infrequently. For this reason (and to keep the amount of material to be learned manageable), we decided not to emphasize confidence intervals in subsequent chapters of this book, which are mainly on significance testing in various types of research situations.

STANDARD DEVIATION OF THE DISTRIBUTION OF SAMPLE MEANS, HYPOTHESIS TESTS ABOUT MEANS OF SAMPLES, AND CONFIDENCE INTERVALS AS DESCRIBED IN RESEARCH ARTICLES

As we have noted several times, research in which there is a known population mean and standard deviation is quite rare in psychology. We have asked you to learn about this situation mainly because it is a building block for understanding hypothesis testing in common research situations. In the rare case in which research with a known population distribution is conducted, it is often described as involving a **Z test,** because it is the Z score that is checked against the normal curve.

Z test

Here is an example. As part of a larger study, Wiseman (1997) administered 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 different from a known population distribution based on a large U.S. study of university students that had been conducted earlier by Russell et al. (1980). Wiseman reported:

> . . . the 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 they studied (the Israeli group) and the population (the data from Russell). However, in the steps of computing each Z (the sample's score on the distribution of means), they would only have used the standard deviation for 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 indicate 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.* For example, Foertsch and Gernsbacher (1997) conducted a study to examine the effect of using the pronoun *they* to avoid fixing the gender of the person referred to, though this use is traditionally considered grammatically improper. Foertsch and Gernsbacher hypothesized that using *they* in this way would not have much effect on reading time. Consider the sentence "A truck driver should never drive while sleepy, even if she may be struggling to make a delivery on time, because many accidents are caused by drivers who fall asleep at the wheel." In their study, the researchers measured the reading time for this version of the sentence, as well as for two other versions, one substituting *he* for *she* and one substituting *they* for *she.* In this sentence, the antecedent (the first clause) was about a truck driver, a stereotypically masculine profession. In other sentences they used, the antecedents were stereotypically feminine (a nurse) or neutral (a runner). Here are some of their results:

> For masculine antecedents, *she* clauses ($M = 59.5$, $SE = 2.05$) were read significantly more slowly than *he* clauses ($M = 54.8$, $SE = 1.77$) or *they* clauses ($M =$

> 55.3, $SE = 1.77$) . . . For feminine antecedents, *he* clauses ($M = 58.7$, $SE = 1.66$) were read significantly more slowly than either *she* clauses ($M = 52.9$, $SE = 1.64$) or *they* clauses ($M = 52.7$, $SE = 1.67$) . . ." (p. 108)

This report gives you the pattern of means and a clear idea of the accuracy of these means as estimates of the population means. Consider the implications of the first standard error (2.05). Knowing this tells us that the mean reading time of sentences with masculine antecedents for *she* clauses is more than 2 standard errors higher than the reading time for either *he* or *they* clauses.

When researchers report the standard error of a result, they also give you information to figure the confidence interval. For example, let's assume a normal curve and figure the 95% confidence interval for masculine antecedents sentences with *she* clauses. Since the *SE* (which is another name for the standard deviation of the distribution of means) is 2.05, the upper 95% confidence limit is the mean plus the result of 1.96 times 2.05. This comes out to 59.5 plus 1.96×2.05, which is 63.52. The lower limit is 55.48. Thus we are 95% confident that the interval from 55.48 to 63.52 includes the true population mean.

Standard errors are also often shown in research articles as lines that go above the tops of the bars in a bar graph. These lines that go above the main bars are themselves called *standard error bars*. For example, Figure 7-10, reproduced from Foertsch and Gernsbacher's article, shows the same results we listed above (plus some additional ones).

As we noted, confidence intervals are rarely reported directly in research articles in psychology, though they may become more common in the future. Here is one example we were able to locate in the current literature. Chiu, Hong, and Dweck (1997) conducted a study that focused on the tendency of some individuals to believe that people's traits are fixed; Chiu et al. labeled these individuals as "entity theorists" because they see other people as fixed entities. In particular, the researchers wanted to test the theory that entity theorists would be more likely to take a single event as evidence that the person has a fixed trait. As part of the study, they described to the research participants a situation in which one person behaved in a more friendly way than

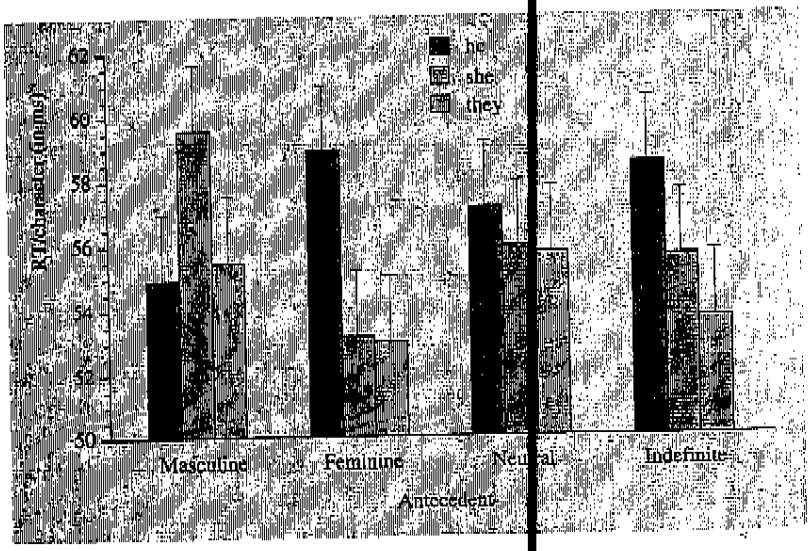

FIGURE 7-10
Effects of antecedent type (masculine, feminine, neutral, or indefinite) and pronoun (*he*, *she*, or *they*) on per-character reading time (RT) when sentences were used nonreferentially (Experiment 1). [From Foertsch, J., & Gernsbacher, M. A. (1997), fig. 1. In search of gender neutrality: Is singular *they* a cognitively efficient substitute for generic *he*? *Psychological Science*, 8, 108. Copyright, 1997, by the American Psychological Society. Reprinted with permission.]

another, and then asked them which person would be more likely to be friendly in the future.

Chiu et al. reported one of their findings about entity theorists as follows: "For them, if one person was found to be friendlier than another person in a particular situation, this relation would more likely than not generalize to a completely different situation" (p. 23). The statistical support for this was described as follows: "Entity theorists' overall prediction [of how likely the person was to be friendly] was significantly greater than .50 (95% CI = .5583 ± .0348)" (p. 23). This means that we can be 95% confident that in the population the true rating would be somewhere between .5235 and .5931, all figures above the .50 you would expect if the entity theorists were choosing at chance levels. By contrast, Chiu et al. found that individuals who were not entity theorists predicted significantly less than .50—with a confidence interval from .3648 to .4902.

SUMMARY

When studying a sample of more than one individual, the comparison distribution in the hypothesis-testing process is a distribution of means of all possible samples of the number of scores being studied. It can be thought of as describing what the result would be of (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.

The distribution of means has the same mean as the population of individuals. However, it has a smaller variance because the means of samples are less likely to be extreme than individuals' scores. (Extremes in any one sample are likely to be balanced by middle scores or extremes in the other direction.) Specifically, its variance 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 population of individuals follows a normal curve or (b) the samples are each of 30 or more scores.

Hypothesis tests involving a single sample of more than one individual and a known population are conducted in the same way as the hypothesis tests of Chapter 6 (where the studies were of a single individual compared to a population of individuals). The main exception is that the comparison distribution is a distribution of means.

The best point estimate for the population mean is the sample mean. You can determine an interval estimate of the population mean based on the distribution of means. When 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 95% confidence interval is an estimate of a range of values which you are 95% confident includes the true population mean.

An aspect of the current debate surrounding significance tests is whether researchers should replace them with confidence intervals. Proponents of confidence intervals argue that they provide additional information, put the focus on estimation, and reduce misuses common with significance tests. However, confidence intervals are rarely used in psychology research articles, in part

due to tradition and unfamiliarity with them, and the awkwardness of describing them. In addition, opponents of relying exclusively on confidence intervals argue that they cannot be used in some advanced procedures, estimation is not always the goal, and they are subject to misuses of their own.

The kind of hypothesis test described in this chapter is seldom used in research practice. (You have learned it as a stepping stone.) The standard deviation of the distribution of means, often referred to as the "standard error" (SE), is occasionally 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 below the top of each bar.

Key Terms

confidence interval (CI)
confidence limits
distribution of means
interval estimate
mean of a distribution of means (μ_M)
95% confidence interval
99% confidence interval
point estimate
shape of the distribution of means
standard deviation of a distribution of means (σ_M)
standard error of the mean (SE)
variance of a distribution of means (σ_M^2)
Z test

Practice Problems

These problems involve computation (with the assistance of a calculator). Most real-life statistics problems are done on a computer. But even if you have a computer, do these by hand to ingrain the method in your mind.

For practice in using a computer to solve statistical problems, refer to the computer section of each chapter of the *Student's Study Guide and Computer Workbook* that accompanies this text.

All data are fictional (unless an actual citation is given).

Answers to Set I problems are given at the back of the book.

SET I

1. Explain why the standard deviation of the distribution of means is generally smaller than the standard deviation of the distribution of the population of individuals.

2. For a population of individuals that has a standard deviation of 10, what is the standard deviation of the distribution of means for samples of size (a) 2, (b) 3, (c) 4, (d) 5, (e) 10, (f) 20, and (g) 100?

3. For each of the examples in Problem 2, compute the 95% confidence interval (that is, the upper and lower confidence limits). 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.

4. A particular population of individuals has a mean of 40, a standard deviation of 6, and follows a normal curve. For each of the following samples indicate whether it would be less likely than 5% to be randomly selected from this population: (a) a sample of 10 with a mean of 44, (b) a sample of 1 with a mean of 48, (c) a sample of 81 with a mean of 42, and (d) a sample of 36 with a mean of 42. For each part, (1) show the calculations for how you arrived at your answer and (2) include a diagram of the distributions involved.

5. 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. 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 efficacy of the course? (a) Carry out the steps of hypothesis testing (use the .01 level). (b) Compute the 99% confidence interval. (c) Explain your answer to someone who is familiar with the general logic of hypothesis testing, the normal curve, Z scores, and probability but is not familiar with the idea of a distribution of means or confidence intervals.

6. A large number of people viewed 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 *not at fault* = 0 to *completely at fault* = 10. The distribution of ratings under ordinary conditions follows a normal curve, $\mu = 5.5$, and $\sigma = .8$. Sixteen randomly selected individuals are tested in a condition in which the wording of the question is changed. In the changed condition, 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,