Chapter Outline

At this point, you may think you know all about hypothesis testing. Here's a surprise: What you know will not help you much as a psychologist. Why? The procedures for testing hypotheses described up to now were, of course, absolutely necessary prerequisites for what you will now learn. However, these procedures involved comparing a group of scores to a known population. In real research practice, you are often comparing two or more groups of scores to each other, without any direct information about populations. For example, you may have two scores for each of several people, such as scores on an anxiety test before and after psychotherapy or number of familiar versus unfamiliar words recalled in a memory experiment. Or you might have one score per person for two groups of people, such as an experimental group and a control group in a study of the effect of sleep loss on problem solving.

These kinds of research situations are among the most common in psychology, where the only information available is from the samples. Nothing is known about the populations that the samples come from. In particular, the researcher does not know the variance of the populations involved, which is a crucial ingredient in Step 2 of the hypothesis-testing process (determining the characteristics of the comparison distribution).

In this chapter, we first consider the solution to the problem of not knowing the population variance. We begin with a special hypothesis-testing situation, comparing the mean of a single sample to a population with a known mean but an unknown variance. Then, having seen how this problem of not knowing the population variance is handled, we go on to consider the situation in which there is no known population at all—the situation in which all we have are two scores for each of a number of people.

t tests

The hypothesis-testing procedures you learn in this chapter, in which the population variance is unknown, are examples of what are called ***t* tests.** The *t* test is sometimes called "Student's *t*" because its main principles were originally developed by William S. Gosset, who published his articles under the name "Student" (see Box 9-1).

Box 9-1
William S. Gosset, Alias "Student": Not a Mathematician, but a "Practical Man"

William S. Gosset graduated from Oxford in 1899 with a degree in mathematics and chemistry. It happened that in the same year the Guinness brewers in Dublin, Ireland, were seeking a few young scientists to take a first-ever scientific look at beer making. Gosset took one of these jobs and soon had immersed himself in barley, hops, and vats of brew.

The problem was how to make beer less variable, and especially to find the cause of bad batches. A proper scientist would say, "Conduct experiments!" But a business such as a brewery could not afford to waste money on experiments involving large numbers of vats, some of which any brewer worth his hops knew would fail. So Gosset was forced to contemplate the probability of, say, a certain strain of barley producing terrible beer when the experiment could consist of only a few batches of each strain. Adding to the problem was that he had no idea of the variability of a given strain of barley—perhaps some fields planted with the same strain grew better barley. (Does this sound familiar? Poor Gosset, like today's psychologists, had no idea of his population's variance.)

Gosset was up to the task, although at the time only he knew that. To his colleagues at the brewery, he was a professor of mathematics and not a proper brewer at all. To his statistical colleagues, mainly at the Biometric Laboratory at University College in London, he was a mere brewer and not a proper mathematician. In short, Gosset was the sort of scientist who was not above applying his talents to real life.

In fact, he seemed to revel in real life: raising pears, fishing, golfing, building boats, skiing, cycling (and lawn bowling, after he broke his leg by driving his car, a two-seater Model T Ford that he called "The Flying Bedstead," into a lamppost). And especially he reveled in simple tools that could be applied to anything, simple formulas that he could compute in his head. (A friend described him as an expert carpenter but claimed that Gosset did almost all of his finer woodwork with nothing but a penknife!)

So Gosset discovered the *t* distribution and invented the *t* test—simplicity itself (compared to most of statistics)—for situations when samples are small and the variability of the larger population is unknown. Most of his work was done on the backs of envelopes, with plenty of minor errors in arithmetic that he had to weed out later. Characteristically, he published his paper on his "brewery methods" only when editors of scientific journals demanded it. To this day, most statisticians call the *t* distribution "Student's *t*" because Gosset wrote under the anonymous name "Student" so that the Guinness brewery would not have to admit publicly that it sometimes brewed a bad batch!

References: Peters (1987); Stigler (1986); Tankard (1984).

INTRODUCTION TO THE *t* TEST: THE *t* TEST FOR A SINGLE SAMPLE

We begin with the following situation: You have scores for a single sample and you want to compare this to a population for which you know the mean but not the variance. Hypothesis testing in this situation is called a ***t* test for a single sample.** (It is also called a "one-sample *t* test.") The *t* test for a single sample works basically the same way as you learned in Chapter 7. There are only two important new wrinkles: First, since you don't know the population variance, you have to estimate it. Second, when you have to estimate the population variance, the shape of the comparison distribution is slightly different from a normal curve.

t test for a single sample

An Example

Suppose that your college newspaper reports an informal survey showing that students at your school study an average of 2.5 hours each day. However, you think that the students in *your* dormitory study much more than that. You randomly pick 16 students from your dormitory and ask them how much they study each day. (We will assume that they are all honest and accurate.) Your result is that these 16 students study an average of 3.2 hours per day. Should you conclude that the students in your dormitory study more than the college average? Or should you conclude that your results are so close to the college average that the small difference of .7 hours might well be due to your having accidentally picked 16 of the more studious residents in your dormitory?

Step 1 of the hypothesis-testing process is to restate the problem in terms of hypotheses about populations. There are two populations:

Population 1: The kind of students who live in your dormitory
Population 2: The kind of students at your college generally

The research hypothesis is that Population 1 students study more than Population 2 students; the null hypothesis is that Population 1 students do not study more than Population 2 students. So far the problem is no different from Chapter 7.

Step 2 is determining the characteristics of the comparison distribution. Its mean will be 2.5, the figure the survey found for students at your college generally (Population 2).

The next part of Step 2 is finding the variance of the distribution of means. With the current example, we face a new kind of problem. So far, you have always known the variance of the population of individuals. Using that variance, you then figured the variance of the distribution of means. In this example, the variance of number of hours studied for the college as a whole was not reported in the newspaper article. So you phone the paper. Unfortunately, the reporter did not calculate the variance, and the original survey results are no longer available. What to do?

Basic Principle of the *t* Test: Estimating the Population Variance From the Sample Scores

If you do not know the variance of the population of individuals, you can estimate it from what you do know: the scores of the people in your sample.

In the logic of hypothesis testing, the group of people we study are considered to be a random sample from a particular population. The variance of this sample ought to reflect the variance of that population. If the population has a lot of spread (there is a lot of variance in the scores), then a sample randomly selected from that population should have a lot of spread; if the population is very compact, with little spread, there should not be much spread in the sample either. Thus, it should be possible to use the spread of the scores in the sample to make an informed guess about the spread of the scores in the population. That is, we could compute the variance of the sample's scores and that should be similar to the variance of the scores in the population. (See Figure 9-1.)

There is, however, one small hitch. The variance of a sample will generally be slightly smaller than the variance of the population the sample comes from. For this reason, the variance of the sample is a **biased estimate** of the population variance.

biased estimate

Why is the sample's variance slightly smaller than the population's? The variance is based on deviations from the mean. A population's variance is based on deviations from that population's mean. However, the variance of a sample is based on deviation's from that sample's mean. A sample's mean is the optimal balance point for its scores. Thus, deviations of a sample's scores from its mean will be smaller than deviations from any other number. The mean of the sample generally is not exactly the same as the mean of the population it comes from. Consequently, deviations of a sample's scores from

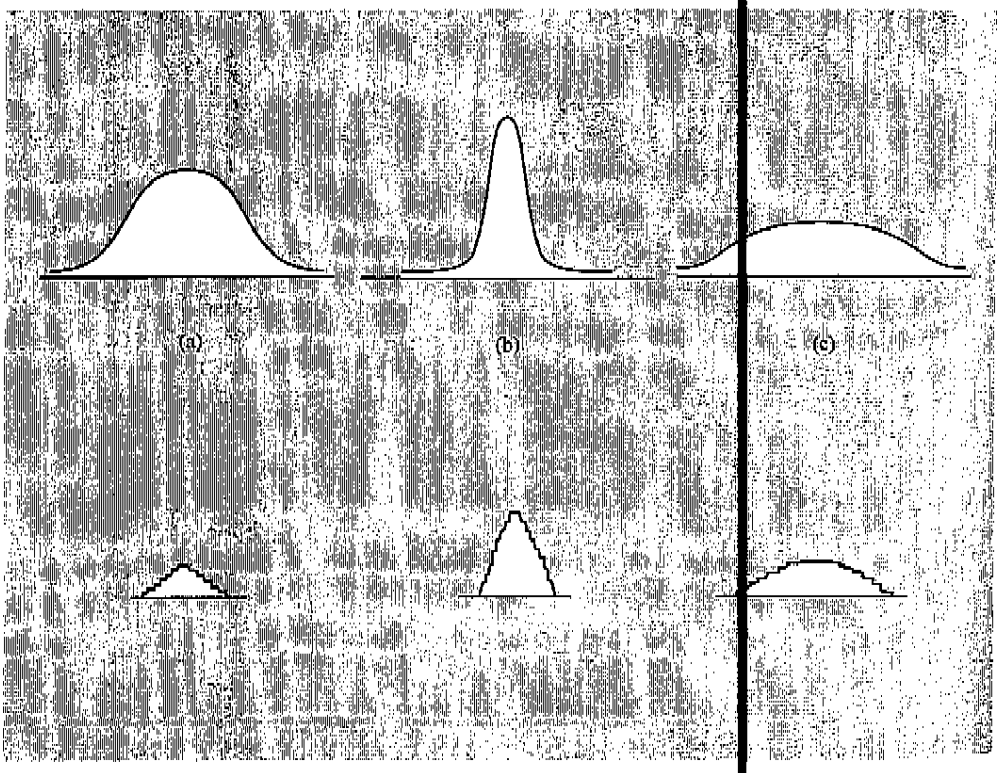

FIGURE 9-1
Variance in samples and the populations they are taken from.

the sample's mean will generally be smaller than deviations of that sample's scores from the population mean.

Suppose you knew the mean of the population the sample comes from and used this mean to compute the deviation for each score in the sample. The variance calculated in this way would be an **unbiased estimate of the population variance**.

unbiased estimate of the population variance

Unfortunately, you do not know the mean of the population the sample comes from. The sample comes from Population 1. In the present situation, you only know the mean of Population 2. But the means of the two populations are the same only if the null hypothesis is true—and that is what we are testing. (Regardless of whether the null hypothesis is true, we *do* assume that both populations have the same variance.)

Fortunately, you can compute an unbiased estimate of the population variance. What we do is make a correction in figuring the variance based on the sample scores that exactly accounts for the degree to which a sample's mean tends to vary from the true population mean. You compute this unbiased estimate by slightly changing the ordinary variance formula. The ordinary way to figure the variance is to take the sum of the squared deviation scores and divide this by the number of scores. In the changed procedure, you still take the sum of the squared deviation scores, but you divide this sum by the number of scores *minus 1*. Dividing by a slightly smaller number makes the result of dividing (the variance) slightly bigger.

It turns out that dividing by the number of scores minus 1 increases the resulting variance just enough to make it an unbiased estimate of the population variance. "Unbiased," incidentally, does not mean that your estimate will be exactly the true population variance. It only means that an estimate is equally likely to be too high as it is to be too low. (The biased estimate—the sample variance computed in the usual way—will be systematically too low.)

The symbol for the unbiased estimate of the population variance is S^2. The formula is the usual formula, but with the division by $N-1$ instead of by N:

$$S^2 = \frac{\Sigma(X-M)^2}{N-1} = \frac{SS}{N-1}, \tag{9-1}$$

The estimated population standard deviation is the square root of the estimated population variance:

$$S = \sqrt{S^2} \tag{9-2}$$

Let us return to our example of hours spent studying and compute the estimated population variance using the sample's 16 scores. First, we compute the sum of squared deviation scores. (Subtract the sample's mean from each of the scores, square those deviation scores, and add them.) Let us presume you do this and it comes out to 9.6 ($SS = 9.6$). To get the estimated population variance, you divide this sum of squared deviation scores by the number of scores in the sample minus 1. There are 16 in the sample, so the number in the sample minus 1 is 15. The result is .64. That is, 9.6/15 is .64. In terms of the formula,

$$S^2 = \frac{\Sigma(X-M)^2}{N-1} = \frac{SS}{N-1} = \frac{9.6}{16-1} = \frac{9.6}{15} = .64$$

Degrees of Freedom

degrees of freedom

The number you divide by (the number of scores minus 1) to figure the estimated population variance has a special name. It is called the **degrees of freedom** because it is the number of scores in a sample that is "free to vary." This is a somewhat complicated notion. The basic idea is that, when figuring the variance, you first have to know the mean. If you know the mean and all the scores in the sample but one, you can figure out the one you don't know with a little arithmetic. (If you are mathematically adventurous, try this out with some examples to see how it works.) Thus, once you know the mean, one of the scores in the sample is not free to have any possible value. So the degrees of freedom is the number of scores minus 1. In terms of a formula,

$$df = N - 1 \tag{9-3}$$

df

where ***df*** is the degrees of freedom. In our example, $df = 16 - 1 = 15$. (In some situations, which you will learn about in later chapters, the degrees of freedom are figured a bit differently. This is because in these situations the number of scores free to vary is different. For all the situations in this chapter, $df = N - 1$.)

The formula for computing the estimated population variance is often written using *df* instead of $N - 1$:

$$S^2 = \frac{\Sigma(X - M)^2}{df} = \frac{SS}{df} \tag{9-4}$$

Determining the Standard Deviation of the Distribution of Means From an Estimated Population Variance

Once you have estimated the population variance, computing the standard deviation of the comparison distribution involves the same procedures you learned in Chapter 7. That is, think of the comparison distribution as a distribution of means. As before, we can figure its variance as the variance of the population of individuals divided by the sample size. The only difference is that instead of knowing the variance of the population of individuals we have had to estimate it. As usual, the standard deviation of the distribution of means is the square root of its variance. Stated as formulas,

$$S_M^2 = \frac{S^2}{N} \tag{9-5}$$

$$S_M = \sqrt{S_M^2} \tag{9-6}$$

Note that when we are using an estimated population variance, the symbols for the variance and standard deviation of the distribution of means use S, instead of σ.

In our example, the sample size was 16, and the estimated population variance we just worked out was .64. The variance of the distribution of means,

based on that estimate, will be .04. That is, .64 divided by 16 equals .04. The standard deviation is .2, the square root of .04. In terms of the formulas,

$$S_M^2 = \frac{S^2}{N} = \frac{.64}{16} = .04$$

$$S_M = \sqrt{S_M^2} = \sqrt{.04} = .2$$

Be careful. To find the variance of a distribution of means, you always divide the population variance by the sample size. This is true whether the population variance is known or only estimated. In our example, you divided the population variance, which you had estimated, by 16. It is only when making the estimate of the population variance that you divide by the sample size minus 1. That is, the degrees of freedom are used only when estimating the variance of the population of individuals.

The Shape of the Comparison Distribution When Using an Estimated Population Variance: The *t* Distribution

In Chapter 7, we said that so long as it is reasonable to assume that the population distribution follows a normal curve, the shape of the distribution of means will also follow a normal curve. This changes when we are doing hypothesis testing using an estimated population variance. When we are using an estimated population variance, we have less true information and there is more room for error. The mathematical effect is that extreme means are slightly more likely than would be found in a normal curve. Further, the smaller your sample size, the bigger this tendency. This is because you are estimating the population variance on the basis of less information.

What is the result of all this when doing hypothesis testing using an estimated variance? The result is that the distribution of means (your comparison distribution) will not follow an exact normal curve. Instead, the comparison distribution follows a mathematically defined curve called a ***t* distribution.** ***t* distribution**

Actually, there are many *t* distributions. They vary in shape according to the degrees of freedom for the sample used in estimating the population variance. (However, for any particular degrees of freedom, there is only one *t* distribution.) Generally, all *t* distributions look to the eye like a normal curve—bell-shaped, completely symmetrical, and unimodal. A *t* distribution differs subtly in having heavier tails (that is, slightly more scores at the extremes). Figure 9-2 shows the shape of a *t* distribution compared to a normal curve.

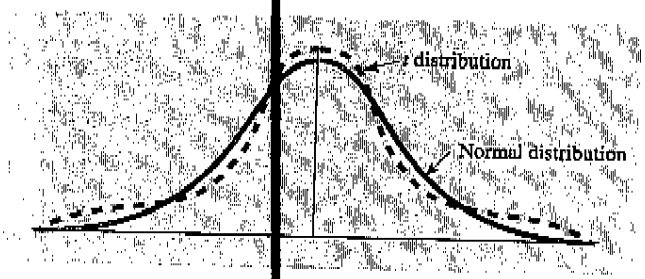

FIGURE 9-2
The *t* distribution compared to the normal curve.

This subtle difference in shape affects how extreme a score you need to reject the null hypothesis. To reject the null hypothesis you need to be in an extreme section of the normal curve, such as the top 5%. However, if there are more extreme scores, the point where the top 5% begins is further out on the curve. Thus, it takes a more extreme sample mean to get significance when using a t distribution than when using a normal curve.

Just how much the t distribution differs from the normal curve depends on the degrees of freedom in estimating the population variance. The t distribution differs most from a normal curve when the estimate of the population variance is based on a very small sample, so that the degrees of freedom are low. For example, using the normal curve, the cutoff for a one-tailed test at the .05 level is 1.64. On a t distribution with 7 degrees of freedom (that is, with a sample size of 8), the one-tailed, 5% cutoff is 1.895. If the population variance estimate is based on a larger sample, say a sample of 25 (so that $df = 24$), the cutoff is 1.711. If your sample size is infinite, the t distribution is the same as the normal curve. (Of course, if your sample size were infinite, it would include the entire population!) But even with sample sizes of 30 or more, the t distribution is nearly identical to the normal curve.

Before going on to learn how you actually find the cutoff using a t distribution, let us first return briefly to our example of the number of hours that students at your dorm study each night. We finally have everything we need to complete Step 2 about the characteristics of the comparison distribution. We have already seen that the distribution of means will have a mean of 2.5 hours and a standard deviation of .2. Based on what we have just discussed, we can now add that the shape of the comparison distribution will be a t distribution with 15 degrees of freedom.[1]

Determining the Cutoff Sample Score for Rejecting the Null Hypothesis: Using the *t* Table

Step 3 of the hypothesis-testing process is determining the cutoff for rejecting the null hypothesis. There is a different t distribution for any particular number of degrees of freedom. However, to avoid taking up pages and pages with tables for each possible t distribution, a simplified table is used that gives only the crucial cutoff points. We have included such a ***t* table** in Appendix B (Table B-2).

t table

[1]Statisticians make a subtle distinction in this situation between the comparison distribution and the distribution of means. We have avoided this distinction here and in later chapters in order to greatly simplify the discussion of what is already fairly difficult. If you are interested, the distinction can be understood as follows: The general procedure of hypothesis testing, as we introduced it in Chapter 7, can be described as comparing a Z score to your sample's mean, where $Z = (M - \mu)/\sigma_M$ and where $\sigma_M = \sqrt{\sigma^2/N}$, and then comparing this Z score to a cutoff Z score from the normal curve table. We described this process as using the distribution of means as your comparison distribution.

Statisticians would say that actually you are comparing your computed Z score to a distribution of Z scores (which is simply a standard normal curve). Similarly, in the case of a t test, statisticians think of the procedure as computing a t score (like a Z score but calculated using an estimated standard deviation)where $t = (M - \mu)/S_M$, where $S_M = \sqrt{S^2/N}$—and then comparing your computed t score to a cutoff t score from a t distribution table. Thus, according to the formal statistical logic, the comparison distribution is a distribution of t scores, not of means.

In the present example, you have a one-tailed test (you are interested in whether students in your dorm study *more* than students in general at your college). You will probably want to use the 5% significance level because the cost of a Type I error (mistakenly rejecting the null hypothesis) is not great. You have 16 participants, making 15 degrees of freedom for the estimate of the population variance.

Table 9-1 shows a portion of a *t* table like Table B-2. Find the column for the .05 significance level for one-tailed tests, then move down this column to the row for 15 degrees of freedom. The crucial cutoff number is 1.753. This means that you will reject the null hypothesis if your sample's mean is 1.753 or more standard deviations above the mean on the comparison distribution. (If you were using a known variance you would have found your cutoff from a normal curve table. The *Z* score needed to reject the null hypothesis based on the normal curve would have been 1.645.)

One other point about using the *t* table. In the full *t* table in the appendix, there are rows for each degree of freedom from 1 through 30, then for every five degrees of freedom (35, 40, 45, etc.) up to 100. Suppose your study involves degrees of freedom in between two values. To be safe, you should use the nearest degrees of freedom *below* yours that is given on the table. For example, if you were doing a study in which there were 43 degrees of freedom, you would use the row in the table for 40 *df*.

Determining the Score of the Sample Mean on the Comparison Distribution: The *t* Score

Step 4 of the hypothesis-testing process is determining your sample's score on the comparison distribution. In previous chapters, this has meant finding the *Z* score on the comparison distribution—the number of standard deviations it

TABLE 9-1
Cutoff Scores for *t* Distributions with 1 Through 17 Degrees of Freedom (Showing Cutoff for Hours Studied Example)

	One-Tailed Tests			Two-Tailed Tests		
df	*.10*	*.05*	*.01*	*.10*	*.05*	*.01*
1	3.078	6.314	31.821	6.314	12.706	63.657
2	1.886	2.920	6.965	2.920	4.303	9.925
3	1.638	2.353	4.541	2.353	3.182	5.841
4	1.533	2.132	3.747	2.132	2.776	4.604
5	1.476	2.015	3.365	2.015	2.571	4.032
6	1.440	1.943	3.143	1.943	2.447	3.708
7	1.415	1.895	2.998	1.895	2.365	3.500
8	1.397	1.860	2.897	1.860	2.306	3.356
9	1.383	1.833	2.822	1.833	2.262	3.250
10	1.372	1.813	2.764	1.813	2.228	3.170
11	1.364	1.796	2.718	1.796	2.201	3.106
12	1.356	1.783	2.681	1.783	2.179	3.055
13	1.350	1.771	2.651	1.771	2.161	3.013
14	1.345	1.762	2.625	1.762	2.145	2.977
15	1.341	1.753	2.603	1.753	2.132	2.947
16	1.337	1.746	2.584	1.746	2.120	2.921
17	1.334	1.740	2.567	1.740	2.110	2.898

is from the mean on the distribution of means. You do exactly the same thing when your comparison distribution is a *t* distribution. The only difference is that in the past, when the comparison distribution was a normal curve, the score we computed on it was called a *Z* score. Now, we are using a *t* distribution as our comparison distribution. Thus, the score we compute on it we call a ***t* score.** In terms of a formula,

t score

$$t = \frac{M - \mu}{S_M} \tag{9-7}$$

In the example, your sample's mean of 3.2 is .7 hours from the mean of the distribution of means. This amounts to 3.5 standard deviations from the mean (that is, .7 hours divided by the standard deviation of .2 hours equals 3.5). In other words, the *t* score in the example is 3.5. In terms of the formula,

$$t = \frac{M - \mu}{S_M} = \frac{3.2 - 2.5}{.2} = \frac{.7}{.2} = 3.5$$

Determining Whether to Reject the Null Hypothesis

Step 5 of hypothesis testing is comparing the scores from Steps 3 and 4 to decide whether to reject the null hypothesis. This step is exactly the same with a *t* test as it was in previous chapters. You compare the cutoff score from Step 3 with the sample's score on the comparison distribution from Step 4. In our example, the cutoff *t* score was 1.753 and the actual *t* score for our sample was 3.5. Conclusion: Reject the null hypothesis; the research hypothesis that students in your dorm study more than students in the rest of the college is supported.

Figure 9-3 shows the distributions for this example.

Summary of Hypothesis Testing When the Population Variance Is Not Known

Hypothesis testing when the population variance is not known is exactly the same as in Chapter 7, with four exceptions: (a) Instead of the population variance being known in advance, it is estimated from the sample (using the formula for the unbiased estimate, $S^2 = SS/df$); (b) instead of the comparison distribution following a normal curve, it is a *t* distribution with *df* equal to the number of scores in your sample minus 1; (c) instead of looking up the significance cutoff point on a normal curve table, you use a *t* table; and (d) the score of your sample on the comparison distribution, instead of being called a *Z* score, is called a *t* score. Table 9-2 systematically compares the two situations.

Another Example of a Single-Sample *t* Test

Consider another fictional example. Suppose a researcher was studying the psychological effects of a devastating flood in a small rural community. Specifically, the researcher was interested in whether people felt more or less hopeful after the flood. The researcher randomly selects 10 people to complete a short questionnaire. The key item on the questionnaire asks these

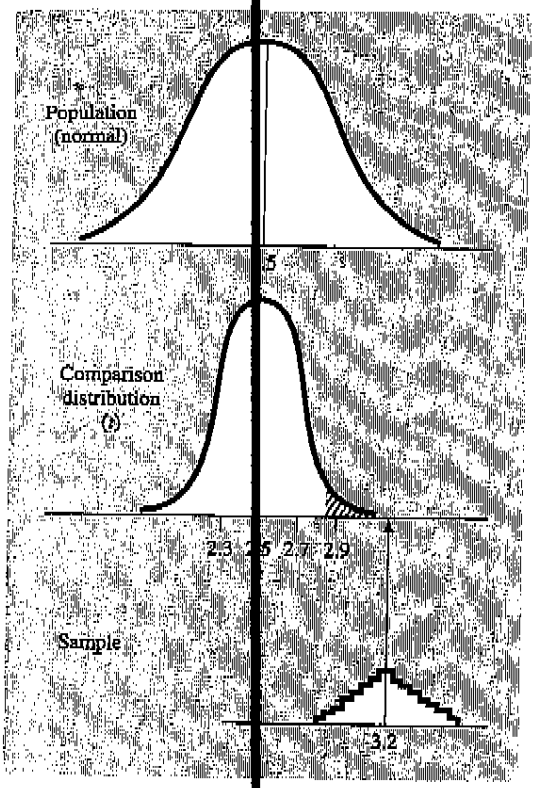

FIGURE 9-3
Distributions involved in the hours studied example.

TABLE 9-2
Hypothesis Testing Involving a Single Sample Mean When Population Variance Is Unknown (*t* Test) Compared to When Population Variance Is Known

Steps in Hypothesis Testing	Difference From When Population Variance Is Known
1. Restate the question as a research hypothesis and a null hypothesis about the populations.	No difference in method.
2. Determine the characteristics of the comparison distribution:	
Population mean	No difference in method.
Population variance	Estimate from the sample.
Standard deviation of the distribution of sample means	No difference in method (but based on estimated population variance).
Shape of the comparison distribution	Use the *t* distribution with $df = N - 1$.
3. Determine the significance cutoff.	Use the *t* table.
4. Determine your sample's score on the comparison distribution.	No difference in method (but called a *t* score).
5. Compare the scores in Steps 3 and 4 to decide whether to reject the null hypothesis.	No difference in method.

individuals to rate how hopeful they feel using a 7-point scale from *extremely unhopeful* (1) to *neutral* (4) to *extremely hopeful* (7). Table 9-3 shows the results and computation for the *t* test for a single sample; Figure 9-4 shows the distributions involved.

The researcher was interested in whether the responses would be consistently above or below the midpoint on the scale (4). Here are the steps of hypothesis testing.

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

Population 1: People who experienced the flood
Population 2: People who are neither hopeful nor unhopeful

The research hypothesis is that the two populations will score differently. The null hypothesis is that they will score the same.

2. Determine the characteristics of the comparison distribution. If the null hypothesis is true, the mean of both population distributions is 4. However, the variance of these population distributions is not known; it must be estimated from the sample. As shown in Table 9-3, the sum of the squared deviations from the sample's mean is 32.10. Thus, the estimated population variance is 3.57; that is, 32.10 divided by 9 degrees of freedom (10 – 1) equals 3.57.

TABLE 9-3
Data and Analysis for a Single-Sample *t* Test for a Study of 10 People's Ratings of Hopefulness Following a Devastating Flood (Fictional Data)

	Rating	Difference From the Mean	Squared Difference From the Mean
	(X)	$(X - M)$	$(X - M)^2$
	5	.3	.09
	3	–1.7	2.89
	6	1.3	1.69
	2	–2.7	7.29
	7	2.3	5.29
	6	1.3	1.69
	7	2.3	5.29
	4	– .7	.49
	2	–2.7	7.29
	5	.3	.09
Σ:	47	0	32.10

$M = \Sigma X/N = 47/10 = 4.7.$
$df = N - 1 = 10 - 1 = 9.$
$\mu = 4.0.$
$S^2 = SS/df = 32.10/(10 - 1) = 32.10/9 = 3.57.$
$S_M^2 = S^2/N = 3.57/10 = .36.$
$S_M = \sqrt{S_M^2} = \sqrt{.36} = .60.$
t with $df = 9$ needed for 1% significance level, two-tailed = ±3.250.
Actual sample $t = (M - \mu)/S_M = (4.7 - 4)/.6 = .7/.6 = 1.17.$

Decision: Do not reject the null hypothesis.

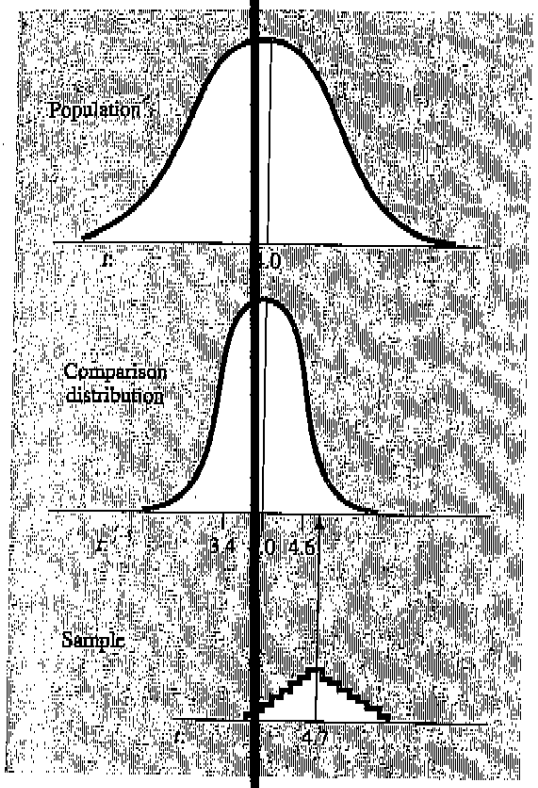

FIGURE 9-4
Distributions involved in the example of how hopeful individuals felt following a devastating flood.

The distribution of means will have a mean of 4 (the same as the population mean). Its variance is the estimated population variance divided by the sample size—3.57 divided by 10 equals .36. The square root of this, the standard deviation of the distribution of means, is .60.

3. Determine the cutoff sample score on the comparison distribution where the null hypothesis should be rejected. The researcher wants to be very cautious about mistakenly concluding that the flood made a difference. Thus, she decides to test the hypothesis at the .01 level. The hypothesis was nondirectional (that is, no specific direction of difference from the mean of 4 was specified; either result would have been of interest), so the researcher uses a two-tailed test. The researcher looks up the cutoff on Table 9-1 (or Table B-2 in Appendix B), for a two-tailed test and 9 degrees of freedom. The figure on the table is 3.250. Consequently, to reject the null hypothesis the researcher needs a t of 3.250 or higher or a t of –3.250 or lower.

4. Determine the sample's score on the comparison distribution. The sample's mean of 4.7 is .7 scale points from the null hypothesis mean of 4.0. That makes it 1.17 standard deviations on the comparison distribution from that distribution's mean (.7/.6 = 1.17); $t = 1.17$.

5. Compare the scores from Steps 3 and 4 to decide whether to reject the null hypothesis. The t of 1.17 is not as extreme as the needed t of ±3.250. Therefore, the researcher cannot reject the null hypothesis. The study is inconclusive. (If the researcher had used a larger sample, giving more power, the result might have been quite different.)

TABLE 9-4
Steps for Conducting a *t* Test for a Single Sample

1. Restate the question as a research hypothesis and a null hypothesis about the populations.
2. Determine the characteristics of the comparison distribution.
   a. The mean is the same as the known population mean.
   b. The standard deviation is computed as follows:
      i. Compute the estimated population variance: $S^2 = SS/df$.
      ii. Compute the variance of the distribution of means: $S_M^2 = S^2/N$.
      iii. Compute the standard deviation: $S_M = \sqrt{S_M^2}$.
   c. The shape will be a *t* distribution with $N - 1$ degrees of freedom.
3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.
   a. Determine the degrees of freedom, desired significance level, and number of tails in the test (one or two).
   b. Look up the appropriate cutoff in a *t* table.
4. Determine your sample's score on the comparison distribution: $t = (M - \mu)/S_M$.
5. Compare the scores from Steps 3 and 4 to decide whether or not to reject the null hypothesis.

Summary of Steps for Conducting a *t* Test for a Single Sample

Table 9-4 summarizes the steps of hypothesis testing when you have scores from a single sample and a population with a known mean but an unknown variance.

THE *t* TEST FOR DEPENDENT MEANS

So far we have considered examples where you know the population mean but not its variance. This type of research situation is fairly rare. Usually, you do not even know the population's mean! We turn now to one common research situation in which you know neither the population mean nor its variance. This kind of situation involves studies in which there are two scores for each of several people. For example, a psychophysiologist might measure the pattern of EEG activity ("brain waves"), comparing the EEG for each person while doing abstract tasks versus concrete tasks. This kind of research setup where each person is measured more than once is called a **repeated-measures design.** (It is also known as a "within-subjects design." See Appendix A for a summary of the major types of research designs.)

repeated-measures design

In one widely used repeated-measures design you measure the same individuals before and after some psychological or social intervention. For example, an organizational psychologist might measure days missed from work for 80 workers before and after a new health promotion program was introduced.

In this common situation of a repeated-measures design, where each person is measured twice, the hypothesis-testing procedure used is called a

***t* test for dependent means.** It has the name "dependent means" because the means for each group of scores (e.g., before scores and after scores) are dependent on each other in that they are both from the same people. (In Chapter 10, we consider the situation in which a researcher compares scores from two different groups of people, a research design analyzed by a "*t* test for independent means.")

t test for dependent means

The *t* test for dependent means is exactly the same as the *t* test for a single sample, except that (a) you use something called difference scores and (b) you assume that the population mean is 0. Let us turn now to each of these new features.

Difference Scores

With a repeated-measures design, our sample includes two scores for each person instead of just one. The way we handle this is to make the two scores per person into one score per person. We do this magic by creating **difference scores:** For each person you subtract one score from the other.

difference scores

Consider the EEG example. For each person the psychophysiologist would do a subtraction: the person's EEG measure during the abstract task minus the person's EEG measure during the concrete task. This gives a single abstract-minus-concrete difference score for each person. Similarly, consider the absence-from-work example. The organizational psychologist would do the following subtraction for each person: the number of days missed after the program minus the number of days missed before the program. This would result in an after-minus-before difference score for each employee.

When the two scores are a before score and an after score, we usually take the after score minus the before score. This gives a measure of change. In other situations, such as the EEG example, it really doesn't matter which you subtract from which—so long as you do it the same way for each person in the sample.

Once you have the difference score for each person in the study, you do all the rest of the hypothesis-testing procedure using the difference scores. That is, you treat the study as if there were a single sample of scores—scores which in this situation happen to be difference scores.[2]

Population of Difference Scores With a Mean of 0

So far in this book, you have always known the mean of Population 2 (the population you are contrasting your sample with). For example, in the college dormitory survey of hours studied, we knew that the mean of the population of students at the college overall was 2.5 hours. However, now we are using difference scores, and we usually do not know the mean of the population of difference scores.

[2]You can also use a *t* test for dependent means in a situation in which you have scores from pairs of research participants. You consider each pair as if it were one person and compute a difference score for each pair. For example, suppose you have 30 married couples and you are comparing ages of husbands and wives to see if husbands are consistently older than wives. You could compute for each couple a difference score of husband's age minus wife's age. You would then carry out the rest of the hypothesis testing the same way as any other *t* test for dependent means. When used in this way, the *t* test for dependent means is sometimes called a *t test for matched pairs* or a *paired t test.*

The solution is as follows: Ordinarily, the null hypothesis in a repeated-measures design is that there is no difference between the two groups of scores. For example, the null hypothesis in the psychophysiology study is that EEG activity will be the same when doing abstract or concrete tasks. Similarly, the null hypothesis in the health promotion study is that absences from work will be the same before and after the health promotion program is introduced. Thus, when using difference scores, we usually compare a research hypothesis of a predicted difference to a null hypothesis of no difference.

Here is the key point: What does "no difference" mean? That is, what does it mean to say that in the population there is on the average no difference between the two scores for each person? It is the same as saying the mean of the population of the difference scores is 0. In other words, saying that there is no difference between the two scores is equivalent to saying that the average of the difference scores is zero.

Therefore, when working with difference scores, we assume an artificial comparison population of difference scores that has a population mean of 0.

Example of a *t* Test for Dependent Means

Olthoff (1989) tested the communication quality of engaged couples 3 months before and again 3 months after marriage. One group studied was 19 couples who had received ordinary premarital counseling from the ministers who were going to marry them. (To keep the example simple, we will focus on just this one group, and on only the husbands in the group. Scores for

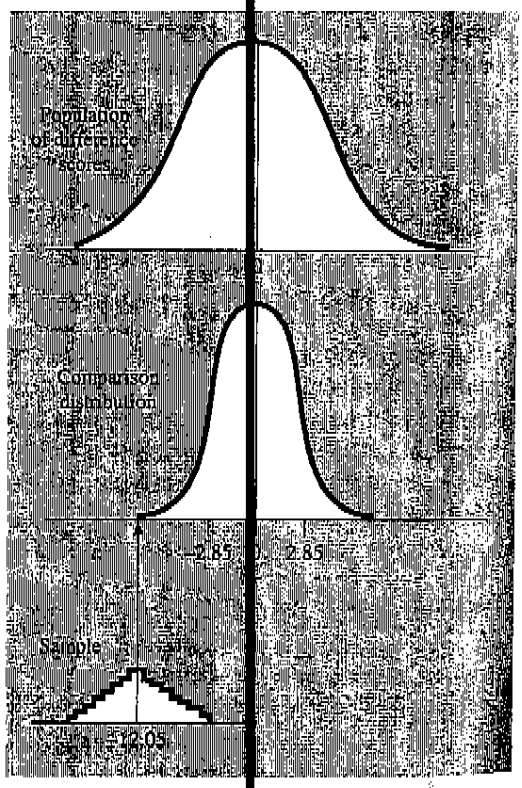

FIGURE 9-5
Distributions for the Olthoff (1993) example of a *t* test for dependent means.

wives were similar, though somewhat more varied, making it a more complicated example for learning the *t*-test procedure.)

The scores for the 19 husbands are listed in the "Before" and "After" columns in Table 9-5, followed by the entire *t*-test analysis. (The distributions involved are shown in Figure 9-5.) The mean of the before scores was 116.316 and the mean of the after scores was 104.263. More important, however, we have also figured the difference scores. The mean of the difference scores is −12.05. On the average, these husbands' communication quality decreased by about 12 points.

Is this decrease significant? In other words, how likely is it that this sample of change scores is a random sample from a population of change scores whose mean is 0? Let's carry out the hypothesis-testing procedure.

TABLE 9-5
***t* Test Analysis for Communication Quality Scores Before and After Marriage for 19 Husbands Who Received No Special Communication Training**

Husband	Communication Quality		Difference (After − Before)	Deviation of Differences From the Mean of Differences	Squared Deviation
	Before	*After*			
A	126	115	− 11	1.05	1.1
B	133	125	− 8	4.05	16.4
C	126	96	− 30	−17.95	322.2
D	115	115	0	12.05	145.2
E	108	119	11	23.05	531.3
F	109	82	− 27	−14.95	233.5
G	124	93	− 31	−18.95	359.1
H	98	109	11	23.05	531.3
I	95	72	− 23	−10.95	119.9
J	120	104	− 16	− 3.95	15.6
K	118	107	− 11	1.05	1.1
L	126	118	− 8	4.05	16.4
M	121	102	− 19	− 6.95	48.3
N	116	115	− 1	11.05	122.1
O	94	83	− 11	1.05	1.1
P	105	87	− 18	− 5.95	35.4
Q	123	121	− 2	10.05	101.0
R	125	100	− 25	−12.95	167.7
S	128	118	− 10	2.05	4.2
Σ:	2,210	1,981	−229		2,772.9

For difference scores:
$M = -229/19 = -12.05$.
$\mu = 0$ (assumed as a no-change baseline of comparison).
$S^2 = SS/df = 2{,}772.9/(19-1) = 154.05$.
$S_M^2 = S^2/N = 154.05/19 = 8.11$.
$S_M = \sqrt{S_M^2} = \sqrt{8.11} = 2.85$.
t with $df = 18$ needed for 5% level, two-tailed $= \pm 2.101$.
$t = (M - \mu)/S_M = (-12.05 - 0)/2.85 = -4.23$.

Decision: Reject the null hypothesis.

Note. Data from Olthoff (1989).

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

Population 1: Husbands who receive ordinary premarital counseling
Population 2: Husbands whose communication quality does not change from before to after marriage

The research hypothesis is that Population 1 is different from Population 2—that husbands who receive ordinary premarital counseling (such as the husbands Olthoff studied) *do* change in communication quality from before to after marriage. The null hypothesis is that the populations are the same—that the husbands who receive ordinary premarital counseling *do not* change in their communication quality from before to after marriage.

Notice that we have no actual information about Population 2 husbands. The husbands in the study are a sample of Population 1 husbands. If the research hypothesis is correct, Population 2 husbands may not even really exist. For the purposes of hypothesis testing, we set up Population 2 as a kind of straw man comparison group. That is, we set up a comparison group for purposes of the analysis of husbands who, if measured before and after marriage, would show no change.

2. Determine the characteristics of the comparison distribution. If the null hypothesis is true, the mean of the population of difference scores is 0. The variance of the population of difference scores can be estimated from the sample of difference scores. As shown in Table 9-5, the sum of squared deviations of the difference scores from the mean of the difference scores is 2,772.9. With 19 husbands in the study, there are 18 degrees of freedom. Dividing the sum of squared deviation scores by the degrees of freedom gives an estimated population variance of 154.05.

The distribution of means (from this population of difference scores) will have a mean of 0, the same as the population mean. Its variance will be the estimated population variance (154.05) divided by the sample size (19), which gives 8.11. The standard deviation is the square root of 8.11, which is 2.85. Because Olthoff was using an estimated population variance, the comparison distribution is a *t* distribution. The estimate of the population variance was based on 18 degrees of freedom, so this comparison distribution is a *t* distribution for 18 degrees of freedom.

3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected. Olthoff used a two-tailed test because there was no clear reason for predicting either an increase or a decrease in communication quality. Using the .05 significance level and 18 degrees of freedom, Table B-2 shows that to reject the null hypothesis you need a *t* score at or above +2.101 or at or below −2.101.

4. Determine the sample's score on the comparison distribution. Olthoff's sample had a mean difference score of −12.05. That is, the mean was 12.05 points below the mean of 0 on the distribution of means. The standard deviation of the distribution of means that we computed was 2.85. Thus, the mean of the difference scores of −12.05 is 4.23 standard deviations below the mean of the distribution of means. So Olthoff's sample of difference scores has a *t* score of −4.23.

5. Compare the scores from Steps 3 and 4 to decide whether to reject the null hypothesis. The *t* of −4.23 for the sample of difference scores is more extreme than the needed *t* of ±2.101. Thus, we can reject the null

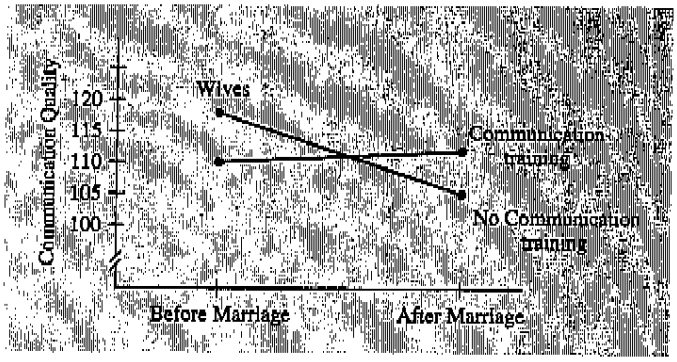

FIGURE 9-6
Communication skills of wives given premarital communications training and wives not given such training. (Based on Olthoff, 1989.)

hypothesis. This suggests that Olthoff's husbands are from a population in which husbands' communication quality is different after marriage from what it was before (it is lower).

Olthoff's actual study was more complex. You may be interested to know that they found that the wives also showed this decrease in communication quality after marriage. But a group of similar engaged couples who were given special communication-skills training by their ministers (much more than the usual short session) had no significant decline in marital communication quality after marriage (see Figure 9-6). In fact, there is now a great deal of research showing that marital quality of all kinds on the average declines (e.g., Karney & Bradbury, 1997) and that intensive communication skills training can be very helpful in reducing or eliminating this decline (Markman et al., 1993).

Another Example of a *t* Test for Dependent Means

Here is another example. A researcher is interested in the effect of noise on hand-eye coordination in surgeons. The researcher gives nine surgeons a standard test of hand-eye coordination under both quiet and noisy conditions—not while doing surgery, of course. The prediction is that surgeons' coordination is better under quiet conditions. (Ideally, any effects of practice or fatigue from taking the hand-eye coordination test twice would be equalized by testing half the surgeons under noisy conditions first, and half under quiet conditions first. See Appendix A for a discussion of such "counterbalancing.")

Table 9-6 shows the results for this fictional study. It also shows the calculation of difference scores and all the other calculations for the *t* test for dependent means. Figure 9-7 shows the distributions involved. Here are the steps of hypothesis testing:

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

Population 1: Surgeons like those tested in this study
Population 2: Surgeons whose coordination is the same under quiet and noisy conditions

The research hypothesis is that Population 1's mean difference scores (quiet minus noisy) is greater than Population 2's. That is, the research hypothesis is

TABLE 9-6
***t* Test for a Study of Hand-Eye Coordination in Which Nine Surgeons Are Measured Under Noisy and Quiet Conditions (Fictional Data)**

Surgeon	Conditions		Difference	Deviation	Squared Deviation
	Quiet	*Noisy*			
1	18	12	6	6 – 2 = 4	16
2	21	21	0	–2	4
3	19	16	3	1	1
4	21	16	5	3	9
5	17	19	–2	–4	16
6	20	19	1	–1	1
7	18	16	2	0	0
8	16	17	–1	–3	9
9	20	16	4	2	4
Σ:	170	152	18	0	60

For difference scores:
$M = 18/9 = 2.0$.
$\mu = 0$ (assumed as a no-change baseline of comparison).
$S^2 = SS/df = 60/(9 - 1) = 60/8 = 7.5$.
$S_M^2 = S^2/N = 7.50/9 = .83$.
$S_M = \sqrt{S_M^2} = \sqrt{.83} = .91$.
t for $df = 8$ needed for 1% significance level, one-tailed = 2.897.
$t = (M - \mu)/S_M = (2.00 - 0)/.91 = 2.20$.

Decision: Do not reject the null hypothesis.

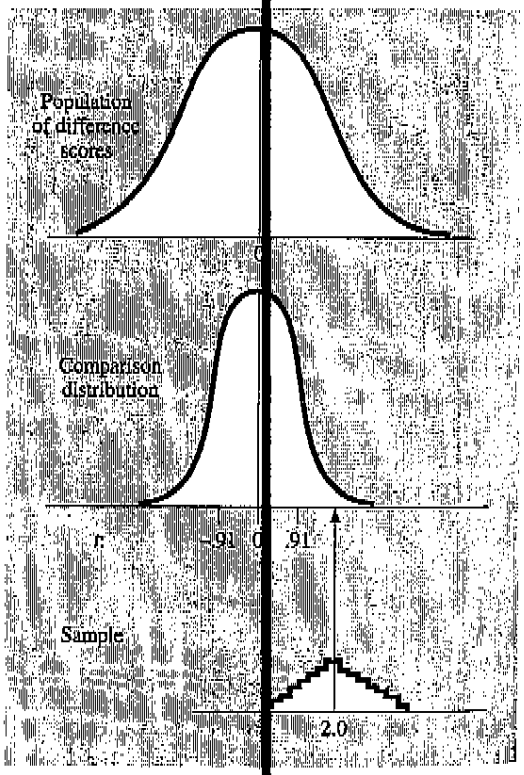

FIGURE 9-7
Distributions for fictional study of hand-eye coordination under noisy and quiet conditions.

that surgeons perform better under quiet conditions. The null hypothesis is that Population 1's difference in performance is not higher than Population 2's. That is, the null hypothesis is that surgeons do no better under quiet conditions.

2. Determine the characteristics of the comparison distribution. If the null hypothesis is true, the mean of the population of difference scores is 0. What is the variance of this population of difference scores? Estimating from the sample of difference scores, it is the sum of the squared deviations of the difference scores from the mean of the difference scores, divided by the degrees of freedom. This is shown in Table 9-6 to be 7.5. The comparison distribution is a distribution of means. Its variance is the variance of the distribution of individuals (in this case an estimated variance) divided by the sample size: 7.5/9 = .83. The standard deviation of the distribution of means is .91 (the square root of .83). The shape of the comparison distribution will be a *t* distribution with 8 degrees of freedom.

3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected. This is a one-tailed test because there was a reasonable basis for predicting the direction of the difference. We will suppose that the researcher wanted to be conservative and used the 1% significance level. With 8 degrees of freedom, Table B-2 shows that a *t* score of at least 2.897 is needed to reject the null hypothesis.

4. Determine the sample's score on the comparison distribution. The sample's mean difference of 2 is 2.20 standard deviations (of .91 each) above the mean of 0 on the distribution of means.

5. Compare the scores from Steps 3 and 4 to decide whether to reject the null hypothesis. The sample's *t* score of 2.20 is less extreme than the cutoff *t* of 2.897. Thus, you cannot reject the null hypothesis. The experiment is inconclusive. (Incidentally, had the researcher set the significance level at .05, this result would have been significant.)

A Third Example of a *t* Test for Dependent Means

A developmental psychologist is studying infants' responsiveness to strangers, using a new type of measure. He is able to measure 10 infants at 3 months of age and then again at 4 months. His prediction is that there will be an increase. Table 9-7 shows the results of this fictional study, along with the calculation of difference scores and all the other calculations for the *t* test for dependent means. Figure 9-8 shows the distributions involved. Here are the steps of hypothesis testing:

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

Population 1: Infants like those in this study
Population 2: Infants whose responsiveness to strangers is the same at 3 months and at 4 months of age

The research hypothesis is that Population 1's mean difference score (of responsiveness to strangers at 4 months minus responsiveness at 3 months) is greater than Population 2's. The null hypothesis is that Population 1's mean difference score is not greater than Population 2's.

TABLE 9-7
***t* Test for a Study of Responsiveness to Strangers of 10 Infants Measured at 3 and 4 Months of Age (Fictional Data)**

Infant	Age		Difference	Deviation	Squared Deviation
	3 months	*4 months*			
1	10.4	10.8	.4	.26	.07
2	12.6	12.1	−.5	−.64	.41
3	11.2	12.1	.9	.76	.58
4	10.9	11.4	.5	.36	.13
5	14.3	13.9	−.4	−.54	.29
6	13.2	13.5	.3	.16	.03
7	9.7	10.9	1.2	1.06	1.12
8	11.5	11.5	0.0	−.14	.02
9	10.8	10.4	−.4	−.54	.29
10	13.1	12.5	−.6	−.74	.55
Σ:	117.7	119.1	1.4	0	3.49

For difference scores:
$M = 1.4/10 = .14$.
$\mu = 0$.
$S^2 = SS/df = 3.49/(10 - 1) = 3.49/9 = .39$.
$S_M^2 = S^2/N = .39/10 = .039$.
$S_M = \sqrt{S_M^2} = \sqrt{.039} = .20$.
t for $df = 9$ needed for 5% significance level, one-tailed = 1.833.
$t = (M - \mu)/S_M = (.14 - 0)/.20 = .70$.

Decision: Do not reject the null hypothesis.

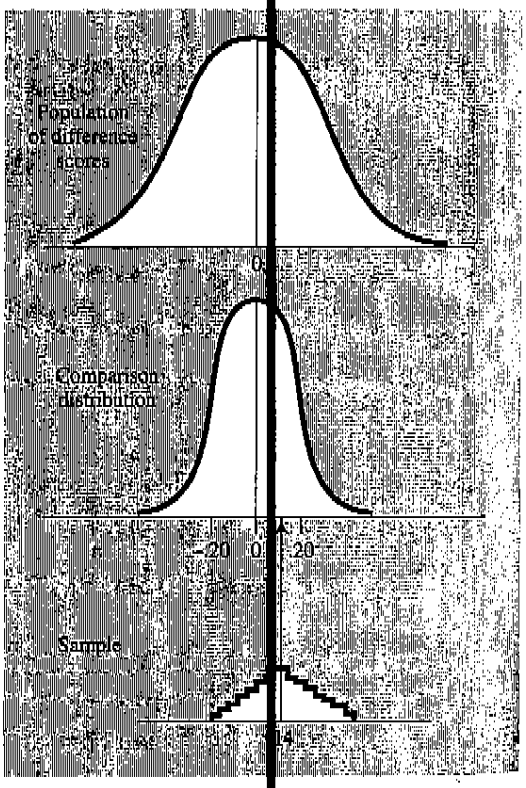

FIGURE 9-8
Distributions for fictional study of infants' responsiveness to strangers at 3 months and 4 months of age.

2. Determine the characteristics of the comparison distribution. Its population mean is 0 difference. The estimated population variance is shown in Table 9-7 to be .39. The comparison distribution will be a t distribution for 9 degrees of freedom with a mean of 0 and a standard deviation of .20.

3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected. This is a one-tailed test (because there was a reasonable basis for predicting the direction of the difference). Using the 5% significance level and 9 degrees of freedom, Table B-2 shows that a t score of at least 1.833 is needed to reject the null hypothesis.

4. Determine the sample's score on the comparison distribution. The sample's mean change of .14 is .70 standard deviations (of .20 each) on the distribution of means above that distribution's mean of 0.

5. Compare the scores from Steps 3 and 4 to decide whether to reject the null hypothesis. The sample's t of .70 is less extreme than the needed t of 1.833. Thus, you cannot reject the null hypothesis. The study is inconclusive.

Summary of Steps for Conducting a *t* Test for Dependent Means

Table 9-8 summarizes the steps in conducting a t test for dependent means. Optional computational formulas making it easier to carry out a t test for dependent means by hand when you have a large number of difference scores are given in the chapter appendix.

TABLE 9-8
Steps for Conducting a *t* Test for Dependent Means

1. Restate the question as a research hypothesis and a null hypothesis about the populations.
2. Determine the characteristics of the comparison distribution.
   a. Make each person's two scores into a difference score. Do all the rest of the steps using these difference scores.
   b. Compute the mean of the difference scores.
   c. Assume a population mean of 0: $\mu = 0$.
   d. Compute the estimated population variance of difference scores: $S^2 = SS/df$.
   e. Compute the variance of the distribution of means of difference scores: $S_M^2 = S^2/N$.
   f. Compute the standard deviation of the distribution of means of difference scores: $S_M = \sqrt{S_M^2}$.
   g. The shape is a t distribution with $df = N - 1$.
3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.
   a. Determine the desired significance level and whether to use a one-tailed or a two-tailed test.
   b. Look up the appropriate cutoff in a t table.
4. Determine your sample's score on the comparison distribution: $t = (M - \mu)/S_M$.
5. Compare the scores from Steps 3 and 4 to decide whether to reject the null hypothesis.

ASSUMPTIONS OF THE *t* TEST

As we have seen, when using an estimated population variance, the comparison distribution is a *t* distribution. However, the comparison distribution will be exactly a *t* distribution only if the distribution of individuals follows a normal curve. Otherwise, the comparison distribution will follow some other (usually unknown) shape.

Thus, strictly speaking, a normal population is a requirement within the logic and mathematics of the *t* test. A requirement like this for a hypothesis-testing procedure is called an **assumption.** A normal population distribution is said to be an assumption of the *t* test. The effect of this assumption is that if the population distribution is not normal, it is technically wrong to use the *t* test.

assumption

Unfortunately, you usually don't know whether the population is normal. This is because when doing a *t* test, usually all you have to go on are the scores in your sample. Fortunately, as we saw in Chapter 5, distributions in psychology research quite often approximate a normal curve. (This also applies to distributions of difference scores.) Also, statisticians have found that in practice, you get reasonably accurate results with the *t* test even when the population is rather far from normal. In other words, the *t* test is said to be *robust* over moderate violations of the assumption of a normal population distribution. How statisticians figure out the **robustness** of a test is an interesting topic, which is described in Box 10-1 in Chapter 10.

robustness

There is one reasonably common situation in which using a *t* test for dependent means is likely to give seriously distorted results. This is when you are doing a one-tailed test and the population is highly skewed (is very asymmetrical, with a much longer tail on one side than the other).

How do you know when your population is highly skewed? One situation is where the sample of difference scores is highly skewed. If the sample is highly skewed, it is likely that the population the sample comes from is highly skewed. Another situation is where you have reason to think there is a floor or ceiling effect, making the distribution skewed because scores on one side can't go any higher or lower. When you have reason to think that conducting a *t* test would seriously violate the normal curve assumption and give distorted results, there are several alternatives to the *t* test you can use. You will learn about these alternatives in Chapter 15.

EFFECT SIZE AND POWER FOR THE *t* TEST FOR DEPENDENT MEANS

Effect Size

You figure the effect size for a study using a *t* test for dependent means the same way as in Chapter 8. It is the difference between the population means divided by the population standard deviation: $(\mu_1 - \mu_2)/\sigma$. However, when using difference scores, the mean of Population 2 is usually 0 (that is, with difference scores, $\mu_2 = 0$). This simplifies the situation:

$$d = \frac{(\mu_1 - 0)}{\sigma} = \frac{\mu_1}{\sigma} \qquad (9\text{-}8)$$

Remember when using this formula that μ_1 is for the predicted mean of the population of difference scores and σ is for the standard deviation of the populations of difference scores.

The effect size conventions for a *t* test for dependent means are the same as you learned for the situation we considered in Chapter 8: A small effect size is .20, a medium effect size is .50, and a large effect size is .80.

Consider an example. A sports psychologist plans a study on attitudes toward teammates before versus after a game. She will administer an attitude questionnaire twice, once before and once after a game. Suppose that the smallest before-after difference that would be of any importance is 4 points on the questionnaire. Also suppose that based on related research, the researcher figures that the standard deviation of difference scores on this attitude questionnaire is about 8 points. Thus, $\mu_1 = 4$ and $\sigma = 8$. Applying the effect size formula: $d = \mu_1/\sigma = 4/8 = .50$. In terms of the effect size conventions, her planned study has a medium effect size.

If you want to estimate the effect size after you have conducted a study, you divide the actual mean of the difference scores in your sample by the estimated standard deviation of the population of difference scores:

$$d = \frac{M}{S} \tag{9-9}$$

Remember, both M and S in this formula are for difference scores. Also note that S is the standard deviation of the population of individuals (that is, in this situation, of individual's difference scores). It is not the same as S_M, the standard deviation of the distribution of means (of difference scores).

Consider our first example of a *t* test for dependent means, the study of husbands' change in communication quality. In that study, the mean of the difference scores was –12.05. The estimated population standard deviation of the difference scores would be 12.41. That is, we computed the estimated variance of the difference scores (S^2) to be 154.05; $\sqrt{S^2} = 12.41$. Therefore, the effect size is computed as $d = M/S = -12.05/12.41 = -.97$. This is a very large effect size. (The negative sign for the effect size means that the large effect was a decrease.)

Power

Table 9-9 gives the approximate power at the .05 significance level for small, medium, and large effect sizes and one- or two-tailed tests. In the sports psychology example, the researcher expected a medium effect size ($d = .50$). If she planned to conduct the study using the .05 level, two-tailed, with 20 participants, the study would have a power of .59. This means that, if the research hypothesis is in fact true and has a medium effect size, there is a 59% chance that this study will come out significant.

The power table (Table 9-9) is also useful when you are reading about a nonsignificant result in a published study. Suppose that a study using a *t* test for dependent means had a nonsignificant result. The study tested significance at the .05 level, two-tailed, and had 10 participants. Should you conclude that there is in fact no difference at all in the populations? Probably not. Even assuming a medium effect size, Table 9-9 shows that there is only

TABLE 9-9
Approximate Power for Studies Using the *t* Test for Dependent Means in Testing Hypotheses at the .05 Significance Level

Difference Scores in Sample (N)	Effect Size		
	Small ($d = .20$)	*Medium* ($d = .50$)	*Large* ($d = .80$)
Two-tailed test			
10	.09	.32	.66
20	.14	.59	.93
30	.19	.77	.99
40	.24	.88	*
50	.29	.94	*
100	.55	*	*
One-tailed test			
10	.15	.46	.78
20	.22	.71	.96
30	.29	.86	*
40	.35	.93	*
50	.40	.97	*
100	.63	*	*

*Power is nearly 1.

a 32% chance of getting a significant result in this study. Now consider another study that was not significant. This study also used the .05 significance level, two-tailed, but had 100 research participants. Table 9-9 tells you that there would be a 63% chance of the study's coming out significant if there were even a true small effect size in the population. If there were a medium effect size in the population, the table indicates that there is almost a 100% chance that this study would have come out significant. Thus, in this study with 100 participants, we could conclude from the results of this study that in the population there is probably no difference at all or at most a very small difference.

To keep Table 9-9 simple, we have given power figures for only a few different numbers of participants (10, 20, 30, 40, 50, and 100). This should be adequate for the kinds of rough evaluations you need to make when evaluating results of research articles.[3]

[3]Cohen (1988, pp. 28–39) provides more detailed tables, in terms of numbers of participants, levels of effect size, and significance levels. If you use his tables, note that the d referred to is actually based on a t test for independent means (the situation we consider in Chapter 10). To use these tables for a t test for dependent means, first multiply your desired effect size by 1.4. For example, if your effect size is .30, for purposes of using Cohen's tables, you would consider it to be .42 (that is, $.30 \times 1.4 = .42$). The only other difference from our table is that Cohen describes the significance level by the letter a (for "alpha level"), with a subscript of either 1 or 2, referring to a one- or two-tailed test. For example, a table that refers to "$a_1 = .05$" at the top means that this is the table for $p < .05$, one-tailed.

Planning Sample Size

Table 9-10 gives the approximate number of research participants needed to have 80% power for small, medium, and large effect sizes using one- and two-tailed tests, for the .05 significance levels. (Eighty percent is a common figure used by researchers for the minimum power to make a study worth doing.) Suppose you plan a study in which you expect a large effect size and will use the .05 significance level, two-tailed. The table shows you would only need 14 participants to have 80% power. On the other hand, a study using the same significance level, also two-tailed, but in which you expect only a small effect size would need 196 participants in your study for 80% power.[4]

The Power of Studies Employing the *t* Test for Dependent Means

Studies using difference scores (that is, studies using a repeated-measures design) often have considerably larger effect sizes for the same amount of expected difference between means than other kinds of research designs. If effect sizes are larger, then power is larger. That is, testing each of a group of participants twice (once under one condition and once under a different condition) usually produces a high power type of study. In particular, this kind of study gives more power than dividing the participants up into two groups and testing each group once (one group tested under one condition and the other tested under another condition). In fact, studies using difference scores usually have even more power than those in which you have twice as many participants, but tested each only once.

Why do repeated-measures designs have so much power? The reason is that the standard deviation of difference scores is usually quite low. (The standard deviation of difference scores is what you divide by to get the effect size when using difference scores.) In a repeated-measures design, the only variation is in the difference scores. Variation among participants on each testing's scores are not part of the variation involved in the analysis. This is because difference scores are all comparing participants to themselves. William S. Gosset, who essentially invented the *t* test (see Box 9-1), made much of the higher power of repeated-measures studies in a historically interesting controversy over an experiment about milk, which is described in Box 9-2.

TABLE 9-10
Approximate Number of Research Participants Needed to Achieve 80% Power for the *t* Test for Dependent Means in Testing Hypotheses at the .05 Significance Level

	Effect Size		
	Small (*d* = .20)	*Medium* (*d* = .50)	*Large* (*d* = .80)
Two-tailed	196	33	14
One-tailed	156	26	12

[4]More detailed tables, giving needed numbers of participants for levels of power other than 80% (and also for effect sizes other than .20, .50, and .80 and for other significance levels) are provided in Cohen (1988, pp. 54–55). However, see footnote 3 in this chapter about using these tables.

Box 9-2
The Power of Studies Using Difference Scores
How the Lanarkshire Milk Experiment Could Have Been Milked for More

In 1930 a major health experiment was conducted in Scotland involving 20,000 schoolchildren. Its main purpose was to compare the growth of a group of children who were assigned to drink milk regularly to those who were in a control group. The results were that those who drank milk showed more growth.

However, William Gosset, a contemporary statistician (see Box 9-1), was appalled at the way the experiment was conducted. It had cost about £7,500, which in 1930 was a huge amount of money, and was done wrong! Large studies such as this were very popular among statisticians in those days because they seemed to imitate the large numbers found in nature. Gosset, by contrast, being a brewer, was forced to use very small numbers in his studies—experimental batches of beer were too costly. And he was often chided by the "real statisticians" for his small sample sizes. But Gosset argued that no number of participants was large enough when strict random assignment was not followed. And in this study, teachers were permitted to switch children from group to group if they took pity on a child whom they felt would benefit from receiving milk! (See Appendix A for a discussion of random assignment to groups.)

However, even more interesting in light of the present chapter, Gosset demonstrated that the researchers could have obtained the same result with 50 pairs of identical twins, flipping a coin to determine which of each pair was in the milk group (and sticking to it). Of course, the statistic you would use is the *t* test as taught in this chapter—the *t* test for dependent means.

More recently, the development of power analysis, which we introduced in Chapter 8, has thoroughly vindicated Gosset. It is now clear just how surprisingly few participants are needed when a researcher can find a way to set up a repeated-measures design in which difference scores are the basic unit of analysis. (In this case, each *pair* of twins would be one "participant.") As Gosset could have told them, studies that use the *t* test for dependent means can be extremely sensitive.

References: Peters (1987); Tankard (1984).

CONTROVERSIES AND LIMITATIONS

The main controversies about the *t* test have to do with its relative advantages and disadvantages in comparison to various alternatives—alternatives that we will discuss in some detail in Chapter 15. (These same issues also arise over the procedures we will cover in Chapters 10–13.) There is, however, one consideration that we want to comment on now. It relates to all research designs in which the same participants are tested before and after some experimental intervention. (This is the kind of situation that the *t* test for dependent means is often used to evaluate.)

Simply measuring a group of people before and after an experimental procedure, without any kind of control group that does not undergo the procedure, may have high power, but it is a weak research design in terms of the clarity of conclusions it can produce (Cook & Campbell, 1979). As described in more detail in Appendix A, even if such a study produces a significant

difference, it leaves many alternative explanations for why that difference occurred. For example, the research participants might have matured or improved during that period anyway, or perhaps other events happened in between, or the participants not getting benefits may have dropped out. It is even possible that the initial test itself caused changes that otherwise might not have occurred.

Note, however, that the difficulties of research that tests people before and after some intervention are shared only slightly with the kind of study in which participants are tested under two conditions, such as noisy versus quiet, with half tested first under one condition and half tested first under the other condition.

t TESTS AS DESCRIBED IN RESEARCH ARTICLES

Research articles usually describe *t* tests in a fairly standard format that provides the degrees of freedom, the *t* score, and the significance level. For example, "$t(24) = 2.80$, $p < .05$" tells you that the researcher used a *t* test with 24 degrees of freedom, obtained a *t* score of 2.80, and the result was significant at the .05 level. Whether a one- or two-tailed test was used may also be noted. (If it is not noted, assume the researcher used a two-tailed test.) Usually the means, and sometimes the standard deviations, are given for each testing. Rarely is the standard deviation of the difference scores reported.

Had our student in the dormitory example reported the results in a research article, it would have been something like this: "The sample from my dormitory studied a mean of 3.2 hours (SD = .80). Based on a single-sample *t* test (one-tailed), this was significantly different from the known mean of 2.5 for the college as a whole, $t(15) = 3.50$, $p < .01$." The researchers in our fictional flood victims example might have written up their results as follows: "The reported hopefulness of our sample of flood victims ($M = 4.7$, $SD = 1.89$) was not significantly different from the midpoint of the scale (4.0), $t(9) = 1.17$."

As we noted earlier, psychologists only rarely use a *t* test for a single sample. We introduced this *t* test mainly as a stepping-stone to the more widely used *t* test for dependent means. Nevertheless, one does sometimes see the *t* test for a single sample in research articles. For example, Weller and Weller (1997) conducted a study of the tendency for the menstrual cycles of women who live together to become synchronized. For their statistical analysis, they compared scores on a measure of synchronization of pairs of women living together for women in their study (Population 1) versus the degree of synchronization of these pairs of women that would be expected by chance (Population 2). That is, they created a kind of artificial population that has a mean of what you would expect if there were no synchronization. They analyzed their results with "one-sample *t* tests" (p. 147). The results are shown in Table 9-11. Each row of the table is a separate single-sample *t* test. The first row is a test comparing the synchrony scores of 6.32 for the 30 roommate sister pairs (their sample from what we would call Population 1) to an expected synchrony score of 7.76 (what we would call the mean of Population 2). The row shows those figures plus the difference of 1.44, the standard deviation of 3.40 of this difference, the *t* score of 2.27, and the *p* level of .011. Notice that their *t* column was actually written as "$t(1)$." This is not standard and

certainly does not mean that their *t* distribution had one degree of freedom. We presume they meant that this was a single-sample *t* test.

As we have said, the t test for dependent means is much more common. Olthoff (1989) might have reported his result in the example we used as: "There was a significant decline in communication quality, dropping from 116.32 before marriage to 104.26 after marriage, $t(18) = 2.76$, $p < .05$, two-tailed." The researcher in the fictional surgeons study could have written the following: "The mean performance for the quiet group was 18.89, while the performance for the noisy group was 16.89. This difference was not statistically significant at the .01 level, even with a one-tailed test, $t(8) = 2.20$." As another example, Holden et al. (1997) compared mothers' reported attitudes towards corporal punishment of their children from before to 3 years after having their first child. "The average change in the women's prior-to-current attitudes was significant, $t(107) = 10.32$, $p < .001$ . . ." (p. 485). (The change was that they felt more negatively about corporal punishment after having their child.)

Researchers also often present the means of the groups in a table. For example, Pezdek and her colleagues (1997) reminded each of a group of college students of several events that supposedly happened to them as 8-year-olds. The students were asked to describe the event in some detail. These descriptions were rated for number of words recalled and number of idea units recalled. The students were also asked to rate each event for how clearly they recalled it and for how confident they were it happened. Some of the events had actually happened and some were ones that could have happened but did not. (The researchers had contacted the mothers of the students in advance with the permission of the students.) As is typical in such research, many of the students incorrectly recalled having experienced the false events. Here are their results:

> To investigate potential differences between memories for true versus false events, we compared various characteristics of the memories for the 13 subjects who recalled at least one false event. Two-tailed significance tests were conducted on these data, and the results are presented in [Table 9-12]. Compared with recall of false events, recall of true events employed significantly more

TABLE 9-11
Menstrual Synchrony and Expected Scores (by Days)

Group/month	*N*	Synchrony score	Expected score	Difference	*SD*	*t*(1)	*p*
Roommates–sisters							
Month 1	30	6.32	7.76	1.44	3.40	2.27	.011
Month 2	30	6.24	7.76	1.52	3.08	2.66	.004
Month 3	29	7.40	7.76	0.36	3.08	0.57	.28
Close friends–roommates							
Month 1	39	5.73	7.75	2.02	3.84	3.25	<.000
Month 2	39	6.01	7.75	1.74	4.25	2.52	.006
Month 3	31	7.44	7.75	0.31	4.61	0.88	.19
Families							
Month 1	18	5.80	7.70	1.90	2.74	2.86	<.000
Month 2	18	6.09	7.70	1.61	1.89	3.52	<.000
Month 3	17	7.19	7.70	0.51	2.71	0.75	.23

Note. Data from Weller, A., & Weller, L. (1997), tab. 1. Menstrual synchrony under optimal conditions: Bedouin families. *Journal of Comparative Psychology, 111,* 143–151. Copyright, 1997, by the American Psychological Association. Reprinted with permission.

TABLE 9-12
Means (and standard deviations) for Measures Comparing Recall of True and False Events in Experiment 1

	Recalled Event	
Measure	True	False
Number of words recalled ***	27.79 (8.81)	15.42 (7.69)
Number of idea units recalled**	6.33 (2.53)	3.23 (1.55)
Clarity rating[a]***	6.90 (0.17)	4.00 (0.18)
Confidence rating[a]**	6.88 (0.21)	5.00 (0.21)

[a]The rating scale ranged from 1 (low) to 10 (high).
*$p < .02$, two-tailed; **$p < .01$, two-tailed; ***$p < .001$, two-tailed.
Note. Data from Pezdek, K., Finger, K., & Hodge, D. (1997), tab. 2. Planting false childhood memories: The role of event plausibility. *Psychological Science, 8*, 439. Copyright,1997, by the American Psychological Society. Reprinted with permission.

> words, $t(12) = 4.54$, $p < .001$, and more idea units, $t(12) = 3.43$, $p < .01$. Thus, the recall output for true versus false events could be differentiated in terms of the number of new details provided for each; there were almost twice as many details provided for true as false events. Compared with recalled false events, recalled true events were also associated with significantly higher ratings of clarity, $t(12) = 3.99$, $p < .01$, and confidence, $t(12) = 2.73$, $p < .02$. (p. 438)

Notice in this example, they never referred to the name of the significance test. However, you know it is a *t* test because they use *t* in describing the results. You can tell it was a *t* test for dependent means because they are comparing each participant's score on memory of true events to his or her score on memory of false events. That is, the comparison is between two scores from each participant.

SUMMARY

The standard five steps of hypothesis testing are used when the variance of the population is not known. However, in this situation you must estimate the population variance from the scores in the sample, using a formula that divides the sum of squared deviation scores by the degrees of freedom ($df = N - 1$). Also, when the variance is not known, the comparison distribution of means is a *t* distribution (with cutoffs given in a *t* table). A *t* distribution has slightly heavier tails than a normal curve (just how much heavier depends on how few degrees of freedom). Finally, in this situation the number of standard deviations from the mean that a sample's mean is on the *t* distribution is called a *t* score.

A *t* test for dependent means is used in studies where each participant has two scores, such as a before score and an after score. In this *t* test, you first figure a difference score for each participant, then carry out the usual five steps of hypothesis testing with the modifications described in the paragraph above and making Population 2 a population of difference scores with a mean of 0 (no difference).

An assumption of the *t* test is that the population distribution is a normal curve. However, even when it is not, the *t* test is usually fairly accurate. The

main exception for the *t* test for dependent means is when the population of difference scores is highly skewed and you are using a one-tailed test.

The effect size of a study using a *t* test for dependent means is the mean of the difference scores divided by the standard deviation of the difference scores. Power and needed sample size for 80% power can be looked up in special tables. The power of studies using difference scores is usually much higher than that of studies using other designs with the same number of participants.

Research methodologists point out that research involving a single group tested before and after some intervening event, without a control group, permits many alternative explanations of any observed changes.

t tests are reported in research articles using a standard format—for example, "$t(24) = 2.80$, $p < .05$."

Key Terms

assumption
biased estimate
degrees of freedom (*df*)
difference scores
repeated-measures design
robustness
t distribution
t score
t table
t tests
t test for a single sample
t test for dependent means
unbiased estimate of the population variance (S^2)

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. In each of the studies below, a single sample's mean is being compared to a population with a known mean but an unknown variance. For each study, decide whether the result is significant.

	Sample Size (N)	Population Mean (μ)	Estimated Population Variance (S^2)	Sample Mean (M)	Tails	Significance Level (α)
(a)	64	12.40	9.00	11.00	1 (low predicted)	.05
(b)	49	1,006.35	317.91	1,009.72	2	.01
(c)	400	52.00	7.02	52.41	1 (high predicted)	.01

2. Suppose that a candidate running for sheriff claims that she will reduce the average time of emergency response to less than 30 minutes, which is thought to be the average response time under the current sheriff. There are no past records, so the actual standard deviation of such response times cannot be determined. Thanks to this campaign, she is elected sheriff, and careful records are now kept. The response times for the first month are 26, 30, 28, 29, 25, 28, 32, 35, 24, and 23 minutes.

Using the 5% level of significance, did she keep her promise? (a) Go through the five steps of hypothesis testing. (b) Illustrate your answer with a histogram of the distribution of the sample and sketches of the population distribution and the distribution of means, showing the *t* score and cutoff points for significance. (c) Explain your answer to someone who has never taken a course in statistics.

3. For each of the following studies using difference scores, determine if the mean difference is significantly different from 0. Also compute the effect size. (If *df* is not given in the table, use the *t* for the nearest lower *df* value.)

	Number of Difference Scores in Sample	Mean of Difference Scores in Sample	Estimated Population Variance of Difference Scores	Tails	Significance Level
(a)	20	1.7	8.29	1 (high predicted)	.05
(b)	164	2.3	414.73	2	.05
(c)	15	−2.2	4.00	1 (low predicted)	.01