CHAPTER 10

t-TESTS—COMPARING TWO MEANS

The *t*-test is probably the most widely used statistical test for the comparison of two means because it can be used with very small sample sizes. You may see it referred to as *Students' t-test.* That doesn't mean that only students use it; rather, it was worked out by a mathematician who used Student as his pen name.

If you use the *t*-test with large-size samples, the values of *t* critical and *z* critical will be almost identical. As the sample size increases, the values of *t* and *z* become very close (the value of *t* will always be somewhat larger than the *z* value). Since the values are so close, many researchers use the *t*-test regardless of sample size.

You will soon see that, aside from using a different distribution, the procedure for obtaining the *t* value is the same as that for the *z* score. We will go through the procedure first for Case I and then for Case II studies.

CASE I STUDIES

You will remember that a Case I study compares a sample mean with an established population mean. The null hypothesis for a Case I *t*-test would be that there is no difference between our sample mean and the mean already established for the population. Suppose that you have an experiment that requires that your 16 *S*s have normal language-learning aptitude. So, you give them all the MLAT (Modern Language Aptitude Test) with the hope that they will place in the central area of the normal distribution of means. The formula for the computation is

$$t_{observed} = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

You can see from the top half of the formula that we want to find the difference between the sample mean and the population mean (your *S*s' mean on the MLAT and the normed population mean for the MLAT). Then the bottom half says that we must divide that difference by the standard error of means. The standard error of means, again, is computed as

$$s_{\bar{X}} = \frac{s_X}{\sqrt{N}}$$

So the $t_{observed}$ formula could also be written as

$$t_{observed} = \frac{\bar{X} - \mu}{s_X / \sqrt{N}}$$

This should be familiar to you, for it is exactly the same as the Case I z score formula (see preceding chapter).

Let's say that your 16 *S*s have a $\bar{X}$ of 75 on the MLAT and an s of 6. Let's pretend that the published mean for the test is 80 (we're sure it isn't; this is just hypothetical). Plugging the data into the formula, we get

$$\begin{aligned} t_{observed} &= \frac{\bar{X} - \mu}{s_X / \sqrt{N}} \\ &= \frac{75 - 80}{6 / \sqrt{16}} \\ &= \frac{-5}{1.5} \\ &= -3.33 \end{aligned}$$

At this point, you must think of yourself as one of a number of *E*s each of whom is trying to find out if their *S*s have normal language aptitude abilities. Each of these *E*s has a sample of 16 *S*s just like you. You visualize your t value as just a member of all the t values found by these other *E*s. It is sort of a "family" of t values. The family forms the t-distribution for the number of *S*s in the sample. All of you have gathered samples from 16 *S*s; so your family of t values is based on the number of *S*s in the sample. The task, now, is to decide whether or not the obtained value of t really fits in that distribution.

Before we can make that decision, we need to add one more concept: the concept of degrees of freedom. You have undoubtedly noticed that we often use $N - 1$ rather than N when we want to find some average. The formula for standard deviation, for example, requires us to divide the total for deviation from the mean not by N but by $N - 1$. The reason we so often use $N - 1$ rather than N is that we are using our statistics to estimate the population parameter. Our sample size is small while the population size is large. It also stands to reason that the dispersion of scores in our sample will be larger than the dispersion of scores in the population. If we divide our sample values by N, we will get the average for our sample, but we will not get a value which is the best estimate of the population parameter. Mathematicians have decided that the best way to use sample averages as estimates of population parameters is to use $N - 1$, which is related to degrees of freedom.

The concept of degrees of freedom is very important in all hypothesis testing. It refers to the number of quantities that can vary if others are given. For

example, if you know that $A + B = C$, you know that A and B are free to vary. You can put any number in the A and B slots and call the sum C. But if you change C to a number so that $A + B = 200$, then only one of the numbers for A and B can vary. As soon as you fill in one of the slots, the other one is fixed ($A + 50 = 200$ or $50 + B = 200$). So we say there is one degree of freedom. Only one of the two quantities is free to vary; the other is fixed. If our formula were $A + B + C = D$, and we say that D is 100, then any two of the values for A, B, and C can vary but the third is fixed. So then there are two degrees of freedom. To find the degrees of freedom for the sample, you can subtract $N - 1$. If you have 20 *S*s in a sample and wish to divide their scores and use that average as an estimate of the population, you must divide not by N but by the degrees of freedom, $N - 1$.

The number of degrees of freedom is important in the *t*-test, for it determines the shape of the frequency distribution for *t* values. For any given number of degrees of freedom there is a particular *t*-distribution with its own set of critical values for significance. In the example of students who have been tested on the MLAT, there were 16 *S*s in the sample group. The data for the 16 *S*s belong to the *t*-distribution with 15 degrees of freedom (16 − 1 = 15 d.f.). You must place your sample $\overline{X}$ in the *t*-distribution formed from the family of an infinite number of samples all with 15 d.f. If you had 5 *S*s, you would have to place it in comparison with the *t*-distribution for four degrees of freedom. If you had 25 *S*s, you would compare it with the *t*-distribution with 24 d.f. The *t critical* value, the *t* value you must obtain in order to claim statistical significance, will vary according to the number of degrees of freedom, and thus the size of the samples that make up the distribution.

The *t*-distribution table in the Appendix allows us to compare our observed value of *t* with the appropriate family in the *t*-distribution table. The rows down the side of the table relate to the separate *t*-distributions, each with a unique number of degrees of freedom. In the example, we had 16 *S*s; so our d.f. row is marked 15. The columns of numbers across the page give us the probability levels. If our hypothesis is directional, we will use the top row (α) to locate our already selected .05 or .01 level of significance. If the hypothesis is non-directional (two-tailed), we would use the row labeled 2-tailed as the guide to our .05 or .01 level.

Let's say that we had already selected the .05 level of significance for rejecting the null hypothesis. Our obtained value for *t* was −3.33. We find row 15 for the degrees of freedom and then check across to where 15 intersects with the column labeled two-tailed .05. The *t* value at the intersection is 2.13. We can reject the null hypothesis because our *t* value is greater than 2.13. (It makes no difference whether the obtained value is positive or negative in reading the table. Since the distribution is symmetrical, the minus quantities would be the same. To reproduce both sides of the distribution would take up unnecessary space.) You can be quite sure that the group is unusual in their language-learning ability; they are much worse than most learners. The score places them at the far left tail of the distribution.

If the difference between your mean and that of the population had resulted in a t value of +3.33 instead, it would be at the far right tail of the distribution. With a positive t value of +3.33, your group would not be typical but better than most in language-learning aptitude.

The t value obtained in the example data allows you to reject the null hypothesis. You are not safe in assuming that your *S*s are typical in their language-learning aptitude.

CASE II STUDIES

Case II studies require a comparison of two means for two groups drawn from the population. The process of making this comparison is similar to that used in Case II studies for the z-distribution. But, once again, we will use a t-distribution instead of a z-distribution because we have so few *S*s in our sample groups.

Let's assume that we believe role-play and group problem-solving promote oral proficiency. We have constructed an oral interview measure (or, better yet, found an established measure) to test oral proficiency. We then select a random sample of 72 ESL students from our schools and randomly assign them to two groups of 36 *S*s each. One group becomes the experimental group and receives role-play and problem-solving activities; the other group is the control group which receives some placebo treatment. At the end of the semester, we administer the oral interview and obtain the following data (e identifies the experimental group; c the control group):

$$\bar{X}_e = 62 \qquad s_e = 12$$
$$\bar{X}_c = 55 \qquad s_c = 15$$

We believe that the special instruction does result in higher scores for the experimental group and we wish we could make an alternative hypothesis that would be one-tailed and directional. However, we want to be extra hard on our predictions, so we decide to make no prediction as to the direction of the difference. Our null hypothesis is:

H_0 = the two samples are from the same population; the difference between the two sample means which represent population means is zero ($\mu_1 - \mu_2 = 0$)

This prediction says we expect that any difference between our two groups falls well within the normal differences found for any two means in the population. If we can reject this hypothesis, we must have a high enough t value to be sure that such a large difference is not due to chance.

Now that we have the null hypothesis, we can set our acceptance level at .05, and try to reject the hypothesis. The formula is exactly the same as the one we used for z scores in Case II studies:

$$t_{obs} = \frac{\bar{X}_e - \bar{X}_c}{s_{(\bar{X}_e - \bar{X}_c)}}$$

$s_{(\bar{X}_e - \bar{X}_c)}$ — standard error of differences between means

If the difference between your mean and that of the population had resulted in a t value of +3.33 instead, it would be at the far right tail of the distribution. With a positive t value of +3.33, your group would not be typical but better than most in language-learning aptitude.

The t value obtained in the example data allows you to reject the null hypothesis. You are not safe in assuming that your *S*s are typical in their language-learning aptitude.

CASE II STUDIES

Case II studies require a comparison of two means for two groups drawn from the population. The process of making this comparison is similar to that used in Case II studies for the z-distribution. But, once again, we will use a t-distribution instead of a z-distribution because we have so few *S*s in our sample groups.

Let's assume that we believe role-play and group problem-solving promote oral proficiency. We have constructed an oral interview measure (or, better yet, found an established measure) to test oral proficiency. We then select a random sample of 72 ESL students from our schools and randomly assign them to two groups of 36 *S*s each. One group becomes the experimental group and receives role-play and problem-solving activities; the other group is the control group which receives some placebo treatment. At the end of the semester, we administer the oral interview and obtain the following data (e identifies the experimental group; c the control group):

$$\bar{X}_e = 62 \qquad s_e = 12$$
$$\bar{X}_c = 55 \qquad s_c = 15$$

We believe that the special instruction does result in higher scores for the experimental group and we wish we could make an alternative hypothesis that would be one-tailed and directional. However, we want to be extra hard on our predictions, so we decide to make no prediction as to the direction of the difference. Our null hypothesis is:

H_0 = the two samples are from the same population; the difference between the two sample means which represent population means is zero ($\mu_1 - \mu_2 = 0$)

This prediction says we expect that any difference between our two groups falls well within the normal differences found for any two means in the population. If we can reject this hypothesis, we must have a high enough t value to be sure that such a large difference is not due to chance.

Now that we have the null hypothesis, we can set our acceptance level at .05, and try to reject the hypothesis. The formula is exactly the same as the one we used for z scores in Case II studies:

$$t_{obs} = \frac{\bar{X}_e - \bar{X}_c}{s_{(\bar{X}_e - \bar{X}_c)}}$$

$s_{(\bar{X}_e - \bar{X}_c)}$ ← standard error of differences between means

The subscripts *e* and *c* refer, again, to experimental and control. The top part of the formula is always the easy part. We already know there is a difference of 7 between the $\overline{X}$ of 62 on the oral interview for our experimental group and the $\overline{X}$ of 55 for the control group. Now we need to work out the standard error of differences between the means.

The formula for the standard error of differences between the means gives us a ruler for the difference in means if we repeated this experiment over and over with different 36-member classes. That ruler is corrected for the size of our classes to estimate the difference for the population:

$$s_{(\overline{X}_e - \overline{X}_c)} = \sqrt{\left(\frac{s_e}{\sqrt{n_1}}\right)^2 + \left(\frac{s_c}{\sqrt{n_2}}\right)^2}$$

$$= \sqrt{\left(\frac{12}{\sqrt{36}}\right)^2 + \left(\frac{15}{\sqrt{36}}\right)^2}$$

$$= \sqrt{\left(\frac{12}{6}\right)^2 + \left(\frac{15}{6}\right)^2}$$

$$= \sqrt{4 + 6.25}$$

$$= \sqrt{10.25}$$

$$= 3.2$$

Now that we have the standard error of differences between the means, we can find the *t* value:

$$t_{obs} = \frac{\overline{X}_e - \overline{X}_c}{s_{(\overline{X}_e - \overline{X}_c)}}$$

$$= \frac{62 - 55}{3.2}$$

$$= \frac{7}{3.2}$$

$$= 2.19$$

At this point, all we need is the critical value for *t* when the sample size is 36 and we have two groups. Each group had 36 *S*s; one of the scores is predictable given the other 35. So each group has 35 d.f. Since there are two groups, the total d.f. $(n_1 - 1 + n_2 - 1)$ is 70. Again, we can turn to the *t*-distribution table to find out whether we are justified in rejecting the null hypothesis. We find that our number of d.f., 70, is not listed but falls between 60 and 120. We choose 60 as being the more conservative estimate, and check across to the .05 column. The *t* value needed for our selected significance level of .05 is 2.000. Fortunately, our *t* value is enough above *t* critical that we are quite safe in rejecting the null

hypothesis. Our two groups have scored differently on the final test of oral proficiency. The difference is statistically significant. This is support for our claim that our method of using role-play and problem-solving promotes oral proficiency.

Let's work through one more example to be sure that the procedure is clear. Let's say that we have the same two classes as before (where *S*s have been randomly selected and randomly assigned to control or experimental groups). Each of the groups has been given a unit of instruction on how to use the library. However, we have also given the experimental group some special instruction on how best to ask native speakers of English for information and help. Both groups are given an assignment to find answers to 15 questions by locating the information in the library. The data for the two groups follow:

$$\begin{array}{ll} \bar{X}_e = 88 \text{ minutes} & s_e = 28 \text{ minutes} \\ \bar{X}_c = 102 \text{ minutes} & s_c = 20 \text{ minutes} \\ N_e = 36 & N_c = 30 \end{array}$$

Notice that on the day we gave the problem, six people were absent from the control group.

Our formula for the t value is

$$\begin{aligned} t_{obs} &= \frac{\bar{X}_e - \bar{X}_c}{s_{(\bar{X}_e - \bar{X}_c)}} \\ &= \frac{88 - 102}{\sqrt{(28/\sqrt{36})^2 + (20/\sqrt{30})^2}} \\ &= \frac{88 - 102}{\sqrt{(4.67)^2 + (3.65)^2}} \\ &= \frac{-14}{\sqrt{21.81 + 13.32}} \\ &= \frac{-14}{\sqrt{35.13}} \\ &= \frac{-14}{5.93} \\ &= -2.36 \end{aligned}$$

To check to see whether this observed value of t is statistically significant or not, we again check the t-distribution table. This time we had 36 *S*s in one group and 30 in the other. This gives us a total of 64 d.f. ($36 - 1 = 35$; $30 - 1 = 29$; $35 + 29 = 64$). Again, our t value is high enough that we can safely reject the null hypothesis. The t-test supports our claim that the instruction actually helped our *S*s.

ASSUMPTIONS UNDERLYING T-TESTS

Every statistical test has certain assumptions which have to be met if we plan to use them in our research. In Case I studies, we assume that there is random selection of subjects. In Case II studies, we assume that: (1) the subject is assigned to one (and only one) group in the experiment; (2) the scores on the independent variable are continuous and that there are only two levels to the variable (i.e., only two means); (3) the variances of the scores in the populations are equal, and the scores are normally distributed.

The *t*-test is a fairly robust test; so we don't have to be terribly concerned about normal distribution of the means. However, the literature in Applied Linguistics abounds with violations of the basic assumption on the number of comparisons that can be made between means using the *t*-test. If means are to be cross-compared, you *cannot* use a *t*-test. That is, you cannot compare Group 1 and 2, 1 and 3, and then 2 and 3, etc. If you try to use the *t*-test for such multiple comparisons, you make the likelihood of being able to reject the null hypothesis very easy. [You can check this out as follows: When you set the probability level at .05 and do multiple *t*-tests, the value of probability increases according to the following formula: $\alpha = 1 - (1 - \alpha)^c$. c refers to the number of comparisons. If you make four comparisons, your actual level is: $\alpha = 1 - (1 - .05)^4 = 1 - (.95)^4 = 1 - .82 = .18$. So your significance level is .18, not .05.]

The *t*-test is one of the most frequently used statistical procedures in our field. It is most often used to compare two groups. You might wonder why we go to all this trouble. Why can't we just look at the $\overline{X}$ of the experimental group and the $\overline{X}$ of the control group and see whether they look different or not? Why do we need all the rest?

Consider the last example problem. The mean time for the experimental group to find the information was 88 minutes and the mean for the control group was 102 minutes. Obviously, the experimental group was faster. Let's consider what would happen, though, if there were 10 *S*s in each group rather than 36 and 30. If you do the comparisons, you will find that the *t* value is 1.59. If we played by the rules and made a nondirectional, two-tailed hypothesis, we could not reject the null hypothesis. That is, the difference between the two means would no longer be considered great enough to allow us to cite the evidence as support for our claim about the special instruction. We cannot simply look at the mean scores of two groups and conclude that they are the same or different.

MATCHED T-TEST

In our examples so far we have compared two means obtained from two independent groups of *S*s. However, it is often the case that the two means we want to compare come from the same *S*s. For example, we may give our students a pretest and a posttest and hope to be able to compare the two means. Or we may give our *S*s two different tasks and hope to compare their performance on the tasks. This gives us paired data where each person has two

scores and we want to determine whether the difference between the two mean scores is significant.

Another instance of paired data is when *S*s have been matched on the basis of some particular variable. For example, suppose you thought it important that your two sample subject groups be matched for language proficiency. You don't trust random selection as a means of matching the groups to start with. So you select one hundred subjects and give them a language proficiency test. Out of the 100 *S*s, you then select 30 who have the same scores. Finally, you randomly assign one member of each pair to the experimental group and the other to the control group. In this case, the *S*s in your experiment are matched for one variable, language proficiency.

When you have paired data (either the same *S* and two scores or matched *S*s on one measure), you will need to use a *t*-test which is appropriate for sets of paired data.

The procedure for matched *t*-test is similar to the *t*-test for independent samples. The difference is more conceptual than computational. In the matched *t*-test, our *N* is the number of pairs rather than the number of observations. Also the standard error of the difference between means will be calculated by dividing not by the number of observations but rather by the number of pairs minus one (the degrees of freedom for pairs).

Suppose that you wanted to show that foreign students assign subject status to whatever noun most immediately precedes the verb in English sentences. This would lead them to misinterpret sentences such as *Roger promised Russ to help the teacher*, so that Russ is expected to do the helping. They will interpret sentences of the *Roger asked Russ to help the teacher* accurately. So you work out a variety of such sentences and present them to your second language learners. Then you categorize the data so that each *S* has a total score for Type 1 sentences and for Type 2 sentences. You expect that the *S*s will do better on Type 2 than Type 1 sentences. The data are shown in Table 10.1. The first step

Table 10.1. Total scores on sentence comprehension

Subject number	Type 1	Type 2	D	D^2
1	47	45	−2	4
2	50	53	3	9
3	40	44	4	16
4	38	49	11	121
5	48	48	0	0
6	41	50	9	81
7	32	45	13	169
8	31	35	4	16
9	33	30	−3	9
10	40	54	14	196
	$\Sigma X = 400$	$\Sigma X = 453$	ΣD 53	$\Sigma D^2 = 621$
	$\overline{X} = 40$	$\overline{X} = 45.3$		

is to find the difference between each pair of scores. These appear to the right under the column labeled D (for difference). These difference scores are then squared in the next column. Each column is added, and the total appears below. These values are then plugged into the matched t-test formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{D}}}$$

The top half of the t formula will, as always, give us the difference between our two obtained means. The denominator, the standard error of differences between two means, will be adjusted to account for the fact that the means are from paired data. Since the formula is adjusted for pairs, we will use the symbol $s_{\bar{D}}$ so that we will remember we are working with pairs of means.

The formula for $s_{\bar{D}}$, the standard error of differences between two means, is

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

$s_{\bar{D}}$ is the standard deviation of the differences. We can easily find it by plugging in the values from our table for differences between means.

$$s_D = \sqrt{\frac{\Sigma D^2 - (1/n)(\Sigma D)^2}{n-1}}$$

The n in the formula now refers to number of pairs (*not* number of individual observations). The standard deviation of the differences is then adjusted for the number of pairs. Our data fit into the formula as follows:

$$s_D = \sqrt{\frac{621 - (1/10)(2{,}809)}{10-1}}$$

$$= \sqrt{\frac{340.1}{9}}$$

$$= \sqrt{37.79}$$

$$= 6.15$$

We can now calculate $s_{\bar{D}}$, the standard error of differences between two means:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

$$= \frac{6.15}{\sqrt{10}}$$

$$= 1.95$$

Now we have the denominator. All we need to do is divide the difference we found between our two sentence types by the denominator to obtain the t value.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{D}}}$$
$$= \frac{40 - 45.3}{1.95}$$
$$= \frac{-5.3}{1.95}$$
$$= -2.72$$

To check the significance of this t value, we use the same t-distribution table as for the regular t-test. A value of -2.72 with 9 d.f. is significant at the .05 level. Therefore, you have evidence to support the claim that your *S*s are simply assigning subject status to the noun that immediately precedes the verb.

The second type of matched-pairs t-test involves, as we said, working with two groups of *S*s who have been previously matched. For example, suppose that you wanted to compare two different approaches to teaching spelling, one based on contrasts between the sound-symbol correspondences in the first language

Table 10.2. Gain scores on spelling test

Matched pair	Gain scores			
	Experimental	Control	D	D^2
A	5	3	2	4
B	7	7	0	0
C	2	4	−2	4
D	6	5	1	1
E	7	5	2	4
F	4	3	1	1
G	8	4	4	16
H	9	6	3	9
I	2	6	−4	16
J	6	5	1	1
	ΣX 56	ΣX 48	ΣD 8	ΣD^2 56
	n 10	n 10		
	$\bar{X}$ 5.6	$\bar{X}$ 4.8		

and English and the other based on regular rules of English spelling. Since you are concerned that random selection of *S*s may not guarantee that you will have *S*s with equal spelling abilities in your two groups, you first give a test of general spelling ability. On the basis of this test you manage to find 10 pairs of matched scores. Each pair of *S*s attained the same score on the test. Then you randomly assign one member of each pair to the experimental group and the other to the control group. This gives you 10 pairs to compare after the treatment. After the period of instruction, you again give the spelling test and calculate the gains made by each subject. The gain scores for each matched pair then form the raw data for the study; see Table 10.2. Again, our first task is to find the denominator. The formula for $s_{\bar{D}}$ is the first step:

$$s_D = \sqrt{\frac{\Sigma D^2 - (1/n)(\Sigma D)^2}{n-1}}$$

(Remember, once again, that n means number of pairs.)

$$= \sqrt{\frac{56 - (1/10)(8)^2}{9}}$$

$$= \sqrt{\frac{56 - 6.4}{9}}$$

$$= \sqrt{\frac{49.6}{9}}$$

$$= \sqrt{5.51}$$

$$= 2.35$$

We then divide this by the square root of the number of pairs:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

$$= \frac{2.35}{\sqrt{10}}$$

$$= .74$$

Now we can calculate the t value for the difference between the pairs of means:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{D}}}$$

$$= \frac{5.6 - 4.8}{.74}$$

$$= \frac{.80}{.74}$$

$$= 1.08$$

When we check the observed t value of 1.08 in the t-distribution table, we find that we need at least a value of 1.83 before we can safely reject the null hypothesis at the .05 level of significance. Therefore, we cannot reject the null hypothesis. We have to assume that the two spelling programs did not produce different results.

When you read the results section of many studies, you will find that two mean scores look quite different but turn out to have t values which are not large enough to reject the null hypothesis. Sometimes the value comes very close to the previously decided upon probability level. For example, you might decide on a .01 level as the level at which you will reject the null hypothesis for some

research you wish to do. After you have carried out the research and obtained your *t* value, you check the table of *t*-distributions. Suppose you found that your *t* value missed the critical value for .01 by 0.02. It still is significant at the .05 level. You are not allowed to change your significance level at this point. You made the decision on some reasoned basis before you did the calculations. You cannot change that decision after the fact. Many researchers try to get around this by reporting a *trend*. They usually say that their *t* value does not allow them to reject the null hypothesis but that there was a *trend* in the expected direction. Reports of trends are legitimate. They also tell us that the differences found might be important, worth our consideration, even though researchers may not want to take the chance of being wrong in rejecting the null hypothesis.

The *t*-test is an excellent statistical procedure to use in comparing two means. However, before using the *t*-test (*and* when reading research reports where the *t*-test has been used), you should first check to make certain that it is the appropriate procedure for the research. You should keep the following cautions in mind: (1) Each *S* must be assigned to one (and only one) group in the experiment if you wish to use the regular *t*-test formula. If the experiment is one which compares each *S*'s performance on two different tests, then you must use the matched *t*-test formula. (2) The scores on the independent variable should be measured on an interval scale. (3) You *must not* do multiple *t*-tests, comparing mean 1 with mean 2 and mean 1 with mean 3 and mean 2 with mean 3, etc. If you wish to make cross-comparisons, you must use the ANOVA (analysis of variance) procedure which will be discussed in the next chapter. Finally, (4) the variances of scores in the population are assumed to be equal and scores are assumed to be normally distributed.

Even though we observe these warnings, there may be problems in the use and interpretation of the *t*-test procedure. If we draw a random sample of foreign students and randomly assign them to two groups, we can assume that the two groups are from the same population. When we are doing an experiment to evaluate the effectiveness of some teaching treatment, there is no problem because we randomly select and randomly assign *S*s to the two groups. We believe that the two groups are truly the same (except for the treatment). However, if the groups are not randomly selected, we need to be certain that they are truly equivalent groups before we begin the teaching treatment. If they are neither randomly selected nor equivalent, then we cannot use a *t*-test to compare the groups following the treatment. Any differences between the groups could be due to preexisting differences. We might try to work with gain scores (pretest and posttest gains) rather than final scores to get around this problem. This, however, is risky. We know that lower-level groups will almost always make larger gains than high groups; they have more room for improvement. (In fact, many companies that guarantee results concentrate on low groups because they know that more dramatic results can be obtained there.) In any case, a *t*-test is not appropriate for such experiments unless the groups are equivalent to begin with. (A covariance procedure could be used instead.)

A final problem is not a problem with the requirements of the *t*-test itself but rather a problem in interpretation. When we want to know that two groups are (or are not) different, it is legitimate to use the *t*-test to discover the statistical probability of the difference. We often compare foreign students and native speakers using *t*-tests because we truly do not know whether the two groups will perform differently. We are not absolutely sure that foreign students will judge short stories in the same way that native speakers do; we are not sure whether they will judge the politeness of apologies in the same way, etc. However, we can be sure that they will perform differently on language tests (unless the foreign students are near-native in English proficiency). Therefore, when we run an experiment comparing native speaker and foreign student performance on some small segment of language—say verb complementation—we are bound to discover significant differences. These differences may have little to do with verb complementation but rather reflect the learn's general language problems. In other words, large differences are to be expected in such research but the differences may be as much due to intervening variables related to general language learning as to the variable tested. Strong claims must therefore be tempered by common sense in interpreting *t*-test findings.

We don't want to be overly cautious in restricting the use of *t*-test procedures. The *t*-test is one of the most useful statistical procedures (and one of the most frequently used procedures) for research in Applied Linguistics. However, it is also a procedure open to problems in interpretation. For this reason, apply caution before using it yourself and interpret the findings with care.

ACTIVITIES

Since *t*-tests are somewhat more difficult than what we've had to do so far, we'll do a step at a time. First just for *t* values for Case I studies:

1. Hypothesis: A sample of 25 reading scores having a $\overline{X}$ of 73 and an s of 7 come from a population having a μ of 78.

$$t_{obs} = \frac{\overline{X} - \mu}{s / \sqrt{n}} = \frac{\quad}{\quad} = \quad$$

d.f. = _____ t critical = _____

Can you reject the null hypothesis?

Now, let's add the requirement that you compute the standard error of the differences between the means in a Case II study:

2. You are teaching a secondary ESL class in Manila. You believe your students are a typical and random sample of high school ESL students in the Philippines. Some *S*s seem to make greater progress during the year than others, and you wonder if it has to do with type of motivation. Hypothesis: There is no difference between *S*s with instrumental motivation for language learning and *S*s with integrative motivation for language learning on language proficiency. The measure is gain scores from pre- to posttest on cloze passages. The integrative group $\overline{X}$ is 75, n = 12, s = 6. The instrumental motivation group $\overline{X}$ is 87, the group size is 10, and the s is 7. The top half of the formula is easy to do so do that first: 75 − 87 = _____. Now work out the standard error of differences between the means using the following formula:

$$s_{(\bar{X}_{inst} - \bar{X}_{integ})} = \sqrt{\left(\frac{s_{inst}}{\sqrt{n_1}}\right)^2 + \left(\frac{s_{integ}}{\sqrt{n_2}}\right)^2}$$

3. The elementary school at which you teach has a Spanish instruction program for Anglo children. The children are an ordinary, random sample of schoolchildren in the community. Some of them appear to be acquiring much more Spanish than others. You have read that a good short-term memory (STM) is very important in initial language learning. You test the children using some recognized measure of STM and then look at the students' language proficiency scores. The children who have poor language proficiency scores have a $\bar{X}$ of 11 on the STM test, the $n = 9$ and the $s = 4$. The children with high language proficiency scores have a $\bar{X}$ of 13.5, $n = 15$, and $s = 5$. Hypothesis: There is no difference in STM for the two groups.

   First, write the formula for the standard error of differences between the means, insert the data, and do the calculations. Next, write the t-test formula for comparison of two means, insert the data, and compute the t value. Note the number of degrees of freedom. Then check the observed t value with the critical value for a significance level of .05. Can you reject the null hypothesis?

   Now, let's try to work all the way from raw data to the final step. Remember that hereafter you will probably do all this on the computer. We're asking you to do it once before you let the machine do it for you.

4. You have decided to test the use of a special set of science readings instead of your regular ESL reading materials in your English class offered in the science stream of secondary school in Hong Kong. You have pretest reading scores on your students. Your control group will receive regular ESL reading materials, and you have pretest scores for them too. At the end of the year you measure their gain in reading comprehension scores. Are you justified in using a one-tailed rather than a two-tailed test?

   If you chose the .05 level of significance, what could you conclude about the effectiveness of the science readings?

Data:

Experimental Group			Control Group		
X	$X - \bar{X}$	$(X - \bar{X})^2$	X	$X - \bar{X}$	$(X - \bar{X})^2$
49			41		
32			88		
49			54		
54			50		
60			45		
41			62		
32			12		
20			63		
54			30		
67			29		

$\bar{X}_E$ 46.8 $\Sigma(X - \bar{X})^2$ _____ $\bar{X}_C$ 47.4 $\Sigma(X - \bar{X})^2$ _____

$\sqrt{\ } = s_E$ $\sqrt{\ } = s_C$

$\text{d.f.}_E =$ 9 $\text{d.f.}_C =$ 9

Formula for s:

$s_E =$ _____ $s_C =$ _____ s_D

Give the formula for standard error of differences between the means and do the computations. Then plug all the above information into the formula for the t-test.

$t_{obs} =$ _____ 18 18 $\frac{D}{s_D}$

What is the critical value for rejecting the null hypothesis at .05? At .01? Can you reject the null hypothesis at the .05 level?