CHAPTER 9

Two-Sample Tests

If you have any interest in knowing how to statistically demonstrate that there is a significant difference between your control group and your experimental group, or in the before and after effects of your educational program, then this chapter should help. In fact, the statistical tests you are about to learn are (arguably) the most common tests reported in professional journals!

In the last chapter, you learned how to evaluate hypotheses for tests when you had one sample and known population parameters. While those tests are powerful, population parameters can be difficult to obtain. Here we introduce the two-sample tests, where you will compare two samples that came from the same population, rather than comparing a single sample to a population. The samples may be completely independent from one another (between-groups design) or related in some way (within-groups design).

Independent or between-groups designs are those in which subjects are randomly selected from a population and are randomly assigned to either the control or experimental conditions. Subjects only serve in one condition.

Dependent or within-groups designs are those in which subjects are randomly selected from a population and serve in more than one condition (such as "before" vs. "after" some treatment) or subjects are matched into pairs and one subject in each pair serves in each condition. Because either the same subjects or subjects that are similar to one another in some significant way serve in the within-groups design experiments, the amount of variation due to nuisance factors is minimized in these designs. When the variability that is not of interest to the researchers is minimized, the power of the experiment is increased. Remember from Chapter 7 that power is the (highly desirable) property that measures the likelihood that a given experimental design will be able to detect a real effect if a real effect exists. Thus, within-groups designs are more powerful than between-groups designs, and we will introduce the test to analyze those designs first. We provide examples for each statistical test, so you learn how to calculate the tests, as well as how to interpret the results, what they look like in computer software output, and how you present the results for professional publications. Next, we return to the concepts of power and effect size, which are topics that are critical for interpretation of your results and are often required for publication.

Notice that with each chapter, we are now logically building the flowchart for choosing the appropriate test. The order of the chapters is meant to provide a logical extension to your statistical knowledge and to allow you to make sense of the myriad tests used to analyze data.

PAIRED *t* TEST (CORRELATED GROUPS *t* TEST)

The paired *t* test (also called the "correlated groups" *t* test) is used when you have two samples and a within-groups design. This design is also called a dependent or repeated-groups design. Both the name of the statistical test and the name of the research design can vary a great deal from book to book and between different statistical software packages. You can navigate this confusion by having a conceptual understanding of what the test is doing. This statistical test requires that you have met one of the following *experimental design* conditions:

1. You have two measures on the same subjects ("before" and "after" measures are common). See the example in Table 9.1.

or

2. You have two separate samples but the subjects in each are *individually* matched so that there are similar subjects in each group (but not the same subjects in each group). For example, you might match

subjects on age and sex, so that you have a 36-year-old woman in your control group and a 36-year-old woman in your experimental group, a 28-year-old man in your control group and a 28-year-old man in your experimental group, and so on. This can also be done by placing one identical twin in the control group and the other twin in the experimental group or by any matching of *individuals* that is an attempt (see Table 9.2). Note that the matching must be pairwise, so that you can literally compare the scores of the twins side by side. You'll see why this is important when you see the formula for the paired t test.

Table 9.1

Subject	Score Before Treatment	Score After Treatment
1	50	55
2	52	58
3	44	48
4	42	41
5	49	56

Example of "Before" and "After" Pairing Using the Same Subjects in Each "Paired" Sample

Table 9.2

Twin Pair	Twin 1 = Control Group	Twin 2 = Experimental Group
A	10	8
B	12	10
C	21	19
D	18	15

Example of a "Paired" Design in Which the Actual Subjects in Each Sample Are Different but Are "Matched" for Characteristics That They Have in Common (Genetics in This Example)

Paired *t* Test Calculation

The calculation of the paired t test statistic comes from a modification of the single-sample t test. However, now we first calculate a difference score for each pair of scores in our two samples and treat those difference scores as a single sample that will be compared to the mean difference score (μ_D) of the null hypothesis population. The mean difference score of the null hypothesis population is assumed to be zero ($\mu_D = 0$)—that is, no difference between our samples or no effect of our independent variable (see Table 9.3). The mean difference score for the paired samples is a measure of any effect of our treatment. If our treatment does not have an effect, then there will not be a difference between the two groups, and the mean difference score will be zero (or close to it), like the mean difference score of the null hypothesis population. However, if the treatment does have an effect, it will increase or decrease the scores from the control condition and therefore produce a mean difference score greater or less than zero.

Thus, we can calculate the sum of the difference scores ($\sum D$), the sum of the squared difference scores ($\sum D^2$), and the mean difference score ($\bar{D}$) and the standard deviation of the difference scores (s_D). These difference scores become our single sample of raw scores that we contrast with a null hypothesis population mean of no difference between subjects ($\mu_D = 0$), and we estimate the standard deviation of the population difference scores (σ_D) based on the sample difference scores, just as we did in the single-sample t test. Because we are using an estimate of the population difference scores based on the sample but assume that we know the population mean (= 0), we use the t distribution to evaluate our t obtained value.

For comparison, we first present the formula for the single-sample t test with which we are familiar. Then we present the modification of this formula where we have replaced the mean of the sample with the mean of the difference scores (our new sample) and the standard error of the sample means with the standard error of the mean differences. The idea is that the formula for the paired t test is really just the formula for the single-sample t

Table 9.3

Twin Pair	Twin 1 = Control	Twin 2 = Experimental	Difference Score
A	10	8	2
B	12	10	2
C	21	19	2
D	18	15	3
			$\sum D = 9$ $\bar{D} = 9/4 = 2.25$

Example of Difference Score Calculations for a Paired t Test

test if you consider the difference scores to be your "single sample." That's the secret of the paired t test. Also, remember that n in the paired t test formula refers to the number of difference scores or the number of pairs of data points, not the total number of data points.

Formula for a Single-Sample *t* Test (Review)

$$t_{\text{obtained}} = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Paired *t* Test Formula

$$t_{\text{obtained}} = \frac{\bar{D} - \mu_D}{s_{\bar{D}}} = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}} = \frac{\bar{D} - \mu_D}{\sqrt{\frac{SS_D}{n(n-1)}}}.$$

Remember that under the null hypothesis, $\mu_D = 0$, so our formula becomes

$$t_{\text{obtained}} = \frac{\bar{D}}{S_{\bar{D}}} = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}} = \frac{\bar{D}}{\sqrt{\frac{SS_D}{n(n-1)}}}.$$

A Quick Example: A One-Tailed Paired t *Test*

A researcher interested in employee satisfaction and productivity measured the number of units produced by employees at a plant before and after a company-wide pay raise occurred. The researcher hypothesized that production would be higher after the raise compared to before the raise. Assume that the difference scores are normally distributed and let $\alpha = 0.05$.

Null hypothesis: There is no difference in the number of units produced before and after the raise, or the number of units was higher before the raise.

Alternative hypothesis: The number of units produced was higher after the raise.

★ **Step 1:** Compute the probability of the mean differences of this sample given that the sample comes from the null hypothesis population of difference scores where $\mu_D = 0$.

Calculate the difference scores and the intermediate numbers for the SS formula:

Participants	Before	After	Difference Score	D^2
1	7	7	***0***	***0***
2	4	5	***−1***	***1***
3	8	9	***−1***	***1***
4	8	9	***−1***	***1***
5	6	6	***0***	***0***
6	6	6	***0***	***0***
7	5	5	***0***	***0***
8	5	4	***+1***	***1***
9	7	7	***0***	***0***
			$\sum D = -2$ $n = 9$ $\bar{D} = 2/9 = -0.2222222$	$\sum D^2 = 4$

Calculate the standard deviation of the sample:

$$SS_D = \sum D^2 - \frac{\left(\sum D\right)^2}{n} = 4 - \frac{(-2)^2}{9} = 4 - 0.4444444 = 3.5555555\cdot$$

$$s_D = \sqrt{\frac{SS_D}{(n-1)}} = \sqrt{\frac{3.5555}{9-1}} = \sqrt{0.4444375} = 0.6666614\,.$$

Apply the formula to our example:

$$t = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}.$$

$$t = \frac{-0.2222222-0}{\frac{0.6666614}{\sqrt{9}}} = \frac{-0.2222}{0.22222046} = -0.9999079\,.$$

★ **Step 2:** Evaluate the probability of obtaining this score due to chance.

Evaluate the *t*-obtained value based on alpha (α) = 0.05 and a one-tailed hypothesis. To evaluate your *t*-obtained value, you must use the *t* distribution (Table B in the Appendix) as you did in the last chapter. To determine your *t*-critical value, you need to know your alpha level (0.05), the number of tails you are evaluating (one in this case), and your degrees of freedom ($n - 1 = 9 - 1 = 8$). Compare the t-critical value with your t-obtained value. When $\alpha = 0.05$, your degrees of freedom are equal to 8, and your hypothesis is one-tailed, you should use 1.86 as your t-critical value.

To reject the null hypothesis for a t test, the t-obtained must be equal to, *or more extreme than,* the t-critical value. Be sure to also check that the effect is in the correct direction (correct based on the hypothesis).

$$|t_{\text{obtained}}| \geq |t_{\text{critical}}|$$

$|-0.99| < |-1.86|$, so we *fail to reject* the null hypothesis.

How should we interpret these data in light of the effect of the raise on productivity?

These results suggest that more than 5% of the time, you would obtain this number of units regardless of whether it was after a raise. Thus, it is likely that the difference in these production values (before and after the raise) comes from the normal null hypothesis population of difference scores. However, remember that there is a chance that there is a real effect of raises on productivity that we have not detected in this analysis.

Complete Example

A sociologist is interested in the decay of long-term memory compared to the number of errors in memory that an individual made after 1 week and after 1 year for a specific crime event. Participants viewed a videotape of a bank robbery and were asked a number of specific questions about the video 1 week after viewing it. They were asked the same questions 1 year after seeing the video. The number of memory errors was recorded for each participant at each time period. The researchers asked whether or not there was a significant difference in the number of errors in the two time periods. Assume that the difference scores are normally distributed and let $\alpha = 0.05$.

Null hypothesis: There is no difference in the number of errors made at 1 week and at 1 year.

Alternative hypothesis: There is a difference in the number of errors made at 1 week and at 1 year.

★ **Step 1:** Compute the probability of the mean differences of this sample given that the sample comes from the null hypothesis population of difference scores where $\mu_D = 0$.

Calculate the difference scores and the intermediate numbers for the SS formula:

Subject	One Week	One Year	Difference Score	D^2
1	5	7	***–2***	***4***
2	4	5	***–1***	***1***
3	6	9	***–3***	***9***
4	8	9	***–1***	***1***
5	6	6	***0***	***0***
6	5	6	***–1***	***1***
7	4	5	***–1***	***1***
8	5	4	***+1***	***1***
9	7	7	***0***	***0***
			$\sum D = -8$ $n = 9$ $\bar{D} = -8/9 = -0.8888$	$\sum D^2 = 18$

Calculate the standard deviation of the sample:

$$SS_D = \sum D^2 - \frac{\left(\sum D\right)^2}{n} = 18 - \frac{(-8)^2}{9} = 18 - 7.1111 = 10.8888.$$

$$S_D = \sqrt{\frac{SS_D}{(n-1)}} = \sqrt{\frac{10.8888}{9-1}} = \sqrt{1.36110} = 1.16666.$$

Apply the formula to our example:

$$t = \frac{\bar{D} - \mu_D}{\frac{s_D}{)n}}.$$

$$t = \frac{-0.8888 - 0}{\frac{1.16666}{\sqrt{9}}} = \frac{-0.8888}{0.388887} = -2.2855.$$

★ **Step 2:** Evaluate the probability of obtaining this score due to chance.

Evaluate the *t*-obtained value based on alpha (α) = 0.05 and a two-tailed hypothesis. To evaluate your *t*-obtained value, you must use the *t* distribution (Table B) as you did in Chapter 8. To determine your *t*-critical value, you need to know your alpha level (0.05), the number of tails you are evaluating (two in this case), and your degrees of freedom ($n - 1 = 9 - 1 = 8$). Compare the *t*-critical value with your *t*-obtained value. When $\alpha = 0.05$, your degrees of freedom are equal to 8, and your hypothesis is two-tailed, you should use 2.306 as your *t*-critical value.

To reject the null hypothesis for a two-tailed test, the absolute value of *t* obtained must be equal to, or more extreme than, the *t*-critical value.

$$|t_{\text{obtained}}| \geq |t_{\text{critical}}|$$

$|-2.2855| \leq |2.306|$, so we *fail to reject* the null hypothesis.

How should we interpret these data in light of the effect of time on the number of memory errors?

These results suggest that more than 5% of the time, you would obtain this number of memory errors regardless of whether it was after 1 week or 1 year. Thus, it is likely that these memory error differences come from the normal null hypothesis population of difference scores. However, remember that there is a chance that there is a real effect of time on memory errors that we have not detected in this analysis.

Results if you use Microsoft Excel to calculate the *t* test:

	One Week	**One Year**
Mean	5.555556	6.444444
Variance	1.777778	3.027778
Observations	9	9
Pearson Correlation	0.742315	
Hypothesized Mean Difference	0	
Df	**8**	
t Stat	**−2.28571**	
P(T<=t) one-tail	0.025804	
t Critical one-tail	1.859548	
P(T<=t) two-tail	0.051609	
t Critical two-tail	**2.306006**	

t-Test: Paired Two Sample for Means

This is the output that you get from Excel when you type in these data for 1 week and 1 year. Note that Excel calls this test "*t*-Test: Paired Two Sample for Means." This wording is slightly different from what we have been using, but it is describing the same analysis. We have bolded the numbers that are comparable to the numbers we just manually calculated or looked up in the table. Note that the *t*-obtained value (Excel calls it "t Stat") is identical to ours, except they have carried more digits (−2.28571). The "t Critical two-tail" is also the same (2.306006), but again they have carried more digits. The degrees of freedom are the same (8). But they have also provided you with the estimated probability of obtaining a *t* value (t Stat) of −2.28571, which is 0.051609. Since $0.051609 > 0.05$, we clearly cannot reject our null hypothesis. Thus, we fail to reject the null hypothesis. These results suggest that 5.1609% of the time you would obtain this number of memory errors regardless of whether it was after 1 week or 1 year. Knowing the estimated probability for each and every *t*-obtained value (not just the *t*-critical values) is one of the major advantages of using a computer to calculate your analyses.

Knowing the calculated probability in this case ($p = .052$) raises another issue: .052 is pretty close to our critical probability of .050. Our results suggest that there is a 5.16% chance that any differences were due to chance, as opposed to our critical level of 5.00%. Many researchers would present these results as "significant" even though the probability technically exceeds the critical level. The point is that there is nothing magical about the critical alpha value of 0.05, 0.01, or whatever value is chosen: Nothing special happens between $p = .052$ and $p = .049$. We have to use some common sense and knowledge of the system being studied to decide whether we are convinced that a real effect was demonstrated. This is an important point in critical thinking about statistics.

Results if you use SPSS to calculate the *t* test:

		Paired Differences					**t**	**df**	**Sig. (2-tailed)**
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	ONEWEEK—ONEYEAR	**−.8889**	**1.1667**	**.3889**	−1.7857	7.890E-03	**−2.286**	**8**	.052

Paired Samples Test

We have placed the numbers that are comparable to our manual calculations in bold. Once again, you see that calculated probability (SPSS calls Sig. 2-tailed) is greater than our alpha level of 0.05, and thus we must assume that these results could occur by chance and not necessarily as a result of time since the event.

How would these results be reported in a scientific journal article?

If the journal required American Psychological Association (APA) format, the results would be reported in a format something like this:

> There was no significant difference in the number of errors made at 1 week and at 1 year, $t(8) = -2.29$, $p > .05$.

This formal sentence includes the dependent variable (number of errors), the independent variable (1 week vs. 1 year), as well as a statement about statistical significance, the symbol of the test (t), the degrees of freedom (8), the statistical value (−2.29), and the estimated probability of obtaining this result simply due to chance (> .05).

Another Complete Example

An animal behaviorist is concerned about the effects of nearby construction on the nesting behavior (trips to nest per hour) of endangered dusky seaside sparrows in Florida. She knows that the quality of the nesting territory's habitat will also influence this nesting behavior, so she picks seven pairs of nests, each with the same territory quality (say, density of seed plants), one nest of the pair near the construction and one in an undisturbed location, for a total of 14 nest observations. Assume that the difference scores are normally distributed and let $\alpha = 0.05$.

Null hypothesis: There is no difference in the rate of nest visits made at "construction" nests and "undisturbed" nests.

Alternative hypothesis: There is a difference in the rate of nest visits between the two locations. No specific direction of the difference is suggested.

★ **Step 1:** Compute the probability of the mean differences of this sample given that the sample comes from the null hypothesis population of difference scores where $\mu_D = 0$.

Calculate the difference scores and the intermediate numbers for the SS formula:

Matched Pair	Undisturbed	Construction	Difference Score	D^2
1	5.4	3.2	***2.2***	***4.84***
2	4.1	3.5	***0.6***	***0.36***
3	9.7	7.1	***2.6***	***6.76***
4	8.4	6.8	***1.6***	***2.56***
5	6.0	6.4	***–0.4***	***0.16***
6	6.0	4.5	***1.5***	***2.25***
7	7.9	7.6	***0.3***	***0.09***
			$\sum D = 8.4$ $n = 7$ $\bar{D} = 8.4/7 = 1.2$	$\sum D^2 = 17.02$

Rate of Nest Visits Per Hour

Calculate the standard deviation of the sample:

$$SS_D = \sum D^2 - \frac{\left(\sum D\right)^2}{n} = 17.02 - \frac{(8.4)^2}{7} = 17.02 - 10.08 = 6.94.$$

$$S_D = \sqrt{\frac{SS_D}{(n-1)}} = \sqrt{\frac{6.94}{7-1}} = \sqrt{1.15667} = 1.07548.$$

Apply the formula to our example:

$$t = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}.$$

$$t = \frac{1.2-0}{\left(\frac{1.07548}{\sqrt{7}}\right)} = \frac{1.2}{0.40649} = 2.95210.$$

★ **Step 2:** Evaluate the probability of obtaining this score due to chance.

Evaluate the *t*-obtained value based on alpha (α) = 0.05 and a two-tailed hypothesis. To evaluate your *t*-obtained value, you must use the *t* distribution (Table B) as you did in Chapter 8. To determine your *t*-critical value, you need to know your alpha level (0.05), the number of tails you are evaluating (two in this case), and your degrees of freedom ($n - 1 = 7 - 1 = 6$). Compare the *t*-critical value with your *t*-obtained value. When $\alpha = 0.05$, your degrees of freedom are equal to 6, and your hypothesis is two-tailed, you should use ±2.447 as your *t*-critical value.

To reject the null hypothesis for a two-tailed test, the absolute value of *t* obtained must be equal to, or more extreme than, the *t*-critical value.

$$|t_{\text{obtained}}| \geq |t_{\text{critical}}|$$

$|2.952| \geq |2.447|$, so we *reject* the null hypothesis.

How should we interpret these data in light of the effect of construction on the rate of nest visits?

These results suggest that *less than 5%* of the time, you would obtain this rate of nest visits regardless of whether it was near or not near to the construction site. Thus, it is likely that the rate of nest visits near the construction site *does not* come from the same underlying population of scores as the nest site visits away from the construction site, and therefore, the difference scores in this example do not represent a null hypothesis population of difference scores. However, remember that there is a chance (however small) that there is, in reality, no real effect of the construction on nest site visits, and our conclusion is in error.

Results if you use Microsoft Excel to calculate the *t* test:

	Undisturbed	**Construction**
Mean	6.785714	5.585714
Variance	3.784762	3.284762
Observations	7	7
Pearson Correlation	0.838487	
Hypothesized Mean Difference	0	
Df	**6**	
t Stat	**2.952067**	
P(T<=t) one-tail	0.012772	
t Critical one-tail	1.943181	
P(T<=t) two-tail	0.025544	
t Critical two-tail	**2.446914**	

t-Test: Paired Two Sample for Means

This is the output that you get from Excel when you type in these data for undisturbed and construction sites. Once again, we have bolded the numbers that are comparable to the numbers we just manually calculated or looked up in the table. Note again that the *t*-obtained value is identical to ours, except they have carried more digits. The "t Critical two-tail" is also the same, but again they have carried more digits. The degrees of freedom are the same (6). But they have also provided the calculated probability of obtaining a *t* value (t Stat) of 2.952067, which is 0.025544. Since $0.025544 < 0.05$, we can reject our null hypothesis. Thus, we reject the null hypothesis. These results suggest that 2.5544% of the time, you would be making a mistake in concluding that there was a "real effect" of the construction on nest site visits. Traditionally, we are willing to accept a 5% probability of making such a mistake. Again, knowing the calculated probability for each *t*-obtained value (not just the *t*-critical values) is one of the major advantages of using a computer to calculate your analyses.

Results if you use SPSS to calculate the *t* test:

		Paired Differences					**t**	**df**	**Sig. (2-tailed)**
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	UNDISTUR—CONSTRUC	**1.2000**	**1.0755**	**.4065**	.2053	2.1947	**2.952**	**6**	.026

Paired Samples Test

Once again, we have placed the numbers that are comparable to our manual calculations in bold. And once again, you see that calculated probability (.026: SPSS calls it "Sig. 2-tailed") is less than our alpha level of 0.05, and thus we must assume that these results are unlikely due to chance and thus are more likely a result of time since the event.

How would these results be reported in a scientific journal article?

If the journal required APA format, the results would be reported in a format something like this:

> There was a significant difference in the rate of nest visits between the undisturbed location ($M = 6.79$, $SD = 1.94$) and the construction location ($M = 5.59$, $SD = 1.81$), $t(6) = 2.95$, $p = .03$.

This formal sentence includes the dependent variable (rate of nest visits), the independent variable (two different locations), the direction of the effect as evidenced by the reported means, as well as a statement about statistical significance, the symbol of the test (t), the degrees of freedom (6), the statistical value (2.95), and the estimated probability of obtaining this result simply due to chance (.03).

When is it appropriate to use a paired *t* test?

1. When you have two samples and a within-groups design
2. When the sampling distribution is normally distributed, which, as you should recall, is satisfied when
   a. The sample size is greater or equal to 30 ($n \geq 30$) or
   b. The null hypothesis population is known to be normally distributed.
3. When the dependent variable is on an interval or ratio scale

INDEPENDENT *t* TEST

In the previous section, we described a situation where your two conditions contain either the same subjects or subjects that have been individually matched on an important characteristic that might potentially influence the outcome of your results and is not interesting to you (age or body weight are potential examples of characteristics that you might match subjects on). This is referred to as a within-groups design. Now we turn to the situation where you actually have two completely different (independent) groups of subjects that you want to compare to determine if they are significantly different from one another: a between-groups design. The classic example of this is when you have a sample and you randomly assign half of your subjects to the control condition and the other half to the experimental treatment condition. In this situation, we wish to compare the means of the two conditions/groups. We can no longer assume that we know a population mean (as we did when we assumed that the $\mu_D = 0$ in the paired t test), and we must develop a new sampling distribution.

Sampling Distribution of the Difference Between the Means

To test for the potential statistical significance of a true difference between sample means, we need a sampling distribution of the difference between sample means $(\bar{X}_1 - \bar{X}_2)$. This would be a sampling distribution that will provide us with the probability that the difference between our two sample means $(\bar{X}_1, \bar{X}_2)$ differs from the null hypothesis population of sample mean differences: a population in which there is no difference between samples or, restated, the independent variable has no effect. The sampling distribution of the difference between the means can be created by taking all possible sample sizes of n_1 and n_2, calculating the sample means, and then taking the difference of those means. If you do this repeatedly for all of the possible combinations of your sample sizes, then you end up with a family of distributions of differences between the two means when they are randomly drawn from the same null hypothesis population. Choice of the specific distribution to be used in a problem depends on the degrees of freedom, as always.

The sampling distribution of the difference between sample means has the following characteristics:

1. If the null hypothesis population of scores is normally distributed, then the population of differences between the sample means will also be normally distributed.
2. The mean of the sampling distribution of the difference between sample means $\left(\mu_{\bar{X}_1 - \bar{X}_2}\right)$ will be equal to $\mu_1 - \mu_2$ (just as $\mu_{\bar{X}} = \mu$).
3. The standard deviation of the sampling distribution of the difference between sample means will be equal to the square root of the sum of each sample variance, or $\sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}}$.

Estimating Variance From Two Samples

Consider for a moment that the formulas we've used in the past contained one measure of variability. In the case of the single-sample z test, we used the population standard deviation, while with the single-sample t test, we estimated the population standard deviation with the sample standard deviation. When we performed the paired t test earlier in this chapter, we calculated the standard deviation of the sample of difference scores, so we were able to come up with one measure of sample variability. Now we have two independent samples, and each has a sample standard deviation. This forces us to determine which standard deviation best represents the true variability. In fact, what we do is to estimate the true population variability (or variance, σ^2) by taking the average variance (s^2) of our samples but weighted by their respective sample sizes. Remember, as we learned in earlier chapters, sample size or degrees of freedom affects the accuracy of our variance estimates, so an estimate from a sample with a large sample size would be more accurate than an estimated variance from a smaller sample. So we need to weight our average variance by the respective sample sizes of each sample. In using this approach, we are going to make a new assumption—that the sample variances are estimating the same underlying population variance, the variance of the null hypothesis population. Later in this chapter, we will have to make sure that our two sample variances are the same, within the bounds of random sampling error. This is referred to as the homogeneity of variance assumption.

Formula for Weighted Variance:

$$s_w^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}.$$

Substituting the equation for degrees of freedom and for variance:

$$= \frac{(n_1 - 1)(SS_1/n_1 - 1) + (n_2 - 1)(SS_2/n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}.$$

Rearranging to simplify:

$$= \frac{SS_1 + SS_2}{n_1 + n_2 - 2}.$$

We have shown the formula in three different ways. The first way is the most intuitive way to present the average variance of the two samples when it is weighted by the sample size or, more specifically, by the appropriate degrees of freedom for each sample. The degrees of freedom are used because we are estimating the population variation from a sample, and thus one degree of freedom is lost each time we do that (one for each sample). The second formula actually plugs in the appropriate formulas for variance and degrees of freedom into the first formula, and the last formula is created by algebraic rearrangement into a simplified version. The first formula may be the best one to use if you are obtaining each sample variance from your calculator directly from these raw data or if you are given either variance or standard deviation of each sample in a problem. The last formula would be best if you have already calculated the sums of squares (SS) for each group.

Derivation of the Independent *t* Test Formula From the Single-Sample *t* Test Formula

Formula for a single-sample *t* test (for review):

$$t_{\text{obtained}} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Conceptual formula for an independent *t* test:

$$t_{\text{obtained}} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1 - \bar{X}_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1 - \bar{X}_2}}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}.$$

Calculation (practical) formula to use for an independent *t* test:

$$t_{\text{obtained}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_w^2(1/n_1 + 1/n_2)}}.$$

The formula for an independent *t* test is derived by assuming the mean of the sampling distribution or differences between means is zero for the null hypothesis population and by using the average variance divided by each sample size. The square root in the denominator of the independent *t* test formula is there not only to take the square root of the sample size (as you did when you calculated a single-sample *t* test) but also because you are working in squared units (variance), and we must take the square root of the variance to get back to the standard deviation.

Variations of the Independent *t* Test Formula

All-purpose formulas:

$$t_{\text{obtained}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_w^2(1/n_1 + 1/n_2)}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)(1/n_1 + 1/n_2)}}.$$

To create the second variation of the formula, we simply substituted the formula for s_w^2 directly into the independent *t* test formula.

Formula to be used only when $n_1 = n_2$:

$$t_{\text{obtained}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{SS_1 + SS_2}{n(n-1)}\right)}}.$$

A Quick One-Tailed Example of an Independent t Test

A researcher breeds rats for nine generations but only breeds the rats that perform very well in a maze (few errors) to each other (called "maze-bright" rats) and also breeds rats that perform very poorly in a maze (many errors) to one another ("maze-dull" rats). After nine generations, is there a significant decrease in the number of errors of the maze-bright rats?

Null hypothesis: There is no difference in the number of errors made by the maze-bright and the maze-dull rats or there are more errors in the maze-bright rats.

Alternative hypothesis: There is a significant decrease in the number of errors of the maze-bright rats.

★ **Step 1:** Compute the probability that each of the sample means comes from the null hypothesis population of differences between means.

Calculate the means and the intermediate numbers for the SS formula:

Maze-Bright Rats	Maze-Dull Rats
2	6
3	4
4	5
3	3
4	6
2	
$\sum X_{\text{BRIGHT}} = 18$	$\sum X_{\text{DULL}} = 24$
$\left(\sum X_{\text{BRIGHT}}\right)^2 = 18^2 = 324$	$\left(\sum X_{\text{DULL}}\right)^2 = 24^2 = 576$
$\sum X^2_{\text{BRIGHT}} = 58$	$\sum X^2_{\text{DULL}} = 122$
$\bar{X}_{\text{BRIGHT}} = \frac{18}{6} = 3$	$\bar{X}_{\text{DULL}} = \frac{24}{5} = 4.8$

Number of Errors in a Maze by Ninth-Generation Rats

Calculate the SS of each sample:

$$SS_{\text{BRIGHT}} = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 58 - \frac{(18)^2}{6} = 58 - \frac{324}{6} = 4.$$

$$SS_{\text{DULL}} = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 122 - \frac{(24)^2}{5} = 122 - \frac{576}{5} = 6.8.$$

Apply the SS formula of the independent t test to our example:

$$t_{\text{obtained}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)(1/n_1 + 1/n_2)}}.$$

$$t = \frac{3 - 4.8}{\sqrt{\left(\frac{4 + 6.8}{6 + 5 - 2}\right)(1/6 + 1/5)}} = \frac{-1.8}{\sqrt{(1.2)(.3666)}} = -2.714.$$

★ **Step 2:** Evaluate the probability of obtaining this score due to chance.

Evaluate the t-obtained value based on alpha (α) = 0.05 and a one-tailed hypothesis. To evaluate your t-obtained value, you must use the t distribution (Table B). To determine your t-critical value, you need to know your alpha level (0.05), the number of tails you are evaluating (one in this case), and your degrees of freedom (*df*). The degrees of freedom for an independent t test are $(n_1 - 1) + (n_2 - 1)$ or $n_1 + n_2 - 2$. Thus, the *df* for this problem are $6 + 5 - 2 = 9$. Compare the t-critical value with your t-obtained value. When $\alpha = 0.05$, your degrees of freedom are equal to 9, and your hypothesis is one-tailed, you should use 1.833 as your t-critical value.

$-2.7414 > -1.833$, so we *reject* the null hypothesis.

How should we interpret these data in light of the effect of time on the number of memory errors?

These results suggest that less than 5% of the time, you would obtain this difference in the number of errors if the breeding had no effect. Thus, it is not very likely that these error differences come from the normal null hypothesis population. However, there is a chance that you could get a difference this large between two means that is purely due to chance, but that chance is less than 5%. Note that you can refer back to the means to determine that the effect was in the correct direction. Specifically, maze-bright rats made 3 errors on average and maze-dull rats made 4.8 errors on average. Not only is the *t*-obtained value more extreme than the *t*-critical value, but the direction of the effect is as the researcher predicted.

Complete Example

A researcher breeds rats for nine generations but only breeds the rats that perform very well in a maze (few errors) to each other (maze-bright rats) and also breeds rats that perform very poorly in a maze (many errors) to one another (maze-dull rats). After nine generations, is there a significant difference in the number of errors in the two groups of rats?

Null hypothesis: There is no difference in the number of errors made by the maze-bright and the maze-dull rats or the differences are due to chance.

Alternative hypothesis: There is a difference in the number of errors made by each group.

Evaluate the *t*-obtained value based on alpha (α) = 0.05 and a two-tailed hypothesis. To evaluate your *t*-obtained value, you must use the *t* distribution (Table B). To determine your *t*-critical value, you need to know your alpha level (0.05), the number of tails you are evaluating (two in this case), and your degrees of freedom (*df*). The degrees of freedom for an independent *t* test are $(n_1 - 1) + (n_2 - 1)$ or $n_1 + n_2 - 2$. Thus, the *df* for this problem are $6 + 5 - 2 = 9$. Compare the *t*-critical value with your *t*-obtained value. When $\alpha = 0.05$, your degrees of freedom are equal to 9, and your hypothesis is two-tailed, you should use 2.262 as your *t*-critical value.

$|-2.714| \geq |2.262|$, so we *reject* the null hypothesis.

How should we interpret these data in light of the effect of time on the number of memory errors?

These results suggest that *less than 5%* of the time, you would obtain this difference in the number of errors if the breeding had no effect. Thus, it is not very likely that these error differences come from the normal null hypothesis population. However, there is a chance that you could get a difference this large between two means that is purely due to chance, but that chance is less than 5%.

Results if you use Microsoft Excel to calculate the *t* test:

	Bright	**Dull**
Mean	**3**	**4.8**
Variance	0.8	1.7
Observations	**6**	**5**
Pooled Variance	1.2	
Hypothesized Mean Difference	0	
Df	**9**	
t Stat	**–2.713602101**	
P(T<=t) one-tail	0.011928192	
t Critical one-tail	1.833113856	
P(T<=t) two-tail	0.023856384	
t Critical two-tail	**2.262158887**	

t-Test: Two-Sample Assuming Equal Variances

This is the output you get from Excel when you type in these data for maze-bright and maze-dull rats. Note that Excel calls this test "*t*-Test: Two-Sample Assuming Equal Variances." This wording is slightly different from what we have been using, but it is describing the same analysis, and that will be even clearer after we have discussed the assumptions of the independent *t* test. We have bolded the numbers that are comparable to the numbers we just manually calculated or looked up in the table. Note that the *t*-obtained value (Excel calls it "t Stat") is identical to ours, except they have carried more digits (–2.713602101). The "t Critical two-tail" is also the same (2.262158887), but again they have carried more digits. The degrees of freedom are the same (9). In addition, they have provided the calculated probability of obtaining a *t* value (t Stat) of –2.713602101, which is 0.023856384. Since 0.023856384 **<** 0.05, we clearly *reject* our null hypothesis. These results suggest that 2.3856384% of the time, you would obtain this difference by chance if breeding had no effect. Again, knowing the calculated probability for each and every *t*-obtained value (not just the *t*-critical values) is one of the major advantages of using a computer to calculate your analyses.

Results if you use SPSS to calculate the *t* test:

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
ERRORS	Equal variances assumed	1.255	.292	**–2.714**	**9**	.024	**–1.8000**	.6633	–3.3005	–.2995
	Equal variances not assumed			–2.616	6.903	.035	–1.8000	.6880	–3.4315	–.1685

Independent Samples Test

We have placed the numbers that are comparable to our manual calculations in bold. Once again, you see that calculated probability (SPSS calls it Sig. 2-tailed) is less than our alpha level of 0.05, and thus we must assume that these results would be *unlikely* simply due to chance.

How would these results be reported in a scientific journal article?

If the journal required APA format, the results would be reported in a format something like this:

> There was a significant difference in the number of errors made by maze-dull (M = 4.8, SD = 1.30) and maze-bright rats (M = 3.0, SD = 0.89), $t(9) = -2.71$, p = .02.

This formal sentence includes the dependent variable (number of errors), the independent variable (two groups of rats), as well as a statement about statistical significance, the direction of the effect as evidence by the means, the symbol of the test (t), the degrees of freedom (9), the statistical value (–2.71), and the estimated probability of obtaining this result simply due to chance (.02).

Another Complete Example

Psychology experiments frequently need to be very careful about the effects of past experience on the outcome of their studies. One common population for psychology research is college students enrolled in psychology courses. But it is traditional to use only freshmen or sophomores who are enrolled in introductory courses so that the researcher avoids the possibility of a student anticipating the study's goals after having had more advanced psychology courses. But does classroom experience in psychology really affect the outcome of the experiment? A researcher develops a simple assessment for response to visual stimuli, and the number of correct (matching) results is scored for a group of five sophomore introductory psychology course students and six psychology seniors. Is there a significant difference in the scores of the two groups of students?

Null hypothesis: There is no difference in the scores of introductory and advanced psychology students or the differences are due to chance.

Alternative hypothesis: There is a difference in the scores of each group.

★ **Step 1:** Compute the probability that each of the sample means comes from the null hypothesis population of differences between means.

Calculate the means and the intermediate numbers for the SS formula:

Introductory	Advanced
9	8
8	9
7	8
9	6
6	10
9	
$\sum X_{\text{Intro}} = 39$	$\sum X_{\text{Adv}} = 50$
$(\sum X_{\text{Intro}})^2 = 39^2 = 1521$	$(\sum X_{\text{Adv}})^2 = 50^2 = 2500$
$\sum X^2_{\text{Intro}} = 311$	$\sum X^2_{\text{Adv}} = 426$
$\bar{X}_{\text{Intro}} = \frac{39}{5} = 7.80$	$\bar{X}_{\text{Adv}} = \frac{50}{6} = 8.33$

Scores for Introductory and Advanced Students

Calculate the SS of each sample:

$$SS_{\text{Intro}} = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 311 - \frac{(39)^2}{5} = 311 - \frac{1521}{5} = 6.80.$$

$$SS_{\text{Adv}} = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 426 - \frac{(50)^2}{6} = 426 - \frac{2500}{6} = 9.33.$$

Apply the SS formula of the independent t test to our example:

$$t_{\text{obtained}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)(1/n_1 + 1/n_2)}}.$$

$$t = \frac{7.80 - 8.33}{\sqrt{\left(\frac{6.80 + 9.33}{5 + 6 - 2}\right)(1/5 + 1/6)}} = \frac{-0.53}{\sqrt{(1.79222)(.36667)}} = -0.6538.$$

★ **Step 2:** Evaluate the probability of obtaining this score due to chance.

Evaluate the t-obtained value based on alpha (α) = 0.05 and a two-tailed hypothesis. To evaluate your t-obtained value, you must use the t distribution (Table B). To determine your t-critical value, you need to know your alpha level (0.05), the number of tails you are evaluating (two in this case), and your degrees of freedom (*df*). The degrees of freedom for an independent t test are $(n_1 - 1) + (n_2 - 1)$ or $n_1 + n_2 - 2$. Thus, the *df* for this problem are $5 + 6 - 2 = 9$. Compare the t-critical value with your t-obtained value. When $\alpha = 0.05$, your degrees of freedom are equal to 9, and your hypothesis is two-tailed, you should use 2.262 as your t-critical value.

$|-0.6538| < |2.262|$, so we *fail to reject* the null hypothesis.

How should we interpret these data in light of the effect of experience on scores?

These results suggest that *greater than 5%* of the time, you would obtain this difference in the scores if experience level had no effect. Thus, it is likely that these differences come from the normal null hypothesis population. In terms of our experiment, freshmen and seniors react the same way: Classroom experience does not have an effect on this response to visual stimuli. However, there is always a chance that there is a significant real effect that you did not detect.

Results if you use Microsoft Excel to calculate the *t* test:

	Introductory	Advanced
Mean	**7.800000**	**8.333333**
Variance	1.700000	1.866667
Observations	**5**	**6**
Pooled Variance	1.792593	
Hypothesized Mean Difference	0	
Df	**9**	
t Stat	**–0.657843**	
P(T<=t) one-tail	0.263552	
t Critical one-tail	1.833114	
P(T<=t) two-tail	0.527105	
t Critical two-tail	**2.262159**	

t-Test: Two-Sample Assuming Equal Variances

This is the output you get from Excel when you type in these data for introductory and advanced psychology students. We have again bolded the numbers that are comparable to the numbers we just manually calculated or looked up in the table. Note again that the *t*-obtained value is identical to ours, except they have carried more digits. The "t Critical two-tail" is also similar to the table except for differences due to rounding (2.262158887), and the same as in the first example since the degrees of freedom are the same (9). In addition, they have provided the calculated probability of obtaining a *t* value (t Stat) of –0.657843, which is 0.527105. Since 0.527105 > 0.05, we clearly *fail to reject* our null hypothesis. These results suggest that 52.7105% of the time, you would obtain this difference in errors if psychology experience had no effect.

Results if you use SPSS to calculate the *t* test:

		Levene's Test for Equality of Variances		**t-test for Equality of Means**						
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
DATA	Equal variances assumed	.008	.929	**–.658**	**9**	.527	–.5333	.8107	–2.3673	1.3007
	Equal variances not assumed			–.661	8.785	.526	–.5333	.8069	–2.3655	1.2989

Independent Samples Test

We have placed the numbers that are comparable to our manual calculations in bold. Once again, you see that that calculated probability (SPSS calls it Sig. 2-tailed) is greater than (*much* greater than) our alpha level of 0.05, and thus we must assume that these results would be *likely* simply due to chance.

How would these results be reported in a scientific journal article?

If the journal required APA format, the results would be reported in a format something like this:

> There is no significant difference between the scores of introductory and advanced psychology students, $t(9) = -.66$, $p = .53$.

This formal sentence includes the dependent variable (scores), the independent variable (introductory vs. advanced psychology students), as well as a statement about statistical significance, the symbol of the test (t), the degrees of freedom (9), the statistical value (–.66), and the probability of obtaining this result simply due to chance (.53).

When is it appropriate to use an independent *t* test?

1. When you have two samples and a between-groups design
2. When the sampling distribution is normally distributed. Again, this is satisfied when
   a. The sample size is greater or equal to 30 ($N \geq 30$) or
   b. The null hypothesis population is known to be normally distributed.
3. When the dependent variable is on an interval or ratio scale
4. When the variances of the two groups are the same or are homogeneous. The homogeneity of variance assumption (HOV) requires that the variances of the underlying populations are equal or, in practical terms, not significantly different from one another. We can test this assumption by examining the sample variances. For now, we will use a simple rule of thumb to decide whether the variances are similar enough to be considered homogeneous: If the $\frac{\text{larger variance}}{\text{smaller variance}} \leq 4$, then the HOV assumption is met (okay) and you can proceed with the independent *t* test. Note that there are more formal ways to test this assumption besides this rule of thumb. Interested students should explore information on Levene's test for more information. However, if the assumption is not met, you cannot proceed with the independent *t* test without some corrections or alterations. One possible solution is to run an unequal variances test rather than an independent *t* test. For example, Welch's *t* test is an adaptation of Student's *t* test intended for use with two samples having possibly unequal variances.

POWER AND TWO-SAMPLE TESTS: PAIRED VERSUS INDEPENDENT DESIGNS

Because high sample variability decreases power, and more variability would be expected from between-groups designs where different individuals are in each condition or group, the independent *t* test is less powerful than the paired *t* test. The paired *t* test gains power by controlling nuisance factors such as the age and temperament of a subject since the subjects are either the same in both conditions or match on characteristics that might influence the outcome of the study but are not of interest to the researcher. Thus, the researcher is more likely to control variability in the within-groups design.

There are several major drawbacks to the within-groups (or paired) design. One occurs when subjects drop out of the experiment prior to experiencing all of the conditions. This creates a situation where the researcher has one data point for the subject but lacks the other data point and thus cannot calculate a difference score for the subject. There are two options in this situation. The first option is to drop those subjects with missing data from your analysis, but this reduces your power significantly if you have a large number of missing data points relative to your sample size. The second option is to treat the two samples as independent and calculate an independent t test, and thus equal ns are not required and difference scores are not the basis for the test.

A second major drawback of the within-groups design is what generally is called experience effects: any research design in which novelty or experience with the assessment tool would bias the results. For example, any study in which the participants must be naive to the assessment to be able to provide a meaningful response would preclude the use of the within-groups design. You would require that both your control or pretest group and your treatment group be different individuals who had never before experienced the assessment itself.

EFFECT SIZES AND POWER

Effect size is a standardized measure of the difference between two (or more) group means; it is the difference in means divided by the shared standard deviation of two or more groups. If testing a sample with no known population parameters, you can estimate the effect size by using Hedges's g:

$$\text{Hedges's } g = \frac{\bar{X}_1 - \bar{X}_2}{S_p}.$$

Where S_p (when sample sizes are equal) $= \sqrt{\dfrac{s_x^2 + s_y^2}{2}}$.

Where S_p (when sample sizes are unequal) $= \sqrt{\dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}}$.

Once you have calculated your effect size, you can use a free program to calculate the power of your test. One that we have often recommended to students is "G*Power," which can be downloaded at www.psycho.uni-duesseldorf.de/aap/projects/gpower/.

The program can be used a priori to predict adequate sample size, based on desired effect size and desired power, or post hoc to calculate power based on calculated effect size and standard error.

A priori, after typing in the desired effect size (based on previous research or current hypotheses) and desired power (usually 0.8 or your research won't be funded), total sample size, along with t_{crit} and actual power, appears in the middle and bottom of the screen.

Post hoc power is determined by inputting effect size and sample size and hitting "Calculate." This process is especially useful after you have completed a preliminary study and are proposing a full study. Given the effect size of your pilot study, you can predict how many subjects are necessary to obtain significant results with a specific likelihood (e.g., 80%).

Other Resources for Calculating Power

If you have an experimental design that does not fit with the options for calculating power with G*Power, you can use the following website to find more free resources on the Internet for calculating statistics, including power: http://statpages.org/.

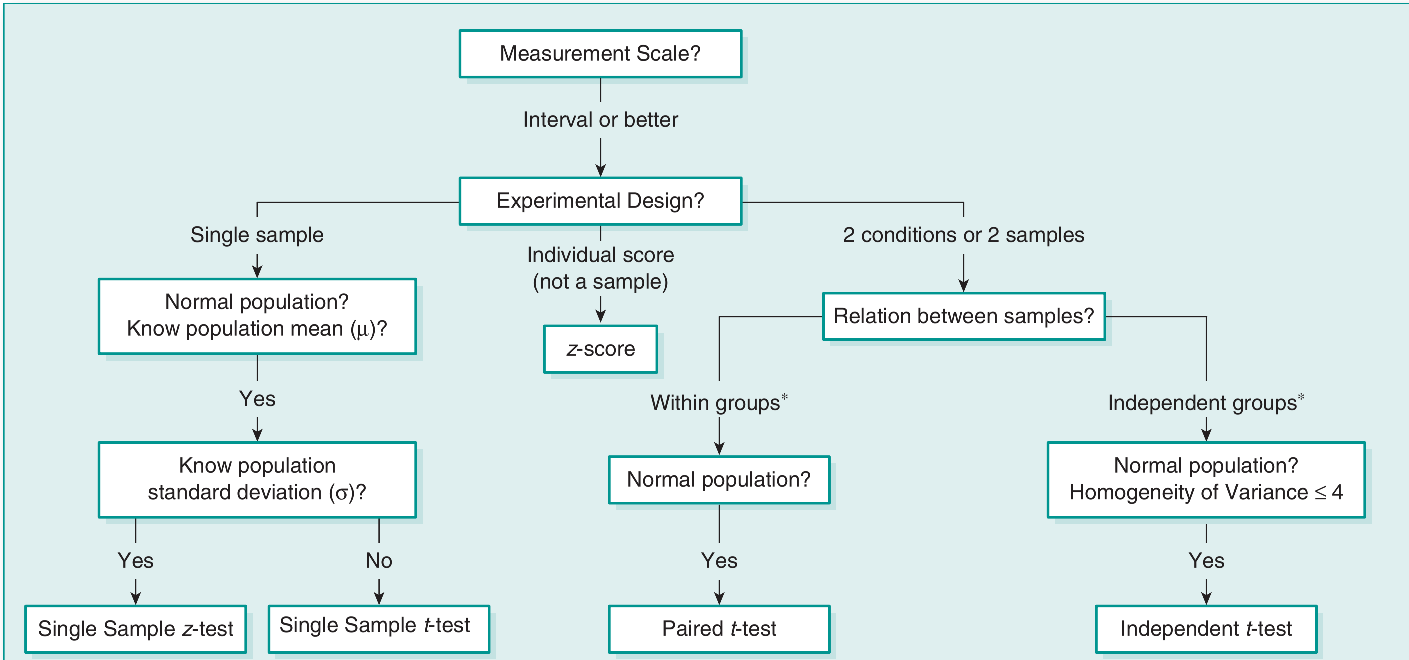
Notice that the flowchart (see page 175) does not yet tell you what to do if you do not meet the assumptions of your statistical tests. In fact, there are some options when you have two samples, but you do not meet the assumptions for parametric tests. The paired *t* test can be replaced with a nonparametric test called the Wilcoxon signed ranks test, while the independent *t* test can be replaced with a nonparametric test called the Mann-Whitney *U* test. We will discuss these nonparametric options in greater detail in Chapter 14. Another option is to perform a log transformation on your data to determine if the mathematical transformation will result in a normal distribution.

SUMMARY

In this chapter, you have covered statistical tests for designs in which you have two samples and lack information about the underlying population. Because we rarely know population parameters such as the mean and standard deviation (μ and σ), the two-sample tests tend to be used more often than the one-sample tests. Two-sample tests are less powerful because we are forced to estimate characteristics of the population, whereas one-sample tests rely on known measures of the population. In the next chapter, we will extend the concept of an independent *t* test or two-sample between-groups design to a situation where we have three samples and a between-groups design.

Excel Step-by-Step: Step-by-Step Instructions for Using Microsoft Excel 2003 or 2007 to Run *t* Tests

1. Your first step will be to open Microsoft Excel and type the raw data into a spreadsheet (data listed on page 176). It is helpful to type the column headers so that your output will be labeled later. Note that the participants who received caffeine in the first treatment are different



1. Within Groups
   a. subjects matched—married, partners, twins, etc., or individually matched for a reason that might affect the I.V.
   b. before vs. after conditions with the same subjects, measured both times
2. Independent Groups
   a. In one group/condition, NOT allowed in the other, subjects randomly assigned to be in the experiment in one group OR the other
   b. Sample sizes can be unequal (IF unequal, definitely independent *t*-test, but if equal, make decision based on experimental design info)
3. Power
   a. Single z > Single t, Paired > Independent

Overview of Single-Sample and Two-Sample Tests: Expanded Flowchart from the participants who received caffeine in the second treatment (between-groups design).

Caffeine First Times	Caffeine Second Times
30	18
32	15
30	17
33	20
32	18

2. Once your data are entered, you will need to calculate an independent *t* test.

For Excel 2003: To find the *t* test you will need, you can go to the built-in data analysis function. You'll find the option under the "Tools" menu, and at the bottom of the list that pops up, you should see "Data Analysis."

If you do not see the "Data Analysis" option under the Tools menu, select "Add Ins" under the Tool menu. Check the box next to "Analysis TookPak" and click OK. Follow any further instructions that the computer gives you.

For Excel 2007: To get to the data analysis option, click on the "DATA" tab and the "Data Analysis" tool will be in the "Analysis" section to the far right of the screen. Once you find the Data Analysis tool, the rest of the instructions are the same.

3. After you click on "Data Analysis," a list of possible statistical tests will pop up. Page down the list until you find the appropriate test. We want to do an independent *t* test, which Excel calls "*t*-Test: Two Sample Assuming Equal Variances." After you have selected the appropriate test, the program will take you through the steps to complete the test. Note that Excel also has an unequal variances test. You can use this test if you meet the assumptions of normality but do not meet the variance assumption.

Here is example output from an independent *t* test:

	Control Group	Alcohol Group
Mean	2.7	4
Variance	1.566666667	1.555555556
Observations	10	10
Pooled Variance	1.561111111	
Hypothesized Mean Difference	0	

	Control Group	Alcohol Group
df	18	
t Stat	–2.326544946	
P(T<=t) one-tail	0.015932925	
t Critical one-tail	1.734063062	
P(T<=t) two-tail	0.031865849	
t Critical two-tail	2.100923666	

t-Test: Two-Sample Assuming Equal Variances

SPSS Step-by-Step: Step-by-Step Instructions for Using SPSS to Run an Independent *t* Test

1. Your first step will be to open SPSS and select the option that allows you to type in new data.
2. This will open a page called "Variable View." To confirm that, look at the tab at the bottom left of the page. There should be two tabs, and one will say "Variable View" (the one you are in now) and the other will say "Data View."
3. Now you need to establish your variables for SPSS. Make a variable that codes for your independent groups by typing in the word *group* in the first box of row 1. Now name your dependent variable and call it *dv* and type that into box 1 in row 2. By default, SPSS will consider each of these variables to be numeric, and for these purposes, all of the default codes will work perfectly. However, keep in mind that this is where you can change some of your options to allow for alphabetical data, define coding in your variables, and so on. For example, you could define your group coding to be 1 for the control group and 2 for the experimental treatment group.
4. Click on the Data View tab now. You should see that the variable names you entered in Variable View have now appeared at the top of this spreadsheet. Now you can enter your raw data:

Group	Dv
2	18
2	15

(Continued)

(Continued)

Group	Dv
2	17
2	20
2	18
1	30
1	32
1	30
1	33
1	32

5. From the SPSS menu, you should now select "Analyze," then "Compare means," and finally "Independent t-test." This will open up a new pop-up window with your variables listed on the left-hand side. Select your dependent variable and use the arrow in the middle of the pop-up to move it to the right-hand side of the pop-up box as your "test variable." Now move your "group" variable the same way over to the "group variable" box.

6. Now you will need to define the groups for your grouping variable. Click this option and type 1 for Group 1 and 2 for Group 2.

7. Once your groups are defined and you are back to the independent *t* test pop-up box, you can just hit "OK" to run the test.

8. You should get output for an independent *t* test assuming equal variances and for an independent *t* test where you do not need to assume equal variances. SPSS will also include Levene's test, which tells you whether or not your variances are homogeneous. If they are homogeneous, you should report the independent *t* test assuming equal variances. If your variables are not homogeneous and are heterogeneous, then you report the unequal variances *t* test. In addition to the *t* value and significance level (*p* value), the output will also include descriptive statistics such as the difference between your means, standard error of the means, your degrees of freedom, and the confidence interval.

CHAPTER 9 HOMEWORK

Provide a short answer for the following questions.

1. Why are paired (or correlated) designs more powerful than independent designs?
2. What are the assumptions for the paired t test?
3. What are the assumptions for the independent t test?
4. List three ways that you can meet the assumption of normality.
5. Why does the independent t test require the assumption of homogeneity of variance but the paired t test does not require this assumption?
6. How should you handle a situation where you have paired design and two conditions but there are some missing data for one of your two conditions?
7. What do you do if you do not meet the assumptions of the paired t test or the independent t test?
8. A researcher records the number of positive early childhood memories that can be recalled by five individuals who grew up in military families to the number of memories of individuals who grew up in nonmilitary families. The number of memories is normally distributed in each group. Using $\alpha = 0.05$ (two-tailed), what do you conclude?

Military family	18	25	17	20	23
Nonmilitary family	20	23	26	30	28

9. A sociologist is interested in whether or not race affects the likelihood that the average person will "shoot" a potential criminal in a computer simulation. Participants are required to make quick decisions about whether to "shoot" or not, and they are shown a variety of images of people. Some of the images are of people with a weapon and some of them are people holding nonviolent objects. Eight participants are randomly sampled for the study. The psychologist records the number of errors (shooting someone holding a nonviolent object) the participants made based on race (African American or Caucasian). The number of errors is normally distributed. The following data are recorded.

Participant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
African American	28	29	25	30	25	27	28	24
Caucasian	25	28	22	30	26	24	25	22

Using $\alpha = 0.05$ (two-tailed), what do you conclude?

10. Professor Jones is intensely curious about differences in testing situations and wondered if students tended to make better scores on her tests depending on whether the test was taken on a Monday morning or a Friday morning. Her exams have always been normally distributed. From a group of 19 similarly talented students, she randomly selected some to take a test on Friday and others to take it on Monday. The scores by groups were as follows:

Monday	Friday
89.8	87.3
90.2	87.6
98.1	87.3
91.2	91.8
88.9	86.4
90.3	86.4
99.2	93.1
94.0	89.2
88.7	90.1
83.9	

Using $\alpha = 0.05$ (two-tailed), what do you conclude?

For Questions 11 to 12, you should choose the most appropriate and powerful test. Support your answers and list assumptions you are making. Do not try to perform the calculations.

11. Extensive research has been done on the subject of birth order. Data on this research show that first-born children develop different characteristics than later-born children. For example, first-born children tend to be more responsible and self-disciplined than later-born children. A researcher is interested in finding out if first-born children tend to be more confident and have higher self-esteem than later-born siblings. A random sample of 31 first-born children and 35 later-born children were given a self-esteem test. The standard deviation for the first-born children is 1.34, and the standard deviation for the later-born children is 2.15. Test whether birth order affects self-esteem.

12. A researcher tested a new medicine to see if it would be effective in lowering blood pressure. Two samples of participants from a normally distributed population were matched for medical history and initial blood pressure readings. Fifteen randomly selected participants were run through the

experimental condition in which they received the new drug. The other 15 randomly selected individuals participated in the control group and received a placebo. Participants receiving the drug showed a lower blood pressure. Test whether the drug had a statistically significant effect on blood pressure. Variances are homogeneous.

Answer Questions 13 to 19 using the following story problem.

A marketer is interested in how an antismoking campaign affects the smoking habits of teenagers. The researcher samples 50 students from a local area high school and asks them how many cigarettes they smoked. After the antismoking campaign has run for a year, the researcher polls the same 50 students and records the exact number of cigarettes smoked after the campaign.

13. State the null hypothesis for this study.
14. State the alternative hypothesis for this study.
15. Is this a directional or nondirectional hypothesis?
16. Is this research an independent groups (between subjects) or a repeated measures/paired design (within subjects)? Why?
17. What are the independent and dependent variables?
18. Which type of measurement scale do the data from this study represent (e.g., nominal, ordinal, interval, or ratio)? Why?
19. What kind of statistical test should be used to test the hypothesis (hint: think of what we have been doing in class lately)?
20. Which of the following are assumptions underlying the use of the paired t test?
    - A. The variance of the population is known
    - B. The sampling distribution is normal
    - C. Data are interval or ratio
    - D. All of the above
    - E. A and B
    - F. B and C
    - G. A and C
21. A drug and alcohol researcher is interested in studying the effects of alcohol on learning ability of college seniors. She randomly assigns 10 students to an "alcohol group" and another 10 students to a control group. The students in

the alcohol group all receive 8 oz of alcohol prior to being tested. Then all of the students are run through a learning assessment and the number of errors is recorded. Assume the data below are normally distributed.

Control Group	**Alcohol Group**
3	5
2	3
4 $\sum x = 27$	7 $\sum x = 38$
1 $\sum x^2 = 87$	2 $\sum x^2 = 166$
2 Mean = 2.7	4 Mean = 3.8
3 $s_{n-1} = 1.2517$	5 $s_{n-1} = 1.5492$
3	4
1	3
5	2
3	3

Perform the statistical test and state whether or not you can reject the null hypothesis.

The following homework questions should be answered with the online data set provided for this chapter via the textbook's website.

22. Produce a table of descriptive statistics using Microsoft Excel or SPSS.
23. Interpret the descriptive statistics produced by Excel. Do you meet the assumption of normality?
24. Do you meet the assumption of homogeneity of variance?
25. Analyze the data set using an independent t test and indicate if you have a statistically significant result. Explain how you evaluated the output.