14

Testing a Hypothesis about Two Independent Means

How can you test the null hypothesis that two population means are equal, based on the results observed in two independent samples?

- Why can't you use a one-sample *t* test?
- What assumptions are needed for the two-independent-samples *t* test?
- Can you prove that the null hypothesis is true?
- What is power, and why is it important?

All flavors of social scientists are agonizing over the effects of Internet use. One day you're told that e-mail helps you to connect to friends and family and makes you a happy, social person. Several days later the news is bad. Internet users don't spend time with their families, they're depressed and addicted. Evaluating the effects of the Internet on society will, no doubt, keep faculty and graduate students occupied for many years to come. (Note that medical researchers still argue over whether chocolate, cheese, and red wine, which have been consumed for centuries, are good or bad for you, in moderation of course.)

You, too, can participate in Internet research by using the General Social Survey to test hypotheses about differences between those who use the Internet and those who don't. You already found that Internet users appear to be better educated and younger. In this chapter, you'll test hypotheses about television-viewing behavior in Internet users and non-users. You'll determine whether Internet use is related to hours of television viewing.

You'll learn how to test whether two population means are equal, based on the results observed in two independent samples—one from each of the populations of interest. The statistical technique you'll use is called the **two-independent-samples t test**. You can use the two-independent-samples *t* test to see if in the population men and women have the same scores on a test of physical dexterity or if two treatments for high cholesterol result in the same mean cholesterol levels.

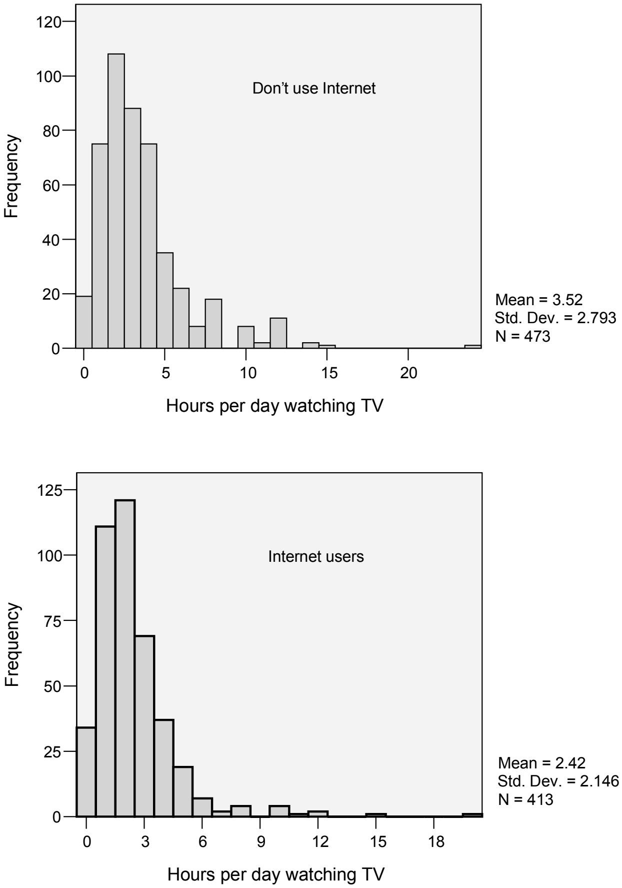
▶ This chapter uses the *gssnet.sav* data file. For instructions on how to obtain the independent-samples *t* test output shown in this chapter, see "How to Obtain an Independent-Samples T Test" on p. 293.

Examining Television Viewing

The first step of any statistical analysis is to examine the data carefully. You want to make sure that the values are plausible. You also want to examine the shape of the distribution so that you can select an appropriate statistical test for testing hypotheses of interest. Figure 14.1 contains histograms of the number of hours of television viewing per day reported by Internet users and non-users. (The GSS question is, "On the average day, about how many hours do you personally watch television?".) You see that both distributions have a tail toward large values, indicating that there are people who report watching television for many hours each day. Some of these values raise statistical concerns as well as concerns about the sanity of some of our fellow citizens. There are people who report watching television for 24 hours a day. You know that isn't possible. It may be that people are reporting how many hours they have the television turned on. "Watch television" is not a very well-defined term. If you have the television on while you're doing homework, are you studying or watching television? It's probably the case that you're doing some of both. When you tally the number of hours you've spent studying for a test, the television time will probably be counted as study time. To an interviewer from the General Social Survey, you might more honestly report it as television time.

Figure 14.1 Histograms of hours spent watching television

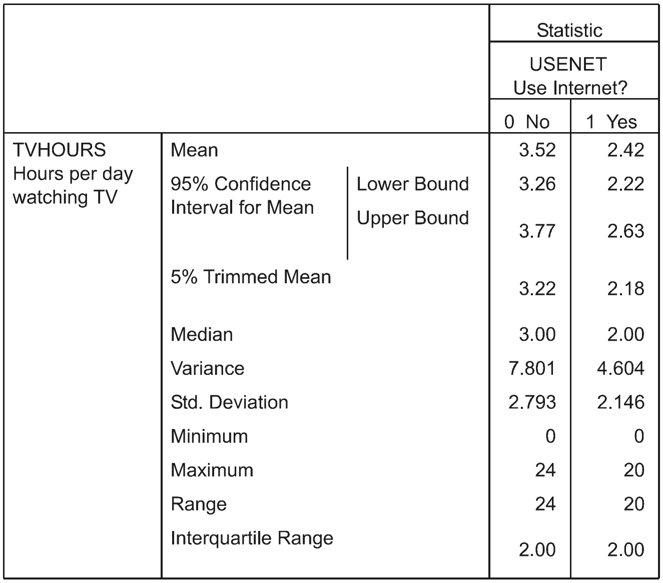
In the Explore dialog box, select tvhours as the dependent variable and usenet as the factor variable. In the Plots dialog box, select Histogram.



We'll proceed with our analyses, assuming that the reported television times are correct. However, we will examine the impact of the outlying points on the results of the analyses. If our conclusions change when the suspect data values are removed, we'll have to consider other approaches to analyzing the data.

From the descriptive statistics in Figure 14.2, you see that Internet users reported an average of 2.42 hours of television viewing per day compared to 3.52 hours for those who don't use the Internet. Internet users, on average, report watching television for about an hour a day less than those who don't use the Internet. Notice that the 5% trimmed means, which are calculated by removing the top and bottom 5% of the values, are 0.2 hours less for both groups than the arithmetic means. Removing those very large values makes both means smaller.

Figure 14.2 Descriptive statistics for hours spent watching television



			Statistic	
			USENET Use Internet?	
			0 No	1 Yes
TVHOURS Hours per day watching TV	Mean		3.52	2.42
	95% Confidence Interval for Mean	Lower Bound	3.26	2.22
		Upper Bound	3.77	2.63
	5% Trimmed Mean		3.22	2.18
	Median		3.00	2.00
	Variance		7.801	4.604
	Std. Deviation		2.793	2.146
	Minimum		0	0
	Maximum		24	20
	Range		24	20
	Interquartile Range		2.00	2.00

? *Why is the number of Internet users and non-users much smaller than in earlier chapters?* The General Social Survey doesn't ask all people all questions. Everyone is asked core questions (about age, education, and income, for example) and a certain number of specialized questions. This is done so that the interviews don't become unbearably long. Only two-thirds of the sample were asked questions about television viewing. The people who are asked any particular question are still a random sample from the United States adult population. ■■■

You know that even if the average hours of television viewing in the population are the same for Internet users and non-users, the sample means will not be equal. Different samples from the same population result in different sample means and standard deviations. To determine if observed sample differences among groups might reflect differences in the population, instead of just sample-to-sample variability, you need to determine if the observed sample means are unusual if the population means are equal. You need to figure out how often you would see a difference of at least 1.1 hours between the two independent groups of Internet users and non-users when there is no difference between the two groups in the population.

? *What do you mean by independent groups?* Samples from different groups are called **independent** if there is no relationship between the people or objects in the different groups. For example, if you select a random sample of males and a random sample of females from a population, the two samples are independent. That's because selecting a person for one group in no way influences the selection of a person for another group. The two groups in a paired design are not independent, since either the same people or closely matched people are in both groups. ■■■

Since you have means from two independent groups, you can't use the one-sample *t* test to test the null hypothesis that two population means are equal. That's because you now have to cope with the variability of two sample means: the mean for Internet users and the mean for those who don't use the Internet. When you test whether a single sample comes from a population with a known mean, you have to worry only about how much individual means from the same population vary. The population value to which you compare your sample mean is a fixed, known number. It doesn't vary from sample to sample. You assumed that the value of 205 mg/dL for the cholesterol of the general population is an established norm based on large-scale studies. Similarly, the value of 40 hours for a work week is a commonly held belief.

The two-independent-samples *t* test is basically a modification of the one-sample *t* test that incorporates information about the variability of the two independent-sample means. The standard error of the mean difference is no longer estimated from the variance and number of cases in a single group. Instead, it is estimated from the variances and sample sizes of the two independent groups.

Distribution of Differences

In the one-sample *t* test, you looked at the distribution of all possible sample means from a population. You saw that the amount in which sample means vary depends on the standard deviation of the values and on the sample size. For the same population, means calculated from large samples vary less than means calculated from small samples. For the same sample size, means calculated from a population with a lot of variability will vary more than means calculated from a population with less variability.

When you want to test hypotheses about two independent population means, you have to look at the distribution of all possible *differences* between the two sample means. Fortunately, the Central Limit Theorem works for differences of sample means as well as for the sample means themselves. So, if your data are samples from approximately normal populations or your sample size is large enough so that the Central Limit Theorem holds, the distribution of differences between two sample means is also approximately normal.

Standard Error of the Mean Difference

If two samples come from populations with the same mean, the mean of the distribution of differences is 0. However, that's not enough information to determine if the observed sample results are unusual. You also need to know how much the sample differences vary. The standard deviation of the difference between two sample means, the **standard error of the mean difference**, tells you that. When you have two independent groups, you must estimate the standard error of the mean difference from the standard deviations and the sample sizes in each of the two groups.

? *How do I estimate the standard error of the difference?*
The formula is

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^{\ 2}}{n_1} + \frac{S_2^{\ 2}}{n_2}}$$

where $S_1^{\ 2}$ is the variance for the first sample and $S_2^{\ 2}$ is the variance for the second sample. The sample sizes for the two samples are n_1 and n_2. If you look carefully at the formula, you'll see that the standard error of the mean difference depends on the standard errors of the two sample means. You square the standard error of the mean for each of the two groups. Next you sum them, and then take the square root. ■■■

Computing the T Statistic

Once you've estimated the standard error of the mean difference, you can compute the *t* statistic the same way as in the previous chapters. You divide the observed mean difference by the standard error of the difference. This tells you how many standard error units from the population mean of 0 your observed difference falls. That is,

$$t = \frac{\left(X_1 - X_2\right) - 0}{S_{\bar{X}_1 - \bar{X}_2}}$$

Equation 14.1

If your observed difference is unlikely when the null hypothesis is true, you can reject the null hypothesis.

? *How is this different from the one-sample* t *test?* The idea is exactly the same. What differs is that you now have two independent-sample means, not one. So you estimate the standard error of the mean difference based on two sample variances and two sample sizes. ■■■

Output from the Two-Independent-Samples T Test

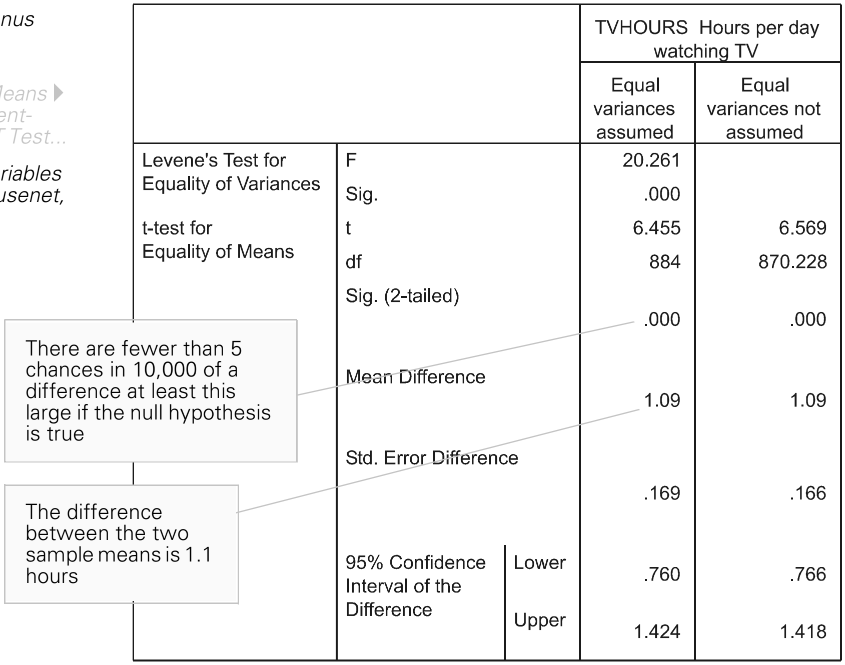
Figure 14.3 shows the results from SPSS of testing the null hypothesis that the average hours of daily television viewing is the same in the population for Internet users and non-users.

Figure 14.3 T test output

From the menus choose:

Analyze
Compare Means ▸
Independent-Samples T Test...

Select the variables tvhours and usenet, as shown in Figure 14.10.



			TVHOURS Hours per day watching TV	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		20.261	
	Sig.		.000	
t-test for Equality of Means	t		6.455	6.569
	df		884	870.228
	Sig. (2-tailed)		.000	.000
	Mean Difference		1.09	1.09
	Std. Error Difference		.169	.166
	95% Confidence Interval of the Difference	Lower	.760	.766
		Upper	1.424	1.418

There are fewer than 5 chances in 10,000 of a difference at least this large if the null hypothesis is true

The difference between the two sample means is 1.1 hours

In the output, there are two slightly different versions of the *t* test. One makes the assumption that the variances in the two populations are equal; the other does not. This assumption affects how the standard error of the mean difference is calculated. You'll learn more about this distinction later in this chapter.

Consider the column labeled *Equal variances not assumed*. You see that for the observed difference of 1.1 hours, the *t* statistic is 6.57. (To calculate the *t* statistic, divide the observed difference of 1.1 hours by 0.17, the standard error of the difference estimate when the two population variances are not assumed to be equal.) The degrees of freedom for the *t* statistic are 870.

The observed two-tailed significance level is less than 0.0005. This tells you that fewer than 5 times in 10,000 would you expect to see a sample difference of 1.1 hours or larger when the two population means are equal. Since this is less than 5%, you reject the null hypothesis that Internet users and non-users watch the same average number of hours of television each night. Your observed results are very unusual if the null hypothesis is true.

Confidence Intervals for the Mean Difference

Take another look at Figure 14.3. The 95% confidence interval for the true difference is from 0.77 hours to 1.42 hours. This tells you it's likely that the true mean difference is anywhere from 45 minutes to 85 minutes. Since your observed significance level for the test that the two population means are equal was less than 5%, you already knew that the 95% confidence interval does not contain the value of 0. (Remember, only likely values are included in a confidence interval. Since you found 0 to be an unlikely value, it won't be included in the confidence interval.)

To calculate a 99% confidence interval, specify 99 in the T Test Options dialog box (see Figure 14.12).

? *If I compute a 99% confidence interval for the true mean difference, will it also not include 0?* Since the observed significance level is less than 0.01, you know that the 99% confidence interval will not include the value of 0. The 99% confidence interval for the mean difference extends from to 0.66 hours to 1.53 hours. ■■■

Testing the Equality of Variances

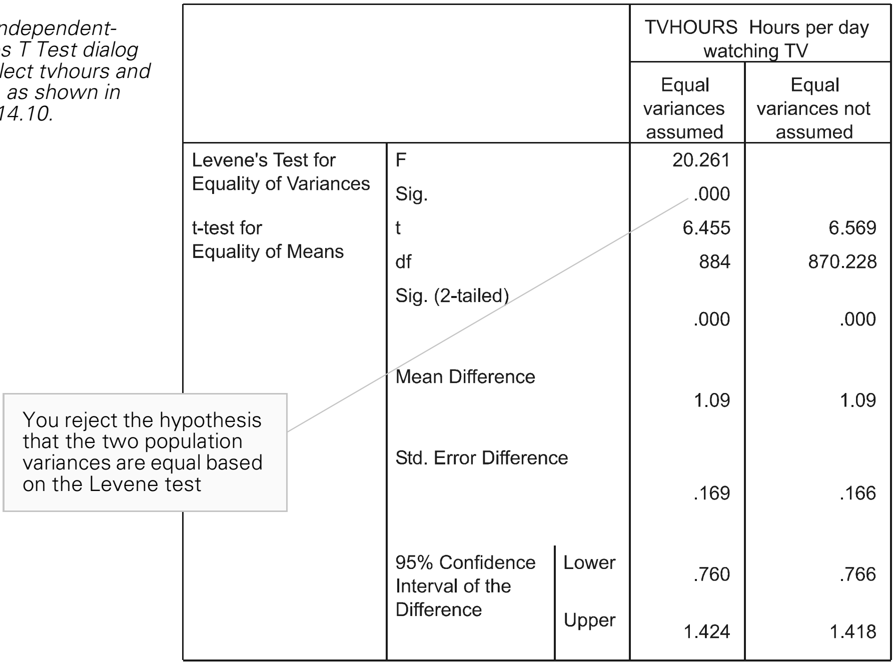
There are two slightly different *t* values in Figure 14.3. That's because there are two different ways to estimate the standard error of the difference. One of them assumes that the variances are equal in the two populations from which you are taking samples, and the other one does not.

In Figure 14.2, you saw that the observed standard deviation for Internet users was somewhat smaller than the standard deviation for non-users. You can test the null hypothesis that the two samples come from populations with the same variances using the Levene test, shown in

Figure 14.4. If the observed significance level for the Levene test is small, you reject the null hypothesis that the two population variances are equal.

In the Independent-Samples T Test dialog box, select tvhours and usenet, as shown in Figure 14.10.

Figure 14.4 Levene test for equality of variances



			TVHOURS Hours per day watching TV	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		20.261	
	Sig.		.000	
t-test for Equality of Means	t		6.455	6.569
	df		884	870.228
	Sig. (2-tailed)		.000	.000
	Mean Difference		1.09	1.09
	Std. Error Difference		.169	.166
	95% Confidence Interval of the Difference	Lower	.760	.766
		Upper	1.424	1.418

You reject the hypothesis that the two population variances are equal based on the Levene test

For this example, you reject the equal variances hypothesis, since the observed significance level for the Levene test is less than 0.005. That means you should use the results labeled *Equal variances not assumed* in Figure 14.4. Notice that the estimate of the standard error of the difference is not the same in the two columns. This affects the *t* value and confidence interval. When you use the estimate of the standard error of the difference that does not assume that the two variances are equal, the degrees of freedom for the *t* statistic are calculated based on both the sample sizes and the standard deviations in each of the groups. This is an approximation and the result is usually not an integer. If the equal variance *t* test is used, the degrees of freedom are just the sum of the two sample sizes minus 2. In this example, both *t* tests give very similar results, but that's not always the case.

? *Why do you get different numbers for the standard error of the mean difference depending on the assumptions you make about the population variances?* If you assume that the two population variances are equal, you can compute what's called a pooled estimate of the variance. The idea is similar to that of averaging the variances in the two groups, taking into account the sample size. The formula for the pooled variance is

$$S^2 = \frac{(n_1 - 1)S_1^{\ 2} + (n_2 - 1)S_2^{\ 2}}{(n_1 - 1) + (n_2 - 1)}$$

It is this pooled value that is substituted for both $S_1^{\ 2}$ and $S_2^{\ 2}$ in Equation 14.1. If you do not assume that the two population variances are equal, the individual sample variances are used in Equation 14.1. ■■■

Effect of Outliers

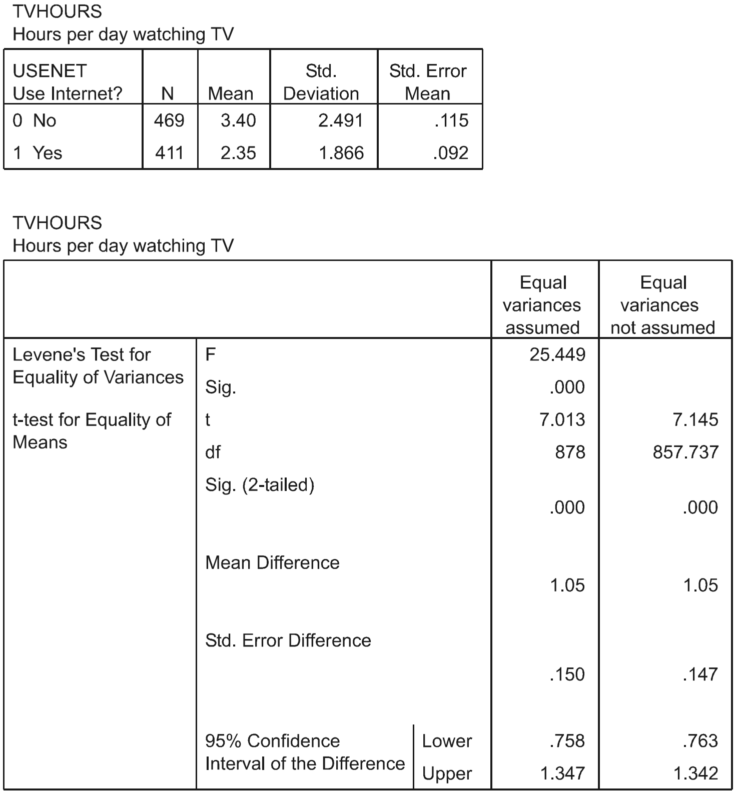
You know from Figure 14.1 that some people reported watching television for very long periods of time, including 24 hours a day. Although you know that a person can't actually watch television for 24 hours a day, some of the other large values are possible, although not particularly believable.

? *Why does the GSS report values that appear to be suspect?* General Social Survey interviewers are trained to conduct interviews in accordance with good survey practice. Their role is to record answers, not to influence or challenge them. Imagine the bias that would be introduced if interviewers were allowed to use their personal judgment to determine which answers they felt were plausible and which were not. With all that gray hair, are you sure you're 35? You graduated from college even though you don't understand a simple question? Do you really earn that much and live in this dump?! Besides creating a hostile environment, challenging answers would result in data that are hopelessly contaminated by interviewer styles and prejudices. The General Social Survey reports the responses given under well-controlled circumstances; users of the data must decide how they will deal with questionable answers and inconsistencies. ■■■

Since the arithmetic mean is affected by data values that are far removed from the rest, you want to make sure that your analysis of differences between the two means is not unduly influenced by the outlying points. This is particularly troublesome for small data sets because a single case can make a big difference in the mean. If removal of the people who watch television 24 hours a day makes the significant difference between Internet users and non-users disappear, you've got a problem. There is no single correct solution for dealing with outliers. A variety of strategies may be useful, such as using statistical techniques that aren't affected by strange data points (some are discussed in Chapter 18), analyzing the data with and without the strange values and seeing whether the results change, or capping the outlying values to decrease their influence.

Figure 14.5 shows the results of the *t* test when people who watch television for more than 12 hours a day are removed from the analysis. The average difference between the two groups has changed only slightly. It went from 1.09 hours to 1.05 hours. You can still reject the null hypothesis that average television-viewing time is the same for the two groups. It's reassuring that your conclusions don't change.

Figure 14.5 T test output when television hours greater than 12 are removed



TVHOURS
Hours per day watching TV

USENET Use Internet?	N	Mean	Std. Deviation	Std. Error Mean
0 No	469	3.40	2.491	.115
1 Yes	411	2.35	1.866	.092

TVHOURS
Hours per day watching TV

			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		25.449	
	Sig.		.000	
t-test for Equality of Means	t		7.013	7.145
	df		878	857.737
	Sig. (2-tailed)		.000	.000
	Mean Difference		1.05	1.05
	Std. Error Difference		.150	.147
	95% Confidence Interval of the Difference	Lower	.758	.763
		Upper	1.347	1.342

From the Data Editor menus choose:

Data
Select Cases...

Select If condition is satisfied. Enter tvhours <= 12 in the Numeric Expression box. Repeat the analysis as shown in Figure 14.10. (For more information, see Appendix B.)

Introducing Education

From the previous analysis, you can conclude that Internet users, on average, watch fewer hours of television per day than non-users. You are 95% confident that the true difference is between 0.8 and 1.4 hours. You may be tempted to rush out and publish the finding that Internet use decreases television-viewing time. But, as a skilled researcher, you know that you must draw conclusions carefully.

You can't say that Internet use *causes* people to watch less television. Causation is very difficult to show in a non-experimental setting. Just because two variables are related doesn't mean that one causes the other. You didn't randomly assign people to be Internet users or non-users, so the two groups may differ in many important ways besides Internet use. This is a serious problem in many observational studies. For example, if you find that people who exercise have lower cholesterol levels, you can't conclude that exercise decreases cholesterol. You know that people who exercise are different from people who don't. They may have healthier diets, smoke less, and be exemplary in other ways. You can't attribute the lower cholesterol to exercise, since it might be due to any or all of the other uncontrolled differences between the groups. If you randomly assign people to exercise or no-exercise programs, you stand a better chance of isolating the impact of exercise.

? *Isn't it misguided to classify people into just two groups: Internet users and non-users?* Absolutely. People use the Internet in different locations, for different purposes, and for different amounts of time. People who have Internet access only at work may be quite different from people with cable modems or DSL service at home. People who use the Internet for one hour a week should not be grouped together with those who use it for long periods of time. You're analyzing only two groups in this chapter, since you're learning about a statistical tests for two independent samples. Many different, and better, criteria for forming groups can be considered (for example, heavy Internet use at home versus light Internet use at home versus no Internet use at home). ■■■

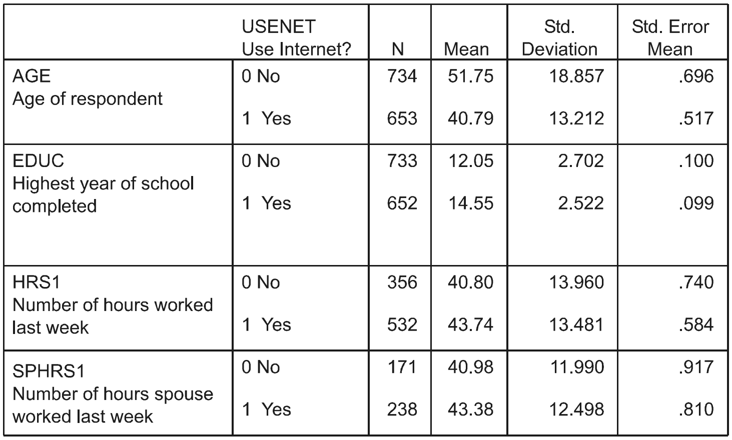
In previous chapters, you saw that in the GSS sample, Internet users are younger and better educated than non-users. That may explain some of the observed differences in television-viewing habits. For example, if people with more education watch less television in general and are more likely to use the Internet, you'll find that Internet use and television viewing are related. Their relationship is explained by education. If you don't include education in your analysis of Internet use, it becomes a *lurking* variable.

? *A lurking variable?!* That's standard statistical jargon for a variable that affects the response you are studying but is not included in your analysis. Age, education, gender, and income are all likely to be lurking variables if you ignore them when looking at differences between Internet users and non-users ■■■

You can start your investigation of the possible effects of age, education, and hours worked per week on television viewing by using the two-independent-samples *t* test to test whether the population values for average age, years of education, and hours worked last week by the respondent and respondent's spouse differ for the two groups. Figure 14.6 shows descriptive statistics for the two groups. Based on the *t* tests in Figure 14.7, you reject the null hypothesis that in the population the two groups have the same average age, education, and hours worked. Internet users are significantly younger, better educated, and work more hours per week. You don't reject the null hypothesis that the average hours worked by the spouses of Internet users and non-users is the same, since the observed significance level is greater than 0.05.

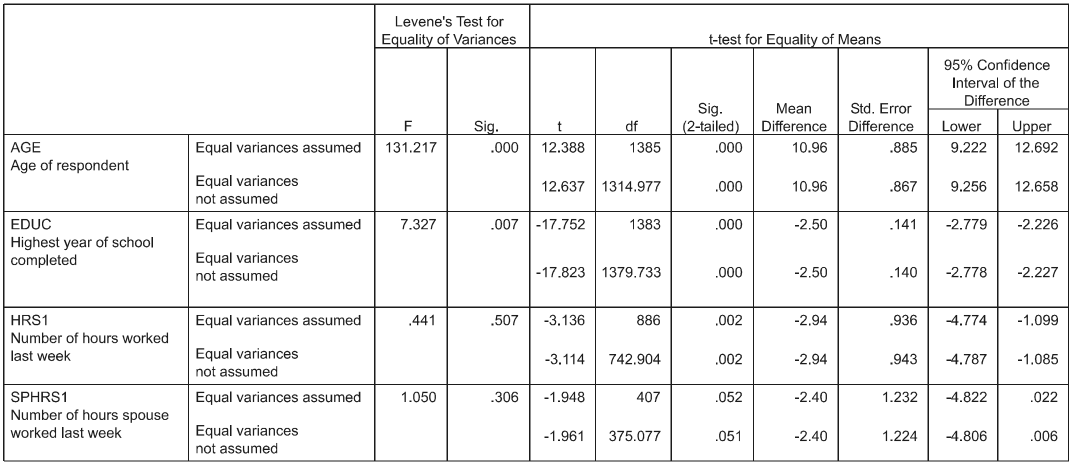
Figure 14.6 Descriptive statistics for age, education, and hours worked

In the Independent-Samples T Test dialog box, select age, educ, hrs1, and sphrs1.



	USENET Use Internet?	N	Mean	Std. Deviation	Std. Error Mean
AGE Age of respondent	0 No	734	51.75	18.857	.696
	1 Yes	653	40.79	13.212	.517
EDUC Highest year of school completed	0 No	733	12.05	2.702	.100
	1 Yes	652	14.55	2.522	.099
HRS1 Number of hours worked last week	0 No	356	40.80	13.960	.740
	1 Yes	532	43.74	13.481	.584
SPHRS1 Number of hours spouse worked last week	0 No	171	40.98	11.990	.917
	1 Yes	238	43.38	12.498	.810

Figure 14.7 T tests for age, education, and hours worked



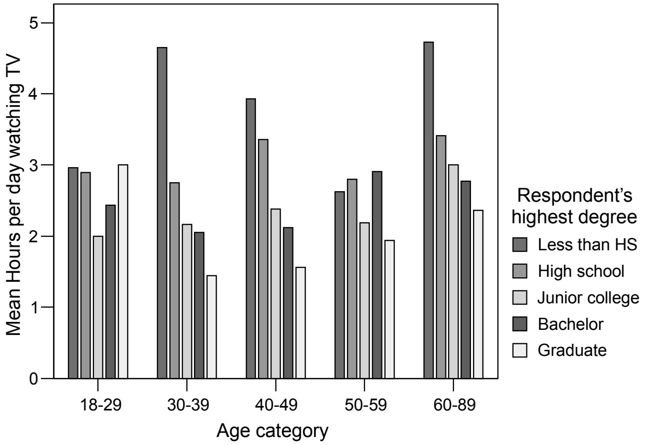
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
AGE Age of respondent	Equal variances assumed	131.217	.000	12.388	1385	.000	10.96	.885	9.222	12.692
	Equal variances not assumed			12.637	1314.977	.000	10.96	.867	9.256	12.658
EDUC Highest year of school completed	Equal variances assumed	7.327	.007	-17.752	1383	.000	-2.50	.141	-2.779	-2.226
	Equal variances not assumed			-17.823	1379.733	.000	-2.50	.140	-2.778	-2.227
HRS1 Number of hours worked last week	Equal variances assumed	.441	.507	-3.136	886	.002	-2.94	.936	-4.774	-1.099
	Equal variances not assumed			-3.114	742.904	.002	-2.94	.943	-4.787	-1.085
SPHRS1 Number of hours spouse worked last week	Equal variances assumed	1.050	.306	-1.948	407	.052	-2.40	1.232	-4.822	.022
	Equal variances not assumed			-1.961	375.077	.051	-2.40	1.224	-4.806	.006

? *Why is the difference in average hours worked by the spouses, 2.4 hours, not statistically significant when the difference in average hours worked by respondents, 2.9 hours, is so highly significant ($p < 0.002$)?* Whether a difference is statistically significant depends on the magnitude of the difference, the variability in the two groups, and the sample sizes in the groups. The difference between 2.9 hours and 2.4 hours is not very great, and the standard deviations of the spouses are actually smaller than the standard deviations for the respondents. The sample sizes are the reason that one of the differences is highly significant and the other is not. Only married people with working spouses are asked how many hours the spouse worked. There are only 238 working Internet spouses and 171 working non-Internet spouses, compared to 532 Internet users and 356 non-users. If the sample included more spouses who worked, a difference of 2.4 hours might well be significant. Remember that when you don't reject the null hypothesis, you haven't shown that the hours worked by spouses are equal. You just didn't have enough evidence to reject the null hypothesis. ■■■

Figure 14.8 is a bar chart of the average hours of television watched when groups are formed by age and education. You see that for most age groups, as education increases, television viewing decreases. In fact, for four out of the five age groups, those without a high school diploma report the largest number of hours of television viewing. For all age groups except those 18–29, people with graduate degrees report the least television viewing. (There are only six people with graduate degrees in the age group 18–29, so you don't have much confidence in the estimate of the average hours of television viewing for such a small group.) Within an education category, there is no clear age effect, although the oldest groups report watching the most hours of television.

In the Define Clustered Bar Summaries for Groups of Cases dialog box, select Other statistic and move tvhours into the Variable box. Select agecat for Category Axis and degree for Define Clusters by.

Figure 14.8 Bar chart of television hours for education and age groups



Since television hours are related to education and education is related to Internet use, it's very likely that some of the differences in television viewing between Internet users and non-users is due to differences in education. You can't look at overall differences between the two groups; you must look at differences within an educational category. That is, you must control for the effects of education. For example, you must compare television viewing for college graduates who are Internet users and non-users.

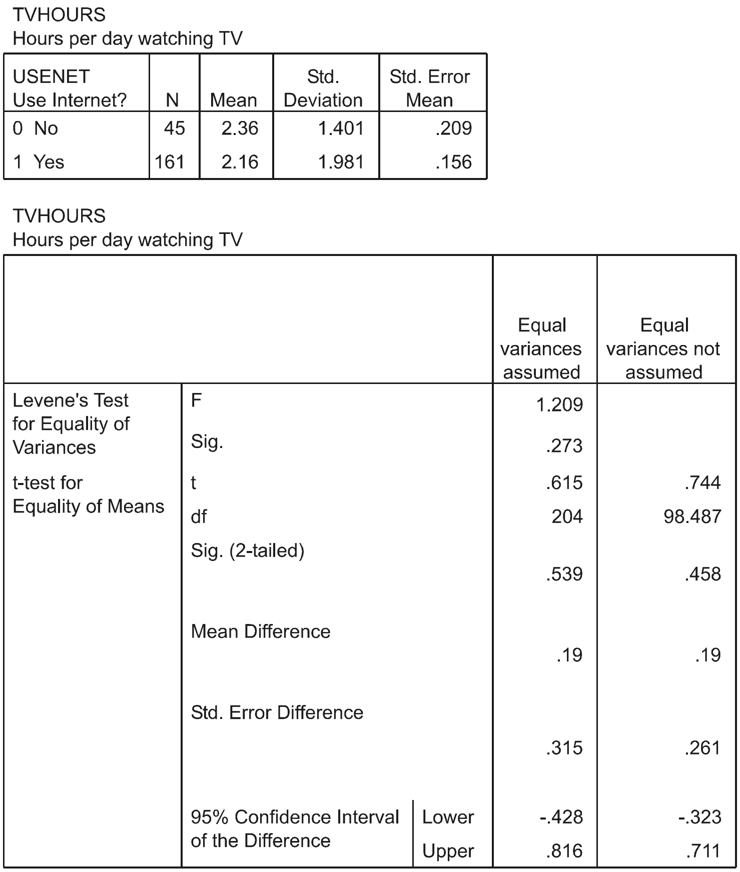
Figure 14.9 is the two-independent-samples *t* test when the comparison of average hours of television viewing is restricted to those with at least a college degree. You see that the average hours of television viewing for those who don't use the Internet is 2.36 hours compared to 2.16 hours for those who do. The difference is 0.2 hours, or about 12 minutes a day.

That’s quite a change from the 1.1 hours you saw when all respondents were included in the analysis. The observed significance level for this difference is 0.5, so you can no longer reject the null hypothesis that average hours of television viewing is the same for the two groups. Controlling the average hours of television use for college education has greatly decreased the difference between the two groups.

Figure 14.9 Comparison of television hours for college graduates

Select cases with degree greater than 2. Repeat the analysis as shown in Figure 14.10. Activate the pivot table and choose:

Pivot
Transpose Rows and Columns



TVHOURS
Hours per day watching TV

USENET Use Internet?	N	Mean	Std. Deviation	Std. Error Mean
0 No	45	2.36	1.401	.209
1 Yes	161	2.16	1.981	.156

TVHOURS
Hours per day watching TV

			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		1.209	
	Sig.		.273	
t-test for Equality of Means	t		.615	.744
	df		204	98.487
	Sig. (2-tailed)		.539	.458
	Mean Difference		.19	.19
	Std. Error Difference		.315	.261
	95% Confidence Interval of the Difference	Lower	-.428	-.323
		Upper	.816	.711

Based on these latest findings, what can you conclude? Can you conclude that Internet use and television viewing are not related? Of course not. You’ve looked at only one of several education groups and ignored possible differences in age between the two groups. What you can conclude is that the study of Internet usage and activities such as television viewing, spending time with one's family, and the like, is a very complex undertaking. There are many factors that influence these behaviors, and chances are excellent that the distribution of these factors for Internet users and non-users is different.

? *So what should I do, just give up my promising research?* Of course not. Just because a problem is complex doesn't mean it can't be solved. There are many statistical techniques that help you untangle relationships between variables. You'll learn about some of them in subsequent chapters. Just remember, fancy statistical techniques are not a substitute for careful thought or good experimental design. ■■■

Can You Prove the Null Hypothesis?

You may have noticed that when phrasing the results of a hypothesis test, we've always been careful to say that you either rejected the null hypothesis or did not reject it. The phrase "You've proved the null hypothesis" has never been used. The reason is that you can't prove the null hypothesis. Think about it. Do you really believe that you've shown that the average hours of television viewing are exactly the same for college graduates who use the Internet and those who do not? Of course not. Your sample results are compatible with many values besides 0 for the population difference. Even if you observed a sample difference of 0, that wouldn't tell you that the true difference is 0. Sample differences of 0 are compatible with values other than 0 for the true population difference.

From the 95% confidence interval, you see that values anywhere in the range of –0.43 to 0.82 years are plausible for the average difference in hours of television viewing between the two groups. If you increase your sample size, your confidence interval will become shorter; that is, it will include fewer "plausible" values, but it will never exclude all values except 0. What all this means is that based on the results of a hypothesis test, you should never claim that you've shown that the null hypothesis is really true. All you can claim is that your sample results are not unusual if the null hypothesis is true.

Sometimes a legal analogy is drawn to statistical hypothesis testing. The null hypothesis is compared to the presumption of innocence in a legal case. Failure to find a defendant guilty doesn't prove innocence. All it says is that there was not enough evidence to establish guilt.

Interpreting the Observed Significance Level

When the observed significance level is small, you reject the null hypothesis and conclude that the two population means appear to be unequal. However, you know that there's a chance that your conclusion is wrong. It's possible that the null hypothesis is true, and your observed difference is one of the remote events that can occur. In fact, that's what the observed significance level tells you—how often you would expect to see a difference at least as large as the one you observed when the null hypothesis is true.

When your observed significance level is too large for you to reject the null hypothesis that the means are equal, two explanations are possible. The first explanation is that really there is no difference between the two population means (a conjecture you can't prove), or that there is a small difference, and you can't detect it. For example, if there really is an average difference of five minutes of television viewing between the Internet users and the non-users, it's of little consequence that you can't identify it. Differences of $10 in annual income or one point on a standardized test are rarely important.

The second explanation is more troublesome: there is an important difference between the two groups, but you did not detect it. How can this happen? One reason you did not reject the null hypothesis when it is false may be that the sample size is small and the observed result doesn't appear to be unusual. Remember, when the sample size is small, many outcomes are compatible with the null hypothesis being true. For example, if you flip a coin only three times, you'll never be able to exclude the possibility that it is a fair coin.

? *Why not?* If a coin is fair and you flip it three times, the probability of observing either three heads or three tails is 0.25. That means that one out of four times when you flip a fair coin three times, you will get all heads or all tails. So even if you're flipping a coin that has two heads or two tails, the results you get (all heads or all tails) are compatible with the null hypothesis that the coin is fair. The outcome is not unlikely if the null hypothesis is true. If you flip the coin 10 times, getting all heads or all tails would be very unusual. ■■■

Even if there is a large difference in television viewing between the Internet users and the non-users, you may not be able to detect it if you have five cases in each of the two groups. That's because the observed sample difference may be compatible with many population values, including 0.

Your ability to reject the null hypothesis when it is false also depends on the variability of the observed values. If your observed values have a lot of variability, the range of plausible values for the true population difference will be broad. For the same sample size, as the variability decreases, the range of plausible population values does, too. Remember, when you calculate a *t* statistic, you divide the observed difference by the standard error of the difference. The standard error of the difference depends on both the sample variances and the sample sizes.

Power

Power is the statistical term used to describe your ability to reject the null hypothesis when it is false. It is a probability that ranges from 0 to 1. The larger the power, the more likely you are to reject the null hypothesis when it is false. Power depends on how large the true difference is, your sample size, the variance of the difference, and the significance level at which you are willing to reject the null hypothesis. Although a detailed discussion of power is beyond the scope of this book, let's consider a simple example because power is so important in data analysis.

? *Why does power depend on all of these factors?* All of these factors are involved in the computation of the *t* statistic, which is what determines whether you reject the null hypothesis. You'll have a large *t* statistic if the numerator is large and the denominator is small. The numerator of the *t* statistic is the difference between the two sample means. So, the larger the population difference, the more likely it is that you will have a large numerator for the *t* statistic. The denominator of the *t* statistic depends on the variances of the groups and the number of cases in each group. You'll have a small denominator if the sample variances are small and the sample sizes are large. ■■■

Monitoring Death Rates

You're the CEO of a large hospital chain that is under increasing pressure to monitor quality. Insurance companies and business coalitions want to see if you're doing a good job before they sign contracts with your organization. Although there are many components that contribute to hospital quality, death rates of hospitalized patients are of major concern to all involved. Since death rates depend on type of disease, you have to examine death rates separately for patients with different diseases. Assume that a 10% death rate is the norm for the condition you want to study. In order to compare your hospital's performance to the norm, how many hospitalized cases of the disease should you include in your sample?

? *Is it fair to compare a hospital's observed mortality rate to a national or state norm?* Comparing mortality rates is tricky. However, with the increasing emphasis on cost and quality of medical care, it's being done more and more often. If hospitals have different patient mixes—that is, if sicker or more complicated cases are concentrated in particular types of hospitals—then it's not fair to expect all hospitals to have the same mortality rates. Sophisticated statistical techniques are used to adjust observed mortality rates for differences in patient characteristics. ■■■

To answer the sample size question, look at Table 14.1. Each of the columns of the table corresponds to a possible true death rate for your hospital. That's the value you would get if you looked at all patients with the diagnosis of interest treated at your hospital chain. It's not the same as the death rate you observe in a sample of patients. Each of the rows of the table corresponds to a different sample size. The entry in each of the cells of the table is the power, that is, the probability that you reject the null hypothesis that the true rate for your hospital is 10%, the norm. The criterion used to reject the null hypothesis is a two-tailed observed significance level of 0.05 or less.

Table 14.1 Probability of rejecting the null hypothesis† using a two-tailed significance level of 0.05

With a true 2% death rate and a sample size of 20, you have only a 36% chance of rejecting the false null hypothesis

	Your hospital's true death rate					
Sample size	2%	5%	15%	20%	30%	40%
20	.36	.14	.10	.24	.64	.90
50	.72	.28	.19	.52	.95	*
80	.90	.41	.27	.72	*	*
100	.95	.49	.33	.81	*	*
160	.99	.68	.49	.95	*	*
200	*	.78	.58	.98	*	*
300	*	.92	.75	*	*	*
400	*	.97	.86	*	*	*
500	*	.99	.92	*	*	*
800	*	*	.99	*	*	*

†The null hypothesis is that the death rate is 10%.

*Indicates probabilities > 0.99.

What is the null hypothesis that I'm testing here? The null hypothesis is that your hospital's death rate for a particular condition is the same as the norm—10%. The norm is not a value you estimate. It's a value that was established previously on the basis of large-scale studies. This is the same situation as in Chapter 10 where you tested whether a new treatment for a disease has the same cure rate as the established treatment. You can use the same binomial test to test this hypothesis. ■■■

Look at the column labeled 2%. If your hospital's true mortality rate is 2% and you count the number of deaths in a random sample of 20 cases, there is only a 36% chance that you will correctly reject the null hypothesis that the true mortality rate for your hospital is 10%, using an observed significance level of 0.05 or less for rejecting the null hypothesis. That means that two out of three times, you will fail to reject the null hypothesis when in fact it's false. You'll fail to identify your hospital as an exceptional performer when it really is. If you increase the sample size to 80, there is a 90% chance that you will correctly reject the null hypothesis. With a sample size of 80, you're much more likely to detect your hospital's good performance. As you can see in Table 14.1, for each of the hypothetical hospital rates (the columns), as you increase your sample size, your power increases.

Now look at the row that corresponds to a sample size of 100. You see that the power is 95% when the true hospital value is 2%. That means that 95% of the time when you take a sample of 100 cases and your hospital's true death rate is 2%, you will correctly reject the null hypothesis that your hospital's true rate is 10%. However, if your true value is 5%, meaning that your hospital's mortality is half of the national average, your probability of detecting that difference is only 49%. Similarly, if your true hospital rate is 15%, a 50% increase over the population rate, with a sample size of 100, you stand only a 33% chance of detecting it. (With a sample of 500 patients, you stand more than a 90% chance of detecting it.) From the table, you see that the larger the difference from the population rate, the easier it is to detect.

Why aren't the power values the same for 5% and 15% true death rates? They are both 5% different from the hypothetical rate of 10%. The reason for this is that if you have a population in which only 5% of the cases die, the variability of the sample death rates will be smaller than if you have a population in which 15% of the cases die. If the population probability of dying is 0, whenever you take a sample there's only one possible outcome—everyone's alive. You're not going to see any variability from sample to sample. If the population probability of dying is 5%, you'll see somewhat more variability, but not that much more. Regardless of the sample size that you take, your sample will consist of mostly living people. If the probability of dying increases to 15%, the variability of possible sample outcomes increases as well. The population isn't as homogeneous as before. In fact, at 50%, you'll have the largest possible variability, since that's when your population is most diverse. You expect to see all kinds of sample values. ■■■

In summary, when you fail to reject the null hypothesis, you should be cautious about the conclusions you draw. In particular, if your sample size is small, you may be failing to detect even large differences. That's why, when you design a survey or experiment, you should make sure to include enough cases so that you'll have reasonable power to detect differences that are of interest.

Does Significant Mean Important?

In Table 14.1, you see that as you increase your sample size, you are able to detect smaller and smaller discrepancies from the null hypothesis. For example, if you have many thousands of cases, you might be able to detect that your hospital rate is really 10.1%, compared to the norm of 10%. Or you might reject the null hypothesis that there is no difference in television viewing between Internet users and non-users, based on difference of 10 minutes a day. Just because you are able to reject the null hypothesis doesn't mean you've uncovered an important or large difference. For very large sample sizes, even very small differences may cause you to reject the null hypothesis. Before you conclude that you've found something important, evaluate the observed difference on its own merits. Is a difference of one month of education really a worthwhile finding? Does a difference of one point in a test score really tell you anything?

Summary

How can you test the null hypothesis that two population means are equal, based on the results observed in two independent samples?

- The one-sample *t* test is appropriate only when you want to test hypotheses about one population mean.
- For the independent-samples *t* test, you must have two unrelated samples from normal distributions, or the sample size must be large enough to compensate for non-normality.
- You can't prove that the null hypothesis is true.
- Power is the probability of correctly rejecting a false null hypothesis. If you have small sample sizes, you may be unable to detect even large population differences.

What's Next?

Now that you know how to test hypotheses about two independent means, you're ready to consider a somewhat more complicated problem. How can you tell if *more* than two means are different from each other? That's what Chapter 15 is about.

How to Obtain an Independent-Samples T Test

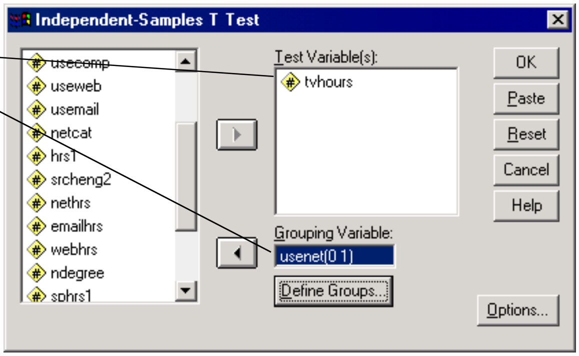
The SPSS Independent-Samples T Test procedure tests the null hypothesis that the population mean of a variable is the same for two groups of cases. It also displays a confidence interval for the difference between the population means in the two groups.

To obtain an independent-samples *t* test, you must indicate the variable(s) whose means you want to compare, and you must specify the two groups to be compared.

- ▶ To open the Independent-Samples T Test dialog box (see Figure 14.10), from the menus choose:

  Analyze
    Compare Means ▶
      Independent-Samples T Test...

Figure 14.10 Independent-Samples T Test dialog box



Test Variable(s): tvhours

Grouping Variable: usenet(0 1)

Make these selections to obtain the output shown in Figure 14.3

In the Independent-Samples T Test dialog box:

- ▶ Select the variable whose mean you want to test and move it into the Test Variable(s) list. You can move more than one variable into the list to test all of them between two groups of cases.

- ▶ Select the variable whose values define the two groups and move it into the Grouping Variable box. Click Define Groups and indicate how the groups are defined.

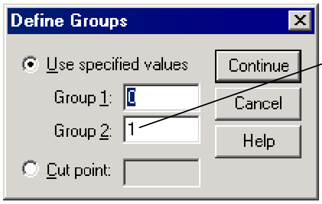
- ▶ Click OK.

For each test variable, SPSS calculates a *t* statistic and its observed significance level.

Define Groups: Specifying the Subgroups

After you move a variable into the Grouping Variable box, click Define Groups to open the Define Groups dialog box, as shown in Figure 14.11.

Figure 14.11 Independent-Samples T Test Define Groups dialog box



● Use specified values
Group 1: 0
Group 2: 1
○ Cut point:

Values correspond to
codes used in the
variable usenet,
0=No, 1=Yes

For a numeric grouping variable, you have the following alternatives:

Use specified values. If each group corresponds to a single value of the grouping variable, select this option and enter the values for Group 1 and Group 2. Other values of the grouping variable are ignored.

Cut point. If one group corresponds to small values of the grouping variable and the other group to large values, select this option and enter a value that separates the groups. Cases that exactly equal the cut point are included in the second group. (If you don't want to remember that, enter a cut point that doesn't occur in your data. To compare codes 1 and 2 with codes 3 and 4, enter a cut point of 2.5 and be sure.)

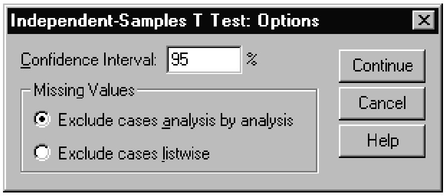
String Grouping Variable

For a string grouping variable, a cut point isn't available. The Define Groups dialog box simply asks for the two values that identify the groups you want to compare.

Options: Confidence Level and Missing Data

In the Independent-Samples T Test dialog box, click Options to open the Independent-Samples T Test Options dialog box, as shown in Figure 14.12. This dialog box allows you to change the confidence level for the confidence interval for the population difference between the means of the two groups and to control the handling of missing values.

Figure 14.12 Independent-Samples T Test Options dialog box



Confidence Interval: 95 %

Missing Values

● Exclude cases analysis by analysis

○ Exclude cases listwise

Confidence Interval. Defines the desired confidence interval (usually 95 or 99).

Missing Values. Two alternatives control the treatment of missing data for multiple test variables.

> **Exclude cases analysis by analysis.** All cases that have valid data for the grouping variable and a test variable are used in the statistics for that test variable.
>
> **Exclude cases listwise.** Only the cases that have valid data for the grouping variable and all specified test variables are used. This ensures that all of the tests are performed using the same cases but doesn't necessarily use all of the available data for each test.

Exercises

Statistical Concepts

1. For the following studies, indicate whether an independent-samples or paired *t* test is appropriate:

   a. You want to study regional differences in consumer spending. You randomly select a sample of consumers in the Midwest and the East and track their spending patterns.

   b. You want to study differences in the spending habits of teenage boys and girls. You select 100 brother-sister pairs and study their spending behavior.

   c. You want to compare error rates for 20 employees before and after they attend a quality improvement workshop.

d. Weight is obtained for each subject before and after Dr. Nogani's new treatment. The hypothesis to be tested is that the treatment has no effect on weight loss.

e. The Jenkins Activity Survey is administered to 20 couples. The hypothesis to be tested is that husbands' and wives' scores do not differ.

f. Subjects are asked their height and then a measurement of height is obtained. The hypothesis to be tested is that self-reported and actual heights do not differ.

g. You want to compare the durability of two types of socks. You have people wear one type of sock on one foot and another type of sock on the other. You see how long it takes for a hole to appear. (Assume they wash them periodically!)

2. You want to compare the average ages of people who buy and who don't buy a product. Suppose you obtain the following results:

		N	Mean	Std. Deviation	Std. Error Mean
AGE	No	100	29.45	15.56	1.56
	Yes	100	38.00	15.49	1.55

			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.052	
	Sig.		.820	
t-test for Equality of Means	t		-3.900	-3.900
	df		198	198.000
	Sig. (2-tailed)		.000	.000
	Std. Error Difference		2.196	2.196
	95% Confidence Interval of the Mean	Lower	-12.88	-12.88
		Upper	-4.22	-4.22

a. Write a short paragraph summarizing your findings.

b. Have you proved that in the population the two groups have different mean ages?

c. What is a plausible range for the true difference?

d. What would happen to the observed significance level if the difference and the standard deviations in the two groups remained the same but the sample sizes were tripled?

3. A market research analyst is studying whether men and women find the same types of cars appealing. He asks 150 men and 75 women to indicate which one of the following types of cars they would be most likely to buy: two-door with trunk, two-door with hatchback, convertible, four-door with trunk, four-door with hatchback, or station wagon. Each of the possible responses is assigned a code number. The analyst runs a *t* test and finds that the average value for males and females appears to differ ($p = 0.008$). He doesn't know how to interpret his output, so he comes to you for advice. Explain to him what his results mean.

4. You are interested in whether the size of a company, as measured by the number of employees, differs between companies that offer pension plans and those that don't. You select a random sample of 75 companies and obtain information about the number of employees and availability of a pension plan. You run a two-independent-samples *t* test and find that there is not a statistically significant difference between the two types of companies ($p = 0.237$). A colleague of yours conducts a similar study. She polls 200 companies and finds a similar difference in the number of employees between the two types of companies. However, she claims that she found a significant difference ($p = 0.002$). Is this possible? Explain.

5. You are interested in whether average family income differs for people who find life exciting and for those who don't. You take a sample of people at a local museum on Sunday afternoon and find that there is a $5,000 difference in income between the two groups. You do a *t* test and find the observed significance level to be 0.03.

   Your friend is also studying the same problem. She takes a sample of people in a department store on Saturday afternoon. She finds a $10,000 difference in family income between the two groups. But when she does a *t* test, she finds an observed significance level of 0.2.

   Discuss these studies, their shortcomings, and possible reasons for the contradictory results.

Data Analysis

Use the *gssnet.sav* data file to answer the following questions:

1. Perform the appropriate analyses to test whether the average number of hours of daily television viewing (variable *tvhours*) is the same for men and women. Write a short summary of your results, including appropriate charts to illustrate your findings. Be sure to look at the distribution of hours of television viewed separately for men and women.

   a. Based on the results you observed, is it reasonable to conclude that in the population, men and women watch the same amount of television?

b. If you found a statistically significant difference between average hours watched by men and women, would you necessarily conclude that men and women do not watch the same amount of television? What other nonstatistical explanations are possible for your findings?

2. Discuss the possible advantages of a paired design to analyze differences between men's and women's television viewing. Discuss the drawbacks as well.

3. Some people claim that they would continue to work if they struck it rich, and others say that they would not (variable *richwork*). Use the two-independent-samples *t* test to identify possible differences between the groups in age, education, television viewing, and so on. Write a paper summarizing your findings.

4. What distinguishes people who believe in life after death from those who do not (variable *postlife*)? Use the available data to identify differences between the two groups. Write a short paper summarizing your results.

5. Consider people who use the Internet and those who don't (*usenet*).

   a. What is the average income for people who use the Internet and those who don't? (Use *rincdol*.) Do you have enough evidence to reject the null hypothesis that the average income is the same for the two groups? Which group makes more money?

   b. Write a short paper discussing differences between Internet users and non-users.

Use the *marathon.sav* data file to answer the following question:

6. If you consider the Chicago marathon runners to be a sample from some population of marathon runners, you can test hypotheses about differences in completion times based on gender and age.

   a. Describe the age and gender composition of those who completed the Chicago marathon.

   b. Test the null hypothesis that men and women have the same average completion times.

   c. Even if men and women of a given age ran as fast, given the age distributions in the two groups, would you expect to reject the null hypothesis that average completion times are the same for men and women? Explain.

   d. Select only runners in the age group 25–39 (*agecat8* equals 3). Test the null hypothesis that the average time to completion is the same for men and women. Summarize the results.

Use the *manners.sav* data file to answer the following questions:

7. Consider the average age for people who think the world would be a better place if people said please and thank you and those who think it would make no difference (*pleasthx*).

a. Make histograms of age for the two groups. Is age approximately normally distributed in the two groups?

b. What assumptions do you need to make in order to test the null hypothesis that the average age is the same in the two groups?

c. What is the average age for people who think the world would be a better place if people said please and thank you? For those who think it would not make a difference?

d. Test the null hypothesis that in the population the average age of those selecting the two responses is equal. What can you conclude?

e. What is the 95% confidence interval for the difference in average ages? Based on the confidence interval, would you reject the null hypothesis that the average difference in age is 10 years? Five years?

Use the *salary.sav* data file to answer the following questions:

8. Use the Select Cases facility to restrict the analysis to clerical workers only (variable *jobcat* equals 1).

a. Test the assertion that male and female clerical workers have the same average starting salaries. Summarize your findings.

b. You are 95% confident that the true difference in average beginning salaries for male and female clerical workers is in what range?

c. How often would you expect to see a difference at least as large in absolute value as the one you observed if, in fact, male and female clerical workers have the same beginning salaries?

d. Evaluate how well your data meet the assumptions needed for a two-independent-sample *t* test.

9. The bank claims that male clerical workers are paid more than female clerical workers because they have more formal education. Do the data support this assertion? Explain.

10. Consider office trainees (variable *jobcat* equals 2). The women trainees have hired you to show that the bank discriminates again women office trainees by paying them less. Analyze the data, and prepare a summary of your findings.

11. The bank has now hired you to refute the claim that they discriminate against women office trainees. What evidence can you come up with to support the bank's position? Write a summary of your findings on behalf of the bank.

Use the *electric.sav* data file to answer the following questions:

12. Write a short report discussing the claim that average diastolic blood pressure in 1958 (variable *dbp58*) is the same for men who were alive in 1968 and for men who were not (variable *vital10*). Include appropriate summary statistics and charts.

13. There is one man with a diastolic blood pressure of 160 mm Hg. Rerun the independent-samples *t* test, excluding him from the analysis. How does your conclusion change? What can you conclude about the effect of outliers on the results of a *t* test?

14. Test the null hypothesis that average diastolic blood pressure is the same for men who smoke and for men who don't smoke. Write a paragraph summarizing your conclusions. Include appropriate charts. Be sure to consider the effect of the outlier on your results.

Use the *schools.sav* data file to answer the following question:

15. Look at schools that are above and below the median percentage of low income for all Chicago schools (variable *medloinc*). Are there differences between the two groups in the school performance variables? Summarize your results.

Use the *renal.sav* data file to answer the following question:

16. Use the independent-samples *t* test to identify differences between cardiac surgery patients who developed acute renal failure (variable *type* equals 1) and those who did not (variable *type* equals 0). Identify the variables in the data file for which a *t* test is appropriate. Summarize your findings.

Use the *buying.sav* data file to answer the following question:

17. Test the null hypothesis that the family buying score (variable *famscore*) is the same when pictures are shown and when they are not (variable *picture*). Test the null hypothesis that the average buying score for the husband (variable *hsumbuy*) is the same with and without pictures. Repeat for the average buying score for wives (variable *wsumbuy*). Summarize your results.