

# Chapter 7

## STATISTICAL SIGNIFICANCE HYPOTHESIS TESTING WHEN COMPARING TWO MEANS

Doing a Study That Tests a Hypothesis of Differences Between Means	146
Design of a Study That Compares Two Conditions	146
Applying the Logic of Hypothesis Testing	147
<i>Reasoning From Reverse:</i>	
<i>The Logic of Testing the Null Hypothesis</i>	149
<i>Steps in Hypothesis Testing</i>	150
Assumptions in Parametric Hypothesis Testing	153
Effects of Violating Parametric Assumptions	154
Testing the Assumptions	155
<i>Testing for Normal Distributions</i>	155
<i>Testing for Homogeneous Variances Among Two or More Groups</i>	157
Comparing Sample and Population Means	159
When the Population Standard Deviation Is Known	159
When the Population Standard Deviation Is Unknown	161
<i>Using SPSS for the One-Sample <math>t</math></i>	163
Comparing the Means of Two Sample Groups: The Two-Sample $t$ Test	164
Using $t$ as a Sampling Distribution of Mean Differences	164
Conducting the Hypothesis Test for the Difference Between Two Sample Means	164
Using SPSS and Excel to Compute the Two-Sample $t$	167
<i>SPSS</i>	167
<i>Excel</i>	168
Effect Size Computations	170
Confidence Intervals for Mean Differences	170

Comparing Means Differences of Paired Scores: The Paired Difference $t$	172
Conducting the Hypothesis Test for Paired Differences	172
Using SPSS and Excel to Compute the Paired Differences $t$	173
<i>SPSS</i>	173
<i>Excel</i>	174
Confidence Intervals for Paired Differences	175
Assessing Power	176

To understand the logic of statistical hypothesis testing, which underlies many other tests, it makes sense to begin with comparisons of two groups. Furthermore, comparisons of two groups (e.g., men and women, people with high communication apprehension and people with low communication apprehension, people who listen to a speech with evidence and people who listen to the same speech without any evidence) on some output measure of interest is common in communication research.

## DOING A STUDY THAT TESTS A HYPOTHESIS OF DIFFERENCES BETWEEN MEANS

Researchers often ask research questions about the means of two groups on some measure of interest. These two groups usually are categories or levels of an independent variable that is measured on the nominal level. To understand this process, it helps to grasp what studies using these tools look like and the general process of hypothesis testing.

### Design of a Study That Compares Two Conditions

Researchers who wish to compare the means of two groups usually are involved in completing experiments or some form of survey research. A **survey** is an “empirical study that uses questionnaires or interviews to discover descriptive characteristics of phenomena” (Reinard, 2001, p. 225). On the other hand, an **experiment** is “the study of the effects of variables manipulated by the researcher, in a situation in which all other variables are controlled, and completed for the purpose of establishing causal relationships” (Reinard, 2001, p. 256). Unlike the work in surveys, in experiments researchers introduce variables that were not already present in the situation (to participants called an “experimental group”), and they withhold those variables from others (participants in a “control group”). For instance, an experimenter might expose one group of people to a message with climax order and another group to a message with anticlimax order. Then, the researcher would compare the dependent variable mean scores of participants from the two groups. In surveys, researchers do not manipulate variables, and they often find that independent variables already are organized into two groups, such as participant sex (men and women), type of cultural background (individualistic or collectivist cultures), or age (old and young). Some of these variables originally were continuous variables, but they have been broken into levels called **variable factors** or just factors.

Once variables are divided into two categories, researchers may posit hypotheses to be tested. For comparisons between means, research hypotheses take the form of comparing the means of two groups, symbolized as

H:  $\mu_1 > \mu_2$  (the dependent variable mean of the first group is higher than the mean of the second group),

H:  $\mu_1 < \mu_2$  (the dependent variable mean of the first group is lower than the mean of the second group), or

H:  $\mu_1 \neq \mu_2$  (the dependent variable mean of the first group is not equal to the mean of the second group).<sup>1</sup>

The first two examples are called directional hypotheses because, not surprisingly, they assert a direction to the differences between means. The last hypothesis is a nondirectional hypothesis because it asserts a difference between groups but not the nature of that difference. In contrast to these research hypotheses is the null hypothesis that states that there is no difference between groups:  $H_0: \mu_1 = \mu_2$ . As will be seen, this null hypothesis is actually what is tested statistically. In the case of a directional hypothesis, rejection of the null hypothesis would have to be accompanied by a finding of mean differences in the predicted direction.

Some, therefore, have suggested that the null hypothesis for the first hypothesis actually be stated as  $H_0: \mu_1 \leq \mu_2$  and that the null hypothesis for the second hypothesis should be stated as  $H_0: \mu_1 \geq \mu_2$ . Though it sometimes may sound curious at first, researchers investigating material hypotheses typically want to *reject* opposing null hypotheses. The approach that includes “directions” in null hypotheses attempts to take account of all relationships that would not support a researcher’s material hypotheses.

Before testing the hypotheses, researchers must examine whether the assumptions underlying the use of the statistics have been satisfied. Finally, after the assumptions have been checked, the primary statistical tools may be employed.

## Applying the Logic of Hypothesis Testing

You might think that testing a research hypothesis is a simple matter of checking to see if the means are in the direction suggested. But because we collect data in samples and try

---

<sup>1</sup>As was explained in Chapter 2, though researchers typically deal with sample data, for conceptual purposes they hypothesize general relationships that may exist in the population as a whole. Hence, hypothesis notation uses Greek letters to represent population characteristics. The standard abbreviation for the population mean is the lowercase Greek letter  $\mu$  (mu). For sample data, the ordinary Roman alphabet (the ABCs) is used. Sometime in the future, statistics books may not use the Greek alphabet notation for hypotheses, but until the change becomes universal, it is helpful to learn the typical notation so that you can understand the meaning of specific concepts you may wish to investigate from other sources in statistics.

## 148 INFERENCE STATISTICS

to make inferences about populations, this approach might not be very helpful. One could imagine a researcher and a skeptic discussing the matter:

Researcher: I have confirmed my research hypothesis that a speech with internal organizers is more easily recalled than a speech without internal organizers.

Skeptic: In the first place, you do not *confirm* or *prove* hypotheses. You can “support” them or find them “tenable,” but that’s it. In the second place, even if we assume that your research design actually manipulated internal organizers without confounding them with other variables, such as language vividness and message length, you still do not have evidence of the impact of internal organizers because you used *sample* data. I’d bet that if you sampled the entire population, you would find no difference at all in recall of the message.

Researcher: But I used random sampling to ensure that my samples would be representative of the population.

Skeptic: That’s just the point. If you used random sampling—and I would like you to tell me how you developed a large enough sampling frame to pull off that trick—then most of the time you would tend to get samples that mirrored the population characteristics of interest.

Researcher: See!

Skeptic: Not exactly. If you sampled at random, in addition to getting results that might reflect the population, you occasionally would get samples that represented extreme results. You do not really know whether your study results reflected the effects of sampling error or whether they identified a real relationship that exists in the population.

Researcher: OK, if you don’t believe that internal organizers increase recall of a message, where is your proof?

Skeptic: Whoa! I don’t have the burden of proof here. You do. Those who assert things must prove them. Besides, how could I be expected to prove a null hypothesis?<sup>2</sup>

Researcher: I guess I’ve wasted my time.

---

<sup>2</sup>Sometimes, though not in this example, you can prove a null hypothesis, but doing so requires two things. First, what is identified must be equally recognized by everyone in a position to make observations, rather than being a matter of personal preference or subjective judgment. Second, the universe must be limited, so that a complete search is possible. So, if a dentist tells you, “You do not have any new cavities,” the dentist is asserting the truth of a null hypothesis. What constitutes a cavity is recognized as the same by everybody with dentistry training (we hope), and the universe in which to search (the teeth in your mouth) is limited. This situation is rare in everyday life. Thus, under most circumstances, it is not possible to prove a null hypothesis.

There is another way.<sup>3</sup> The rest of this chapter will explain this logic of hypothesis testing. Because it underlies all the other tests of significance, it is useful to be sure to be comfortable with this logic before proceeding to specific tests.

### ***Reasoning From Reverse: The Logic of Testing the Null Hypothesis***

Rather than trying to prove the point directly, researchers may use a *process of elimination* to support a hypothesis. Just for the sake of argument, you could assume that the null hypothesis of the skeptic is true. Then, you could ask how improbable it would be to find a difference as large as observed as a result of random sampling error. If random sampling could produce a sample such as yours quite often, then you would decide that the evidence is not good enough to reject the doubts of the skeptic (in other words, you would fail to reject the null hypothesis). Yet, if your sample could be found *quite rarely* due to sampling error, then you would decide that it is unlikely that your sample came from a population described by the null hypothesis. Now you have two options:

1. Decide that your results are, in fact, just the kind of occurrence that happens at random once in a while, even though it is admittedly unlikely—and conclude that this occurrence is one of those random sampling oddities, or
2. Decide that it is so improbable that your results could be found at random from a population defined by the null hypothesis that the null hypothesis must be untrue.

You do not *prove* that your research hypothesis is true. Instead, you show how *improbable* an explanation the null hypothesis is for your results. If the null hypothesis is improbable, what's left? The properly stated research hypothesis is the only alternative. Statistics cannot prove that your research hypothesis is true, but you can use the statistics that follow to show how long the odds are against skeptics who would posit a null hypothesis. As you can see, the null hypothesis is actually the one that researchers test. If they can reject it as improbable, they use a process of elimination to conclude that the research hypothesis is “supported.”<sup>4</sup>

---

<sup>3</sup>There actually is more than one other way. The approach presented here is a standard treatment of significance testing, but you should know that serious scholars have suggested other ways to approach hypothesis testing. Relying on Bayesian probability theory, many are exploring alternatives to the “process of elimination” approach taken here.

<sup>4</sup>The logic of hypothesis testing is not a metaphor. The conditional syllogism is used throughout the process. For instance:

Major premise: If the null hypothesis is true, then no statistically significant differences will be found.

Minor premise: Statistically significant differences were found.

Conclusion: Therefore, the null hypothesis is untrue.

This form of reasoning is known as *modus tollens*. You may notice that finding no significant differences would not prove the null hypothesis to be true. In such a case, the minor premise “no statistically significant differences were found” would commit the fallacy of affirming the consequent. Even with formal logic, it is difficult to make a valid argument to prove a null hypothesis.

### *Steps in Hypothesis Testing*

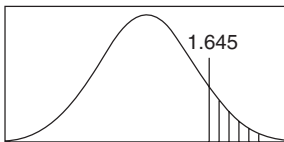
To test a statistical hypothesis, researchers follow several steps. Each will be considered in turn.

- *Determining a decision rule to reject the null hypothesis* is the starting point for assessing a hypothesis. This decision rule is called setting a level of **alpha risk** ( $\alpha$  risk). The researcher announces alpha risk before the research is completed, and it is the decision rule under which null hypotheses are to be rejected. The decision rule (or “alpha” for short) is usually set at a probability of .05 for research in communication studies. So, if a set of results could have been found by random sampling error from a distribution defined by the null hypothesis no more than 5 times out of 100, the researcher agrees to reject the null hypothesis explanation. Of course, this level means that 5 times out of 100 (or 1 time out of every 20 tests), when the researcher claims to have found a significant difference, the effect *really is* just attributable to random sampling error. To make the test understandable, it often is useful to state the null hypothesis explicitly (though in published research, such a feature is rare).
- *Computing a test statistic* is the result of using a statistical formula. In this chapter, the statistics of interest are  $z$  and  $t$ , but many others are available.
- *Finding the critical value* to interpret the meaning of the test statistic requires researchers to look at distributions and tables. Then, based on adjustments for sample sizes and parameters estimated from samples, researchers look at distributions to identify critical regions of interest. The **critical region** of a distribution represents

values that are “critical” to a particular study. They are critical because when a sample statistic falls in that region, the researcher can reject the null hypothesis. (For this reason, the critical region also is called the “region of rejection.” (Vogt, 2005, p. 70)

Researchers look at a distribution for their type of data and then identify the proportion that corresponds to their decision rule for rejecting the null hypothesis. For instance, if a researcher were using the standard normal curve as the underlying probability distribution, and using an alpha risk of .05 as the decision rule to reject the null hypothesis, then 5% of the standard normal curve would have to be identified as the critical region.

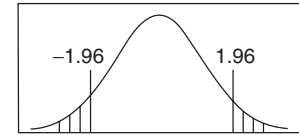
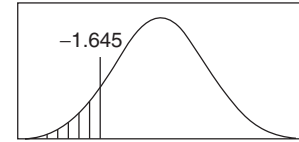
If the research hypothesis is a directional hypothesis, 5% of the distribution that is the critical region would be on one side of the distribution. In the case of the standard normal curve, the last 5% of the distribution begins at 1.645 standard deviations.<sup>5</sup> But which side of the distribution is the location of the critical region? It depends on the hypothesis.



- If the directional hypothesis is stated as  $H: \mu_1 > \mu_2$ , the 5% of the distribution that is the critical region is on the right side, as shown in the diagram on the left.

<sup>5</sup>Though it appears in Table C.1 in this book, the value of 1.645 does not appear in most tables of the standard normal curve and must be interpolated from the surrounding  $z$  values.

- If the directional hypothesis is stated as  $H: \mu_1 < \mu_2$ , the 5% of the distribution that is critical region is on the left side, as shown in the diagram on the right.
- If the research hypothesis is a nondirectional hypothesis, the 5% of the distribution that is the critical region would be divided, with half on one side of the distribution and half on the other side. Hence, for a null hypothesis such as  $H: \mu_1 \neq \mu_2$ , 2.5% of the distribution that is part of the critical region is on the right side, and the remaining 2.5% of the distribution that is part of the critical region is on the left side. Because the last 2.5% of the distribution starts at  $\pm 1.96$  standard deviations, the critical regions can be identified as in the diagram shown in the right.



- *Rejecting or failing to reject the null hypothesis* is the final step in statistical hypothesis testing. If the test statistic falls in any critical region for the particular hypothesis, the researcher applies the decision rule to reject the null hypothesis, and a real relationship or “statistically significant” difference is claimed. The term **statistical significance** is often defined as a relationship that is beyond what might be expected to occur by chance alone, but “statistical significance means that the result was unlikely due to chance; if the null hypothesis is true, an improbable event has occurred” (Johnson, 1995, p. 1999). Either the improbable event can be dismissed, or it can be taken as evidence that the null hypothesis explanation is unpersuasive. Researchers usually claim statistical significance with such claims as “statistically significant differences were found ( $p < .05$ ).” The  $p$  in this statement symbolizes the probability that observed differences could have been found if the null hypothesis were true. In short, the smaller this probability is, the more potent the evidence is against the null hypothesis—and, by a process of elimination, the more tenable is the alternative research hypothesis. It is important to remember

that a  $p$ -value merely indicates the probability of a particular set of data being generated by the null model—it has little to say about size of a deviation from that model (especially in the tails of the distribution, where large changes in effects size cause only small changes in  $p$ -values). (Helberg, 1995, ¶ 28)

Of course, the decision a researcher makes to reject a null hypothesis is a “yes” or “no” option. Sometimes these choices will prove—in the long run—to be sound decisions, and sometimes they will be mistaken. When completing a study, researchers play the odds, but they cannot know for sure that they have decided correctly. The options are found in Table 7.1.

The “Actual Situation” is not known to the researcher at the time of a study, of course. But there are two options: The null hypothesis could be true, or it might be false. In addition, researchers might look at statistical analyses completed and decide that the odds seem to be against the null hypothesis explanation of the data, that the null hypothesis should be rejected.

**Table 7.1**

		<i>Actual Situation</i>	
		<i>H<sub>0</sub> is false</i>	<i>H<sub>0</sub> is true</i>
<i>Researcher's Decision Based on Statistical Testing Decision</i>	Reject H <sub>0</sub>	Correct decision Power	Type I error $\alpha$ risk
	Do not reject H <sub>0</sub>	Type II error $\beta$ risk	Correct decision

If the researcher decides to reject the null hypothesis and the null hypothesis is false, the researcher has made a correct decision. Of course, at the time of the study, the researcher could not know for sure, but the researcher can compute the probability of correctly rejecting the null hypothesis. Called **statistical power**, this term refers to “the probability of rejecting the null hypothesis when it is false—and therefore should be rejected” (Vogt, 2005, p. 242). Of course, if a researcher rejects the null hypothesis and it turns out that the null hypothesis is true, then the researcher has made an incorrect decision. This type of mistake is known as **Type I error**.<sup>6</sup> One cannot know whether a Type I error has occurred at the time the data are first analyzed statistically (though in the long run, researchers usually find out). But researchers can identify the probability that a Type I error might occur. This probability of incorrectly rejecting the null hypothesis is known as **alpha ( $\alpha$ ) risk**. Many students find it useful to think of Type I error as a researcher producing a “false positive” claim of support for the research hypothesis. The researcher thought there was a predicted relationship, but it was a false positive finding.

The researcher could fail to reject the null hypothesis. If the null hypothesis were false—if there actually were differences that went undetected—the researcher would have made a mistake. This type of error is called **Type II error**. Though at the time the study is conducted, the researcher cannot know if Type II error has occurred, the probability that the error might have occurred can be identified as **beta ( $\beta$ ) risk**. Students sometimes find it useful to think of Type II error as a researcher producing a “false negative” claim about the research hypothesis. Though the data *seemed* to suggest the absence of a relationship, it actually was a false negative finding. In general, researchers control beta risk by using large enough sample sizes for statistics to detect relationships in the data. Of course, if the researcher fails to reject the null hypothesis that is true, the decision is a correct one. As you might expect,  $1 - \beta$  is the power of a statistical test.

In passing, it might be mentioned that researchers must have a hypothesis before using these steps of statistical hypothesis testing. It is not appropriate to compose a hypothesis after the data are examined. Similarly, it is not appropriate to set a decision rule after a test statistic has been computed.

<sup>6</sup>One might wonder why they are called Type I error and Type II error (to be discussed later). The reason appears to go all the way back to Aristotle, who identified two types of errors: to say about that which is true that it is untrue, and to say about that which is untrue that it is true.



### Special Discussion 7.1

#### Troubled Language Use in Hypothesis Testing

Reasoning by a process of elimination often has created difficulties for researchers in reporting their findings. One writer (Thompson, 1994, p. 6) explains:

Many of the problems in contemporary uses of statistical significance testing originate in the language researchers use. Several names can refer to a single concept (e.g., “SOS (BETWEEN)” = “SOS(EXPLAINED)” = “SOS(MODEL)” = “SOS(REGRESSION)”), and different meanings are given to terms in different contexts (e.g., “univariate” means having only one dependent variable but potentially many predictor variables, but may also refer to a statistic that can be computed with only a single variable).

Overcoming three habits of language will help avoid unconscious misinterpretations:

- Say “statistically significant” rather than “significant.” Referring to the concept as a phrase will help break the erroneous association between rejecting a null hypothesis and obtaining an important result.
- Don’t say things like “my results approached statistical significance.” This language makes little sense in the context of the statistical significance testing logic. My favorite response to this is offered by a fellow editor who responds, “How did you know your results were not trying to avoid being statistically significant?”
- Don’t say things like “the statistical significance testing evaluated whether the results were ‘due to chance.’” This language gives the impression that replicability is evaluated by statistical significance testing.

## ASSUMPTIONS IN PARAMETRIC HYPOTHESIS TESTING

When comparing two means, a family of statistical tests called parametric tests is used. **Parametric statistics** are methods that “make assumptions about populations from which the samples were drawn” (Reinard, 2001, p. 341). Four major assumptions underlie the use of parametric tests:

- Interval or ratio level measurement of dependent variables;
- Randomization in sampling and any assignment of events to experimental and control conditions;
- Normal probability distribution of dependent variables; and
- Equal (homogeneous) variances of the dependent variable in the population (and the corresponding requirement that sample variances remain equal within the limits of sampling error).

Many of these assumptions are not about sample data; they are inferences about population elements. For instance, the one assumption states that the populations have normal probability distributions. But because population characteristics rarely are known to the researcher, unbiased sample statistics are taken as the next best indicator of these characteristics.

Researchers, therefore, look at sample data to get evidence about whether the assumptions have been met in the population as a whole.

### Effects of Violating Parametric Assumptions

Naturally, scholars are interested in what happens if the assumptions underlying these parametric tests are not satisfied. Some matters, such as the required level of measurement and randomization, seem fairly firm. If one wishes to ask how rarely one's study results could be found by random sampling error, it is vital for researchers to reference distributions with randomness in mind. Though there may be controversy regarding whether many measures used in communication research really are interval level measures, regardless of the way the researcher comes down on the issue, using at least quasi-interval measurement is presumed.<sup>7</sup>

Univariate parametric tests seem to be resistant to the effects of violating the assumption of normality (see classic studies by Boneau [1960] and Hsu and Feldt [1969]). When the sample sizes are at least 15 in each condition, the actual number of Type I errors tends to be off by an average of only  $\pm 1\%$ . If the sample sizes are under 6 per condition, a skewed distribution can lead to more Type I errors than the  $\pm 1\%$  range limit. If there is a nonnormal distribution, researchers may want to know why. According to the central limit theorem (Chapter 4), distributions of means will tend toward normality as sample sizes used to compute the means are increased. Hence, if one still finds nonnormal distributions, it may be that some uncontrolled variables are introducing nonrandom influences that should be isolated and studied. What should researchers do if the distributions are nonnormal and sample sizes do not permit one to believe that assumptions have been satisfied? One option is to use nonparametric test alternatives. This approach, however, frequently reduces statistical power (Hodges & Lehmann, 1956; Tanizaki, 1997) and introduces bias when multiple violations of assumptions exist (Zimmerman, 1998).

Another option is to use transformations of nonnormal data. Yet another option is to employ a "robust" statistic, such as *Yuen's t* (Yuen, 1974).<sup>8</sup> The ideal solution, of course, involves taking steps to avoid the problem of nonnormal distributions—by use of adequate sample sizes and control of extraneous variables when sampling.

Studies of the assumption of homogeneous variances of the dependent variable in the population have revealed that the effect of violating this assumption usually is trivial if sample sizes are equal in comparison groups (Glass, Peckham, & Sanders, 1972). If sample sizes are not equal, the impact still is not great unless the ratio of the largest to the smallest sample size is more than 5 to 4. If the variances are unequal and the largest variance comes from the group with the

---

<sup>7</sup>As was mentioned in Chapter 2, there is some controversy regarding whether most measures in communication studies and the social sciences are interval or quasi-interval data. Over the years, Monte Carlo simulation studies have revealed that the effects of true intervality are relatively unimportant for the sorts of data typically found by social science researchers (Baker, Hardyk, & Petrino, 1966; Borgatta & Bohrnstedt, 1980). For a review of this controversy, see Velleman and Wilkinson (1993).

<sup>8</sup>This method involves trimming data to compute means and winsorizing data to compute a measure of within-groups variance (see Chapter 3). Though it has some effect on overcoming the problem of heterogeneous variances, *Yuen's t* primarily addresses the difficulties of nonnormal distributions. Yet, the method suffers criticisms of winsorizing and trimming generally, including the charge that it may give misleading results because it arbitrarily dismisses actual unexplained variability that probably should not be dropped arbitrarily. Furthermore, methods using trimmed means often have lower power than methods that use standard nonparametric techniques (Keselman & Zumbo, 1997).

largest sample size, Type I error actually would be lower than the announced alpha risk. So, researchers would think that they were rejecting the null hypothesis at the .05 level when they really were rejecting it at the .04 or .03 level. Thus, the test would be increasingly conservative. If the largest variance comes from the group with the smallest sample size, the resulting Type I error would be greater than the announced alpha risk. So, researchers would claim rejecting null hypotheses at the .05 level when, in fact, they were rejecting null hypotheses at a .06, .07, or greater probability level. In other words, the test would be increasingly likely to reject null hypotheses erroneously. It might be mentioned that if the heterogeneity in variances accompanies other violations, the violations could be increasingly important (Lix & Keselman, 1998).

## Testing the Assumptions

There are formal ways to test assumptions of normal distributions and homogeneous variances. These two matters will be examined.

### Testing for Normal Distributions

The assumption of an underlying normal probability distribution in the population is often checked by looking at a plot of sample data. Examining skewness and kurtosis statistics may be all that is necessary. Yet, there are other ways to check on such information. The Lilliefors modification of the Kolmogorov-Smirnov one-sample test (often called the **Lilliefors test** for normality) may be used to test the assumption of normality (Lilliefors, 1967). The Lilliefors test transforms data into  $z$  scores. Then, the cumulative frequency distribution of the data is compared to the cumulative frequency distribution that would be expected based on the  $z$  values. For instance, suppose there were 10 scores on a measure of interpersonal solidarity: 5, 10, 14, 6, 8, 8, 7, 11, 9, 12. Arranging the scores from the lowest to highest  $z$  scores produces the following:

Scores	5	6	7	8	8	9	10	11	12	14
CFD	0.1	0.2	0.3	0.5	0.5	0.6	0.7	0.8	0.9	1
$z$ score	-1.43	-1.08	-0.72	-0.36	-0.36	0	0.39	0.72	1.08	1.79

The CFD is the cumulative frequency distribution of scores (expressed as proportions—with 10 scores, each score is one tenth of the total or 0.1). Because there were two people with scores of 8, they both share the same location on the cumulative distribution. The method involves looking at the discrepancy between the cumulative frequency distribution for the raw and  $z$  scores. The researcher needs to compute the sample mean and standard deviation for the raw data (in this case, the mean is 9 and the standard deviation is 2.79). Then, the researcher looks at the table of the standard normal curve and identifies the proportion of the area that is to the left of the particular  $z$  score. For instance, the score of 5 corresponds to a  $z$  score of -1.43:

$$z = \frac{x - \bar{x}}{s} = \frac{5 - 9}{2.79} = -1.43.$$

Checking the table of the standard normal curve reveals that this  $z$  score identifies the location of the area under the standard normal curve that exceeds .0764 of all scores. As the figure below shows, this value is placed in the row identified as “area below  $z$ .” Differences in cumulative frequency distributions are inserted in a separate row in the table of comparisons.

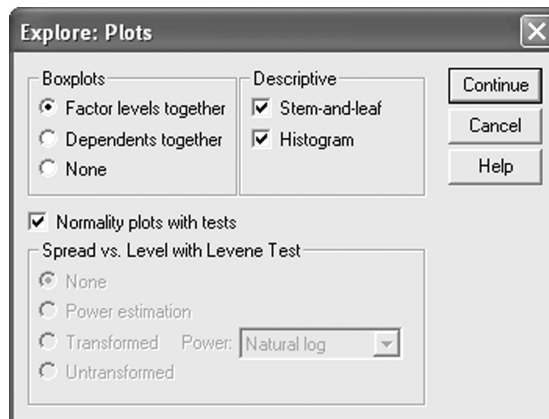
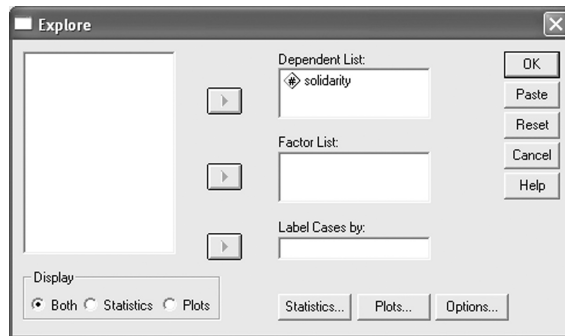
## 156 INFERENCE STATISTICS

Scores	5	6	7	8	8	9	10	11	12	14
CFD	0.1	0.2	0.3	0.5	0.5	0.6	0.7	0.8	0.9	1
z score	-1.43	-1.08	-0.72	-0.36	-0.36	0	0.39	0.72	1.08	1.79
Area below z	0.08	0.14	0.24	0.36	0.36	0.5	0.64	0.76	0.86	0.96
Difference	0.02	0.06	0.06	0.14	0.14	0.1	0.06	0.04	0.04	0.04

The largest difference is .14. The critical values are found in Appendix C.12. For alpha risk of .05, the minimum critical difference is .258. Because the test statistic is smaller than this critical value, the assumption of a normal distribution continues to be tenable. The Lilliefors test has been shown to be powerful, especially when detecting “heavy-tailed” distributions (Young & Seaman, 1990).

To use SPSS to produce such a test, the researcher selects *Descriptive Statistics* from

the *Analyze* menu. On the drop-down menu, the researcher then selects *Explore...* On the dialog box that appears, the researcher moves the dependent variable into the “Dependent List:” field. In this case, the same data from the example above are used. Hence, the “solidarity” variable is moved by use of the arrow button. The box for both plots and statistics is checked, though the researcher might choose only the statistical analysis if desired.



Clicking on the *Plots...* button produces a dialog box in which the specific “Normality plots with tests” choices may be checked.

Among other things, the results of the analysis include the following table. The Kolmogorov-Smirnov value is .14, which corresponds to the results produced in our calculations above. This value is associated with a probability value of .20, which is well above the standard of .05 used to identify statistically significant deviations from normality. Hence, this test did *not* suggest that the assumption of an underlying normal distribution was untenable.

### Tests of Normality

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
solidarity	.140	10	.200 <sup>*</sup>	.979	10	.962

\*This is a lower bound of the true significance.

a. Lilliefors Significance Correction.

A popular alternative is the *Shapiro-Wilk test*, which remains powerful even when sample sizes are as small as 20 (Wilk, Shapiro, & Chen, 1965). This test examines the null hypothesis that the sample distribution is normal. Hence, a statistically significant difference means that the distribution is not normal. In this case, the probability associated with this test was .962, a value suggesting that the assumption of a normal distribution remained tenable for these data.

Another choice is the Anderson-Darling test for normality (T. W. Anderson & Darling, 1954).<sup>9</sup> Though involving more complicated computations, the Anderson-Darling test is more powerful than the Lilliefors modification of the Kolmogorov-Smirnov test (Crown, 2000; Spinelli & Stephens, 1987; Stephens, 1974). Computer programs have been developed for this test (Calzada & Scariano, 2002), and a link to one can be found on this chapter's Web site.

### Testing for Homogeneous Variances Among Two or More Groups

To test the equality of variances, there are several options. One of the most popular is the *F* test (sometimes called  $F_{\max}$ ) for the equality of two variances. The null hypothesis to be tested is  $H_0: \sigma_1^2 = \sigma_2^2$ . This formula (sometimes known as  $F_{\max}$ )<sup>10</sup> takes the largest variance and divides it by the smallest variance:

$$F = \frac{\sigma_{\text{largest}}^2}{\sigma_{\text{smallest}}^2}.$$

Though popular, this measure tends to exaggerate the chances of finding heterogeneous variances as sample sizes increase. An alternative formula is *Levene's test*, which subtracts each score from its cell mean and then performs a test called analysis of variance on the difference scores. Other tests often are suitable options for different sorts of data.<sup>11</sup> In each case, a distribution is referenced to see if the test statistic falls in the critical region corresponding

<sup>9</sup>The discrete Anderson-Darling test statistic is

$$A^* = \left\{ \frac{-1}{n} \left[ \sum_i^n (2i-1) [\ln(p_i) + \ln(1-p_{n+1-i})] \right] - n \right\} \left( 1 + \frac{.75}{n} + \frac{2.25}{n^2} \right).$$

In this formula,  $p_i$  is the cumulative probabilities for each value of the variable (transformed into  $z$  scores). To estimate the significance of the test statistic, the following formula may be used:  $\alpha \approx 3.6789468e^{-\frac{A^*}{.1749916}}$  (see Nelson, 1998).

<sup>10</sup>Another related test, sometimes called Hartley's *H*, is, in fact, the *F* test for which there are equal cell sizes and for which the table of critical values has been simplified.

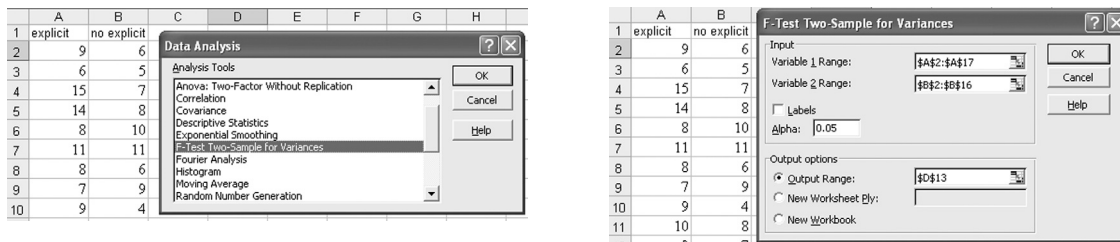
<sup>11</sup>For instance, when sample sizes are unequal and when any of the variances is very small, Cochran's *C* often is indicated.


## 158 INFERENCE STATISTICS

to the decision rule the researcher sets. If the test statistic is located in the critical region, the null hypothesis (of equal variances) is rejected. If not, the assumption continues to “hold.”


For the data found in the two-sample test in the example found in Table 7.4 (in the section “Conducting the Hypothesis Test for the Difference Between Two Sample Means,” pp. 165–166), a test of the assumption of heterogeneous variances is included using the  $F$  test. The test statistic of 1.68 is smaller than the critical value of  $F$  (2.95 with 17 and 14 degrees of freedom at the assigned alpha risk). Hence, the assumption of homogeneous variances cannot be rejected.

The test can be completed by Excel. After the data have been placed in separate columns (perhaps labeled “explicit” and “no explicit” as shown on page 166), the researcher selects *Data Analysis* . . . from the *Tools* menu. In the “Data Analysis” dialog box, the “F-Test Two-Sample for Variances” is highlighted. Then the researcher clicks on the *OK* button.



“The F-Test Two-Sample for Variances” dialog box (above right) includes fields into which the researcher must identify the location of the scores for each of the groups. The step is accomplished by clicking on the  symbol in the “Variable 1 Range:” field. This step puts the researcher on the spreadsheet where data are located. By clicking on the first cell in which data appear for the

F-Test Two-Sample for Variances		
	Variable 1	Variable 2
Mean	9.625	8
Variance	8.383333	5
Observations	16	15
df	15	14
F	1.676667	
P(F<=f) one-tail	0.170421	
F Critical one-tail	2.463004	

first group of scores and highlighting the remaining data in that row (by holding down the left mouse button), the cell range in which the first group’s data are found can be input. Clicking on the  symbol on the drop-down menu returns the researcher to the main dialog box. Then, the process of highlighting the cell range may be completed for the second group of data scores. If the researcher has highlighted cells that contain variable or group labels, the “Labels” box should be checked to prevent attempts to analyze group variable names as data. The researcher also must specify a location in which the output is to be placed. In this case, the researcher has identified the “Output Range:” to begin at cell D13. Clicking the *OK* button causes the output to be produced. As can be seen on the left, the observed  $F$  statistic of 1.67 has an accompanying probability level (“P(F <= f) one-tail”) that is larger than the standard .05 (or smaller) probability necessary to reject the null hypothesis. Hence, the researcher concluded that the assumption of homogeneous variances continues to hold.

		Levene's Test for Equality of Variances	
		F	Sig.
attitude	Equal variances assumed	.030	.863
	Equal variances not assumed		

SPSS also provides ways to examine the assumption of homogeneous variances. Regularly provided as part of the  $t$  test, for independent samples, SPSS reports Levene’s test. Because getting this output will be covered in the section on using SPSS for the two-sample  $t$  test, the individual steps to get this output will not also be presented here. Because Levene’s test is robust to violations of the assumption of

normal distributions, it often is preferred for tests of computing homogeneous variances. In this case, the Levene test produced an  $F$  ratio of .03, which produced a very high significance “Sig.” value. Thus, the null hypothesis of equal variances in the two groups could not be rejected. If the Levene test had indicated statistically significant differences, the researcher would have been invited to use the  $t$  test computation method in which “equal variances [are] not assumed.”

Aside from the desire to determine if heterogeneous variances have affected actual risk levels in hypothesis testing, there is another reason to look at heterogeneous variances. Two categories of influences can be responsible for unequal variances. First, there may be **ceiling or floor effects** in the data. In other words, it may be that there are conditions where the means are so high (or low) that there is not enough room left in the measurement range for the scores to show normal spread.<sup>12</sup> By looking at the means and checking that the cells with the means near the top of their measurement range also have low variances, a ceiling effect may be identified. Similarly, looking at the cells with means near the low end of the measurement range and checking that they also had small variances would reveal floor effects. If more than two groups are compared, a correlation between the cell means and variances could be computed. A high inverse correlation would point to a ceiling effect, and a high direct correlation would suggest a floor effect. Of course, the actual means have to be examined to tell if the ceiling or floor effect actually is present.

Second, if there is no ceiling or floor effect, the heterogeneous variances indicate the presence of **participants by treatments interactions**. This condition suggests that there is at least one other additional variable—and perhaps many more than one—that is mixing nonrandom variation with the chief variables in the study. Thus, there are uncontrolled variables affecting the observed relationships. Researchers would be encouraged to reconsider their studies and to look for additional variables that should be included in future research.

## COMPARING SAMPLE AND POPULATION MEANS

Comparisons of samples with some population mean—such as a historical standard or simply a defining characteristic—can be completed under two conditions. First, the population standard deviation may be known. Second, sample standard deviations may be substituted for population standard deviations.

### When the Population Standard Deviation Is Known

Under many circumstances, if a researcher wishes to compare a mean from a sample against characteristics of a well-defined population, the  $z$  test (using the standard normal curve) would be an appropriate option. In addition to requiring that population means and

---

<sup>12</sup>As Paul R. Cohen (1995) explains:

Technically, a ceiling effect occurs when the dependent variable,  $y$ , is equal in the control and treatment conditions, and both are equal to the best possible value of  $y$ . In practice, we use the term when performance is nearly as good as possible in the treatment and control conditions. Note that “good” sometimes means large (i.e., higher accuracy is better) and sometimes it means small (e.g., low run times are better), so the ceiling can be approached from above or below. A ceiling thus bounds the abstract “goodness” of performance. Floor effects occur when performance is nearly as bad as possible in the treatment and control conditions. Again, poor performance might involve small or large scores, so the “floor” can be approached from above or below. (p. 80)

## 160 INFERENCE STATISTICS

standard deviations be known, the use of  $z$  requires fairly large samples, at least 30 events. The formula for  $z$  is

$$z = \frac{X - \mu}{\sigma}.$$

To adapt to a comparison with means instead of  $X$  scores, one might imagine that one could simply use the following modification:

$$z = \frac{\bar{X} - \mu}{\sigma}.$$

Using this formula would be *incorrect*. The scores compared in the numerator and denominator of the formula are not the same types of data. The sigma ( $\sigma$ ) score in the denominator is the standard deviation of raw scores, but the numerator does not compare *scores*, but *means*. A distribution of means has a much smaller standard deviation than a distribution of scores. The reason is found in the central limit theorem, which was introduced in Chapter 4. This theorem states that a sampling distribution of means tends toward a normal distribution with increased sample sizes regardless of the shape of the parent population. As a result of the central limit theorem, the standard deviation of means ( $\sigma_{\bar{X}}$ , called the “standard error of the mean”) could be computed as

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}.$$

So, a new sample could be compared with a population mean, but the following formula is required:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}},$$

where

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

In this case, the size of the new sample ( $n$ ) is substituted for  $N$ , the number of events in the population.

The  $z$  test also can be used to determine how often a set of findings might occur at random. For instance, one might wonder how unusual it would be to find a sample of 50 people with a mean of 70 or higher from a population in which the mean on a measure of communication apprehension is 65.6 and the standard deviation is 15.3. Inserting these numbers into the formula reveals these results:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{70 - 65.6}{\frac{15.3}{\sqrt{50}}} = \frac{4.4}{2.16} = 2.04.$$

Looking up the value on the  $z$  table (a portion of which is shown below on the left) reveals that the area from the mean to a  $z$  score of 2.04 includes .4793 of the total area. Thus, only .0207 lies above that point. So, we may say that in the population, only 2.07% of the time will one find a random sample of 50 with a mean score of 70 or above. An example of the  $z$  test of statistical significance is found in Table 7.2.

$z$	.00	.01	.02	.03	.04
0.0	.0000	.0040	.0080	.0120	.0160
0.1	.0398	.0438	.0478	.0517	.0557
2.0	.4772	.4778	.4783	.4788	.4793



**Table 7.2** The One-Sample  $z$  Test

The Personal Report of Communication Apprehension has a known population mean (from studies of 52 university samples including more than 25,000 participants) of 65.6 and a standard deviation of 15.3 (see McCroskey, Beatty, Kearney, & Plax, 1985). Yet another study of 64 pharmacy students found an initial communication apprehension mean of 62.14 (Berger & McCroskey, 1982).

The one-sample  $z$  test may be used to test the null hypothesis that there is no difference between the sample mean and the population mean,  $H_0: \mu_{\text{pharmacy students}} = \mu_{\text{population}}$ .<sup>13</sup> If the null hypothesis is tested with a two-tailed (nondirectional) test featuring an alpha risk of .05, the critical value of  $z$  (the point where the critical region begins) would be  $\pm 1.96$ . The test statistic would be computed as follows:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z = \frac{62.14 - 65.6}{\frac{15.3}{\sqrt{64}}} = \frac{-3.46}{1.91} = -1.81$$

Thus, the null hypothesis would not be rejected. One would conclude that there is no significant difference in the mean communication anxiety of pharmacy students and the general population.

<sup>13</sup>The null hypothesis sometimes is stated as  $H_0: \mu = \mu_0$ . In this case,  $\mu_0$  represents a particular assigned population value for purposes of comparison.

## When the Population Standard Deviation Is Unknown

Researchers often have only sample data. Thus, they cannot always use the  $z$  test, because they may not know the population standard deviation. An alternative is to substitute the sample standard deviation,  $s$ , an unbiased estimate of the population standard deviation,  $\sigma$ .

But the  $z$  test also requires sample sizes of at least 30. So, if either the population standard deviation is not known or the sample size is below 30, using  $t$  (or Student's  $t$ )<sup>14</sup> is required. By either design or accident, the symbol in the  $t$  distribution emphasizes that the  $t$  test focuses on testing the difference between two means.<sup>15</sup>

The  $t$  distribution shares many characteristics with the standard normal curve. In fact, as an inspection of the table of critical values of  $t$  will reveal (Appendix C.4), with an infinite sample size, the standard normal curve and the  $t$  distribution are identical. But as sample sizes get

<sup>14</sup>The Student  $t$  distribution has nothing to do with educational research. William Sealy Gosset trained as a mathematician at Oxford University and worked for the Guinness Brewery in Dublin, Ireland. Guinness is the same organization responsible for Guinness Stout Malt Liquor and the famous book of world records. While doing experiments related to temperature, he developed the  $t$  distribution and the  $t$  test. Because Guinness had a policy that prevented employees from publishing under their own names, he published his discovery under the pen name "Student" (1908), and the label has stuck.

<sup>15</sup>There is a song ("Tea for Two") from a 1924 musical called *No, No, Nanette*. Aside from constituting a way for modern students to remember the purpose of the  $t$  test, it has nothing to do with Student's  $t$  test.

162 INFERENCE STATISTICS

smaller and smaller, the  $t$  distribution tends to flatten out. To use the  $t$  distribution, one must identify the **degrees of freedom**, which is calculated as the number of events in a sample minus the number of parameters estimated from sample statistics. By looking at the formula for the test statistic, one may identify the number of  $X$ -bars ( $\bar{X}$ ) used to estimate population means.

The  $t$  test making comparisons of a sample mean and a population mean uses the following formula:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

where  $s_{\bar{X}}$  is equal to  $\frac{s}{\sqrt{n}}$ . This formula differs from the  $z$  test by the use of the *sample* standard deviation,  $s$ , instead of the *population* standard deviation,  $\sigma$ . Thus, the  $t$  test is actually one more building block formula that follows a basic pattern of statistical uses that were identified earlier in this text.

<b>Building Block Formula Box 6: One-Sample <math>z</math> and <math>t</math></b>	
$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$	$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$

Because the one-sample  $t$  formula includes only one  $\bar{X}$  in the numerator, degrees of freedom are equal to  $n - 1$ . These degrees of freedom may be used to enter the table found in Appendix C.4 to find critical values of  $t$ . (A portion of this table is reproduced below.)

degrees of freedom	.10	.05	.025	.01	.005	← Alpha risk for <i>one-tailed</i> tests
	.20	.10	.05	.02	.01	← Alpha risk for <i>two-tailed</i> tests
1	1.078	6.314	12.706	31.821	63.657	Degrees of freedom are computed by taking the number of events in the study and subtracting the
2	1.886	2.920	4.303	6.965	9.925	
...	...					
19	1.328	1.729	2.093	2.539	2.861	

In addition to testing the difference between a sample mean and a standard or a historical mean, the one-sample  $t$  test also is useful when a researcher wishes to examine whether a sample is representative of the population. In particular, researchers may use this test when they

wish to tell if the population and sample means actually are from the same populations.

<b>Table 7.3</b>	The One-Sample $t$ Test
<p>The population mean for class grades of undergraduates at a university is 2.59 (possible range: 0 to 4). A researcher noticed that a group of 17 students taking courses in Intercultural Communication had the following grades in that class: 4, 3, 2, 2, 2, 2, 1, 1, 1, 4, 3, 2, 2, 2, 1, 1, and 1.</p> <p>The mean of this sample is 2, and the standard deviation is 1.0. One may wonder if this sample is unrepresentative of the ordinary population of students. Using the one-sample <math>t</math> test, the researchers may test the null hypothesis <math>H_0: \mu_1 = \mu_0</math>, which states that the mean grade of the sample of students taking Intercultural</p>	

**Table 7.3** (Continued)

Communication classes is equal to the population mean of 2.59 (in fact, the null hypothesis could have been written as  $H_0: \mu_1 = 2.59$ ). With a sample of 17, degrees of freedom are  $n - 1$  or 16. Using a two-tailed  $t$  test with alpha risk at .05, the critical value of  $t$  is 2.120. Using the one-sample  $t$  test, one would find:

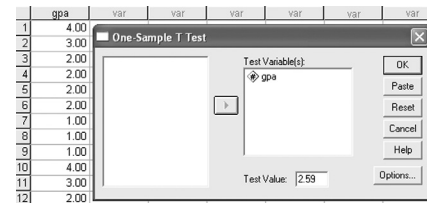
$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$z = \frac{2. - 2.59}{\frac{1}{\sqrt{17}}} = \frac{-.59}{\frac{1}{4.1231}} = \frac{-.59}{.2425} = -2.433.$$

Because the test statistic is greater than the critical value (remember, the negative value does not mean subtraction, but a location on the  $t$  distribution), the null hypothesis would be rejected. One would conclude that the sample is not representative of the population. Thus, researchers would want to determine why and explore possible explanations.

### Using SPSS for the One-Sample $t$

Though Excel does not have built-in functions that permit the direct computation of the one-sample  $t$  test, SPSS has such an option. To use the SPSS package for this application of the  $t$  test, the researcher starts by clicking on the *Analyze* menu followed by selecting *Compare Means* from the drop-down menu that appears. Then, the *One-Sample T Test...* option is selected. In the “One-Sample T Test” dialog box, the researcher selects the sample measure of interest and uses the arrow key to move it to the “Test Variable(s):” field. As an example, we may use the same data as employed in Table 7.3, in which case the variable “gpa” is selected for analysis. In this example, the population mean against which comparisons are made is 2.59. Hence, this value is entered into the “Test Value:” field. To execute the program, the researcher clicks the *OK* button.



The output shows that the probability of finding a difference such as that observed here by random sampling error is only .027, or 2.7 chances out of a hundred.

#### One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
gpa	17	2.0000	1.00000	.24254

#### One-Sample Test

	Test Value = 2.59					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
gpa	-2.433	16	.027	-.59000	-1.1042	-.0758

Because this probability is below the .05 usually employed as a decision rule, most researchers would reject the null hypothesis and conclude that there is a difference between this sample mean and the traditional population mean.

## COMPARING THE MEANS OF TWO SAMPLE GROUPS: THE TWO-SAMPLE $t$ TEST

Researchers often do not have a population standard against which to make comparisons, but they often have control groups to help them draw conclusions. If researchers wish to compare two sample groups, the two-sample  $t$  test is appropriate.

### Using $t$ as a Sampling Distribution of Mean Differences

To make this comparison, the null hypothesis takes the form  $H_0: \mu_1 = \mu_2$ , which tests that the dependent variable mean of the first group is equal to the mean of the second group. Actual alternative research hypotheses may be directional or nondirectional.

In addition to the assumptions of parametric statistics generally, the two-sample  $t$  test also assumes **independence**. This assumption means that the events in the sample are unaffected by each other. In many cases, this sort of thing is quite reasonable, but in some cases, it is not. For example, some researchers sample college classrooms. If these classes are required in an academic major, students probably interact with each other and may discuss things that happen in their classes, such as a new teaching approach or a study in which they are participating. Thus, the samples of student responses from such classes may not be completely independent.

### Conducting the Hypothesis Test for the Difference Between Two Sample Means

To examine a hypothesis about two means, such as  $H_1: \mu_1 > \mu_2$ , the researcher must state a null hypothesis for direct testing. Then, it is useful for the researcher to test the assumption of homogeneous variances before computing the actual  $t$  test statistic. As we have seen:

- If sample sizes are equal in the two groups, the result of heterogeneous variances on Type I error rate is negligible. If variances are equal (within the limits of sampling error), the so-called pooled standard deviation ( $s_p$ ) or “equal variances” model may be used.
- But if sample sizes are unequal, a significant heterogeneity in variances requires the researchers to use the “separate variance” (also called unequal variances  $t$ ) method of conducting the independent samples  $t$  test. Using the pooled standard deviation when assuming equal variances, the formula for  $t$  is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The formula looks a lot like the one-sample  $t$  test formula. With equal sample sizes, the pooled standard deviation ( $s_p$ ) is simply the square root of the average of the variances. With unequal sample sizes, the following formula for the pooled standard deviation is used:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - 1 + n_2 - 1}}$$

Because there are two sample sizes, instead of dividing the standard deviation estimate by  $\sqrt{n}$ , the pooled standard deviation is multiplied by the square root of the

fractions  $\frac{1}{n}$  (equivalent to dividing a term by  $n$ ).<sup>16</sup> An alternative formula for  $t$  with unequal variances is the separate variance estimate in which the variance for the control group is used as the measure of variance in the denominator of the  $t$  statistic. In addition, the following is a popular formula (used in Excel, for instance) employed when the variances are unequal:

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

To determine if a statistically significant difference exists between the two means, researchers enter the  $t$  distribution with degrees of freedom that adjust sample sizes for the number of population parameters estimated from sample means.

- If the variances are equal between the two groups, degrees of freedom are equal to  $n - 2$  (because there are two sample means in the numerator of the  $t$  formula).
- If the variances are significantly different, the formula for degrees of freedom is:

$$d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

This formula usually yields a number with a decimal point. So, the result must be rounded to a whole number.

#### Building Block Formula Box 7: Independent Samples $t$ Test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

**Table 7.4** Independent Samples  $t$  Test

A researcher wondered whether it would be more persuasive for a speaker to include an explicit statement of the advocated position even when the audience was initially hostile to the topic. Thus, as part of a pilot study, a randomly selected group of 16 individuals was given a message with an explicit statement of the persuasive proposition. A control group of 15 individuals was given the message with the explicit statement omitted. The chief dependent variable was attitude toward the topic, measured on a set of interval level scales with possible scores ranging from 3 (most negative attitude) to 21 (most positive attitude).

The study hypothesis was  $H: \mu_{\text{explicit statement of proposition}} > \mu_{\text{no explicit statement of proposition included}}$ . The null hypothesis was  $H_0: \mu_{\text{explicit statement of proposition}} = \mu_{\text{no explicit statement of proposition included}}$ . The researcher tested the null hypothesis with alpha risk of .05.

(Continued)

<sup>16</sup>Multiplying a value by the fraction  $\frac{1}{n}$  produces the same result as dividing a value by  $n$ .

**Table 7.4** (Continued)

The following data were collected:

<i>Explicit Statement Group</i>	<i>No Explicit Statement Group</i>
9	6
6	5
15	7
14	8
8	10
11	11
8	6
7	9
9	4
10	8
9	7
10	12
9	8
4	10
12	9
13	
Mean = 9.63	8
Variance = 8.38	5

- Using a one-tailed  $t$  test, with  $n - 2$  degrees of freedom ( $31 - 2 = 29$ ), the critical value of  $t$  is 1.699.

- Computing  $t$  assuming equal variances, the following computations are made:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{9.63 - 8}{2.6 \sqrt{\frac{1}{16} + \frac{1}{15}}}$$

$$t = \frac{1.63}{2.6 \sqrt{.13}}$$

$$t = \frac{1.63}{2.6 * .36} = \frac{1.63}{.94} = 1.73$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - 1 + n_2 - 1}}$$

$$s_p = \sqrt{\frac{(16 - 1)8.38 + (15 - 1)5}{16 - 1 + 15 - 1}}$$

$$s_p = \sqrt{\frac{195.7}{29}} = \sqrt{6.75} = 2.6$$

Because 1.73 is greater than the critical value (1.699), the null hypothesis is rejected.

- To determine the effect size in terms of a correlation, the following formula is used:

$$r = \sqrt{\frac{t^2}{t^2 + \text{degrees of freedom}}}$$

Inserting the information from this pilot study, the following computations may be completed:

$$r = \sqrt{\frac{2.99}{2.99 + 29}} = \sqrt{.09} = 0.3.$$

Thus, using the interpretation guides for correlations found in Chapter 5, this degree of relationship is equivalent to a “slight,” “moderate,” or “medium” relationship.

Test the assumption of homogeneous variances.

- Testing  $H_0: \sigma_1^2 = \sigma_2^2$  at alpha risk of .05:

$$F = \frac{S_{\text{largest}}^2}{S_{\text{smallest}}^2}$$

$$F = \frac{9.63}{5} = 1.68$$

- Critical value:

$$d.f. : = \frac{n_{\text{largest}} - 1}{n_{\text{smallest}} - 1} = \frac{15}{14}$$

$$\text{Critical } F_{(15,14)} \text{ with } \alpha/2: 2.949$$

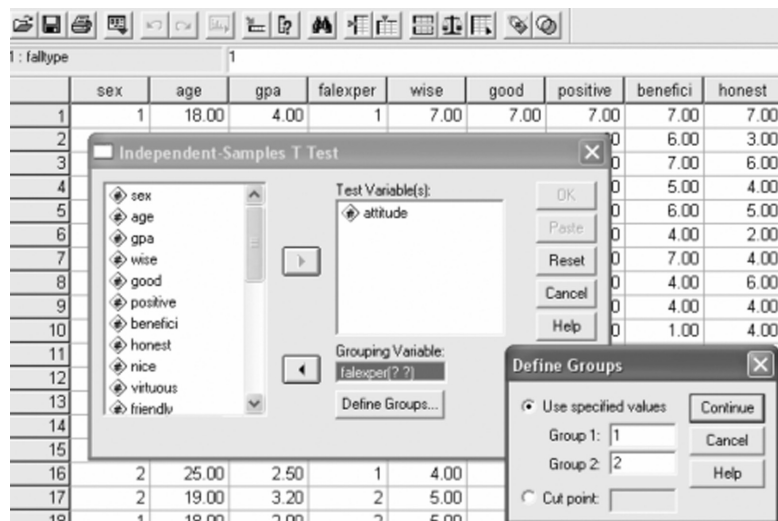
Therefore, no significant heterogeneity of variances was claimed, and the assumption of homogeneous variances was maintained. (When Excel computes this test, it uses a one-tailed probability, in this case involving a critical value of 2.463.)

## Using SPSS and Excel to Compute the Two-Sample $t$

Computers have made easy work of completing statistical hypothesis testing for comparing two sample means. The methods in both SPSS and Excel will be reviewed here.

### SPSS

For the pooled standard deviation and separate standard deviation methods, the  $t$  test can be completed in one step. From the *Analyze* menu, select *Compare Means*. Several alternative  $t$  tests are provided: *Means. . .* includes descriptive statistics and various ways of testing the equality means; *One-Sample T Test. . .*; *Independent-Samples T Test. . .*; *Paired-Samples T Test. . .*; and *One-Way Anova. . .* To illustrate the use of the program here, the choice of an “Independent Samples T Test” will be made. In the dialog box that emerges, the researcher highlights the dependent variables of interest and transfers them to the fields marked “Test Variable(s):”. Separate  $t$  tests will be completed for each of the variables listed as a test variable.



To identify the two groups, a categorization variable is highlighted and then transferred to the box marked “Grouping Variable:”. It is assumed that the two groups are identified by taking such values as 1 or 2 in this variable. But sometimes researchers want to use grouping variables that originally had three or more groups. After the *Define Groups. . .* button is clicked, another dialog box opens and the researcher indicates the values used to identify group one and group two. Sometimes researchers take a continuous variable and break it down into two categories or groups. For instance, a researcher might want everyone with an IQ score equal to or below 100 to be in the first group and all those with higher scores in the second group. In such a case, the researcher would have clicked on the “Cut Point” radio button and entered a number in the field that became active. Participants with scores above that point are placed in the first group, and the rest are identified as members of the second group.

## 168 INFERENCE STATISTICS

**Group Statistics**

	FALEXPER	N	Mean	Std. Deviation	Std. Error Mean
ATTITUDE	1.00	23	19.8696	6.0250	1.2563
	2.00	29	22.3793	3.8952	.7233

variances. A “Sig.” value of .05 or smaller for the Levene test is taken as evidence that the two samples did not have equal variances. If such is found, then the  $t$  test that is based on “Equal variances not assumed” is used. The  $t$  test in the example below shows no significant difference in the means at the .05 level (.075 is found). Of course, these results are for a two-tailed  $t$  test. If a one-tailed test were used, then the results would have been statistically significant (.075 divided by two would have produced a probability of .0375).

After the researcher is done, the *Continue* and *OK* buttons are clicked and the following output at the left appears. The first portion of the output provides simple descriptive statistics about the sample.

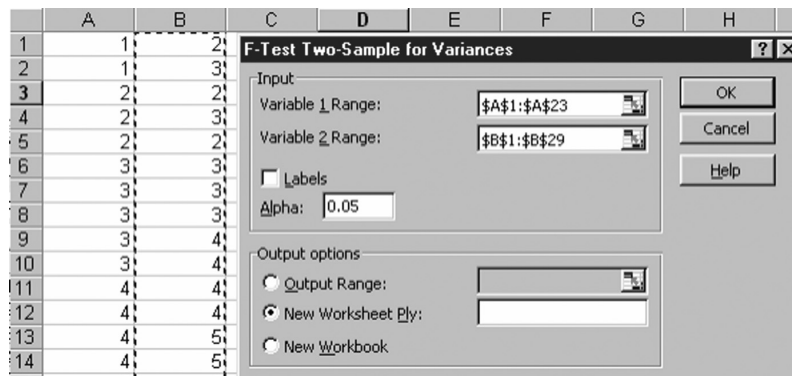
The next section of the output provides information about the Levene test of homogeneous

**Independent Samples 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
ATTITUDE	Equal variances assumed	3.606	.063	-1.817	50	.076	-2.5097	1.3812	-5.2839	.2844
	Equal variances not assumed			-1.731	35.903	.092	-2.5097	1.4498	-6.4500	.4305

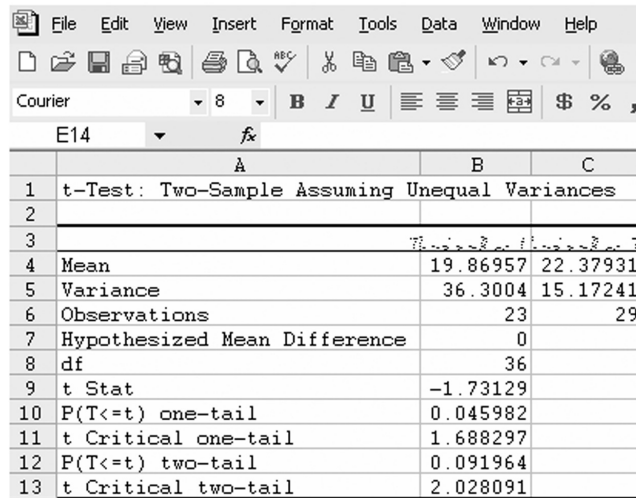
**Excel**

To compute the  $F$  test for homogeneous variances, researchers using Excel select *Data Analysis. . .* from the *Tools* menu. In this example, we will illustrate the use of the  $t$  test when unequal variances are present. Thus, the effort begins with the  $F$  test of differences in variances. In the dialog box that appears, one may select *F-Test Two-Sample for Variances* and click *OK*. Using the highlighting tools, the researcher selects data for the first variable and second variable (indicated in the “Variable 1 Range:” and the “Variable 2 Range:” fields).





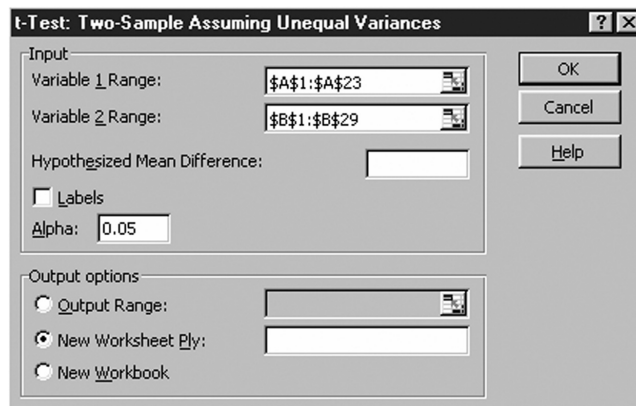
Clicking on the *OK* button reveals the following output.



	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3			
4	Mean	19.86957	22.37931
5	Variance	36.3004	15.17241
6	Observations	23	29
7	Hypothesized Mean Difference	0	
8	df	36	
9	t Stat	-1.73129	
10	P(T<=t) one-tail	0.045982	
11	t Critical one-tail	1.688297	
12	P(T<=t) two-tail	0.091964	
13	t Critical two-tail	2.028091	

As can be seen, the output reveals the results of the *F* test of the difference between variances. Because there is a statistically significant difference in the variances (indicated by a *p* value smaller than .05), the researchers would be invited to use a *t* test for unequal variances.

To use Excel to compute an independent samples *t* test, select *Data Analysis* from the *Tools* menu and then choose *t-Test Two Sample Assuming Unequal Variances*. In the dialog box that appears, select ranges to indicate the scores for variables 1 and 2.



**t-Test: Two-Sample Assuming Unequal Variances**

Input:

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

Labels

Alpha:

Output options:

Output Range:

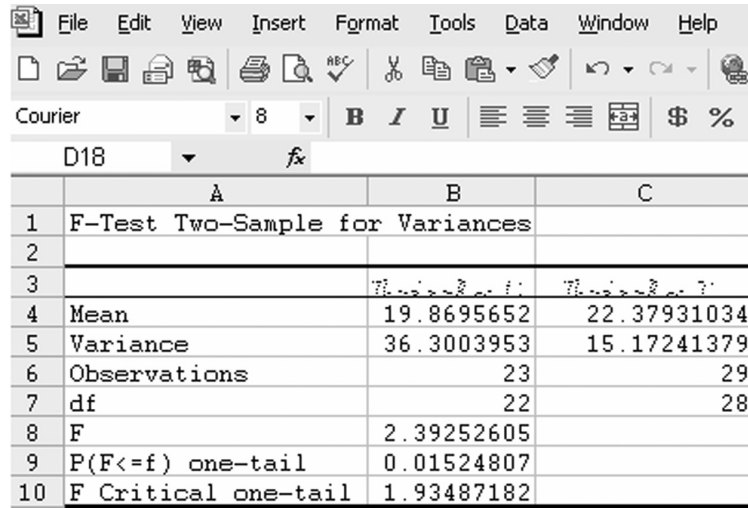
New Worksheet Ply:

New Workbook

Buttons: OK, Cancel, Help

After identifying the variable ranges, click on *OK* to complete the analysis. The result is found on output such as that on page 170. As can be seen, the one-tailed probability level is less than .05, which would indicate a statistically significant difference. If researchers had chosen a two-tailed test, however, the difference would not be statistically significant by usual standards.

## 170 INFERENCE STATISTICS



	A	B	C
1	F-Test Two-Sample for Variances		
2			
3		Mean	Mean
4	Mean	19.8695652	22.37931034
5	Variance	36.3003953	15.17241379
6	Observations	23	29
7	df	22	28
8	F	2.39252605	
9	P(F<=f) one-tail	0.01524807	
10	F Critical one-tail	1.93487182	

## Effect Size Computations

Statistical significance really tells us only how improbable the null hypothesis is when it comes to explaining sample results. But a statistically significant effect may be large or small. To learn if the results are substantial or not, it is useful to look at the size of relationships. To use the language of correlation, researchers may take information from a  $t$  test by use of this formula:

$$r = \sqrt{\frac{t^2}{t^2 + \text{degrees of freedom}}}$$

Squaring this number reveals the proportion of variance in one variable that may be explained by knowledge of variation in the other alone.

## Confidence Intervals for Mean Differences

To say that a mean difference of five points on a scale is beyond what might have been expected to occur by sampling error tells only part of the story. The mean difference is only a single best estimate of the difference, called a **point estimate** because it is a single number. But the true population difference may lie within an interval around the observed differences. An alternative way of using these data employs a **confidence coefficient** or **degree of confidence** followed by an interval (called, not surprisingly, a **confidence interval**) into which the population difference is likely to fall. As shorthand, researchers often report such confidence

intervals with such brief statements as “the 90% confidence interval of the differences in means is 3.5 to 6.5.” The numbers 3.5 and 6.5 are known as the lower and upper **confidence bounds** (or **confidence limits**). The statement indicates that the researcher is 90% “confident” that the difference in means lies somewhere between 3.5 and 6.5 on the measure in use. To be precise, however, a 95% confidence interval means that if a large number of random samples were drawn and confidence intervals were computed, 95% of them would include (or capture) the mean difference parameter of interest. The fact that a 95% confidence interval has been drawn, however, also means that there is a 5% chance that the report made by the researcher is *way off*, not even close.

To compute a confidence interval for differences in means, such as the study found in the independent samples *t* test example in Table 7.4, the following formula may be used:

$$95\% \text{ C.I.} = \bar{X}_1 - \bar{X}_2 \pm (t_{(\alpha/2, d.f.)} * s_{\bar{x}_1 - \bar{x}_2}),$$

where  $\bar{X}_1 - \bar{X}_2$  is the difference in means,  $t_{(\alpha/2, d.f.)}$  is the critical value of *t* found in the *t* table (the  $\alpha$  corresponds to the announced confidence interval [e.g.,  $1 - \alpha$  of .05 corresponds to a 95% degree of confidence]), the  $\alpha/2$  term means that the critical value should be two-tailed (in this case, the two-tailed critical *t* value with alpha risk of .05 and 29 degrees of freedom is 2.045), and  $s_{\bar{x}_1 - \bar{x}_2}$  is the standard error of the mean differences and includes everything in the denominator of the *t* test formula

$$\left( s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right).$$

For the example, the confidence interval may be identified:

$$95\% \text{ C.I.} = 163 \pm (2.045 * .94)$$

$$95\% \text{ C.I.} = 1.63 \pm 1.92$$

In other words, the researcher is 95% confident that the difference between means is somewhere between  $-1.92$  and  $1.63$ . Because the confidence interval includes zero, one would conclude that the mean difference is not significantly different from zero. But the example found a significant difference between the groups. How can this situation exist? The test in the example was a one-tailed test, but the confidence interval was a two-tailed test.

One-tailed confidence intervals can be constructed. One could imagine a person asking only about one side of the confidence interval, such as “what is the *minimum* mean difference improvement that can be expected if the controversial proposition is stated explicitly?” Such a person would not care to know the upper limit, just the *least* improvement to be expected. The difference in the formula is that a one-sided critical *t* value (1.699) is used, such as:

$$95\% \text{ C.I.} = \bar{X}_1 - \bar{X}_2 \pm (t_{(\alpha/1, d.f.)} * s_{\bar{x}_1 - \bar{x}_2})$$

$$95\% \text{ C.I.} = 1.63 \pm (1.699 * .94)$$

$$95\% \text{ C.I.} = 1.63 \pm (1.59)$$

In other words, the researcher is 95% confident that the difference between means is at least 0.04. This number might be taken as the smallest difference for which one might be confident.

## COMPARING MEANS DIFFERENCES OF PAIRED SCORES: THE PAIRED DIFFERENCE $t$

Researchers often give people a pretest, followed by some treatment, and then a posttest. The “before and after” designs gather sample data from every person twice; thus, the samples are not independent. A way to deal with such data is to use the  $t$  test for paired differences.

### Conducting the Hypothesis Test for Paired Differences

Instead of dealing with mean differences among groups, researchers using the paired differences  $t$  test subtract the posttest from the pretest and examine the size of these differences. The null hypothesis in the paired samples  $t$  test is  $H_0: \mu_{\text{difference}} = 0$ . Because the same sample is examined twice, the degrees of freedom for the  $t$  test are based on the number of events, not the number of scores. The degrees of freedom are equal to  $n - 1$ .

The formula for the paired samples  $t$  test is a bit different from those for the other  $t$  tests:

$$t = \frac{\bar{X}_{diff}}{\frac{s_D}{\sqrt{n}}},$$

where  $\bar{X}_{diff}$  is the mean difference between the paired scores (often pretest and posttest; when the research hypothesis speculates that the posttest scores will be higher than the pretest scores, the difference would have a negative sign before it; if the research hypothesis speculates that the posttest scores will be lower than the pretest scores, the difference would have a positive value),  $s_D$  is the standard deviation of the difference scores, and  $n$  is the number of events in the sample (not the number of total scores).

**Table 7.5** Paired Difference  $t$  Test

A researcher explored the hypothesis that students who go through a unit of instruction on communication styles have higher posttest scores on a measure of perceived comfort in communication with difficult people (possible range: 3 to 21) than they did on a pretest. A sample of 19 people was collected, and following the pretest, sample members were given instruction and then posttested. Because the researcher predicted that the posttest scores would be higher than the pretest scores, a negative value is predicted in the one-tailed hypothesis:  $H: \mu_{\text{difference}} < 0$ . The null hypothesis is

$$H_0: \mu_{\text{difference}} = 0.$$

Degrees of freedom are  $n - 1$  or  $19 - 1 = 18$ .

The critical value of  $t$  (one tailed) with alpha risk of .05 is  $-1.734$ .

The following data were collected:

Table 7.5 (Continued)

Pretest	Posttest	Difference
3	5	-2
12	11	1
7	9	-2
20	18	2
18	19	-1
16	18	-2
14	17	-3
15	18	-3
16	18	-2
17	19	-2
7	6	1
8	10	-1
18	17	1
10	14	-4
12	15	-3
14	15	-1
13	15	-2
15	16	-1
16	17	-1
Mean =		-1.37
Standard Deviation =		1.61

$$t = \frac{\bar{X}_{diff}}{\frac{S_D}{\sqrt{n}}}$$

$$t = \frac{-1.37}{\frac{1.61}{\sqrt{19}}} = \frac{-1.37}{\frac{1.61}{4.36}} = \frac{-1.37}{.37} = -3.7$$

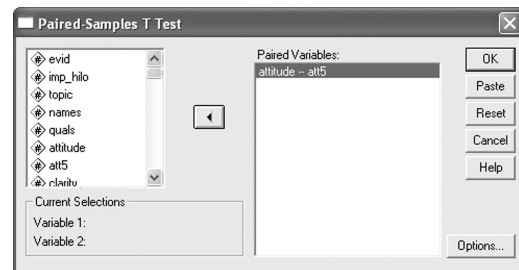
The null hypothesis would be rejected because the test statistic (-3.7) is beyond the critical value of -1.734. (It is helpful to remember that the negative sign is not a symbol for subtraction but instead a way to identify the place where the critical region begins. In this case, the critical region starts at -1.734 and extends out to  $-\infty$ . For these data, the test statistic of -3.7 falls into the critical region.)

## Using SPSS and Excel to Compute the Paired Differences $t$

Both SPSS and Excel include ways to conduct the paired differences  $t$  test. The basic format remains relatively unchanged from that which has been described previously. Hence, only the chief differences will be examined in this brief treatment.

### SPSS

From the *Analyze* menu, researchers select *Compare Means* from the drop-down menu, followed by *Paired-Samples T Test*. . . from the subsequent menu that appears. In the dialog box that appears, two separate variables for each participant must be selected and transferred into the “Paired Variables:” field by highlighting them and moving them with the arrow key.



Clicking the *OK* button causes the program to execute. The output produced by this process is a little different from the previous example. In addition to a set of descriptive statistics, a measure of correlation between the two measures appears. This correlation is not a measure of effect size, but a measure of association between the two sets of scores.

## 174 INFERENCE STATISTICS

## Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 ATTITUDE & ATT5	66	.843	.000



## Paired Samples Test

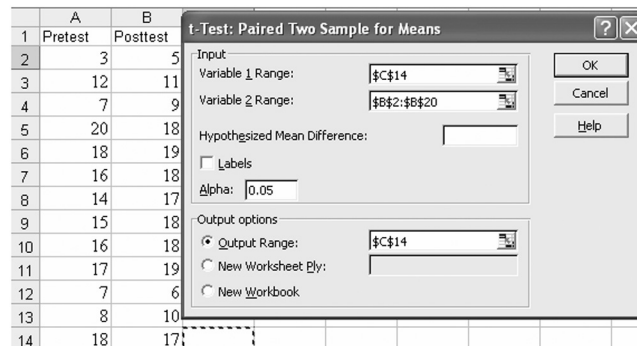
	Paired Difference					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 ATTITUDE - ATT	-.67	3.788	.466	-1.60	.26	-1.430	65	.158

A regular feature of the  $t$  test in SPSS is the presentation of confidence intervals. If the confidence interval contains zero, then no statistically significant difference between the means is revealed. In this case, because the 95% confidence interval extends from  $-1.6$  to  $0.26$ , it includes zero, which indicates no statistically significant difference.

This correlation should be reasonably large when there is no statistically significant difference between two sets of scores. If the difference between the pairs scores is large and the correlation is low, researchers may wish to consider whether they really wished to use the independent samples  $t$  test instead. As can be seen in the output, there was no statistically significant difference in the paired scores.

**Excel**

The paired samples  $t$  test also may be computed with Excel. The researcher begins by selecting *Data Analysis...* from the *Tools* menu. Then, on the drop-down menu that appears, the researcher chooses *t-Test: Paired Two Sample for Means* option. In the “t-Test: Paired Two Sample for Means” dialog box, the researcher clicks the  symbol in the “Variable 1 Range:” field and highlights the cells where the first set of data scores are located. Clicking on the  symbol on the drop-down menu returns the researcher to the main dialog box. Then, the same process can be followed to identify the “Variable 2 Range:” of data.



The “Labels” box should be checked if the researcher has highlighted any cells with variable or group labels. Before clicking on the *OK* button, the researcher also will want to select a location for the placement of output in the “Output Range” field.

The output differs from that of other  $t$  tests by including a measure of correlation between the two measures. In this case, the correlation is quite high, so any differences between the two sets of scores should be modest (exactly the situation found in this case).

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3			
4	Mean	17.4545455	18.1212121
5	Variance	44.5902098	19.4927739
6	Observations	66	66
7	Pearson Correlation	0.84346902	
8	Hypothesized Mean Difference	0	
9	df	65	
10	t Stat	-1.4297963	
11	P(T<=t) one-tail	0.07878232	
12	t Critical one-tail	1.66863629	
13	P(T<=t) two-tail	0.15756465	
14	t Critical two-tail	1.99713668	

## Confidence Intervals for Paired Differences

A confidence interval may be constructed around mean paired differences using the formula

$$95\% \text{ C.I.} = \bar{X}_{diff} \pm (t_{(\alpha/2, d.f.)} * \frac{S_D}{\sqrt{n}}),$$

where  $\bar{X}_{diff}$  is the mean paired differences,

$t_{(\alpha/2, d.f.)}$  is the critical value of  $t$  (with  $\alpha$  corresponding to the announced confidence interval; e.g.,  $1 - \alpha$  of .05 corresponds to a 95% degree of confidence),

$\alpha/2$  means that the critical value should be two-tailed, and

$\frac{S_D}{\sqrt{n}}$  is the standard error of the mean paired differences and includes everything in the denominator of the paired  $t$ -test formula.

For the example of the paired  $t$  test, a 95% confidence interval would be computed as follows:

$$95\% \text{ C.I.} = -1.73 \pm (-1.734 * .37)$$

$$95\% \text{ C.I.} = -1.73 \pm (-.64)$$

Thus, the researcher would claim 95% confidence that the difference between the pretest and posttest is somewhere between  $-1.09$  and  $-2.37$ .

## ASSESSING POWER

**Statistical power** is “the probability of rejecting the null hypothesis when it is false—and therefore should be rejected” (Vogt, 2005, p. 242). Some writers recommend that researchers routinely compute power before they complete studies. This step may help researchers select appropriate sample sizes. Furthermore, when researchers propose a new study, it makes sense to compute power to decide if there are enough possible data to make a study feasible.

There are many ways to increase the power of a statistical significance test. First, the researcher may decide to examine only large differences. Large differences are more easily detected with statistical significance tests than are small differences. A popular rule of thumb has been Cohen’s (J. Cohen, 1988) effect size guidelines. He suggests that an effect size difference of below 0.2 standard deviations is small, up to a 0.5 standard deviation difference is medium, and 0.8 standard deviations or greater is large. For correlations, he suggests that associations below  $r = .1$  are small, in the range of  $r = .3$  are medium, and above  $r = .5$  are large. Yet, these guidelines and the use of “after the fact” power analyses have been questioned in recent years, especially when used as a basis for trying to balance power and sample size issues (see Lenth, 2001). Second, researchers may exercise control to minimize the size of the population standard deviation. Third, the researcher could raise the alpha risk. If alpha risk were raised from .05 to .10, for instance, more null hypotheses are likely to be rejected than when the decision rule is kept at .05. Finally, of course, researchers may increase sample size.

This last option is the one researchers have given their greatest attention (probably because the first three options are difficult to apply). They often wonder how large their sample sizes would have to be for statistical significance to be claimed. They identify the level of power desired, typically .80 or .90; the alpha risk to be used in the statistical significance test; and the smallest size for an effect they would be interested in reporting. Once done, the researchers may use formulae and occasionally tables to determine the power of a test. The power of a test is computed as  $1 - \beta$  where

$$\beta = P\left(Z < \frac{C - \mu_1}{\sigma/\sqrt{n}}\right)$$

and  $C$  = upper confidence limit, such as

$$95\% UL = \mu_0 + \left(1.645 * \frac{\sigma}{\sqrt{n}}\right).$$

Suppose you were doing a study on the trustworthiness of television news anchors. On scales ranging from 5 to 35 points, the traditional mean has been 22 with a standard deviation of 16. You have a sample size of 100 and wish to identify differences in trustworthiness ratings of 25 in comparison with the traditional mean (22) in trustworthiness ratings. What would be the power of a one-tailed test?



### Special Discussion 7.2

#### The Treachery of After-the-Fact Power Analyses

Sometimes researchers compute power *after* they have completed their studies and analyzed most of their data, but such an approach is generally not advised. In the first place, computing power after the fact does not reveal anything that the probability level of the hypothesis test does not (Hoenig & Heisey, 2001). In fact, after-the-fact power statistics are simply proportional to alpha risk: The smaller the alpha risk, the greater the power. In the second place, the effort to show high power in the absence of statistical significance probably is misleading because it often is a veiled effort to “prove” a null hypothesis. When advertisers state that “no product has been shown superior to the ingredients contained in [their product],” they are asserting the truth of a null hypothesis, generally in the absence of evidence. The use of power analyses, however, cannot prove that a null hypothesis is true.

As a superior alternative, some writers have suggested reporting confidence intervals (S. N. Goodman & Berlin, 1994; M. Levine & Ensome, 2001). If one computes a confidence interval around a mean difference (or around a correlation) and finds that the confidence interval is very small, then the failure to find support for a research hypothesis may have some practical value (because it would indicate that only trivial effects remained undetected). A broad confidence interval might indicate a need to increase sample sizes and reduce background variation in future research. Indeed, the National Center for Educational Statistics of the U.S. Department of Education developed a program of statistical standards for researchers in which it stated that one of the preferred options when a null hypothesis is not rejected is to “use a 95% confidence interval to describe the magnitude of the possible difference or effect” (National Center for Educational Statistics, 2002, Standard 5-1-5: 7).

Using the formula on page 176, one would find the following:

$$\begin{aligned}
 C &= \mu_0 + \left( 1.645 * \frac{\sigma}{\sqrt{n}} \right) \\
 &= 22 + \left( 1.645 * \frac{16}{\sqrt{100}} \right) = 22 + (1.645 * 1.6) = 22 + 2.63 = 24.63 \\
 \text{Power} &= 1 - \beta \\
 \text{Power} &= 1 - P \left( Z < \frac{C - \mu_1}{\sigma/\sqrt{n}} \right) \\
 &= 1 - P \left( Z < \frac{24.632 - 25}{16/\sqrt{100}} \right) = 1 - P \left( Z < \frac{-.368}{1.6} \right) = 1 - P(Z < -.23)
 \end{aligned}$$

To compute this beta, we look at the  $z$  table and ask how much area exists below the point identified. On the  $z$  table, the area from 0 to  $-0.23$  standard deviations includes .091 of the total area. The area *below* that point includes .50 (50% of the distribution) minus .091, which comes out to .409. Thus, power is computed as follows:

$$\text{Power} = 1 - .409 = .591.$$

This estimate means that a new mean as large as that identified with such a sample and such population characteristics will be detected as statistically significant 59.1% of the time.

