CHAPTER 7

Comparison of Two Groups

The comparison of two groups is a very common type of analysis in the social and behavioral sciences. A study might compare mean income for men and women having similar jobs and experience. Another study might compare the proportions of Americans and Canadians who favor certain gun control laws. Means are compared for quantitative variables and proportions are compared for categorical variables.

Section 7.1 introduces some basic concepts for comparing groups. Section 7.2 illustrates these for comparing proportions and Section 7.3 for comparing means. The rest of the chapter shows some alternative methods useful for special cases.

7.1 PRELIMINARIES FOR COMPARING GROUPS

Do women tend to spend more time on housework than men? If so, how much more? In Great Britain in 2005, the Time Use Survey[1] studied how a random sample of Brits spend their time on a typical day. For those who reported working full time, Table 7.1 reports the mean and standard deviation of the reported average number of minutes per day spent on cooking and washing up. We use Table 7.1 to present some basic concepts for comparing groups.

TABLE 7.1: Cooking and Washing Up Minutes, per Day, for a National Survey of Men and Women Working Full Time in Great Britain

		Cooking and Washing Up Minutes	
Sex	Sample Size	Mean	Standard Deviation
Men	1219	23	32
Women	733	37	16

Bivariate Analyses with Response and Explanatory Variables

Two groups being compared constitute a ***binary*** variable —a variable having only two categories, sometimes also called ***dichotomous***. In a comparison of mean housework

[1] www.statistics.gov.uk

time for men and women, men and women are the two categories of the binary variable, sex. Methods for comparing two groups are special cases of ***bivariate*** statistical methods —an outcome variable of some type is analyzed for each category of a second variable.

From Section 3.5 (page 55), recall that an outcome variable about which comparisons are made is called a ***response variable***. The variable that defines the groups is called the ***explanatory variable***. In Table 7.1, time spent cooking and washing up is the response variable. The sex of the respondent is the explanatory variable.

Dependent and Independent Samples

Some studies compare means or proportions at two or more points in time. For example, a ***longitudinal study*** observes subjects at several times. An example is the Framingham Heart Study, which every two years since 1948 has observed many health characteristics of more than 5000 adults from Framingham, Massachusetts. Samples that have the same subjects in each sample are called ***dependent samples***.

More generally, two samples are *dependent* when a natural matching occurs between each subject in one sample and a subject in the other sample. Usually this happens when each sample has the same subjects. But matching can also occur when the two samples have different subjects. An example is a comparison of housework time of husbands and wives, the husbands forming one sample and their wives the other.

More commonly, comparisons use ***independent samples***. This means that the observations in one sample are *independent* of those in the other sample. The subjects in the two samples are different, with no matching between one sample with the other sample. An example is Table 7.1. Subjects were randomly selected and then classified on their sex and measured on how much time they spend in various activities. The samples of men and women were independent.

Suppose you plan to analyze whether a tutoring program improves mathematical understanding. One study design administers a math achievement test to a sample of students both before and after they go though the program. The sample of test scores before the program and the sample of test scores after the program are then *dependent*, because each sample has the same subjects.

Another study design randomly splits a class of students into two groups, one of which takes the tutoring program (the *experimental* group) and one of which does not (the *control* group). After the course, both groups take the math achievement test, and mean scores are compared. The two samples are then *independent*, because they contain different subjects without a matching between samples.

These two studies are *experimental*. As mentioned at the end of Section 2.2, many social science studies are instead *observational*. For example, many comparisons of groups result from dividing a sample into subsamples according to classification on a variable such as sex or race or political party. Table 7.1 is an example of this. Such cases are examples of ***cross-sectional*** studies, which use a single survey to compare groups. If the overall sample was randomly selected, then the subsamples are independent random samples from the corresponding subpopulations.

Why do we distinguish between *independent* and *dependent* samples? Because the standard error formulas for statistics that compare means or compare proportions are different for the two types of sample. With dependent samples, matched responses are likely to be associated. In the study about a tutoring program, the students who perform relatively well on one exam probably tend to perform well on the second exam also. This affects the standard error of statistics comparing the groups.

Difference of Estimates and Their Standard Error

To compare two populations, we can estimate the difference between their parameters. To compare population means μ_1 and μ_2, we treat $\mu_2 - \mu_1$ as a parameter and estimate it by the difference of sample means, $\bar{y}_2 - \bar{y}_1$. For Table 7.1, the estimated difference between the population mean daily cooking and washing up time for women and for men equals $\bar{y}_2 - \bar{y}_1 = 37 - 23 = 14$ minutes.

The sampling distribution of the estimator $\bar{y}_2 - \bar{y}_1$ has expected value $\mu_2 - \mu_1$. For large random samples, or for small random samples from normal population distributions, this sampling distribution has a normal shape, as Figure 7.1 portrays.

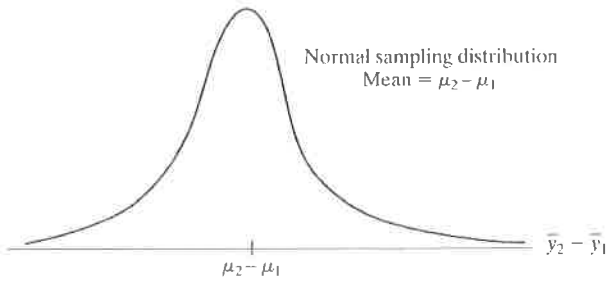

FIGURE 7.1: For Random Samples, the Sampling Distribution of the Difference between the Sample Means $\bar{y}_2 - \bar{y}_1$ Is Approximately Normal about $\mu_2 - \mu_1$

An estimate has a standard error that describes how precisely it estimates a parameter. Likewise, so does the difference between estimates from two samples have a standard error. For Table 7.1, the standard error of the sampling distribution of $\bar{y}_2 - \bar{y}_1$ describes how precisely $\bar{y}_2 - \bar{y}_1 = 14$ estimates $\mu_2 - \mu_1$. If many studies had been conducted in Britain comparing daily cooking and washing up time for women and men, the estimate $\bar{y}_2 - \bar{y}_1$ would not have equaled 14 minutes for each of them. The estimate would vary from study to study. The standard error describes the variability of the estimates from different potential studies of the same size.

The following general rule enables us to find the standard error when we compare estimates from independent samples:

Standard Error of Difference Between Two Estimates

For two estimates from independent samples that have estimated standard errors se_1 and se_2, the sampling distribution of their difference has

$$\text{estimated standard error} = \sqrt{(se_1)^2 + (se_2)^2}.$$

Each estimate has sampling error, and the variabilities add together to determine the standard error of the difference of the estimates. The standard error formula for dependent samples differs from this formula, and Section 7.4 presents it.

Recall that the estimated standard error of a sample mean equals

$$se = \frac{s}{\sqrt{n}},$$

where s is the sample standard deviation. Let n_1 denote the sample size for the first sample and n_2 the sample size for the second sample. Let s_1 and s_2 denote the standard

deviations. The difference $\bar{y}_2 - \bar{y}_1$ between two sample means with independent samples has estimated standard error

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

For example, from Table 7.1, the estimated standard error of the difference of 14 minutes between the sample mean cooking and washing up time for women and men equals

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(32)^2}{1219} + \frac{(16)^2}{733}} = 1.1.$$

For such large sample sizes, the estimate $\bar{y}_2 - \bar{y}_1$ would not vary much from study to study.

From the formula, the standard error of the difference is larger than the standard error for either sample estimate alone. Why is this? In practical terms, $(\bar{y}_2 - \bar{y}_1)$ is often farther from $(\mu_2 - \mu_1)$ than $\bar{y}_1$ is from μ_1 or $\bar{y}_2$ is from μ_2. For instance, suppose $\mu_1 = \mu_2 = 30$ (unknown to us), but the sample means are $\bar{y}_1 = 23$ and $\bar{y}_2 = 37$. Then the errors of estimation were

$$\bar{y}_1 - \mu_1 = 23 - 30 = -7 \quad \text{and} \quad \bar{y}_2 - \mu_2 = 37 - 30 = 7,$$

each estimate being off by a distance of 7. But the estimate $(\bar{y}_2 - \bar{y}_1) = 37 - 23 = 14$ falls 14 from $(\mu_2 - \mu_1) = 0$. The error of size 14 for the difference is larger than the error of size 7 for either mean individually. Suppose a sample mean that falls 7 away from a population mean is well out in the tail of a sampling distribution for a single sample mean. Then a difference between sample means that falls 14 away from the difference between population means is well out in the tail of the sampling distribution for $\bar{y}_2 - \bar{y}_1$.

The Ratio of Parameters

Another way to compare two proportions or two means uses their *ratio*. The ratio equals 1.0 when the parameters are equal. Ratios farther from 1.0 represent larger effects.

In Table 7.1, the ratio of sample mean cooking and washing up time for women and for mean is $37/23 = 1.61$. The sample mean for women was 1.61 times the sample mean for men. This can also be expressed by saying that the mean for women was 61% higher than the mean for women.

The ratio of two proportions is often called the ***relative risk***, because it is often used in public health applications to compare rates for an undesirable outcome for two groups. The ratio is often more informative than the difference when both proportions are close to zero.

For example, according to recent data from the United Nations, the annual gun homicide rate is 62.4 per one million residents in the U.S. and 1.3 per one million residents in Britain. In proportion form, the results are 0.0000624 in the U.S. and 0.0000013 in Britain. The difference between the proportions is $0.0000624 - 0.0000013 = 0.0000611$, extremely small. By contrast, the ratio is $0.000624/0.0000013 = 624/13 = 48$. The proportion of people killed by guns in the U.S. was 48 times the proportion in Britain. In this sense, the effect is large.

Software can form a confidence interval for a population ratio of means or proportions. The formulas are complex, and we will not cover them in this text.

7.2 CATEGORICAL DATA: COMPARING TWO PROPORTIONS

Let's now learn how to compare proportions inferentially. Let π_1 denote the proportion for the first population and π_2 the proportion for the second population. Let $\hat{\pi}_1$ and $\hat{\pi}_2$ denote the sample proportions. You may wish to review Sections 5.2 and 6.3 on inferences for proportions in the one-sample case.

EXAMPLE 7.1 Does Prayer Help Coronary Surgery Patients?

A study used patients at six U.S. hospitals who were to receive coronary artery bypass graft surgery.[2] The patients were randomly assigned to two groups. For one group, Christian volunteers were instructed to pray for a successful surgery with a quick, healthy recovery and no complications. The praying started the night before surgery and continued for two weeks. The response was whether medical complications occurred within 30 days of the surgery. Table 7.2 summarizes results.

TABLE 7.2: Whether Complications Occurred for Heart Surgery Patients Who Did or Did Not Have Group Prayer

Prayer	Complications	No Complications	Total
Yes	315	289	604
No	304	293	597

Is there a difference in complication rates for the two groups? Let π_1 denote the probability for those patients who had a prayer group. Let π_2 denote the probability for the subjects not having a prayer group. From Table 7.2, the sample proportions equal

$$\hat{\pi}_1 = \frac{315}{604} = 0.522, \quad \hat{\pi}_2 = \frac{304}{597} = 0.509. \qquad \blacksquare$$

We compare the probabilities using their difference, $\pi_2 - \pi_1$. The difference of sample proportions, $\hat{\pi}_2 - \hat{\pi}_1$, estimates $\pi_2 - \pi_1$. If n_1 and n_2 are relatively large, the estimator $\hat{\pi}_2 - \hat{\pi}_1$ has a sampling distribution that is approximately normal. See Figure 7.2. The mean of the sampling distribution is the parameter $\pi_2 - \pi_1$ to be estimated.

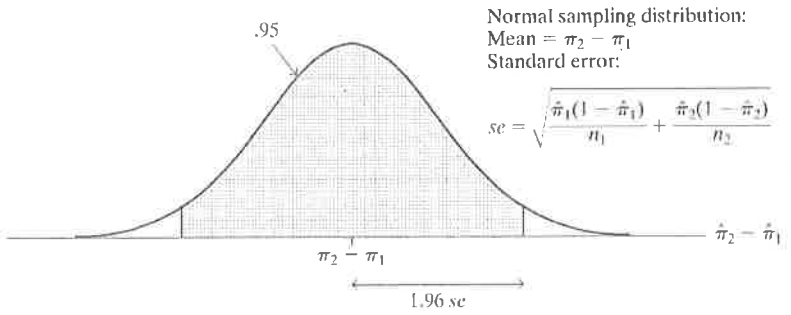

FIGURE 7.2: For Large Random Samples, the Sampling Distribution of the Estimator $\hat{\pi}_2 - \hat{\pi}_1$ of the Difference of Proportions Is Approximately Normal

From the rule in the box in Section 7.1 (page 185), the standard error of the difference of sample proportions equals the square root of the sum of squared

[2] H. Benson et al., *American Heart Journal*, vol. 151, 2006, pp. 934–952.

standard errors of the separate sample proportions. Recall that the estimated standard error of a single sample proportion is

$$se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

Therefore, the difference between two proportions has estimated standard error

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$

For Table 7.2, $\hat{\pi}_2 - \hat{\pi}_1$ has estimated standard error

$$se = \sqrt{\frac{(0.522)(0.478)}{604} + \frac{(0.509)(0.491)}{597}} = 0.0288.$$

For samples of these sizes, the difference in sample proportions would not vary much from study to study.

Confidence Interval for Difference of Proportions

As with a single proportion, the confidence interval takes the point estimate and adds and subtracts a margin of error that is a z-score times the estimated standard error, such as

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm 1.96(se)$$

for 95% confidence.

Confidence Interval for $\pi_2 - \pi_1$

For large, independent random samples, a confidence interval for the difference $\pi_2 - \pi_1$ between two population proportions is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se), \text{ where } se = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$

The z-score depends on the confidence level, such as 1.96 for 95% confidence.

The sample is large enough to use this formula if, for each sample, at least ten observations fall in the category for which the proportion is estimated, and at least ten observations do not fall in that category. Most studies easily satisfy this.

EXAMPLE 7.2 Prayer and Coronary Surgery Patients, Continued

For Table 7.2, we estimate the difference $\pi_2 - \pi_1$ between the probability of complications for the non-prayer and prayer patients. Since $\hat{\pi}_1 = 0.522$ and $\hat{\pi}_2 = 0.509$, the estimated difference equals $\hat{\pi}_2 - \hat{\pi}_1 = -0.013$. There was a drop of 0.013 in the proportion who had complications among those not receiving prayer.

To determine the precision of this estimate, we form a confidence interval. Previously we determined that $se = 0.0288$. A 95% confidence interval for $\pi_2 - \pi_1$ is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm 1.96(se), \text{ or } (0.509 - 0.522) \pm 1.96(0.0288)$$

$$= -0.013 \pm 0.057 \quad \text{or} \quad (-0.07, 0.04).$$

It seems that the difference is close to 0, so the probability of complications is similar for the two groups. ■

Interpreting a Confidence Interval Comparing Proportions

When the confidence interval for $\pi_2 - \pi_1$ contains 0, as in the previous example, it is plausible that $\pi_2 - \pi_1 = 0$. That is, it is believable that $\pi_1 = \pi_2$. Insufficient evidence exists to conclude which of π_1 or π_2 is larger. For the confidence interval for $\pi_2 - \pi_1$ of $(-0.07, 0.04)$, we infer that π_2 may be as much as 0.07 smaller or as much as 0.04 larger than π_1.

When a confidence interval for $\pi_2 - \pi_1$ contains only *negative* values, this suggests that $\pi_2 - \pi_1$ is negative. In other words, we infer that π_2 is *smaller* than π_1. When a confidence interval for $\pi_2 - \pi_1$ contains only *positive* values, we conclude that $\pi_2 - \pi_1$ is positive; that is, π_2 is *larger* than π_1.

Which group we call Group 1 and which we call Group 2 is arbitrary. If we let Group 1 be the nonprayer group rather than the prayer group, then the estimated difference would be $+0.013$ rather than -0.013. The confidence interval would have been $(-0.04, 0.07)$, the negatives of the endpoints we obtained. Similarly, it does not matter whether we form a confidence interval for $\pi_2 - \pi_1$ or for $\pi_1 - \pi_2$. If the confidence interval for $\pi_2 - \pi_1$ is $(-0.07, 0.04)$, then the confidence interval for $\pi_1 - \pi_2$ is $(-0.04, 0.07)$.

The magnitude of values in the confidence interval tells you how large any true difference is. If all values in the confidence interval are near 0, such as the interval $(-0.07, 0.04)$, we infer that $\pi_2 - \pi_1$ is small in practical terms even if not exactly equal to 0.

As in the one-sample case, larger sample sizes contribute to a smaller *se*, a smaller margin of error, and narrower confidence intervals. In addition, higher confidence levels yield wider confidence intervals. For the prayer study, a 99% confidence interval equals $(-0.09, 0.06)$. This is wider than the 95% confidence interval of $(-0.07, 0.04)$.

Significance Tests about $\pi_2 - \pi_1$

To compare population proportions π_1 and π_2, a significance test specifies H_0: $\pi_1 = \pi_2$. For the difference of proportions parameter, this hypothesis is H_0: $\pi_2 - \pi_1 = 0$, *no difference*, or *no effect*.

Under the presumption for H_0 that $\pi_1 = \pi_2$, we estimate the common value of π_1 and π_2 by the sample proportion for the entire sample. Denote this by $\hat{\pi}$. To illustrate, for the data in Table 7.2 from the prayer study, $\hat{\pi}_1 = 315/604 = 0.522$ and $\hat{\pi}_2 = 304/597 = 0.509$. For the entire sample,

$$\hat{\pi} = (315 + 304)/(604 + 597) = 619/1201 = 0.515.$$

The proportion $\hat{\pi}$ is called a *pooled estimate*, because it pools together observations from the two samples.

The test statistic measures the number of standard errors between the estimate and the H_0 value. Treating $\pi_2 - \pi_1$ as the parameter, we test that $\pi_2 - \pi_1 = 0$; that is, the null hypothesis value of the parameter $\pi_2 - \pi_1$ is 0. The estimated value of $\pi_2 - \pi_1$ is $\hat{\pi}_2 - \hat{\pi}_1$. The test statistic is

$$z = \frac{\text{Estimate} - \text{null hypothesis value}}{\text{Standard error}} = \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{se_0}.$$

Rather than use the standard error from the confidence interval, you should use an alternative formula based on the presumption stated in H_0 that $\pi_1 = \pi_2$. We use the

notation se_0, because it is a *se* that holds under H_0. This standard error is

$$se_0 = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

For Table 7.2, the standard error estimate for the test equals

$$se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.515(0.485)\left(\frac{1}{604} + \frac{1}{597}\right)}$$
$$= \sqrt{0.000832} = 0.0288.$$

The test statistic for H_0: $\pi_1 = \pi_2$ equals

$$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0} = \frac{0.509 - 0.522}{0.0288} = -0.43.$$

The P-value depends in the usual way on whether the test is two sided, H_a: $\pi_1 \neq \pi_2$ (i.e., $\pi_2 - \pi_1 \neq 0$), or one sided, H_a: $\pi_1 > \pi_2$ (i.e., $\pi_2 - \pi_1 < 0$) or H_a: $\pi_1 < \pi_2$ ($\pi_2 - \pi_1 > 0$). Most common is the two-sided alternative. Its P-value is the two-tail probability from the standard normal distribution that falls beyond the observed test statistic value. A z-score of -0.43 has two-sided P-value equal to 0.67. There is not much evidence against H_0.

In summary, it is plausible that the probability of complications is the same for the prayer and nonprayer conditions. However, this study does not disprove the power of prayer. Apart from the fact that we cannot accept a null hypothesis, the experiment could not control many factors, such as whether friends and family were also praying for the patients.

The z test for comparing proportions works quite well even for relatively small sample sizes. We'll give detailed guidelines in Section 8.2 when we study a more general test for comparing several groups. For simplicity, you can use the guideline for confidence intervals comparing proportions, namely that each sample should have at least 10 outcomes of each type. In practice, *two-sided* tests are robust and work well if each sample has at least five outcomes of each type.

Contingency Tables and Conditional Probabilities

Table 7.2 is an example of a ***contingency table***. Each row is a category of the explanatory variable (whether prayed for) which defines the two groups compared. Each column is a category of the response variable (whether complications occurred). The ***cells*** of the table contain frequencies for the four possible combinations of outcomes.

The parameters π_1 and π_2 estimated using the contingency table are called ***conditional probabilities***. This term refers to probabilities for a response variable evaluated under two conditions, namely the two levels of the explanatory variable. For instance, under the condition that the subject is being prayed for, the conditional probability of developing complications is estimated to be $315/604 = 0.52$.

This section has considered binary response variables. Instead, the response could have several categories. For example, the response categories might be (No complications, Slight complications, Severe complications). Then we could compare the two groups in terms of the conditional probabilities of observations in each of the three categories. Likewise, the number of groups compared could exceed two.

Chapter 8 shows how to analyze contingency tables having more than two rows or columns.

7.3 QUANTITATIVE DATA: COMPARING TWO MEANS

To compare two population means μ_1 and μ_2, we can make inferences about their difference. You may wish to review Sections 5.3 and 6.2 on inferences for means in the one-sample case.

Confidence Interval for $\mu_2 - \mu_1$

For large random samples, or for small random samples from normal population distributions, the sampling distribution of $(\bar{y}_2 - \bar{y}_1)$ has a normal shape. As usual, inference for means with *estimated* standard errors uses the t distribution for test statistics and for the margin of error in confidence intervals. A confidence interval takes the point estimate and adds and subtracts a margin of error that is a t-score times the standard error.

Confidence Interval for $\mu_2 - \mu_1$

For independent random samples from two groups that have normal population distributions, a confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{y}_2 - \bar{y}_1) \pm t(se), \text{ where } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The t-score is chosen to provide the desired confidence level.

The formula for the degrees of freedom for the t-score, called the *Welch-Satterthwaite approximation*, is complex. The df depends on the sample standard deviations s_1 and s_2 as well as the sample sizes n_1 and n_2. If $s_1 = s_2$ and $n_1 = n_2$, it simplifies to $df = (n_1 + n_2 - 2)$. This is the sum of the df values for separate inference about each group; that is, $df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$. Generally, df falls somewhere between $n_1 + n_2 - 2$ and the minimum of $(n_1 - 1)$ and $(n_2 - 1)$. Software can easily find this df value, the t-score, and the confidence interval.

In practice, the method is robust to violations of the normal population assumption. This is especially true when both n_1 and n_2 are at least about 30, by the Central Limit Theorem. As usual, you should be wary of extreme outliers or of extreme skew that may make the mean unsuitable as a summary measure.

EXAMPLE 7.3 Comparing Housework Time of Men and Women

For Table 7.1 (page 183), on the daily time full-time workers spend cooking and cleaning up, denote the population mean in Britain by μ_1 for men and μ_2 for women. That table reported sample means of 23 minutes for 1219 men and 37 minutes for 733 women, with sample standard deviations of 32 and 16. The point estimate of $\mu_2 - \mu_1$ equals $\bar{y}_2 - \bar{y}_1 = 37 - 23 = 14$. Section 7.1 found that the estimated standard error of this difference equals

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(32)^2}{1219} + \frac{(16)^2}{733}} = 1.09.$$

The sample sizes are very large, so the t-score for the margin of error is essentially the z-score. So the 95% confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{y}_2 - \bar{y}_1) \pm 1.96(se) = 14 \pm 1.96(1.09), \text{ or } 14 \pm 2, \text{ which is } (12, 16).$$

We can be 95% confident that the population mean amount of daily time spent on cooking and washing up is between 12 and 16 minutes higher for women than men. ■

Interpreting a Confidence Interval Comparing Means

The confidence interval (12, 16) contains only positive values. Since we took the difference between the mean for women and the mean for men, we can conclude that the population mean is higher for women. A confidence interval for $\mu_2 - \mu_1$ that contains only positive values suggests that $\mu_2 - \mu_1$ is positive, meaning that μ_2 is larger than μ_1. A confidence interval for $\mu_2 - \mu_1$ that contains only negative values suggests that μ_2 is smaller than μ_1. When the confidence interval contains 0, insufficient evidence exists to conclude which of μ_1 or μ_2 is larger. It is then plausible that $\mu_1 = \mu_2$.

The identification of which is group 1 and which is group 2 is arbitrary, as is whether we estimate $\mu_2 - \mu_1$ or $\mu_1 - \mu_2$. For instance, a confidence interval of (12, 16) for $\mu_2 - \mu_1$ is equivalent to one of $(-16, -12)$ for $\mu_1 - \mu_2$.

Significance Tests about $\mu_2 - \mu_1$

To compare population means μ_1 and μ_2, we can also conduct a significance test of H_0: $\mu_1 = \mu_2$. For the difference of means parameter, this hypothesis is H_0: $\mu_2 - \mu_1 = 0$ (no effect).

As usual, the test statistic measures the number of standard errors between the estimate and the H_0 value,

$$t = \frac{\text{Estimate of parameter} - \text{null hypothesis value of parameter}}{\text{Standard error of estimate}}.$$

Treating $\mu_2 - \mu_1$ as the parameter, we test that $\mu_2 - \mu_1 = 0$. Its estimate is $\bar{y}_2 - \bar{y}_1$. The standard error is the same as in a confidence interval. The t test statistic is

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}, \quad \text{where} \quad se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

EXAMPLE 7.4 Test Comparing Mean Housework for Men and Women

Using the data from Table 7.1 (page 183), we now test for a difference between the population mean cooking and washing up time, μ_1 for men and μ_2 for women. We test H_0: $\mu_1 = \mu_2$ against H_a: $\mu_1 \neq \mu_2$. We've seen that the estimate $\bar{y}_2 - \bar{y}_1 = 37 - 23 = 14$ has $se = 1.09$.

The test statistic equals

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se} = \frac{(37 - 23)}{1.09} = 12.8.$$

With large samples, since the t distribution is essentially the same as the standard normal, $t = 12.8$ is enormous. It gives a P-value that is 0 to many decimal places. We conclude that the population means differ. The sample means show that the difference takes the direction of a higher mean for women. ■

In practice, significance tests are much more common for two-sample comparisons than for one-sample analyses. It is usually artificial to test whether the population mean equals one particular value, such as in testing H_0: $\mu = \mu_0$. However, it is often relevant to test whether a *difference* exists between two population means, such as in testing H_0: $\mu_1 = \mu_2$. For instance, we may have no idea what to hypothesize for the mean amount of housework time for men, but we may want to know whether that mean (whatever its value) is the same as, larger than, or smaller than the mean for women.

Correspondence between Confidence Intervals and Tests

For means, the equivalence between two-sided tests and confidence intervals mentioned in Sections 6.2 and 6.4 also applies in the two-sample case. For example, since the two-sided P-value in Example 7.4 is less than 0.05, we reject H_0: $\mu_2 - \mu_1 = 0$ at the $\alpha = 0.05$ level. Similarly, a 95% confidence interval for $\mu_2 - \mu_1$ does not contain 0, the H_0 value. That interval equals (12, 16).

As in one-sample inference, confidence intervals are more informative than tests. The confidence interval tells us not only that the population mean differs for men and women, but it shows us just how large that difference is likely to be, and in which direction.

7.4 COMPARING MEANS WITH DEPENDENT SAMPLES

Dependent samples occur when each observation in sample 1 matches with an observation in sample 2. The data are often called ***matched pairs*** data because of this matching.

Paired Difference Scores for Matched Samples

Dependent samples commonly occur when each sample has the same subjects. Examples are *longitudinal* observational studies that observe a person's response at several points in time and experimental studies that take *repeated measures* on subjects. An example of the latter is a *cross-over* study, in which a subject receives one treatment for a period and then the other treatment. The next example illustrates.

EXAMPLE 7.5 Cell Phone Use and Driver Reaction Time

A recent experiment[3] used a sample of college students to investigate whether cell phone use impairs drivers' reaction times. On a machine that simulated driving situations, at irregular periods a target flashed red or green. Participants were instructed to press a brake button as soon as possible when they detected a red light. Under the cell phone condition, the student carried out a conversation about a political issue on the cell phone with someone in a separate room. In the control condition, they listened to a radio broadcast or to books-on-tape while performing the simulated driving.

For each student, for a particular condition the outcome recorded in Table 7.3 is their mean response time (in milliseconds) over several trials. Figure 7.3 shows box plots of the data for the two conditions. Student 28 is an outlier under each condition. ■

[3]Data courtesy of David Strayer, University of Utah. See D. Strayer and W. Johnston, *Psych. Science*, vol. 21, 2001, pp. 462–466.

TABLE 7.3: Reaction Times (in Milliseconds) on Driving Skills Task and Cell Phone Use (Yes or No). The difference score is the reaction time using the cell phone minus the reaction time not using it, such as 636 − 604 = 32 milliseconds.

	Cell Phone?				Cell Phone?		
Student	No	Yes	Difference	Student	No	Yes	Difference
1	604	636	32	17	525	626	101
2	556	623	67	18	508	501	−7
3	540	615	75	19	529	574	45
4	522	672	150	20	470	468	−2
5	459	601	142	21	512	578	66
6	544	600	56	22	487	560	73
7	513	542	29	23	515	525	10
8	470	554	84	24	499	647	148
9	556	543	−13	25	448	456	8
10	531	520	−11	26	558	688	130
11	599	609	10	27	589	679	90
12	537	559	22	28	814	960	146
13	619	595	−24	29	519	558	39
14	536	565	29	30	462	482	20
15	554	573	19	31	521	527	6
16	467	554	87	32	543	536	−7

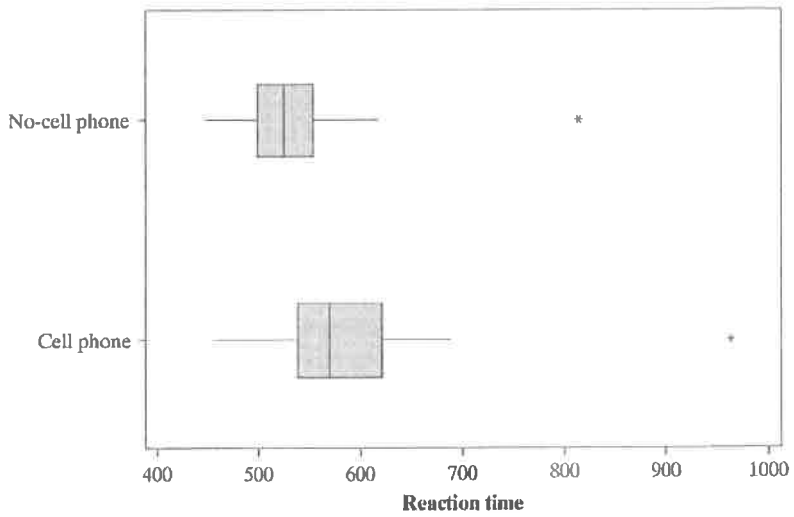

FIGURE 7.3: Box Plots of Observations for the Experiment on the Effects of Cell Phone Use on Reaction Times

For matched-pairs data, each observation in one sample pairs with an observation in the other sample. For each pair, we form

$$\text{Difference} = \text{Observation in sample 2} - \text{Observation in sample 1.}$$

Table 7.3 shows the difference scores for the cell phone experiment.

Let $\bar{y}_d$ denote the sample mean of the difference scores. This estimates μ_d, the population mean difference. In fact, the parameter μ_d is identical to $\mu_2 - \mu_1$, the difference between the population means for the two groups. The mean of the differences equals the difference between the means.

For matched-pairs data, the difference between the means of the two groups equals the mean of the difference scores.

Inferences Comparing Means Using Paired Differences

We can base analyses about $\mu_2 - \mu_1$ on inferences about μ_d, using the single sample of difference scores. This simplifies the analysis, because it reduces a two-sample problem to a one-sample problem.

Let n denote the number of observations in each sample. This equals the number of difference scores. The confidence interval for μ_d is

$$\bar{y}_d \pm t\left(\frac{s_d}{\sqrt{n}}\right).$$

Here, $\bar{y}_d$ and s_d are the sample mean and standard deviation of the difference scores, and t is the t-score for the chosen confidence level, having $df = n - 1$. This confidence interval has the same form as the one Section 6.3 presented for a single mean. We apply the formula to the single sample of n differences rather than the original two sets of observations.

For testing H_0: $\mu_1 = \mu_2$, we express the hypothesis in terms of the difference scores as H_0: $\mu_d = 0$. The test statistic is

$$t = \frac{\bar{y}_d - 0}{se}, \quad \text{where} \quad se = s_d/\sqrt{n}.$$

This compares the sample mean of the differences to the null hypothesis value of 0, in terms of the number of standard errors between them. The standard error is the same one used for a confidence interval. Since this test uses the difference scores for the pairs of observations, it is called a ***paired-difference t test***.

EXAMPLE 7.6 Cell Phones and Driver Reaction Time, Continued

We now analyze the matched-pairs data in Table 7.3 for the driving and cell phone experiment. The mean reaction times were 534.6 milliseconds without the cell phone and 585.2 milliseconds while using it. The 32 difference scores (32, 67, 75, ...) from Table 7.3 have a sample mean of

$$\bar{y}_d = (32 + 67 + 75 + \cdots + (-7))/32 = 50.6.$$

This equals the difference between the sample means of 585.2 and 534.6 for the two conditions. The sample standard deviation of the 32 difference scores is

$$s_d = \sqrt{\frac{(32 - 50.6)^2 + (67 - 50.6)^2 + \cdots}{32 - 1}} = 52.5.$$

The standard error of $\bar{y}_d$ is $se = s_d/\sqrt{n} = 52.5/\sqrt{32} = 9.28$.

For a 95% confidence interval for $\mu_d = \mu_2 - \mu_1$ with $df = n - 1 = 31$, we use $t_{.025} = 2.04$. The confidence interval equals

$$\bar{y}_d \pm 2.04(se) = 50.6 \pm 2.04(9.28), \quad \text{which is} \quad (31.7, 69.5).$$

We infer that the population mean reaction time when using cell phones is between about 32 and 70 milliseconds higher than when not using cell phones. The confidence interval does not contain 0. We conclude that the population mean reaction time is greater when using a cell phone.

Next consider the significance test of H_0: $\mu_d = 0$ (and hence equal population means for the two conditions) against H_a: $\mu_d \neq 0$. The test statistic is

$$t = \frac{(\bar{y}_d - 0)}{se} = \frac{50.6}{9.28} = 5.5,$$

with $df = 31$. The two-tail P-value equals 0.000005. There is extremely strong evidence that mean reaction time is greater when using a cell phone. Table 7.4 shows how SPSS software reports these results for its paired-samples t test option. ■

TABLE 7.4: SPSS Printout for Matched-Pairs Analysis Comparing Driver Reaction Times with and without Cell Phone Use

```
t-tests for Paired Samples

                Number of
Variable         pairs        Mean       SD      SE of Mean
NO-CELL PHONE                534.56    66.45       11.75
                    32
CELL PHONE                   585.19    89.65       15.85

Paired Differences
  Mean        SD     SE of Mean     t-value    df    2-tail Sig
  50.63     52.49      9.28           5.46     31       0.000
    95% CI (31.70, 69.55)
```

Paired-difference inferences make the usual assumptions for t procedures: The observations (the difference scores) are randomly obtained from a population distribution that is normal. Confidence intervals and two-sided tests work well even if the normality assumption is violated (their *robustness* property), unless the sample size is small and the distribution is highly skewed or has severe outliers. For the study about driving and cell phones, one subject was an outlier on both reaction times. However, the difference score for that subject, which is the observation used in the analysis, is not an outlier. The article about the study did not indicate whether the subjects were randomly selected. The subjects in the experiment were probably a volunteer sample, so inferential conclusions are tentative.

Independent versus Dependent Samples

Using dependent samples can have certain benefits. First, known sources of potential bias are controlled. Using the same subjects in each sample, for instance, keeps many other factors fixed that could affect the analysis. Suppose younger subjects tend to have faster reaction times. If group 1 has a lower sample mean than group 2, it is not because subjects in group 1 are younger, because both groups have the same subjects.

Second, the standard error of $\bar{y}_2 - \bar{y}_1$ may be smaller with dependent samples. In the cell phone study, the standard error was 9.3. If we had observed *independent* samples with the same scores as in Table 7.3, the standard error of $\bar{y}_2 - \bar{y}_1$ would have been 19.7. This is because the variability in the difference scores tends to be less than the variability in the original scores when the scores in the two samples are strongly correlated. In fact, for the data in Table 7.3, the correlation (recall Section 3.5) between the no–cell phone reaction times and the cell phone reaction times is 0.81, showing a strong positive association.

7.5 OTHER METHODS FOR COMPARING MEANS*

Section 7.3 presented inference comparing two means with independent samples. A slightly different inference method can be used when we expect similar variability for the two groups. For example, under a null hypothesis of "no effect," we often expect the entire distributions of the response variable to be identical for the two groups. So we expect standard deviations as well as means to be identical.

Comparing Means while Assuming Equal Standard Deviation

In comparing the population means, this method makes the additional assumption that the population standard deviations are equal, that is, $\sigma_1 = \sigma_2$. For it, a simpler *df* expression holds for an *exact* t distribution for the test statistic. Although it seems disagreeable to make an additional assumption, confidence intervals and two-sided tests are fairly robust against violations of this and the normality assumption, particularly when the sample sizes are similar and not extremely small.

The common value σ of σ_1 and σ_2 is estimated by

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{\Sigma(y_1 - \bar{y}_1)^2 + \Sigma(y_2 - \bar{y}_2)^2}{n_1 + n_2 - 2}}.$$

Here, $\Sigma(y_1 - \bar{y}_1)^2$ denotes the sum of squares about the mean for the observations in the first sample, and $\Sigma(y_2 - \bar{y}_2)^2$ denotes the sum of squares about the mean for the observations in the second sample. The estimate s pools information from the two samples to provide a single estimate of variability. It is called the *pooled estimate*. The term inside the square root is a weighted average of the two sample variances. When $n_1 = n_2$, it's the ordinary average. The estimate s falls between s_1 and s_2. With s as the estimate of σ_1 and σ_2, the estimated standard error of $\bar{y}_2 - \bar{y}_1$ simplifies to

$$se = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The confidence interval for $\mu_2 - \mu_1$ has the usual form

$$(\bar{y}_2 - \bar{y}_1) \pm t(se).$$

The t-score comes from the t table for the desired confidence level, with $df = n_1 + n_2 - 2$. The df equals the total number of observations $(n_1 + n_2)$ minus the number of parameters estimated in order to calculate s (namely, the two means, μ_1 and μ_2, estimated by $\bar{y}_1$ and $\bar{y}_2$).

To test H_0: $\mu_1 = \mu_2$, the test statistic has the usual form,

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{se}.$$

Now, se uses the pooled formula, as in the confidence interval. The test statistic has the t distribution with $df = n_1 + n_2 - 2$.

EXAMPLE 7.7 Comparing a Therapy to a Control Group

Examples 5.5 (page 120) and 6.4 (page 151) described a study that used a cognitive behavioral therapy to treat a sample of teenage girls who suffered from anorexia. The study observed the mean weight change after a period of treatment. Studies of that type also usually have a control group that receives no treatment or a standard treatment. Then researchers can analyze how the change in weight compares for the treatment group to the control group.

In fact, the anorexia study had a control group. Teenage girls in the study were randomly assigned to the cognitive behavioral treatment (Group 1) or to the control group (Group 2). Table 7.5 summarizes the results. (The data for both groups are shown in Table 12.21 on page 396.)

TABLE 7.5: Summary of Results Comparing Treatment Group to Control Group for Anorexia Study

Group	Sample Size	Mean	Standard Deviation
Treatment	29	3.01	7.31
Control	26	−0.45	7.99

If H_0 is true that the treatment has the same effect as the control, then we would expect the groups to have equal means and equal standard deviations. For these data, the pooled estimate of the assumed common standard deviation equals

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{28(7.31)^2 + 25(7.99)^2}{29 + 26 - 2}}$$
$$= \sqrt{\frac{3092.2}{53}} = 7.64.$$

Now, $\bar{y}_1 - \bar{y}_2 = 3.01 - (-0.45) = 3.46$ has an estimated standard error of

$$se = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 7.64\sqrt{\frac{1}{29} + \frac{1}{26}} = 2.06.$$

Let μ_1 and μ_2 denote the mean weight gains for these therapies for the hypothetical populations that the samples represent. We test H_0: $\mu_1 = \mu_2$ against H_a: $\mu_1 \neq \mu_2$. The test statistic equals

$$t = \frac{\bar{y}_1 - \bar{y}_2}{se} = \frac{3.01 - (-0.45)}{2.06} = 1.68.$$

This statistic has $df = n_1 + n_2 - 2 = 29 + 26 - 2 = 53$. From the t-table (Table B), the two-sided P-value is $P = 0.10$. There is only weak evidence of better success using the cognitive behavioral therapy.

When $df = 53$, the t-score for a 95% confidence interval for $(\mu_1 - \mu_2)$ is $t_{.025} = 2.006$. The interval is

$(\bar{y}_1 - \bar{y}_2) \pm t(se) = 3.46 \pm 2.006(2.06)$, which is 3.46 ± 4.14, or $(-0.7, 7.6)$.

We conclude that the mean weight change for the cognitive behavioral therapy could be as much as 0.7 pound lower or as much as 7.6 pounds higher than the mean weight change for the control group. Since the interval contains 0, it is plausible that the population means are identical. This is consistent with the P-value exceeding 0.05 in the test. If the population mean weight change is less for the cognitive behavioral group than for the control group, it is just barely less (less than 1 pound), but if the population mean change is greater, it could be nearly 8 pounds greater. Since the sample sizes are not large, the confidence interval is relatively wide. ■

Completely Randomized versus Randomized Block Design

The anorexia study used a *completely randomized* experimental design: Subjects were randomly assigned to the two therapies. With this design, there's the chance that the subjects selected for one therapy might differ in an important way from subjects selected for the other therapy. For moderate to large samples, factors that could influence results (such as initial weight) tend to balance by virtue of the randomization. For small samples, an imbalance could occur.

An alternative experimental design *matches* subjects in the two samples, such as by taking two girls of the same weight and randomly deciding which girl receives which therapy. This matched-pairs plan is a simple example of a ***randomized block design***. Each pair of subjects forms a *block*, and within blocks subjects are randomly assigned to the treatments. With this design, we would use the methods of the previous section for dependent samples.

Inferences Reported by Software

Table 7.6 illustrates the way SPSS reports results of two-sample t tests. The table shows results of two tests for comparing means, differing in terms of whether they assume equal population standard deviations. The t test just presented assumes that $\sigma_1 = \sigma_2$. The t statistic that software reports for the "equal variances not assumed" case is the t statistic of Section 7.3,

$$t = (\bar{y}_2 - \bar{y}_1)/se, \text{ with } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

When $n_1 = n_2$, the "equal variances" and "unequal variances" test statistics are identical. They are usually similar if n_1 and n_2 are close or if s_1 and s_2 are close.

TABLE 7.6: SPSS Output for Performing Two-Sample t Tests

		t-test for Equality of Means				
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
WEIGHT CHANGE	Equal variances assumed	1.68	53	0.099	3.46	2.06
	Equal variances not assumed	1.67	50	0.102	3.46	2.07

If the data show evidence of a potentially large difference in standard deviations (with, say, one sample standard deviation being at least double the other), it is better to use the approximate t test (Section 7.3) that does not make the $\sigma_1 = \sigma_2$ assumption. It can yield a t statistic value much different from the method that assumes $\sigma_1 = \sigma_2$ if s_1 and s_2 are quite different and the sample sizes are unequal.

Many texts and most software present a statistic denoted by F for testing that the population standard deviations are equal. It's not appropriate to conduct this test in order to determine which t method to use. In fact, we don't recommend this test even if your main purpose is to compare variability of two groups. The test assumes that the population distributions are normal, and it is not robust to violations of that assumption.

Effect Size

In Example 7.7, on the anorexia study, is the estimated difference between the mean weight gains of 3.46 large or small in practical terms? Keep in mind that the size of an estimated difference depends on the units of measurement. These data were in pounds, but if converted to kilograms the estimated difference would be 1.57 and if converted to ounces it would be 55.4.

A standardized way to describe the difference divides it by the estimated standard deviation for each group. This is called the ***effect size***. With sample means of 3.01 and -0.45 pounds and an estimated common standard deviation of $s = 7.64$ pounds, the standardized difference is

$$\text{Effect size} = \frac{\bar{y}_1 - \bar{y}_2}{s} = \frac{3.01 - (-0.45)}{7.64} = 0.45.$$

The difference between the sample means is less than half a standard deviation, a relatively small difference. We would obtain the same value for the effect size if we measured these data in different units, such as kilograms or ounces.

A Model for Means

In the second half of this book, we'll learn about advanced methods for analyzing associations among variables. We'll base analyses explicitly on a *model*. For two variables, a ***model*** is a simple approximation for the true relationship between those variables in the population.

Let $N(\mu, \sigma)$ denote a normal distribution with mean μ and standard deviation σ. Let y_1 denote a randomly selected observation from group 1 and y_2 a randomly selected observation from group 2. The hypothesis tested above for comparing means under the assumption $\sigma_1 = \sigma_2$ can be expressed as the model

H_0: Both y_1 and y_2 have a $N(\mu, \sigma)$ distribution.

H_a: y_1 has a $N(\mu_1, \sigma)$ distribution, y_2 has a $N(\mu_2, \sigma)$ distribution, with $\mu_1 \neq \mu_2$.

Under H_0, the population means are equal, with some common value μ. Under H_a, the population means differ. This is a special case of a model Chapter 12 uses for comparing *several* means.

Sampling distributions and resulting inferences are derived under the assumed model structure. But models are merely convenient simplifications of reality. We do not expect distributions to be exactly normal, for instance. One of the key parts of becoming more comfortable using statistical methods is becoming knowledgeable about which assumptions are most important in a model and how to check such

assumptions. Generally, there are benefits to using simpler models. They have fewer parameters to estimate, and inferences can be more powerful. However, when such a model is badly in error, we're better off using a more complex model.

The first significance test we discussed for comparing means used a slightly more complex model,

H_0: y_1 has a $N(\mu, \sigma_1)$ distribution, y_2 has a $N(\mu, \sigma_2)$ distribution.

H_a: y_1 has a $N(\mu_1, \sigma_1)$ distribution, y_2 has a $N(\mu_2, \sigma_2)$ distribution, with $\mu_1 \neq \mu_2$.

Again, under H_0 the population means are equal. But now, no assumption is made about the standard deviations being equal. If there is reason to expect the standard deviations to be very different, or if the data indicate this (with one of the sample standard deviations being at least double the other), then we're better off using analyses based on this model. If the data show that even this model is badly in error, such as when the sample data distributions are so highly skewed that the mean is an inappropriate summary, we're better off using a different model yet. The final section of this chapter presents a model that does not assume normality.

7.6 OTHER METHODS FOR COMPARING PROPORTIONS*

Section 7.2 presented large-sample methods for comparing proportions with independent samples. This section presents methods for comparing proportions with (1) dependent sample and (2) small samples.

Comparing Dependent Proportions

Section 7.4 presented dependent-samples methods for comparing means. The following example illustrates dependent-samples methods for comparing proportions.

EXAMPLE 7.8 Comparing Two Speech Recognition Systems

In recent years there have been impressive improvements in systems for automatically recognizing speech. When you call many service centers these days, before speaking with a human being you are asked to answer various questions verbally, whereas in the past you had to use the telephone dial pad.

Research in comparing the quality of different speech recognition systems often uses as a benchmark test a series of isolated words, checking how often each system makes errors recognizing the word. Table 7.7 shows an example[4] of one such test, comparing two speech recognition systems, called generalized minimal distortion segmentation (GMDS) and continuous density hidden Markov model (CDHMM).

TABLE 7.7: Results of Benchmark Test Using 2000 Words for Two Speech Recognition Systems

	CDHMM		
GMDS	Correct	Incorrect	Total
Correct	1921	58	1979
Incorrect	16	5	21
Total	1937	63	2000

[4]From S. Chen and W. Chen, *IEEE Transactions on Speech and Audio Processing*, vol. 3, 1995, pp. 141–145.

The rows of Table 7.7 are the (correct, incorrect) categories for each word using GMDS. The columns are the same categories for CDHMM. The row marginal counts (1979, 21) are the (correct, incorrect) totals for GMDS. The column marginal counts (1937, 63) are the totals for CDHMM.

We will compare the proportion of correct responses for these two speech recognition systems. The samples are dependent, because the two systems used the same 2000 words. We'll regard these 2000 words as a random sample of the possible words on which the systems could have been tested. Let π_1 denote the population proportion correct for GMDS, and let π_2 denote the population proportion correct for CDHMM. The sample estimates are $\hat{\pi}_1 = 1979/2000 = 0.9895$ and $\hat{\pi}_2 = 1937/2000 = 0.9685$.

If the proportions correct were identical for the two systems, the number of observations in the first row of Table 7.7 would equal the number of observations in the first column. The first cell (the one containing 1921 in Table 7.7) is common to both the first row and first column, so the other cell count in the first row would equal the other cell count in the first column. That is, the number of words judged correctly by GMDS but incorrectly by CDHMM would equal the number of words judged incorrectly by GMDS but correctly by CDHMM. We can test H_0: $\pi_1 = \pi_2$ using the counts in those two cells. If H_0 is true, then of these words, we expect 1/2 to be correct for GMDS and incorrect for CDHMM and 1/2 to be incorrect for GMDS and correct for CDHMM.

As in the matched-pairs test for a mean, we reduce the inference to one about a single parameter. For the population in the two cells just mentioned, we test whether half are in each cell. In Table 7.7, of the $58 + 16 = 74$ words judged correctly by one system but incorrectly by the other, the sample proportion $58/74 = 0.784$ were correct with GMDS. Under the null hypothesis that the population proportion is 0.50, the standard error of the sample proportion for these 74 observations is $\sqrt{(0.50)(0.50)/74} = 0.058$.

From Section 6.3, the z statistic for testing that the population proportion equals 0.50 is

$$z = \frac{\text{sample proportion} - H_0 \text{ proportion}}{\text{standard error}} = \frac{0.784 - 0.50}{0.058} = 4.88.$$

The two-sided P-value equals 0.000. This provides strong evidence against H_0 : $\pi_1 = \pi_2$. Based on the sample proportions, the evidence favors a greater population proportion of correct recognitions by the GMDS system. ■

McNemar Test for Comparing Dependent Proportions

A simple formula exists for this z test statistic for comparing two dependent proportions. For a table of the form of Table 7.7, denote the cell counts in the two relevant cells by n_{12} for those in row 1 and in column 2 and by n_{21} for those in row 2 and in column 1. The test statistic equals

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

When $n_{12} + n_{21}$ exceeds about 20, this statistic has approximately a standard normal distribution when H_0 is true. This test is often referred to as ***McNemar's test***. For smaller samples, use the binomial distribution to conduct the test.

For Table 7.7, the McNemar test uses $n_{12} = 58$, the number of words correctly recognized by GMDS and incorrectly by CDHMM, and $n_{21} = 16$, the number for the reverse. The test statistic equals

$$z = \frac{58 - 16}{\sqrt{58 + 16}} = 4.88.$$

The P-value is 0.000.

Confidence Interval for Difference of Dependent Proportions

A confidence interval for the difference of proportions is more informative than a significance test. For large samples, this is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se),$$

where the standard error is estimated using

$$se = \sqrt{(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n}/n.$$

For Table 7.7, $\hat{\pi}_1 = 1979/2000 = 0.9895$ and $\hat{\pi}_2 = 1937/2000 = 0.9685$. The difference $\hat{\pi}_1 - \hat{\pi}_2 = 0.9895 - 0.9685 = 0.021$. For $n = 2000$ observations with $n_{12} = 58$ and $n_{21} = 16$,

$$se = \sqrt{(58 + 16) - (58 - 16)^2/2000}/2000 = 0.0043.$$

A 95% confidence interval for $\pi_1 - \pi_2$ equals $0.021 \pm 1.96(0.0043)$, or (0.013, 0.029). We conclude that the population proportion correct with the GMDS system is between about 0.01 and 0.03 higher than the population proportion correct with the CDHMM system. In summary, the difference between the population proportions seems to be quite small.

Fisher's Exact Test for Comparing Proportions

The inferences for proportions with independent samples introduced in Section 7.2 are valid for relatively large samples. We next consider small-sample methods.

The two-sided significance test for comparing proportions with z test statistic works quite well if each sample has at least about 5–10 outcomes of each type (i.e., at least 5–10 observations in each cell of the contingency table). For smaller sample sizes, the sampling distribution of $\hat{\pi}_2 - \hat{\pi}_1$ may not be close to normality. You can then compare two proportions π_1 and π_2 using a method called ***Fisher's exact test***, due to the eminent statistician R. A. Fisher.

The calculations for Fisher's exact test are complex and beyond the scope of this text. The principle behind the test is straightforward, however, as Exercise 7.57 shows. Statistical software provides its P-value. As usual, the P-value is the probability of the sample result or a result even more extreme, under the presumption that H_0 is true. For details about Fisher's exact test, see Agresti (2007, pp. 45–48).

EXAMPLE 7.9 Depression and Suicide among HIV Infected Persons

A recent study[5] examined rates of major depression and suicidality for HIV infected and uninfected persons in China. The study used a volunteer sample. In an attempt to

[5] H. Jin et al., *J. Affective Disorders*, vol. 94, 2006, pp. 269–275.

TABLE 7.8: Comparison of HIV-Infected and Uninfected Subjects on Whether Have Ever Attempted Suicide

```
HIV        suicide

          |yes      |no       |  Total
----------+---------+---------+
positive  |      10 |      18 |     28
----------+---------+---------+
negative  |       1 |      22 |     23
----------+---------+---------+
Total            11        40       51

 STATISTICS FOR TABLE OF HIV BY SUICIDE
 Statistic                              Prob
 ------------------------------------------
 Fisher's Exact Test (Left)           0.9995
                     (Right)          0.0068
                     (2-Tail)         0.0075
```

make the sample more representative, subjects were recruited from clinics in two very different regions of China, one urban and one rural. Table 7.8 shows results based on a diagnostic interview asking whether the subject had ever attempted suicide. The table also shows output from conducting Fisher's exact test.

Denote the population proportion who had ever made a suicide attempt by π_1 for those who were HIV positive and by π_2 for those who were HIV negative. Then $\hat{\pi}_1 = 10/28 = 0.36$ and $\hat{\pi}_2 = 1/23 = 0.04$. We test H_0: $\pi_1 = \pi_2$ against H_a: $\pi_1 > \pi_2$. One of the four counts is very small, so to be safe we use Fisher's exact test.

On the printout, the right-sided alternative refers to H_a: $\pi_1 - \pi_2 > 0$; that is, H_a: $\pi_1 > \pi_2$. The P-value = 0.0068 gives very strong evidence that the population proportion attempting suicide is higher for those who are HIV positive. The P-value for the two-sided alternative equals 0.0075. This is not double the one-sided P-value because, except in certain special cases, the sampling distribution (called the ***hypergeometric distribution***) is not symmetric. ■

Small-Sample Estimation Comparing Two Proportions

From Section 7.2, the confidence interval for comparing proportions with large samples is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se), \text{ where } se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

A simple adjustment of this formula so that it works better, even for small samples, adds one observation of each type to each sample. For the data in Table 7.8 for Example 7.9, we replace the cell counts (10, 18, 1, 22) by (11, 19, 2, 23).

Then the adjusted estimates are $\tilde{\pi}_1 = (10 + 1)/(28 + 2) = 0.367$ and $\tilde{\pi}_2 = (1 + 1)/(23 + 2) = 0.080$. The adjusted standard error (using $n_1 = 30$ and $n_2 = 25$) equals 0.108, and a 95% confidence interval is

$$(0.367 - 0.080) \pm 1.96(0.103), \text{ or } 0.287 \pm 0.203, \text{ which is } (0.08, 0.49).$$

Not surprisingly, with such small samples the interval is very wide.

7.7 NONPARAMETRIC STATISTICS FOR COMPARING GROUPS*

We have seen that many statistics have large-sample normal sampling distributions, even when population distributions are not normal. In fact, with random sampling, nearly all parameter estimators have normal distributions for large sample sizes. Small samples, though, often require additional assumptions. For instance, inferences for means using the t distribution assume normal population distributions.

A body of methods exist that make *no* assumption about the shape of the population distribution. These methods are called ***nonparametric***. They contrast with the traditional (so-called *parametric*) methods that assume normal populations. Nonparametric methods are useful, for instance, when the normality assumption for methods using the t distribution is badly violated. They are primarily useful for small samples, especially for one-sided tests, as parametric methods may then work poorly when the normal population assumption is badly violated. They are also useful when the two groups have highly skewed distributions, because then the mean may not be a meaningful summary measure.

Wilcoxon-Mann-Whitney Test

To illustrate, consider the t distribution method for comparing means that assumes normal population distributions with identical standard deviations (Section 7.5). These assumptions are mainly relevant for small samples, say when n_1 or n_2 is less than about 20–30. Most nonparametric comparisons of groups also assume identical shapes for the population distributions, but the shapes are not required to be normal. The model for the test is then,

H_0: Both y_1 and y_2 have the same distribution.

H_a: The distributions for y_1 and y_2 have the same shape, but the one for y_1 is shifted up or shifted down compared to the one for y_2.

The most popular test of this type is called the *Wilcoxon* test. This test is an ordinal-level method, in the sense that it uses only the rankings of the observations. The combined sample of $n_1 + n_2$ measurements are ranked from 1 to $n_1 + n_2$, and the means of the ranks are computed for observations in each sample. The test statistic compares the sample mean ranks. For large samples, a z test statistic has an approximate standard normal distribution. For small samples, an exact P-value is based on how unusual the observed difference between the mean ranks is (under the presumption that H_0 is true) compared to the differences between the mean ranks for all other possible rankings.

Another nonparametric test is the *Mann-Whitney* test. It views all the pairs of observations, such that one observation is from one group and the other observation is from the other group. The test statistic is based on the number of pairs for which the observation from the first group was higher. This test is equivalent to the Wilcoxon test, giving the same P-value. (Frank Wilcoxon developed equivalent tests as Henry Mann and D. R. Whitney at about the same time in the late 1940s.)

For Example 7.5, comparing weight changes for a cognitive behavioral therapy group and a control group in the anorexia study (page 198), the parametric t test had a two-sided P-value of 0.10. The large-sample version of the Wilcoxon-Mann-Whitney test reports similar results, with a two-sided P-value of 0.11.

Some software also can report a corresponding confidence interval for the difference between the population medians. The method assumes that the two population distributions have the same shape, but not necessarily bell shaped. The median weight change was 1.4 pounds for the cognitive behavioral therapy group and -0.35 pound for the control group. Software reports a 95% confidence interval for the difference between the medians of $(-0.6, 8.1)$ pounds.

Effect Size: Proportion of Better Responses for a Group

Section 7.5 mentioned that the size of the difference between two groups is sometimes summarized by the *effect size*, which for two samples is defined as $(\bar{y}_1 - \bar{y}_2)/s$. When the distributions are very skewed or have outliers, the means are less useful and this effect size summary may be inappropriate. A nonparametric effect size measure is the proportion of pairs of observations (one from each group) for which the observation from the first group was higher. If y_1 denotes a randomly selected observation from group 1 and y_2 a randomly selected observation from group 2, then this measure estimates $P(y_1 > y_2)$.

To illustrate, suppose the anorexia study had 4 girls, 2 using a new therapy and 2 in a control group. Suppose the weight changes were

$$\text{Therapy group } (y_1): 4, 10$$
$$\text{Control group } (y_2): 2, 6.$$

There are four pairs of observations, with one from each group:

$$y_1 = 4, y_2 = 2 \text{ (Group 1 is higher)}$$
$$y_1 = 4, y_2 = 6 \text{ (Group 2 is higher)}$$
$$y_1 = 10, y_2 = 2 \text{ (Group 1 is higher)}$$
$$y_1 = 10, y_2 = 6 \text{ (Group 1 is higher).}$$

Group 1 is higher in 3 of the 4 pairs, so the estimate of $P(y_1 > y_2)$ is 0.75. If two observations had the same value, we would count it as y_1 being higher for 1/2 the pair (rather than 1 or 0).

Under H_0 of no effect, $P(y_1 > y_2) = 0.50$. The farther $P(y_1 > y_2)$ falls from 0.50, the stronger the effect. For the full anorexia data set analyzed in Example 7.7 on page 198, the sample estimate of $P(y_1 > y_2)$ is 0.63. The estimated probability that a girl using the cognitive behavioral therapy has a larger weight gain than a girl using the control therapy is 0.63.

When the two groups have normal distributions with the same standard deviation, a connection exists between this effect size and the parametric one, $(\mu_1 - \mu_2)/\sigma$. For example, when $(\mu_1 - \mu_2)/\sigma = 0$, then $P(y_1 > y_2) = 0.50$; when $(\mu_1 - \mu_2)/\sigma = 0.5$, then $P(y_1 > y_2) = 0.64$; when $(\mu_1 - \mu_2)/\sigma = 1$, then $P(y_1 > y_2) = 0.71$; when $(\mu_1 - \mu_2)/\sigma = 2$, then $P(y_1 > y_2) = 0.92$. The effect is relatively strong if $P(y_1 > y_2)$ is larger than about 0.70 or smaller than about 0.30.

Treating Ordinal Variables as Quantitative

Social scientists often use parametric statistical methods for quantitative data with variables that are only ordinal. They do this by assigning scores to the ordered

categories. Example 6.2 (page 149), on political ideology, showed an example of this. Sometimes the choice of scores is straightforward. For categories (liberal, moderate, conservative) for political ideology, any set of equally spaced scores is sensible, such as (1, 2, 3) or (0, 5, 10). When the choice is unclear, such as with categories (not too happy, pretty happy, very happy) for happiness, it is a good idea to perform a sensitivity study. Choose two or three reasonable sets of potential scores, such as (0, 5, 10), (0, 6, 10), (0, 7, 10), and check whether the ultimate conclusions are similar for each. If not, any report should point out how conclusions depend on the scores chosen.

Alternatively, nonparametric methods are valid with ordinal data. The reason is that nonparametric methods do not use quantitative scores, but rather rankings of the observations, and rankings are ordinal information. However, this approach works best when the response variable is continuous (or nearly so), so each observation has its own rank. When used with ordered categorical responses, such methods are often less sensible than using parametric methods that treat the response as quantitative. The next example illustrates.

EXAMPLE 7.10 Alcohol Use and Infant Malformation

Table 7.9 refers to a study of maternal drinking and congenital malformations. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on presence or absence of congenital sex organ malformations. Alcohol consumption was measured as average number of drinks per day.

Is alcohol consumption associated with malformation? One approach to investigate this is to compare the mean alcohol consumption of mothers for the cases where malformation occcurred to the mean alcohol consumption of mothers for the cases where malformation did not occur. Alcohol consumption was measured by grouping values of a quantitative variable. To find means, we assign scores to alcohol consumption that are midpoints of the categories; that is, 0, 0.5, 1.5, 4.0, 7.0, the last score (for ≥6) being somewhat arbitrary. The sample means are then 0.28 for the absent group and 0.40 for the present group, and the t statistic of 2.56 has P-value of 0.01. There is strong evidence that mothers whose infants suffered malformation had a higher mean alcohol consumption.

An alternative, nonparametric, approach assigns ranks to the subjects and uses them as the category scores. For all subjects in a category, we assign the average of the ranks that would apply for a complete ranking of the sample. These are called *midranks*. For example, the 17,114 subjects at level 0 for alcohol consumption share ranks 1 through 17,114. We assign to each of them the average of these ranks, which is the midrank $(1 + 17{,}114)/2 = 8557.5$. The 14,502 subjects at level <1 for alcohol consumption share ranks 17,115 through $17{,}114 + 14{,}502 = 31{,}616$, for a midrank of

TABLE 7.9: Infant Malformation and Mother's Alcohol Consumption

	Alcohol Consumption				
Malformation	0	<1	1–2	3–5	≥6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1
Total	17,114	14,502	793	127	38

Source: Graubard, B. I., and Korn, E. L., *Biometrics*, vol. 43, 1987, pp. 471–476.

(17,115 + 31,616)/2 = 24,365.5. Similarly the midranks for the last three categories are 32,013, 32,473, and 32,555.5. Used in a large-sample Wilcoxon test, these scores yield much less evidence of an effect ($P = 0.55$).

Why does this happen? Adjacent categories having relatively few observations necessarily have similar midranks. The midranks (8557.5, 24,365.5, 32,013, 32,473, 32,555.5) are similar for the final three categories, since those categories have considerably fewer observations than the first two categories. A consequence is that this scoring scheme treats alcohol consumption level 1–2 (category 3) as much closer to consumption level ≥ 6 (category 5) than to consumption level 0 (category 1). This seems inappropriate. It is better to use your judgment by selecting scores that reflect well the distances between categories. ■

Although nonparametric methods have the benefit of weaker assumptions, in practice social scientists do not use them as much as parametric methods. Partly this reflects the large sample sizes for most studies, for which assumptions about population distributions are not so vital. In addition, nonparametric methods for multivariate data sets are not as thoroughly developed as parametric methods. Most nonparametric methods are beyond the scope of this text. For details, see Hollander and Wolfe (1999).

7.8 CHAPTER SUMMARY

This chapter introduced methods for comparing two groups. For quantitative response variables, inferences apply to the difference $\mu_2 - \mu_1$ between population means. For categorical response variables, inferences apply to the difference $\pi_2 - \pi_1$ between population proportions.

In each case, the significance test analyzes whether 0 is a plausible difference. If the confidence interval contains 0, it is plausible that the parameters are equal. Table 7.10

TABLE 7.10: Summary of Comparison Methods for Two Groups, for Independent Random Samples

	Type of Response Variable	
	Categorical	Quantitative
Estimation		
1. Parameter	$\pi_2 - \pi_1$	$\mu_2 - \mu_1$
2. Point estimate	$\hat{\pi}_2 - \hat{\pi}_1$	$\bar{y}_2 - \bar{y}_1$
3. Standard error	$se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$	$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
4. Confidence interval	$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se)$	$(\bar{y}_2 - \bar{y}_1) \pm t(se)$
Significance testing		
1. Assumptions	Randomization ≥10 observations in each category, for each group	Randomization Normal population dist.'s (robust, especially for large n's)
2. Hypotheses	H_0: $\pi_1 = \pi_2$ ($\pi_2 - \pi_1 = 0$) H_a: $\pi_1 \neq \pi_2$	H_0: $\mu_1 = \mu_2$ ($\mu_2 - \mu_1 = 0$) H_a: $\mu_1 \neq \mu_2$
3. Test statistic	$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0}$	$t = \frac{\bar{y}_2 - \bar{y}_1}{se}$
4. P-value	Two-tail probability from standard normal or t (Use one tail for one-sided alternative)	

summarizes the methods for ***independent*** random samples, for which observations in the two samples are not matched. This is the most common case in practice.

- Both for differences of proportions and differences of means, confidence intervals have the form

$$\text{Estimated difference} \pm (\text{score})(se)$$

  using a z-score for proportions and t-score for means. In each case, the test statistic equals the estimated difference divided by the standard error.
- For ***dependent*** samples, each observation in one sample matches with an observation in the other sample. For quantitative variables, we compare means by analyzing the mean of difference scores computed between the paired observations. The ***paired-difference*** confidence interval and test procedures are the one-sample methods of Chapters 5 and 6 applied to the difference scores.
- Another approach for comparing means makes the extra assumption that the normal population distributions have equal standard deviations. This approach pools the standard deviations from the two samples to find a common estimate.
- For comparing proportions, with independent samples the small-sample test is ***Fisher's exact test***. For dependent samples, ***McNemar's test*** compares the number of subjects who are in category 1 in the first sample and category 2 in the second sample to the number of subjects who are in category 2 in the first sample and category 1 in the second.
- ***Nonparametric*** statistical methods make no assumption about the shape of the population distribution. Most such methods use the ranks of the observations.

At this stage, you may feel confused about which method to use for any given situation. It may help if you use the following checklist. Ask yourself, is the analysis about

- Means or proportions (quantitative or categorical response variable)?
- Independent samples or dependent samples?
- Confidence interval or significance test?

PROBLEMS

Practicing the Basics

7.1. An Associated Press story (Feb. 23, 2007) about UCLA's annual survey of college freshmen indicated that 73% of college freshmen in 2006 considered being financially well off to be very important, compared to 42% in 1966 (the first year the survey was done). It also reported that 81% of 18- to 25-year-olds in the U.S. see getting rich as a top goal in life. Are the sample percentages of 42% in 1966 and 73% in 2006 based on independent samples or dependent samples? Explain.

7.2. *Transatlantic Trends* is an annual survey of American and European public opinion (see www.transatlantictrends.org), with a random sample of about 1000 adults from each of 13 European countries each year. In 2002, 38% of Europeans expressed a positive attitude about President George W. Bush's handling of international affairs. In 2006, 18% expressed a positive attitude.

(a) Explain what it would mean for these results to be based on (a) *independent* samples, (b) *dependent* samples.

(b) If we compare results in 2002 and 2006, identify the response variable and the explanatory variable, and specify whether the response variable is quantitative or categorical.

7.3. The National Health Interview Survey (www.cdc.gov/nchs) estimated that current cigarette smokers were 41.9% of American adults in 1965 and 21.5% in 2003.

(a) Estimate the difference between the proportions who smoked in the two years.

(b) Suppose the standard error were reported as 0.020 for each proportion. Find the standard error of the difference. Interpret.

7.4. When a recent Eurobarometer survey asked subjects in each European Union country whether they would be willing to pay more for energy produced from renewable sources than for energy produced from other sources, the proportion answering *yes* varied from a high of 0.52 in Denmark ($n = 1008$) to a low of 0.14 in Lithuania ($n = 1002$). For this survey:

(a) Estimate the difference between Denmark and Lithuania in the population proportion of *yes* responses.

(b) From the $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ formula in Chapter 5, the proportion estimates have $se = 0.0157$ for Denmark and $se = 0.110$ for Lithuania. Use these to find the *se* for the difference estimate in (a). Interpret this *se*.

7.5. The National Center for Health Statistics recently estimated that the mean weight for adult American women was 140 pounds in 1962 and 164 pounds in 2002.

(a) Suppose these estimates had standard errors of 2 pounds each year. Estimate the increase in mean weight in the population from 1962 to 2002, and find and interpret the standard error of that estimate.

(b) Show that the estimated mean in 2002 was 1.17 times the estimated mean in 1962. Express this in terms of the percentage increase.

(c) The estimated mean weights for men were 166 pounds in 1962 and 191 in 2002. Find and interpret the difference and the ratio.

7.6. The U.S. Census Bureau reported that in 2002 the median net worth in the U.S. was estimated to be about $89,000 for white households and $6000 for black households.

(a) Identify the response variable and the explanatory variable.

(b) Compare the groups using a (i) difference, (ii) ratio.

7.7. According to the U.S. Department of Justice, in 2002 the incarceration rate in the nation's prisons was 832 per 100,000 male residents, and 58 per 100,000 female residents.

(a) Find the relative risk of being incarcerated, comparing males to females. Interpret.

(b) Find the difference of proportions incarcerated. Interpret.

(c) Which measure do you think better summarizes these data? Why?

7.8. According to the U.S. National Center for Health Statistics, the annual probability that a male between the ages of 20 and 24 is a homicide victim is about 0.00164 for blacks and 0.00015 for whites.

(a) Compare these rates using the difference of proportions.

(b) Compare these rates using the relative risk.

(c) Which of the two measures seems to better summarize results when both proportions are very close to 0? Explain.

7.9. An Associated Press story (August 7, 2006) about a research study regarding the impact on teens of sexual lyrics in songs reported, "Teens who said they listened to lots of music with degrading sexual messages were almost twice as likely to start having intercourse ... within the following two years as were teens who listened to little or no sexually degrading music." The reported percentages were 51% and 29%.

(a) A 95% confidence interval for the difference between corresponding population proportions was (0.18, 0.26). Explain how to interpret it.

(b) The *P*-value is <0.001 for testing the null hypothesis that the corresponding population proportions are equal. Interpret.

7.10. For a random sample of Canadians, 60% indicate approval of the prime minister's performance. A similar poll a month later has a favorable rating of 57%. A 99% confidence interval for the change in the population proportions is $(-0.07, 0.01)$. Explain why (a) there may have been no change in support, (b) if a decrease in support occurred, it may have been fairly important, (c) if an increase in support occurred, it was probably so small as to be substantively unimportant.

7.11. The College Alcohol Study at the Harvard School of Public Health has interviewed random samples of students at 4-year colleges several times since 1993. Of the students who reported drinking alcohol, the percentage who reported that drinking "to get drunk" is an important reason for drinking was 39.9% of 12,708 students in 1993 and 48.2% of 8783 students in 2001.[6] For comparing results in 1993 and 2001:

(a) Show that the standard error for the estimated difference between the corresponding population proportions in 2001 and in 1993 equals 0.0069.

(b) Show that the 95% confidence interval for the difference is (0.07, 0.10). Interpret.

7.12. In the study mentioned in the previous exercise, the percent who said they had engaged in unplanned sexual activities because of drinking alcohol was 19.2% in 1993 and 21.3% in 2001.

(a) Specify assumptions, notation, and hypotheses for a two-sided test comparing the corresponding population proportions.

[6] *Journal of American College Health*, vol. 50, 2002, pp. 203–217.

(b) The test statistic $z = 3.8$ and the P-value = 0.0002. Interpret the P-value.

(c) Some might argue that the result in (b) reflects *statistical significance* but not *practical significance*. Explain the basis of this argument, and explain why you learn more from the 95% confidence interval, which is (0.009, 0.033).

7.13. For the Time Use Survey reported in Table 7.1 (page 183), of those working full time, 55% of 1219 men and 74% of 733 women reported spending some time on cooking and washing up during a typical day. Find and interpret a 95% confidence interval for the difference in participation rates.

7.14. Table 7.11 summarizes responses from General Social Surveys in 1977 and in 2006 to the question (FEFAM), "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." Let π_1 denote the population proportion who agreed with this statement in 1977, and let π_2 denote the population proportion in 2006.

(a) Show that $\hat{\pi}_1 - \hat{\pi}_2 = 0.30$, with standard error 0.0163.

(b) Show that the 95% confidence interval for $\pi_1 - \pi_2$ is (0.27, 0.33). Interpret.

(c) Explain how results would differ for comparing the proportions who did *not* agree in the two years.

TABLE 7.11

Year	Agree	Disagree	Total
1977	989	514	1503
2006	704	1264	1968

7.15. Refer to the previous exercise on a woman's role. In 2004, of 411 male respondents, 153 (37.2%) replied yes. Of 472 female respondents, 166 (35.2%) replied yes.

(a) Set up notation and specify hypotheses for the hypothesis of no difference between the population proportions of males and of females who would respond yes.

(b) Estimate the population proportion presuming H_0, find the standard error of the sample difference of proportions, and find the test statistic.

(c) Find the P-value for the two-sided alternative. Interpret.

(d) Of 652 respondents having less education than a college degree, 40.0% replied yes. Of 231 respondents having at least a college degree, 25.6% replied yes. Which variable, gender or educational level, seems to have had the greater influence on opinion? In other words, did opinion tend to differ more between men and women or between the most and least educated?

7.16. In a survey conducted by Wright State University, senior high school students were asked if they had ever used marijuana. Table 7.12 shows software output. Treating these observations as a random sample from the population of interest:

(a) State a research question that could be addressed with this output.

(b) Interpret the reported confidence interval.

(c) Interpret the reported P-value.

TABLE 7.12

```
Sample        yes       N        Sample prop
1. Female     445       1120       0.3973
2. Male       515       1156       0.4455

estimate for p(1) - p(2):  -0.0482
95% CI for p(1) - p(2): ( -0.0887, -0.0077)
Test for difference = 0 (vs not = 0):
z = -2.33  P-value = 0.020
```

7.17. A study of compulsive buying behavior (uncontrolled urges to buy) conducted a national telephone survey in 2004 of adults ages 18 and over.[7] Of 800 men, 44 were judged to be compulsive buyers according to the Compulsive Buying Scale. Of 1501 women, 90 were judged to be compulsive buyers. Conduct an inference to analyze whether one sex is more likely than the other to be a compulsive buyer. Interpret.

7.18. Table 7.13 shows results from a recent General Social Survey on two variables, sex and whether one believes in an afterlife (AFTERLIF). Conduct all steps of a significance test, using $\alpha = 0.05$, to compare the population proportions of females and males who would respond *yes* to belief in an afterlife. If you have made an error in your decision, what type of error is it, Type I or Type II?

TABLE 7.13

	Belief in Afterlife		
Sex	Yes	No or Undecided	Total
Female	435	147	582
Male	375	134	509

[7] Koran et al., *Amer. J. Psychiatry*, vol. 163, 2006, p. 1806.

7.19. A GSS reported that the 486 females had a mean of 8.3 close friends (s = 15.6) and the 354 males had a mean of 8.9 close friends (s = 15.5).

(a) A 95% confidence interval for the difference between the population means for males and for females is (-1.5, 2.7). Interpret.

(b) For each sex, does it seem like the distribution of number of close friends is normal? Explain why this does not invalidate the result in (a), but may affect the usefulness of the interval.

7.20. Table 7.14 summarizes the number of hours spent in housework per week by gender, based on the 2002 GSS (variable RHHWORK).

(a) Estimate the difference between the population means for women and men.

(b) Show that the estimated standard error of the sample difference is 0.81. Interpret.

(c) Show that a 99% confidence interval for the difference is (2.3, 6.5). Interpret.

TABLE 7.14

		Housework Hours	
Gender	Sample Size	Mean	Standard Deviation
Men	292	8.4	9.5
Women	391	12.8	11.6

7.21. A 30-month study evaluated the degree of addiction that teenagers form to nicotine once they begin experimenting with smoking.[8] The study used a random sample of 332 seventh-grade students in two Massachusetts cities who had ever used tobacco by the start of the study. The response variable was constructed from the Hooked on Nicotine Checklist (HONC). This is a list of ten questions such as "Have you ever tried to quit but couldn't?" The HONC score is the total number of questions to which a student answered *yes*. The higher the score, the greater the dependence on nicotine. There were 75 smokers and 257 ex-smokers at the end of the study. The HONC means describing nicotine addiction were 5.9 (s = 3.3) for the smokers and 1.0 (s = 2.3) for the ex-smokers.

(a) Find and interpret a point estimate to compare HONC means for smokers and ex-smokers.

(b) Software reports a 95% confidence interval of (4.1, 5.7). Interpret.

(c) Was the HONC sample data distribution for ex-smokers approximately normal? How does this affect inference?

7.22. Refer to Exercise 7.17, on compulsive buying behavior. The total credit card balance had a mean of \$3399 and standard deviation of \$5595 for 100 compulsive buyers and a mean of \$2837 and standard deviation of \$6335 for 1682 other respondents.

(a) Estimate the difference between the means for compulsive buyers and other respondents, and find its standard error.

(b) Compare the population means using a two-sided significance test. Interpret.

7.23. A recent GSS asked, "How many days in the past 7 days have you felt sad?" Software reported sample means of 1.8 for females and 1.4 for males, with a 95% confidence interval comparing them of (0.2, 0.6), a t statistic of 4.8, and a P-value of 0.000. Interpret these results.

7.24. For the 2006 GSS, a comparison of females and males on the number of hours a day that the subject watched TV gave:

Group	N	Mean	StDev	SE Mean
Females	1117	2.99	2.34	0.070
Males	870	2.86	2.22	0.075

(a) Conduct all parts of a significance test to analyze whether the population means differ for females and males. Interpret the P-value, and report the conclusion for α-level = 0.05.

(b) If you were to construct a 95% confidence interval comparing the means, would it contain 0? Answer based on the result of (a), without finding the interval.

(c) Do you think that the distribution of TV watching is approximately normal? Why or why not? Does this affect the validity of your inferences?

7.25. For the 2004 GSS, Table 7.15 shows software output for evaluating the number of hours of TV watching per day by race.

TABLE 7.15

Race	N	Mean	StDev	SE Mean
Black	101	4.09	3.63	0.3616
White	724	2.59	2.31	0.0859

```
Difference = mu (Black) - mu (White)
Estimate for difference :  1.50
95% CI for difference: (0.77, 2.23)
T-Test of difference = 0: T-value = 4.04,
P-value = 0.000
```

[8] J. DiFranza et al., *Archives of Pediatric and Adolescent Medicine*, vol. 156, 2002, pp. 397–403.

(a) Interpret the reported confidence interval. Can you conclude that one population mean is higher? If so, which one? Explain.
(b) Interpret the reported P-value.
(c) Explain the connection between the result of the significance test and the result of the confidence interval.

7.26. A study[9] compared personality characteristics between adult children of alcoholics and a control group matched on age and gender. For the 29 pairs of women, the authors reported a mean of 24.8 on the well-being measure for the children of alcoholics, and a mean of 29.0 for the control group. They reported $t = 2.67$ for the test comparing the means. Assuming that this is the result of a dependent-samples analysis, identify the df for the t test statistic, report the P-value, and interpret.

7.27. A paired-difference experiment[10] dealing with response latencies for noise detection under two conditions used a sample of twelve 9-month-old children and reported a sample mean difference of 70.1 and standard deviation of 49.4 for the differences. In their discussion, the authors reported a t statistic of 4.9 having $P < 0.01$ for a two-sided alternative. Show how they constructed the t statistic, and confirm the P-value.

7.28. As part of her class project, a student at the University of Florida randomly sampled 10 fellow students to investigate their most common social activities. As part of the study, she asked the students to state how many times they had done each of the following activities during the previous year: Going to a movie, going to a sporting event, or going to a party. Table 7.16 shows the data.

(a) To compare the mean movie attendance and mean sports attendance using statistical inference, should we treat the samples as independent or dependent? Why?
(b) For the analysis in (a), software shows results:

```
------------------------------------------------
              N      Mean     StDev   SE Mean
movies       10    13.000    13.174     4.166
sports       10     9.000     8.380     2.650
Difference   10     4.000    16.166     5.112

95% CI for mean difference: (-7.56, 15.56)
T-Test of mean difference = 0 (vs not = 0):
          T-Value = 0.78   P-Value = 0.454
------------------------------------------------
```

Interpret the 95% confidence interval shown.

(c) Show how the test statistic shown on the printout was obtained from the other information given. Report the P-value, and interpret in context.

TABLE 7.16

	Activity		
Student	Movies	Sports	Parties
1	10	5	25
2	4	0	10
3	12	20	6
4	2	6	52
5	12	2	12
6	7	8	30
7	45	12	52
8	1	25	2
9	25	0	25
10	12	12	4

7.29. Refer to the previous exercise. For comparing parties and sports, software reports a 95% confidence interval of $(-3.33, 28.93)$ and a P-value of 0.106.

(a) Interpret the P-value.
(b) Explain the connection between the results of the test and the confidence interval.

7.30. A clinical psychologist wants to choose between two therapies for treating mental depression. For six patients, she randomly selects three to receive therapy A, and the other three receive therapy B. She selects small samples for ethical reasons; if her experiment indicates that one therapy is superior, that therapy will be used on her other patients having these symptoms. After one month of treatment, the improvement is measured by the change in score on a standardized scale of mental depression severity. The improvement scores are 10, 20, 30 for the patients receiving therapy A, and 30, 45, 45 for the patients receiving therapy B.

(a) Using the method that assumes a common standard deviation for the two therapies, show that the pooled $s = 9.35$ and $se = 7.64$.
(b) When the sample sizes are very small, it may be worth sacrificing some confidence to achieve more precision. Show that the 90% confidence interval for $(\mu_2 - \mu_1)$ is (3.7, 36.3). Interpret.
(c) Estimate and summarize the effect size.

7.31. Refer to the previous exercise. To avoid bias from the samples being unbalanced with such small n, the psychologist redesigned the experiment. She

[9] D. Baker and L. Stephenson, *Journal of Clinical Psychology*, vol. 51, 1995, p. 694.
[10] J. Morgan and J. Saffran, *Child Development*, vol. 66, 1995, pp. 911–936.

forms three pairs of subjects, such that the patients matched in any given pair are similar in health and socioeconomic status. For each pair, she randomly selects one subject for each therapy. Table 7.17 shows the improvement scores, and Table 7.18 shows results of using SPSS to analyze the data.

(a) Compare the means by (i) finding the difference of the sample means for the two therapies, (ii) finding the mean of the difference scores. Compare.

(b) Verify the standard deviation of the differences and standard error for the mean difference.

(c) Verify the confidence interval shown for the population mean difference. Interpret.

(d) Verify the test statistic, df, and P-value for comparing the means. Interpret.

TABLE 7.17

Pair	Therapy A	Therapy B
1	10	30
2	20	45
3	30	45

7.32. A study[11] of bulimia among college women considered the effect of childhood sexual abuse on various components of a Family Environment Scale. For a measure of family cohesion, the sample mean for the bulimic students was 2.0 for 13 sexually abused students and 4.8 for 17 nonabused students. Table 7.19 shows software results of a two-sample comparison of means.

(a) Assuming equal population standard deviations, construct a 95% confidence interval for the difference in mean family cohesion for sexually abused students and nonabused students. Interpret.

(b) Explain how to interpret results of significance tests from this printout.

TABLE 7.19

```
Variable: COHESION
ABUSED     N      Mean      Std Dev    Std Error
-------------------------------------------------
yes       13       2.0          2.1         0.58
no        17       4.8          3.2         0.78

Variances        T        DF      P-value
-----------------------------------------
Unequal       2.89      27.5        0.007
Equal         2.73        28        0.011
```

7.33. For the survey of students described in Exercise 1.11, the responses on political ideology had a mean of 3.18 and standard deviation of 1.72 for the 51 nonvegetarian students and a mean of 2.22 and standard deviation of 0.67 for the 9 vegetarian students. When we use software to compare the means with a significance test, we obtain

```
Variances        T        DF      P-value
-----------------------------------------
Unequal      2.915      30.9       0.0066
Equal        1.636      58.0       0.1073
```

Explain why the results of the two tests differ so much, and give your conclusion about whether the population means are equal.

7.34. In 2006, the GSS asked about the number of hours a week spent on the World Wide Web (WWWTIME). The 1569 females had a mean of 4.9 and standard deviation of 8.6. The 1196 males had a mean of 6.2 and standard deviation of 9.9. Use these results to make an inference comparing males and females on WWWTIME in the population, assuming equal population standard deviations.

7.35. Two new short courses have been proposed for helping students who suffer from severe math phobia, scoring at least 8 on a measure of math phobia that falls between 0 and 10 (based on responses to

TABLE 7.18

```
t-tests for Paired Samples

                Number of
Variable        pairs          Mean          SD      SE of Mean
THERAPY A                     20.000      10.000       5.774
                    3
THERAPY B                     40.000       8.660       5.000

Paired Differences
  Mean          SD     SE of Mean    t-value    df    2-tail Sig
  20.0000     5.00       2.887         6.93      2       0.020
  95% CI   (7.58, 32.42)
```

[11] J. Kern and T. Hastings, *J. Clinical Psychology*, vol. 51, 1995, p. 499.

10 questions). A sample of ten such students were randomly allocated to the two courses. Following the course, the drop in math phobia score was recorded. The sample values were

Course A: 0, 2, 2, 3, 3
Course B: 3, 6, 6, 7, 8.

(a) Make an inferential comparison of the means, assuming equal population standard deviations. Interpret your results.
(b) Using software, report and interpret the P-value for the two-sided Wilcoxon test.
(c) Find and interpret the effect size $(\bar{y}_B - \bar{y}_A)/s$.
(d) Estimate and interpret the effect size $P(y_B > y_A)$.

7.36. A GSS asked subjects whether they believed in heaven and whether they believed in hell. Of 1120 subjects, 833 believed in both, 160 believed in neither, 125 believed in heaven but not in hell, and 2 believed in hell but not in heaven.
(a) Display the data in a contingency table, cross classifying belief in heaven (*yes, no*) with belief in hell (*yes, no*).
(b) Estimate the population proportion who believe in heaven and the population proportion who believe in hell.
(c) Show all steps of McNemar's test to compare the population proportions, and interpret.
(d) Construct a 95% confidence interval to compare the population proportions, and interpret.

7.37. A GSS asked subjects their opinions about government spending on health and government spending on law enforcement. For each, should it increase, or should it decrease? Table 7.20 shows results.
(a) Find the sample proportion favoring increased spending, for each item.
(b) Test whether the population proportions are equal. Report the P-value, and interpret.
(c) Construct a 95% confidence interval for the difference of proportions. Interpret.

TABLE 7.20

Health Spending	Law Enforcement Spending	
	Increase	Decrease
Increase	292	25
Decrease	14	9

7.38. A study[12] used data from the Longitudinal Study of Aging to investigate how older people's health and social characteristics influence how far they live from their children. Consider Table 7.21, which shows whether an older subject lives with a child at a given time and then again four years later. The author expected that as people aged and their health deteriorated, they would be more likely to live with children. Do these data support this belief? Justify your answer with an inferential analysis.

TABLE 7.21

First Survey	Four Years Later	
	Yes	No
Yes	423	138
No	217	2690

7.39. A study[13] investigated the sexual orientation of adults who had been raised as children in lesbian families. Twenty-five children of lesbian mothers and a control group of 20 children of heterosexual mothers were seen at age 10 and again at age about 24. At the later time, they were interviewed about their sexual identity, with possible response *Bisexual/Lesbian/Gay* or *Heterosexual*. Table 7.22 shows results, in the form of a SAS printout for conducting Fisher's exact test.
(a) Why is Fisher's exact test used to compare the groups?
(b) Report and interpret the P-value for the alternative that the population proportion identifying as bisexual/lesbian/gay is higher for those with lesbian mothers.

TABLE 7.22

IDENTITY

MOTHER	B/L/G	HETERO	Total
Lesbian	2	23	25
Heterosx	0	20	20
Total	2	43	45

STATISTICS FOR TABLE OF MOTHER BY IDENTITY

Statistic		Prob
Fisher's Exact Test	(Left)	1.000
	(Right)	0.303
	(2-Tail)	0.495

7.40. Refer to the previous problem. The young adults were also asked whether they had ever had a same-gender sexual relationship. Table 7.23

[12] M. Silverstein, *Demography*, vol. 32, 1995, p. 35.
[13] S. Colombok and F. Tasker, *Developmental Psychology*, vol. 32, 1996, pp. 3–11.

shows results. Use software to test whether the probability of this is higher for those raised by lesbian mothers. Interpret.

TABLE 7.23

	Same-Gender Relationship	
Mother	Yes	No
Lesbian	6	19
Heterosexual	0	20

Concepts and Applications

7.41. For the "Student survey" data file (Exercise 1.11 on page 8), compare political ideology of students identifying with the Democratic party and with the Republican
- **(a)** Using graphical and numerical summaries.
- **(b)** Using inferential statistical methods. Interpret.

7.42. Using software with the student survey data set (Exercise 1.11), construct a confidence interval and conduct a test:
- **(a)** To compare males and females in terms of opinions about legalized abortion. Interpret.
- **(b)** To compare the mean weekly time spent watching TV to the mean weekly time in sports and other physical exercise.

7.43. For the data file created in Exercise 1.12, with variables chosen by your instructor, state a research question and conduct inferential statistical analyses. Prepare a report that summarizes your findings. In this report, also use graphical and numerical methods to describe the data and, if necessary, to check assumptions you make for your analysis.

7.44. Exercise 3.6 in Chapter 3 on page 61 showed data on carbon dioxide emissions, a major contributor to global warming, for advanced industrialized nations. Is there a difference between European and non-European nations in their emission levels? Conduct an investigation to answer this question.

7.45. Pose null and alternative hypotheses about the relationship between time spent on the Internet (WWWHR for the GSS) and a binary predictor available at the GSS that you believe may be associated with Internet use. Using the most recent GSS data on these variables at sda.berkeley.edu/GSS, conduct the test. Prepare a short report summarizing your analysis. (*Note*: The GSS Web site enables you to compare means for groups, by clicking on "Comparison of means.")

7.46. Browse one or two daily newspapers such as *The New York Times* (hard copy or online). Copy an article about a research study that compared two groups. Prepare a short report that answers the following questions:
- **(a)** What was the purpose of the research study?
- **(b)** Identify explanatory and response variables.
- **(c)** Can you tell whether the statistical analysis used (1) independent samples or dependent samples, or (2) a comparison of proportions or a comparison of means?

7.47. A recent study[14] considered whether greater levels of TV watching by teenagers were associated with a greater likelihood of committing aggressive acts over the years. The researchers randomly sampled 707 families in two counties in northern New York State and made follow-up observations over 17 years. They observed whether a sampled teenager later conducted any aggressive act against another person, according to a self report by that person or by their mother. Of 88 cases with less than 1 hour per day of TV watching, 5 had committed aggressive acts. Of 619 cases with at least 1 hour per day of TV, 154 had committed aggressive acts. Analyze these data, summarizing your analyses in a short report.

7.48. When asked by the GSS about the number of people with whom the subject had discussed matters of importance over the past six months (variable NUMGIVEN), the response of 0 was made by 8.9% of 1531 respondents in 1985 and by 24.6% of 1482 respondents in 2004. Analyze these data inferentially and interpret.

7.49. A study[15] compared substance use, delinquency, psychological well-being, and social support among various family types, for a sample of urban African-American adolescent males. The sample contained 108 subjects from single-mother households and 44 from households with both biological parents. The youths responded to a battery of questions that provides a measure of perceived parental support. This measure had sample means of 46 ($s = 9$) for the single-mother households and 42 ($s = 10$) for the households with both biological parents. Consider the conclusion, "The mean parental support was 4 units higher for the single-mother households. If the true means were equal, a difference of this size could be expected only 2% of the time. For samples of this size, 95% of the time one would expect this difference to be within 3.4 of the true value."
- **(a)** Explain how this conclusion refers to the results of (i) a confidence interval, (ii) a test.
- **(b)** Describe how you would explain the results of the study to someone who has not studied inferential statistics.

[14] J. G. Johnson et al., *Science*, vol. 295, 2002, pp. 2468–2471.

[15] M. Zimmerman et al., *Child Development*, vol. 66, 1995, pp. 1598–1613.

7.50. The results in Table 7.24 are from a study[16] of physical attractiveness and subjective well-being. A sample of college students were rated by a panel on their physical attractiveness. The table presents the number of dates in the past three months for students rated in the top or bottom quartile of attractiveness. Analyze these data, and interpret.

7.51. A report (12/04/2002) by the Pew Research Center on *What the World Thinks in 2002* reported that "the American public is strikingly at odds with publics around the world in its views about the U.S. role in the world and the global impact of American actions." Conclusions were based on polls in several countries. In Pakistan, in 2002 the percentage of interviewed subjects who had a favorable view of the U.S. was 10%, and the percentage who thought the spread of American ideas and customs was good was 2% ($n = 2032$).

(a) Do you have enough information to make an inferential comparison of the proportions? If so, do so. If not, what else would you need to know?

(b) For a separate survey in 2000, the estimated percentage who had a favorable view of the U.S. was 23%. To compare inferentially the percentages in 2000 and 2002, what more would you need to know?

7.52. A *Time Magazine* article titled "Wal-Mart's Gender Gap" (July 5, 2004) stated that in 2001 women managers at Wal-Mart earned $14,500 a year less, on the average, than their male counterparts. If you were also given the standard errors of the annual mean salaries for male and female managers at Wal-Mart, would you have enough information to determine whether this is a "statistically significant" difference? Explain.

7.53. The International Adult Literacy Survey (www.nifl.gov/nifl/facts/IALS.html) was a 22-country study in which nationally representative samples of adults were interviewed and tested at home, using the same literacy test having scores that could range from 0-500. For those of age 16–25, some of the mean prose literacy scores were UK 273.5, New Zealand 276.8, Ireland 277.7, U.S. 277.9, Denmark 283.4, Australia 283.6, Canada 286.9, Netherlands 293.5, Norway 300.4, Sweden 312.1. The Web site does not provide sample sizes or standard deviations. Suppose each sample size was 250 and each standard deviation was 50. How far apart do two sample means have to be before you feel confident that an actual difference exists between the population means? Explain your reasoning, giving your conclusion for Canada and the U.S.

7.54. Table 7.25 compares two hospitals on the outcomes of patient admissions for severe pneumonia. Although patient status is an ordinal variable, two researchers who analyze the data treat it as an interval variable. The first researcher assigns the scores (0, 5, 10) to the three categories. The second researcher, believing that the middle category is much closer to the third category than to the first, uses the scores (0, 9, 10). Each researcher calculates the means for the two institutions and identifies the institution with the higher mean as the one having more success in treating its patients. Find the two means for the scoring system used by (a) the first researcher, (b) the second researcher. Interpret. (Notice that the conclusion depends on the scoring system. So if you use methods for quantitative variables with ordinal data, take care in selecting scores.)

TABLE 7.25

	Patient Status		
	Died in Hospital	Released After Lengthy Stay	Released After Brief Stay
Hospital A	1	29	0
Hospital B	8	8	14

7.55. From Example 6.4 (page 151) in Chapter 6, for the cognitive behavioral therapy group the sample mean change in weight of 3.0 pounds was significantly different from 0. However, Example 7.7 (page 198) showed it is not significantly different from the mean change for the control group, even though that group had a negative sample mean change. How do you explain this paradox? (*Hint*: From Sections 7.1 and 7.3, how does the *se* value for estimating a difference between two means

TABLE 7.24

	No. Dates, Men			No. Dates, Women		
Attractiveness	Mean	Std. Dev.	n	Mean	Std. Dev.	n
More	9.7	10.0	35	17.8	14.2	33
Less	9.9	12.6	36	10.4	16.6	27

[16]E. Diener et al., *Journal of Personality and Social Psychology*, vol. 69, 1995, pp. 120–129.

compare to the *se* value for estimating a single mean?)

7.56. A survey by the Harris Poll of 2201 Americans in 2003 indicated that 51% believe in ghosts and 31% believe in astrology.

(a) Is it valid to compare the proportions using inferential methods for independent samples? Explain.

(b) Do you have enough information to compare them using inferential methods for dependent samples? Explain.

7.57. A pool of six candidates for three managerial positions includes three females and three males. Table 7.26 shows the results.

(a) Denote the three females by F_1, F_2, F_3 and the three males by M_1, M_2, M_3. Identify the 20 distinct samples of size three that can be chosen from these six individuals.

(b) Let $\hat{\pi}_1$ denote the sample proportion of males selected and $\hat{\pi}_2$ the sample proportion of females. For Table 7.26, $\hat{\pi}_1 - \hat{\pi}_2 = (2/3) - (1/3) = 1/3$. Of the 20 possible samples, show that 10 have $\hat{\pi}_1 - \hat{\pi}_2 \geq 1/3$. Thus, if the three managers were randomly selected, the probability would equal $10/20 = 0.50$ of obtaining $\hat{\pi}_1 - \hat{\pi}_2 \geq 1/3$. In fact, this is the reasoning that provides the one-sided P-value for Fisher's exact test.

(c) Find the P-value if all three selected are male. Interpret.

TABLE 7.26

	Chosen for Position	
Gender	Yes	No
Male	2	1
Female	1	2

7.58. Describe a situation in which it would be more sensible to compare means using dependent samples than independent samples.

7.59. An Associated Press story (Feb. 1, 2007) about a University of Chicago survey of 1600 people of ages 15 to 25 in several Midwest U.S. cities indicated that 58% of black youth, 45% of Hispanic youth, and 23% of white youth reported listening to rap music every day.

(a) True or false: If a 95% confidence interval comparing the population proportions for Hispanic and white youths was (0.18, 0.26), then we can infer that at least 18% but no more than 26% of the corresponding white population listens daily to rap music.

(b) The study reported that 66% of black females and 57% of black males agreed that rap music videos portray black women in bad and offensive ways. True or false: Because both these groups had the same race, inferential methods comparing them must assume dependent rather than independent samples.

7.60. True or false? If a 95% confidence interval for $(\mu_2 - \mu_1)$ contains only positive numbers, then we can conclude that both μ_1 and μ_2 are positive.

7.61. True or false? If you know the standard error of the sample mean for each of two independent samples, you can figure out the standard error of the difference between the sample means, even if you do not know the sample sizes.

In Exercises 7.62–7.64, select the correct response(s). More than one may be correct.

7.62. A 99% confidence interval for the difference $\pi_2 - \pi_1$ between the proportions of men and women in California who are alcoholics equals (0.02, 0.09).

(a) We are 99% confident that the proportion of alcoholics is between 0.02 and 0.09.

(b) We are 99% confident that the proportion of men in California who are alcoholics is between 0.02 and 0.09 larger than the proportion of women in California who are.

(c) At this confidence level, there is insufficient evidence to infer that the population proportions are different.

(d) We are 99% confident that a minority of California residents are alcoholics.

(e) Since the confidence interval does not contain 0, it is impossible that $\pi_1 = \pi_2$.

7.63. To compare the population mean annual incomes for Hispanics (μ_1) and for whites (μ_2) having jobs in construction, we construct a 95% confidence interval for $\mu_2 - \mu_1$.

(a) If the confidence interval is (3000, 6000), then at this confidence level we conclude that the population mean income is higher for whites than for Hispanics.

(b) If the confidence interval is $(-1000, 3000)$, then the corresponding $\alpha = 0.05$ level test of H_0: $\mu_1 = \mu_2$ against H_a: $\mu_1 \neq \mu_2$ rejects H_0.

(c) If the confidence interval is $(-1000, 3000)$, then it is plausible that $\mu_1 = \mu_2$.

(d) If the confidence interval is $(-1000, 3000)$, then we are 95% confident that the population mean annual income for whites is between \$1000 less and \$3000 more than the population mean annual income for Hispanics.

7.64. The Wilcoxon test differs from parametric procedures (for means) in the sense that

(a) It applies directly to ordinal as well as interval response variables.

(b) It is unnecessary to assume that the population distribution is normal.

(c) Random sampling is not assumed.

***7.65.** A test consists of 100 true–false questions. Joe did not study, so on each question, he randomly guesses the correct response.

(a) Find the probability that he scores at least 70, thus passing the exam. (*Hint*: Use the sampling distribution for the proportion of correct responses.)

(b) Jane studied a little and has a 0.60 chance of a correct response for each question. Find the probability that her score is nonetheless lower than Joe's. (*Hint*: Use the sampling distribution of the difference of sample proportions.)

(c) How do the answers to (a) and (b) depend on the number of questions? Explain.

***7.66.** Let y_{i1} denote the observation for subject i at time 1, y_{i2} the observation for subject i at time 2, and $y_i = y_{i2} - y_{i1}$.

(a) Letting $\bar{y}_1$, $\bar{y}_2$, and $\bar{y}_d$ denote the means of these observations, show that $\bar{y}_d = \bar{y}_2 - \bar{y}_1$.

(b) Is the median difference (i.e., the median of the y_i values) equal to the difference between the medians of the y_{i1} and y_{i2} values? Show that this is true, or give a counterexample to show that it is false.

This page intentionally left blank