This page intentionally left blank

CHAPTER 6

Statistical Inference: Significance Tests

An aim of many studies is to check whether the data agree with certain predictions. The predictions typically result from the theory that drives the research. These predictions are *hypotheses* about the study population.

Hypothesis

In statistics, a ***hypothesis*** is a statement about a population. It is usually a prediction that a parameter describing some characteristic of a variable takes a particular numerical value or falls in a certain range of values.

Examples of hypotheses are the following: "For workers in service jobs, the mean income is the same for women and for men," and "There is no difference between Democrats and Republicans in the probabilities that they vote with their party leadership," and "Half or more of adult Canadians are satisfied with their national health service."

A ***significance test*** uses data to summarize the evidence about a hypothesis. It does this by comparing point estimates of parameters to the values predicted by the hypothesis. The following example illustrates concepts behind significance tests.

EXAMPLE 6.1 Testing for Gender Bias in Selecting Managers

A large supermarket chain in Florida selected some employees to receive management training. A group of women employees claimed that males were picked at a disproportionally high rate for such training. The company denied this claim.[1] A similar claim of gender bias was made about promotions and pay for women who work for Wal-Mart.[2] How could the women employees statistically back up their assertion?

Suppose the employee pool for potential selection for management training is half male and half female. Then the company's claim of a lack of gender bias is a hypothesis. It states that, other things being equal, at each choice the probability of selecting a female equals 1/2 and the probability of selecting a male equals 1/2. If the

[1] *Tampa Tribune*, April 6, 1996.

[2] *New York Times*, February 7, 2007.

employees truly are selected for management training randomly in terms of gender, about half the employees picked should be females and about half should be male. The women's claim is an alternative hypothesis that the probability of selecting a male exceeds 1/2.

Suppose that nine of the ten employees chosen for management training were male. We might be inclined to believe the women's claim. However, we should analyze whether these results would be unlikely, if there were *no* gender bias. Would it be highly unusual that 9/10 of the employees chosen would have the same gender if they were truly selected at random from the employee pool? Due to sampling variation, not exactly 1/2 of the sample need be male. How far above 1/2 must the sample proportion of males chosen be before we believe the women's claim? ■

This chapter introduces statistical methods for summarizing evidence and making decisions about hypotheses. We first present the parts that all significance tests have in common. The rest of the chapter presents significance tests about population means and population proportions. We'll also learn how to find and how to control the probability of an incorrect decision about a hypothesis.

6.1 THE FIVE PARTS OF A SIGNIFICANCE TEST

Now let's take a closer look at the significance test method, also called a *hypothesis test*, or *test* for short. All tests have five parts: assumptions, hypotheses, test statistic, *P*-value, and conclusion.

Assumptions

Each test makes certain assumptions or has certain conditions for the test to be valid. These pertain to the following:

- ***Type of data***: Like other statistical methods, each test applies for either quantitative data or categorical data.
- ***Randomization***: Like the confidence interval method of statistical inference, a test assumes that the data were obtained using randomization, such as a random sample.
- ***Population distribution***: For some tests, the variable is assumed to have a particular distribution, such as the normal distribution.
- ***Sample size***: The validity of many tests improves as the sample size increases.

Hypotheses

Each significance test has two hypotheses about the value of a parameter.

Null Hypothesis, Alternative Hypothesis

The ***null hypothesis*** is a statement that the parameter takes a particular value. The ***alternative hypothesis*** states that the parameter falls in some alternative range of values. Usually the value in the null hypothesis corresponds, in a certain sense, to *no effect*. The values in the alternative hypothesis then represent an effect of some type.

Notation for Hypotheses

The symbol H_0 represents the null hypothesis. The symbol H_a represents the alternative hypothesis.

Consider Example 6.1 about possible gender discrimination in selecting management trainees. Let π denote the probability that any particular selection is a male. The company claims that $\pi = 1/2$. This is an example of a null hypothesis, *no effect* referring to a lack of gender bias. The alternative hypothesis reflects the skeptical women employees' belief that this probability actually exceeds 1/2. So the hypotheses are H_0: $\pi = 1/2$ and H_a: $\pi > 1/2$. Note that H_0 has a *single* value whereas H_a has a range of values.

A significance test analyzes the sample evidence about the null hypothesis, H_0. The test investigates whether the data contradict H_0, hence suggesting that H_a is true. The approach taken is the indirect one of *proof by contradiction*. The null hypothesis is presumed to be true. Under this presumption, if the data observed would be very unusual, the evidence supports the alternative hypothesis. In the study of potential gender discrimination, we presume the null hypothesis value of $\pi = 1/2$ is true. Then we determine if the sample result of selecting 9 men for management training in 10 choices would be unusual, under this presumption. If so, then we may be inclined to believe the women's claim. But if the difference between the sample proportion of men chosen (9/10) and the H_0 value of 1/2 could easily be due to ordinary sampling variability, there's not enough evidence to accept the women's claim.

A researcher usually conducts a test to gauge the amount of support for the alternative hypothesis. Thus, H_a is sometimes called the ***research hypothesis***. The hypotheses are formulated *before* collecting or analyzing the data.

Test Statistic

The parameter to which the hypotheses refer has a point estimate. The ***test statistic*** summarizes how far that estimate falls from the parameter value in H_0. Often this is expressed by the number of standard errors between the estimate and the H_0 value.

P-Value

To interpret a test statistic value, we create a probability summary of the evidence against H_0. This uses the sampling distribution of the test statistic, under the presumption that H_0 is true. The purpose is to summarize how unusual the observed test statistic value is compared to what H_0 predicts.

Specifically, if the test statistic falls well out in a tail of the sampling distribution in a direction predicted by H_a, then it is far from what H_0 predicts. We can summarize how far out in the tail the test statistic falls by the tail probability of that value and of more extreme values. These are the possible test statistic values that provide *at least as much evidence against H_0 as the observed test statistic*, in the direction predicted by H_a. This probability is called the ***P-value***.

P-value

The ***P-value*** is the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_a. It is calculated by presuming that H_0 is true. The *P*-value is denoted by *P*.

A small *P*-value (such as $P = 0.01$) means that the data observed would have been unusual, if H_0 were true. *The smaller the P-value, the stronger the evidence is against H_0.*

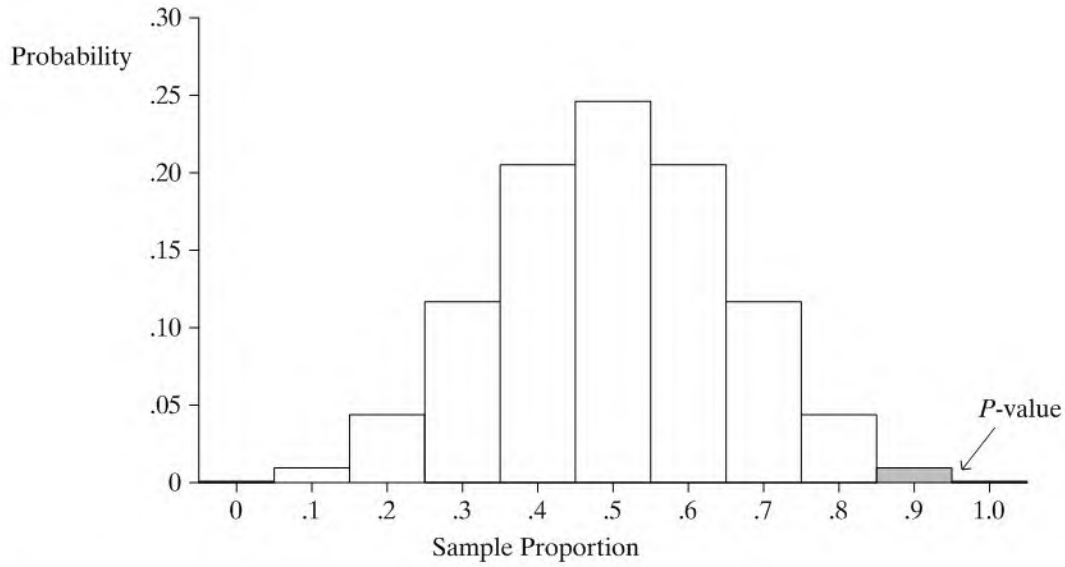

FIGURE 6.1: The *P*-Value Equals the Probability of the Observed Data or Even More Extreme Results. It is calculated under the presumption that H_0 is true, so a very small *P*-value gives strong evidence against H_0.

For Example 6.1 on potential gender discrimination in choosing managerial trainees, π is the probability of selecting a male. We test H_0: $\pi = 1/2$ against H_a: $\pi > 1/2$. One possible test statistic is the sample proportion of males selected, which is $9/10 = 0.90$. The values for the sample proportion that provide this much or even more extreme evidence against H_0: $\pi = 1/2$ and in favor of H_a: $\pi > 1/2$ are the right-tail sample proportion values of 0.90 and higher. See Figure 6.1. A formula from Section 6.7 calculates this probability as 0.01, so the *P*-value equals $P = 0.01$. If the selections truly were random with respect to gender, the probability is only 0.01 of such an extreme sample result, namely, that nine or all ten selections would be males. Other things being equal, this small *P*-value provides considerable evidence against H_0: $\pi = 1/2$ and supporting the alternative H_a: $\pi > 1/2$ of discrimination against females.

By contrast, a moderate to large *P*-value means the data are consistent with H_0. A *P*-value such as 0.26 or 0.83 indicates that, if H_0 were true, the observed data would not be unusual.

Conclusion

The *P*-value summarizes the evidence against H_0. Our conclusion should also *interpret* what the *P*-value tells us about the question motivating the test. Sometimes it is necessary to make a decision about the validity of H_0. If the *P*-value is sufficiently small, we reject H_0 and accept H_a.

Most studies require very small *P*-values, such as $P \leq 0.05$, in order to reject H_0. In such cases, results are said to be *significant at the 0.05 level*. This means that if H_0 were true, the chance of getting such extreme results as in the sample data would be no greater than 0.05.

Making a decision by rejecting or not rejecting a null hypothesis is an optional part of the significance test. We defer discussion of it until Section 6.4. Table 6.1 summarizes the parts of a significance test.

TABLE 6.1: The Five Parts of a Statistical Significance Test

1.	**Assumptions** Type of data, randomization, population distribution, sample size condition
2.	**Hypotheses** Null hypothesis, H_0 (parameter value for "no effect") Alternative hypothesis, H_a (alternative parameter values)
3.	**Test statistic** Compares point estimate to H_0 parameter value
4.	***P*-value** Weight of evidence against H_0; smaller P is stronger evidence
5.	**Conclusion** Report P-value Formal decision (optional; see Section 6.4)

6.2 SIGNIFICANCE TEST FOR A MEAN

For quantitative variables, significance tests usually refer to the population mean μ. The five parts of the significance test follow:

The Five Parts of a Significance Test for a Mean

***1.* Assumptions**

The test assumes the data are obtained using randomization, such as a random sample. The quantitative variable is assumed to have a normal population distribution. We'll see that this is mainly relevant for small sample sizes and certain types of H_a.

***2.* Hypotheses**

The null hypothesis about a population mean μ has the form

$$H_0 : \mu = \mu_0,$$

where μ_0 is a particular value for the population mean. In other words, the hypothesized value of μ in H_0 is a single value. This hypothesis usually refers to *no effect* or *no change* compared to some standard. For example, Example 5.5 in the previous chapter (page 120) estimated the population mean weight change μ for teenage girls after receiving a treatment for anorexia. The hypothesis that the treatment has *no effect* is a null hypothesis, H_0: $\mu = 0$. Here, the H_0 value μ_0 for the parameter μ is 0.

The alternative hypothesis contains alternative parameter values from the value in H_0. The most common alternative hypothesis is

$$H_a : \mu \neq \mu_0, \quad \text{such as} \quad H_a : \mu \neq 0.$$

This alternative hypothesis is called ***two sided***, because it contains values both below and above the value listed in H_0. For the anorexia study, H_a: $\mu \neq 0$ states that the treatment has *some effect*, the population mean equaling some value other than 0.

***3.* Test Statistic**

The sample mean $\bar{y}$ estimates the population mean μ. When the population distribution is normal, the sampling distribution of $\bar{y}$ is normal about μ. This is also approximately true when the population distribution is *not* normal but the random sample size is relatively large, by the Central Limit Theorem.

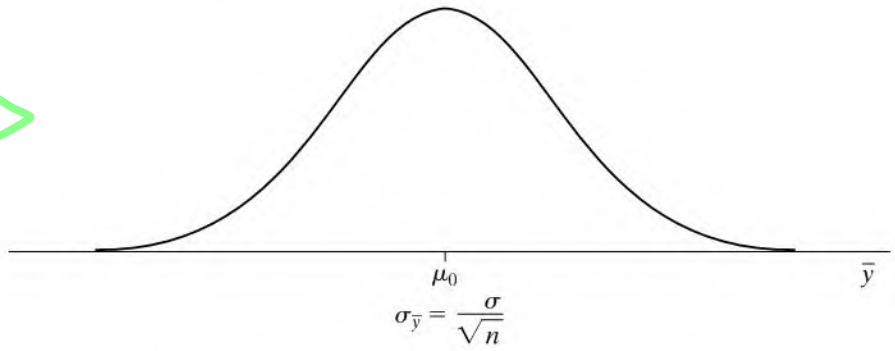

FIGURE 6.2: Sampling Distribution of $\bar{y}$ if H_0: $\mu = \mu_0$ Is True. For large random samples, it is approximately normal, centered at the null hypothesis value, μ_0.

Under the presumption that H_0: $\mu = \mu_0$ is true, the center of the sampling distribution of $\bar{y}$ is the value μ_0, as Figure 6.2 shows. A value of $\bar{y}$ that falls far out in the tail provides strong evidence against H_0, because it would be unusual if truly $\mu = \mu_0$. The evidence about H_0 is summarized by the number of standard errors that $\bar{y}$ falls from the null hypothesis value μ_0.

Recall that the *true* standard error is $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. As in Chapter 5, we substitute the sample standard deviation s for the unknown population standard deviation σ to get the *estimated* standard error, $se = s/\sqrt{n}$. The test statistic is the t-score

$$t = \frac{\bar{y} - \mu_0}{se}, \quad \text{where} \quad se = \frac{s}{\sqrt{n}}.$$

The farther $\bar{y}$ falls from μ_0, the larger the absolute value of the t test statistic. Hence, the larger the value of $|t|$, the stronger the evidence against H_0.

We use the symbol t rather than z because, as in forming a confidence interval, using s to estimate σ in the standard error introduces additional error. The null sampling distribution of the t test statistic is the *t distribution*, introduced in Section 5.3. It looks like the standard normal distribution, having mean equal to 0 but being more spread out, moreso for smaller n. It is specified by its degrees of freedom, $df = n - 1$.

4. *P*-Value

The test statistic summarizes how far the data fall from H_0. Different tests use different test statistics, though, and simpler interpretations result from transforming it to the probability scale of 0 to 1. The P-value does this.

We calculate the P-value under the presumption that H_0 is true. That is, we give the benefit of the doubt to H_0, analyzing how unusual the observed data would be if H_0 were true. The P-value is the probability that the test statistic equals the observed value or a value in the set of more extreme values that provide even stronger evidence against H_0. For H_a: $\mu \neq \mu_0$, the more extreme t values are the ones even farther out in the tails of the t distribution. So the P-value is the two-tail probability that the t test statistic is at least as large in absolute value as the observed test statistic. This is also the probability that $\bar{y}$ falls at least as far from μ_0 *in either direction* as the observed value of $\bar{y}$.

Figure 6.3 shows the sampling distribution of the t test statistic when H_0 is true. A test statistic value of $t = (\bar{y} - \mu_0)/se = 0$ results when $\bar{y} = \mu_0$. This is the t-value most consistent with H_0. The P-value is the probability of a t test statistic value at least as far from this consistent value as the one observed. To illustrate its calculation, suppose $t = 0.68$ for a sample size of 186. (This is the result in Example 6.2) This t-score means that the sample mean $\bar{y}$ falls 0.68 estimated standard errors above μ_0.

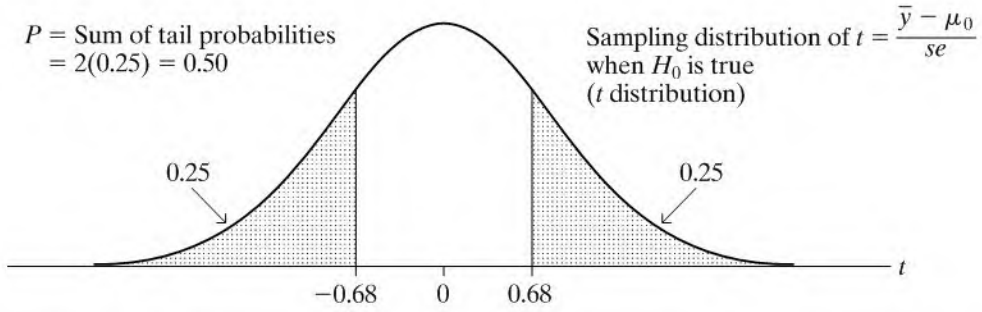

FIGURE 6.3: Calculation of *P*-Value when $t = 0.68$, for Testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$. The *P*-value is the two-tail probability of a more extreme result than the observed one.

The *P*-value is the probability that $t \geq 0.68$ or $t \leq -0.68$ (i.e., $|t| \geq 0.68$). Since $n = 186$, $df = n - 1 = 185$ is large, and the t distribution is nearly identical to the standard normal. From Table A, the probability in one tail above 0.68 is about 0.25, so the two-tail probability is about $P = 2(0.25) = 0.50$.

A more precise *P*-value calculation with the t distribution using software gives *P*-value = 0.4973545. Round such a value, say to 0.50, before reporting it. Reporting the *P*-value with many decimal places, such as 0.4973545, makes it seem as if more accuracy exists than actually does. In practice, the sampling distribution is only *approximately* the t distribution, because the population distribution is not exactly normal as is assumed with the t test.

5. Conclusion

Finally, the study should interpret the *P*-value in context. The smaller *P* is, the stronger the evidence against H_0 and in favor of H_a.

EXAMPLE 6.2 Political Conservatism and Liberalism

Some political commentators have remarked that citizens of the United States are increasingly conservative, so much so that many treat *liberal* as a dirty word. We can study political ideology by analyzing response to certain items on the GSS. For instance, that survey asks (with the POLVIEWS item) where you would place yourself on a seven-point scale of political views ranging from extremely liberal, point 1, to extremely conservative, point 7. Table 6.2 shows the scale and the distribution of responses among the levels for the 2006 survey. Results are shown separately according to the three categories for the variable labelled as RACE in the GSS.

TABLE 6.2: Responses of Subjects on a Scale of Political Ideology

	Race		
Response	Black	White	Other
1. Extremely liberal	10	36	1
2. Liberal	21	109	13
3. Slightly liberal	22	124	13
4. Moderate, middle of road	74	421	27
5. Slightly conservative	21	179	9
6. Conservative	27	176	7
7. Extremely conservative	11	28	2
	$n = 186$	$n = 1073$	$n = 72$

Political ideology is an ordinal scale. Often, we treat such scales in a quantitative manner by assigning scores to the categories. Then we can use quantitative summaries such as means, allowing us to detect the extent to which observations gravitate toward the conservative or the liberal end of the scale.

If we assign the category scores shown in Table 6.2, then a mean below 4 shows a propensity toward liberalism, and a mean above 4 shows a propensity toward conservatism. We can test whether these data show much evidence of either of these by conducting a significance test about how the population mean compares to the moderate value of 4. We'll do this here for the black sample and in Section 6.5 for the entire sample.

1. *Assumptions*: The sample is randomly selected. We are treating political ideology as quantitative with equally spaced scores. The t test assumes a normal population distribution for political ideology. We'll discuss this assumption further at the end of this section.
2. *Hypotheses*: Let μ denote the population mean ideology for black Americans, for this seven-point scale. The null hypothesis contains one specified value for μ. Since we conduct the analysis to check how, if at all, the population mean departs from the moderate response of 4, the null hypothesis is

$$H_0 : \mu = 4.0.$$

The alternative hypothesis is then

$$H_a : \mu \neq 4.0.$$

The null hypothesis states that, on the average, the population response is politically "moderate, middle of road." The alternative states that the mean falls in the liberal direction ($\mu < 4$) or in the conservative direction ($\mu > 4$).

3. *Test statistic*: The 186 observations in Table 6.2 for blacks are summarized by $\bar{y} = 4.075$ and $s = 1.512$. The estimated standard error of the sampling distribution of $\bar{y}$ is

$$se = \frac{s}{\sqrt{n}} = \frac{1.512}{\sqrt{186}} = 0.111.$$

The value of the test statistic is

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{4.075 - 4.0}{0.111} = 0.68.$$

The sample mean falls 0.68 estimated standard errors above the null hypothesis value of the mean. The df value is $186 - 1 = 185$.

4. *P-value*: The P-value is the two-tail probability, presuming H_0 is true, that t would exceed 0.68 in absolute value. From the t distribution with $df = 185$ (or its standard normal approximation), this two-tail probability is $P = 0.50$. If the population mean ideology were 4.0, then the probability equals 0.50 that a sample mean for $n = 186$ subjects would fall at least as far from 4.0 as the observed $\bar{y}$ of 4.075.
5. *Conclusion*: The P-value of $P = 0.50$ is not small, so it does not contradict H_0. If H_0 were true, the data we observed would not be unusual. It is plausible that the population mean response for black Americans in 2006 was 4.0, not leaning toward the conservative or liberal direction. ■

Correspondence between Two-Sided Tests and Confidence Intervals

Conclusions using two-sided significance tests are consistent with conclusions using confidence intervals. If a test says that a particular value is believable for the parameter, then so does a confidence interval.

EXAMPLE 6.3 Confidence Interval for Mean Political Ideology

For the data in Example 6.2, let's construct a 95% confidence interval for the population mean political ideology. With $df = 185$, the multiple of the standard error ($se = 0.111$) is $t_{.025} = 1.97$. Since $\bar{y} = 4.075$, the confidence interval is

$$\bar{y} \pm 1.97(se) = 4.075 \pm 1.97(0.111) = 4.075 \pm 0.219, \quad \text{or} \quad (3.9, 4.3).$$

At the 95% confidence level, these are the plausible values for μ.

This confidence interval indicates that μ may equal 4.0, since 4.0 falls inside the confidence interval. Thus, it is not surprising that the P-value ($P = 0.50$) in testing H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ in Example 6.2 was not small. In fact;

- Whenever the P-value $P > 0.05$ in a two-sided test, a 95% confidence interval for μ necessarily contains the H_0 value of μ.

By contrast, suppose the P-value $= 0.02$ in testing H_0: $\mu = 4.0$. Then a 95% confidence interval would tell us that 4.0 is implausible for μ, with 4.0 falling *outside* the confidence interval.

- Whenever $P \leq 0.05$ in a two-sided test, a 95% confidence interval for μ does not contain the H_0 value of μ.

Section 6.4 discusses further the connection between the two methods. ■

One-Sided Significance Tests

A different alternative hypothesis is sometimes used when a researcher predicts a deviation from H_0 in a particular direction. It has the form

$$H_a : \mu > \mu_0 \quad \text{or} \quad H_a : \mu < \mu_0.$$

The alternative H_a: $\mu > \mu_0$ is used to detect whether μ is *larger* than the particular value μ_0, whereas H_a: $\mu < \mu_0$ is used to detect whether μ is *smaller* than that value. These hypotheses are called ***one sided***. By contrast, the *two-sided* H_a is used to detect *any* type of deviation from H_0. This choice is made before analyzing the data.

For H_a: $\mu > \mu_0$, the P-value is the probability (presuming H_0 is true) of a t-score *above* the observed t-score; that is, to the right of it on the real number line. These t-scores provide more extreme evidence than the observed value in favor of H_a: $\mu > \mu_0$. So P equals the right-tail probability under the t curve, as Figure 6.4 portrays. A t-score of 0.68 results in $P = 0.25$ for this alternative.

For H_a: $\mu < \mu_0$, the P-value is the left-tail probability, (*below*) the observed t-score. A t-score of $t = -0.68$ results in $P = 0.25$ for this alternative. A t-score of 0.68 results in $P = 1 - 0.25 = 0.75$.

EXAMPLE 6.4 Mean Weight Change in Anorexic Girls

Example 5.5 in Chapter 5 (page 120) analyzed data from a study comparing treatments for teenage girls suffering from anorexia. For each girl, the study observed her change in weight while receiving the therapy. Let μ denote the population mean change in

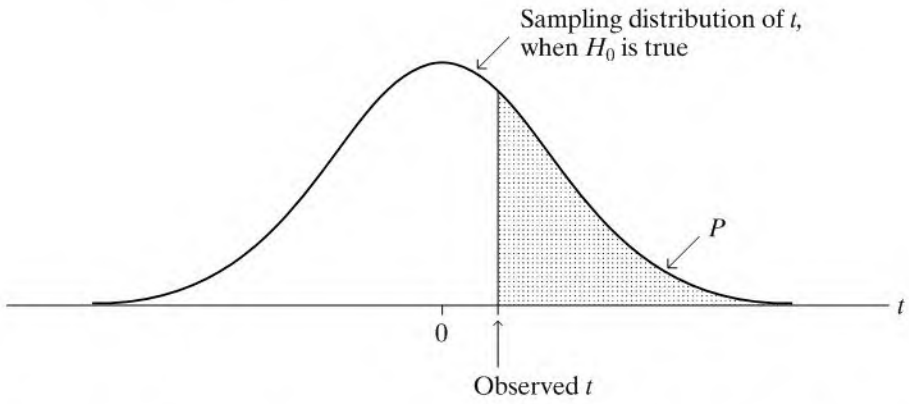

FIGURE 6.4: Calculation of P-value in Testing H_0: $\mu = \mu_0$ against H_a: $\mu > \mu_0$. The P-value is the probability of values to the right of the observed test statistic.

weight for the cognitive behavioral treatment. If this treatment has beneficial effect, as expected, then μ is positive. To test for no treatment effect versus a positive mean weight change, we test H_0: $\mu = 0$ against H_a: $\mu > 0$.

Software (SPSS) used to analyze the data reports:

Variable	Number of Cases	Mean	SD	SE of Mean
CHANGE	29	3.007	7.309	1.357

For the $n = 29$ girls, the sample mean weight change of 3.007 pounds had an estimated standard error of $se = 1.357$. The test statistic equals

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{3.007 - 0}{1.357} = 2.22.$$

For this one-sided H_a, the P-value is the right-tail probability above 2.22. Why do we use the right tail? Because H_a: $\mu > 0$ has values *above* (that is, to the right of) the null hypothesis value of 0. It's the positive values of t that support this alternative hypothesis.

Now, for $n = 29$, $df = n - 1 = 28$. From Table B, $t = 2.048$ yields $P = 0.025$ for the one-sided H_a and $t = 2.467$ yields $P = 0.01$. The observed $t = 2.22$ is between 2.048 and 2.467, so the P-value is between 0.01 and 0.025. Figure 6.5 illustrates. Table B is not detailed enough to provide the exact P-value. When software performs the analysis, the output reports the actual P-value rather than bounds for it. Most software reports the P-value for a two-sided alternative as the

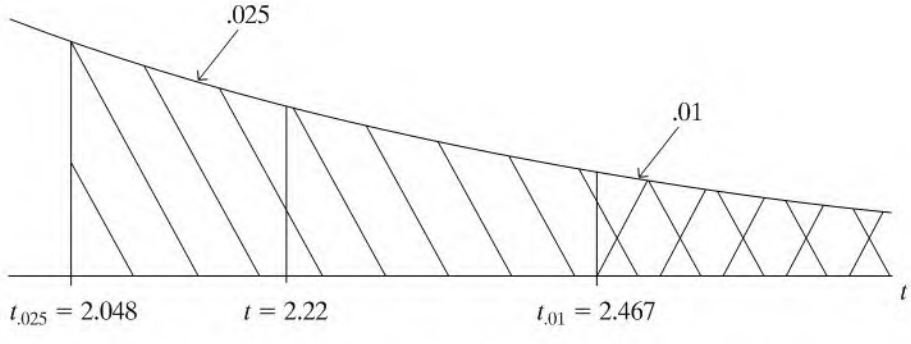

FIGURE 6.5: For $df = 28$, $t = 2.22$ Has a Tail Probability between 0.01 and 0.025

default, unless you request otherwise. SPSS reports results for the two-sided test and confidence interval as

```
Mean        95% CI
          Lower    Upper      t-value    df    2-Tail Sig
3.01      .227     5.787       2.22      28     .035
```

SPSS reports $P = 0.035$ (under "2-Tail Sig"). The one-sided P-value is half this, or about $P = 0.02$. There is relatively strong evidence against H_0. It seems that the treatment has an effect.

The significance test concludes that the mean weight gain was not equal to 0. But the 95% confidence interval of (0.2, 5.8), shown also in the SPSS output, is more informative. It shows just how different from 0 the population mean change is likely to be. The effect could be very small. Also, keep in mind that this experimental study (like many medically oriented studies) had to use a volunteer sample. So these results are highly tentative, another reason that it is silly for studies like this to report P-values to several decimal places. ■

Implicit One-Sided H_0 for One-Sided H_a

From Example 6.4, the one-sided P-value = 0.017. So, if $\mu = 0$, the probability equals 0.017 of observing a sample mean weight gain of 3.01 or greater. Now, suppose $\mu < 0$; that is, the population mean weight change is negative. Then the probability of observing $\bar{y} \geq 3.01$ would be even smaller than 0.017. For example, a sample value of $\bar{y} = 3.01$ is even less likely when $\mu = -5$ than when $\mu = 0$, since 3.01 is farther out in the tail of the sampling distribution of $\bar{y}$ when $\mu = -5$ than when $\mu = 0$. Thus, rejection of H_0: $\mu = 0$ in favor of H_a: $\mu > 0$ also inherently rejects the broader null hypothesis of H_0: $\mu \leq 0$. In other words, one concludes that $\mu = 0$ is false *and* that $\mu < 0$ is false.

The Choice of One-Sided versus Two-Sided Tests

In practice, two-sided tests are more common than one-sided tests. Even if a researcher predicts the direction of an effect, two-sided tests can also detect an effect that falls in the opposite direction. In most research articles, significance tests use two-sided P-values. Partly this reflects an objective approach to research that recognizes that an effect could go in either direction. In using two-sided P-values, researchers avoid the suspicion that they chose H_a when they saw the direction in which the data occurred. That is not ethical.

Two-sided tests coincide with the usual approach in estimation. Confidence intervals are two-sided, obtained by adding and subtracting some quantity from the point estimate. One can form one-sided confidence intervals; for instance, concluding that a population mean is *at least* equal to 7 (i.e., between 7 and ∞). In practice, though, one-sided intervals are rarely used.

In deciding whether to use a one-sided or a two-sided H_a in a particular exercise or in practice, consider the context. An exercise that says "Test whether the mean has *changed*" suggests a two-sided alternative, to allow for increase or decrease. "Test whether the mean has *increased*" suggests the one-sided H_a: $\mu > \mu_0$.

In either the one-sided or two-sided case, hypotheses always refer to population parameters, not sample statistics. So *never* express a hypothesis using sample statistic notation, such as H_0: $\bar{y} = 0$. There is no uncertainty or need to conduct statistical inference about sample statistics such as $\bar{y}$, because we can calculate their values exactly from the data.

The α-Level: Using the *P*-Value to Make a Decision

A significance test analyzes the strength of the evidence against the null hypothesis, H_0. We start by presuming that H_0 is true. We analyze whether the data would be unusual if H_0 were true by finding the P-value. If the P-value is small, the data contradict H_0 and support H_a. Generally, researchers do not regard the evidence against H_0 as strong unless P is very small, say, $P < 0.05$ or $P < 0.01$.

Why do smaller P-values indicate stronger evidence against H_0? Because the data would then be more unusual if H_0 were true. When H_0 is true, the P-value is roughly equally likely to fall anywhere between 0 and 1. By contrast, when H_0 is false, the P-value is more likely to be near 0 than near 1.

In practice, it is sometimes necessary to decide whether the evidence against H_0 is strong enough to reject it. The decision is based on whether the P-value falls below a prespecified cutoff point. It's most common to reject H_0 if $P \leq 0.05$ and conclude that the evidence is not strong enough to reject H_0 if $P > 0.05$. The boundary value 0.05 is called the ***α-level*** of the test.

α-Level

The ***α-level*** is a number such that we reject H_0 if the P-value is less than or equal to it. The α-level is also called the ***significance level***. In practice, the most common α-levels are 0.05 and 0.01.

Like the choice of a confidence level for a confidence interval, the choice of α reflects how cautious you want to be. The smaller the α-level, the stronger the evidence must be to reject H_0. To avoid bias in the decision-making process, you select α *before* analyzing the data.

EXAMPLE 6.5 Adding Decisions to Previous Examples

Let's use $\alpha = 0.05$ to guide us in making a decision about H_0 for the examples of this section. Example 6.2 (page 149) tested H_0: $\mu = 4.0$ about mean political ideology. With sample mean $\bar{y} = 4.075$, the P-value was 0.50. The P-value is not small, so if truly $\mu = 4.0$, it would not be unusual to observe $\bar{y} = 4.075$. Since $P = 0.50 > 0.05$, there is insufficient evidence to reject H_0. It is believeable that the population mean ideology was 4.0.

Example 6.4 tested H_0: $\mu = 0$ about the mean weight gain for teenage girls suffering from anorexia. The P-value was 0.017. Since $P = 0.017 < 0.05$, there is sufficient evidence to reject H_0 in favor of H_a: $\mu > 0$. We conclude that the treatment results in an increase in mean weight. Such a conclusion is sometimes phrased as, "The increase in the mean weight is *statistically significant* at the 0.05 level." Since $P = 0.017$ is *not* less than 0.010, the result is *not* significant at the 0.010 level. In fact, *the P-value is the smallest level for α at which the results are significant*. So, with P-value $= 0.017$, we reject H_0 if $\alpha = 0.02$ or 0.05 or 0.10, but not if $\alpha = 0.015$ or 0.010 or 0.001. ■

Table 6.3 summarizes significance tests for population means.

Robustness for Violations of Normality Assumption

The t test for a mean assumes that the population distribution is normal. This ensures that the sampling distribution of the sample mean $\bar{y}$ is normal (even for small n) and, after using s to estimate σ in finding the se, the t test statistic has the t distribution. As the sample size increases, this assumption of a normal population becomes less

TABLE 6.3: The Five Parts of Significance Tests for Population Means

1. **Assumptions**
   Quantitative variable
   Randomization
   Normal population (robust, especially for two-sided H_a, large n)
2. **Hypotheses**
   H_0: $\mu = \mu_0$
   H_a: $\mu \neq \mu_0$ (or H_a: $\mu > \mu_0$ or H_a: $\mu < \mu_0$)
3. **Test statistic**
   $t = \dfrac{\bar{y} - \mu_0}{se}$ where $se = \dfrac{s}{\sqrt{n}}$
4. ***P*-value**
   In t curve, use
   P = Two-tail probability for H_a: $\mu \neq \mu_0$
   P = Probability to right of observed t-value for H_a: $\mu > \mu_0$
   P = Probability to left of observed t-value for H_a: $\mu < \mu_0$
5. **Conclusion**
   Report P-value. Smaller P provides stronger evidence against H_0 and supporting H_a. Can reject H_0 if $P \leq \alpha$-level.

important. We've seen that when n is roughly about 30 or higher, an approximate normal sampling distribution occurs for $\bar{y}$ regardless of the population distribution, by the Central Limit Theorem (Section 4.5).

From Section 5.3, a statistical method is ***robust*** if it performs adequately even when an assumption is violated. Statisticians have shown that *two-sided* inferences for a mean using the t distribution are robust against violations of the normal population assumption. Even if the population is not normal, two-sided t tests and confidence intervals still work quite well. The test does not work so well for a one-sided test with small n when the population distribution is highly skewed.

Figure 6.6 shows a histogram and a box plot of the data from the anorexia study of Example 6.4 (page 151). Figure 6.6 suggests skew to the right. The box plot highlights (as outliers) six girls who had considerable weight gains. As just mentioned, a two-sided t test works quite well even if the population distribution is skewed. However, this plot makes us wary about using a one-sided test, since the sample size is not large ($n = 29$). Given this and the discussion in the previous subsection about one-sided versus two-sided tests, we're safest with that study to report a two-sided P-value of

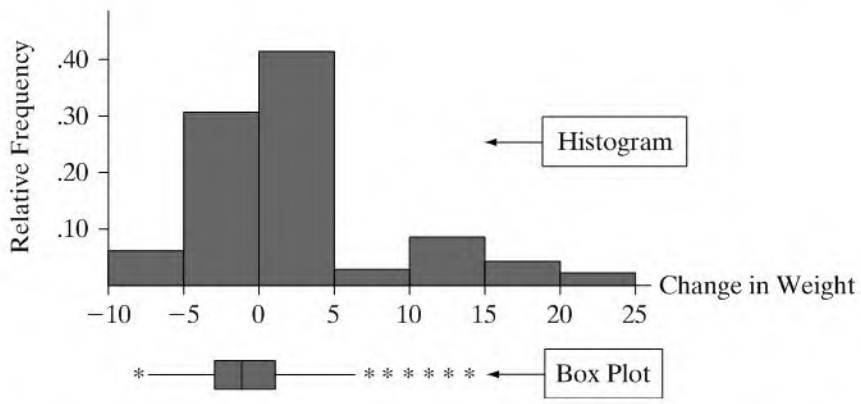

FIGURE 6.6: Histogram and Box Plot of Weight Change for Anorexia Sufferers

0.035. Also, as Example 5.5 (page 120) noted, the median may be a more relevant summary for these data.

6.3 SIGNIFICANCE TEST FOR A PROPORTION

For a categorical variable, the parameter is the population proportion for a category. For example, a significance test could analyze whether a majority of the population support embryonic stem-cell research by testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$, where π is the population proportion π supporting it. The test for a proportion, like the test for a mean, finds a P-value for a test statistic measuring the number of standard errors a point estimate falls from a H_0 value.

The Five Parts of a Significance Test for a Proportion

***1.* Assumptions**

Like other tests, this test assumes the data are obtained using randomization, such as a random sample. The sample size must be sufficiently large that the sampling distribution of $\hat{\pi}$ is approximately normal. For the most common case, in which the H_0 value of π is 0.50, a sample size of at least 20 is sufficient. We'll give a precise guideline in Section 6.7, which presents a small-sample test.

***2.* Hypotheses**

The null hypothesis of a test about a population proportion has form

$$H_0 : \pi = \pi_0, \quad \text{such as} \quad H_0 : \pi = 0.50.$$

Here, π_0 denotes a particular proportion value between 0 and 1, such as 0.50. The most common alternative hypothesis is

$$H_a : \pi \neq \pi_0, \quad \text{such as} \quad H_a : \pi \neq 0.50.$$

This *two-sided* alternative states that the population proportion differs from the value in H_0. The *one-sided* alternatives

$$H_a : \pi > \pi_0 \quad \text{and} \quad H_a : \pi < \pi_0$$

apply when the researcher predicts a deviation in a certain direction from the H_0 value.

***3.* Test Statistic**

From Section 5.2, the sampling distribution of the sample proportion $\hat{\pi}$ has mean π and standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$. When H_0 is true, $\pi = \pi_0$, so the standard error is $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$. We use the notation se_0 to indicate that this is the standard error under the presumption that H_0 is true.

The test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{se_0}, \quad \text{where} \quad se_0 = \sqrt{\pi_0(1 - \pi_0)/n}.$$

This measures the number of standard errors that the sample proportion $\hat{\pi}$ falls from π_0. For large samples, if H_0 is true, the sampling distribution of the z test statistic is the standard normal distribution.

The test statistic has a similar form as in tests for a mean:

Form of Test Statistic

$$z = \frac{\text{Estimate of parameter} - \text{null hypothesis value of parameter}}{\text{Standard error of estimator}}$$

Here, the estimate $\hat{\pi}$ of the proportion replaces the estimate $\bar{y}$ of the mean, and the null hypothesis proportion π_0 replaces the null hypothesis mean μ_0.

4. P-Value

The *P*-Value is a one- or two-tail probability, as in tests for a mean, except using the normal rather than the *t* distribution. For H_a: $\pi \neq \pi_0$, P is the two-tail probability. See Figure 6.7. This probability is double the single-tail probability beyond the observed z-value.

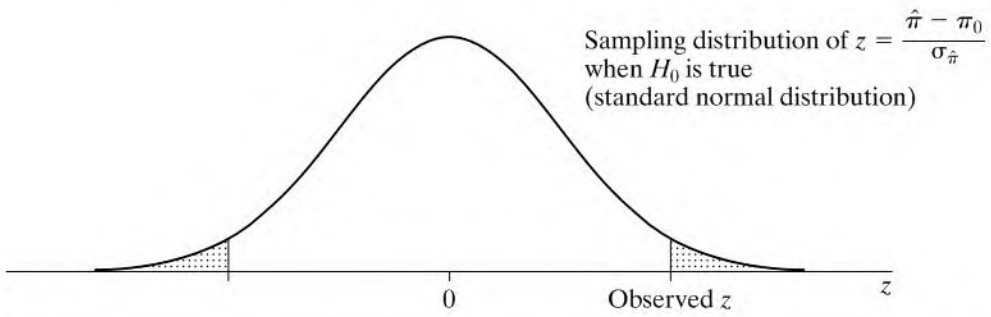

FIGURE 6.7: Calculation of *P*-Value in Testing H_0: $\pi = \pi_0$ against H_a: $\pi \neq \pi_0$. The two-sided alternative hypothesis uses a two-tail probability.

For one-sided alternatives, the *P*-value is a one-tail probability. Since H_a: $\pi > \pi_0$ predicts that the population proportion is *larger* than the H_0 value, its *P*-value is the probability *above* (i.e., to the right) of the observed z-value. For H_a : $\pi < \pi_0$, the *P*-value is the probability *below* (i.e., to the left) of the observed z-value.

5. Conclusion

As usual, the smaller the *P*-value, the more strongly the data contradict H_0 and support H_a. When we need to make a decision, we reject H_0 if $P \leq \alpha$ for a prespecified α-level such as 0.05.

EXAMPLE 6.6 Reduce Services or Raise Taxes?

These days, whether at the local, state, or national level, government often faces the problem of not having enough money to pay for the various services that it provides. One way to deal with this problem is to raise taxes. Another way is to reduce services. Which would you prefer? When the Florida Poll[3] asked a random sample of 1200 Floridians in 2006, 52% said raise taxes and 48% said reduce services.

Let π denote the population proportion in Florida who would choose raising taxes. If $\pi < 0.50$, this is a minority of the population, whereas if $\pi > 0.50$, it is a majority.

[3]www.fiu.edu/orgs/ipor/ffp.

To analyze whether π is in either of these ranges, we test H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$.

The estimate of π is $\hat{\pi} = 0.52$. Presuming H_0: $\pi = 0.50$ is true, the standard error of $\hat{\pi}$ is

$$se_0 = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}} = \sqrt{\frac{(0.50)(0.50)}{1200}} = 0.0144.$$

The value of the test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{se_0} = \frac{0.52 - 0.50}{0.0144} = 1.39.$$

From Table A, the two-tail P-value is $P = 2(0.0823) = 0.16$. If H_0 is true (i.e., if $\pi = 0.50$), the probability equals 0.16 that sample results would be as extreme in one direction or the other as in this sample.

This P-value is not small, so there is not much evidence against H_0. It seems believable that $\pi = 0.50$. With an α-level such as 0.05, since $P = 0.16 > 0.05$, we would not reject H_0. We cannot determine whether those favoring raising taxes are a majority or minority of the population. ■

In the standard error formula, $\sqrt{\pi(1 - \pi)/n}$, note we substituted the null hypothesis value $\pi_0 = 0.50$ for the population proportion π. The parameter values in sampling distributions for tests presume that H_0 is true, since the P-value is based on that presumption. This is why, for tests, we use $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$ rather than the estimated standard error, $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$. With the estimated se, the normal approximation for the sampling distribution of z is poorer. This is especially true for proportions close to 0 or 1. The validity of the P-value is then poorer. By contrast, the confidence interval method does not have a hypothesized value for π, so that method uses the estimated se rather than a H_0 value.

Never "Accept H_0"

In Example 6.6, about raising taxes or reducing services, the P-value of 0.16 was not small. So H_0: $\pi = 0.50$ is plausible. In this case, the conclusion is sometimes reported as, "Do not reject H_0," since the data do not contradict H_0.

We say "Do not reject H_0" rather than "Accept H_0." The population proportion has many plausible values besides the number in H_0. For instance, a 95% confidence interval for the population proportion π who support raising taxes rather than reducing services is

$$\hat{\pi} \pm 1.96\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} = 0.52 \pm 1.96\sqrt{\frac{(0.52)(0.48)}{1200}}, \text{ or } (0.49, 0.55).$$

This interval shows a range of plausible values for π. Even though insufficient evidence exists to conclude that $\pi \neq 0.50$, it is improper to conclude that $\pi = 0.50$.

In summary, H_0 contains a single value for the parameter. When the P-value is larger than the α-level, saying "Do not reject H_0" instead of "Accept H_0" emphasizes that that value is merely one of many believable values. Because of sampling variability, there is a range of believable values, so we can never accept H_0. The

reason "accept H_a" terminology is permissible for H_a is that when the P-value is sufficiently small, the entire range of believable values for the parameter fall within the range of values H_a specifies.

Effect of Sample Size on *P*-Values

In Example 6.6, on raising taxes or cutting services, suppose $\hat{\pi} = 0.52$ had been based on $n = 4800$ instead of $n = 1200$. The standard error then decreases to 0.0072 (half as large), and you can verify that the test statistic $z = 2.77$. This has two-sided P-value $= 0.006$. That P-value provides strong evidence against H_0: $\pi = 0.50$ and suggests that a majority support raising taxes rather than cutting services. In that case, though, the 95% confidence interval for π equals (0.506, 0.534). This indicates that π is quite close to 0.50 in practical terms.

A given difference between an estimate and the H_0 value has a smaller P-value as the sample size increases. The larger the sample size, the more certain we can be that sample deviations from H_0 are indicative of true population deviations. In particular, notice that even a small difference between $\hat{\pi}$ and π_0 (or between $\bar{y}$ and μ_0) can yield a small P-value if the sample size is very large.

6.4 DECISIONS AND TYPES OF ERRORS IN TESTS

When we need to decide whether the evidence against H_0 is strong enough to reject it, we've seen that we reject H_0 if $P \leq \alpha$, for a prespecified α-level. Table 6.4 summarizes the two possible conclusions for α-level $= 0.05$. The null hypothesis is either *rejected* or *not rejected*. If H_0 is rejected, then H_a is accepted. If H_0 is not rejected, then H_0 is plausible, but other parameter values are also plausible. Thus, H_0 is never *accepted*. In this case, results are inconclusive, and the test does not identify either hypothesis as more valid.

TABLE 6.4: Possible Decisions in a Significance Test with α-Level $= 0.05$

	Conclusion	
P-Value	H_0	H_a
$P \leq 0.05$	Reject	Accept
$P > 0.05$	Do not reject	Do not accept

It is better to report the P-value than to indicate merely whether the result is "statistically significant." Reporting the P-value has the advantage that the reader can tell whether the result is significant at any level. The P-values of 0.049 and 0.001 are both "significant at the 0.05 level," but the second case provides much stronger evidence than the first case. Likewise, P-values of 0.049 and 0.051 provide, in practical terms, the same amount of evidence about H_0. It is a bit artificial to call one result "significant" and the other "nonsignificant."

Type I and Type II Errors for Decisions

Because of sampling variability, decisions in tests always have some uncertainty. The decision could be erroneous. There are two types of potential errors, conventionally called *Type I* and *Type II* errors.

Type I and Type II Errors

When H_0 is true, a ***Type I error*** occurs if H_0 is rejected.
When H_0 is false, a ***Type II error*** occurs if H_0 is not rejected.

There are four possible results. These refer to the two possible decisions cross-classified with the two possibilities for whether H_0 is true. See Table 6.5.

TABLE 6.5: The Four Possible Results of Making a Decision in a Test. Type I and Type II Errors Are the Two Possible Incorrect Decisions

		Decision	
		Reject H_0	*Do not reject* H_0
Condition of H_0	H_0 true	Type I error	Correct decision
	H_0 false	Correct decision	Type II error

Rejection Regions

The collection of test statistic values for which the test rejects H_0 is called the ***rejection region***. For example, the rejection region for a test of level $\alpha = 0.05$ is the set of test statistic values for which $P \leq 0.05$.

For two-sided tests about a proportion, the two-tail P-value is ≤ 0.05 whenever the test statistic $|z| \geq 1.96$. In other words, the rejection region consists of values of z resulting from the estimate falling at least 1.96 standard errors from the H_0 value.

The α-Level Is the Probability of Type I Error

When H_0 is true, let's find the probability of Type I error. Suppose $\alpha = 0.05$. We've just seen that for the two-sided test about a proportion, the rejection region is $|z| \geq 1.96$. So the probability of rejecting H_0 is exactly 0.05, because the probability of the values in this rejection region under the standard normal curve is 0.05. But this is precisely the α-level.

The probability of a Type I error is the α-level for the test.

With $\alpha = 0.05$, if H_0 is true, the probability equals 0.05 of making a Type I error and rejecting that (true) H_0. We control P (Type I error) by the choice of α. The more serious the consequences of a Type I error, the smaller α should be. In practice, $\alpha = 0.05$ is most common, just as an error probability of 0.05 is most common with confidence intervals (that is, 95% confidence). However, this may be too high when a decision has serious implications.

For example, consider a criminal legal trial of a defendant. Let H_0 represent innocence and H_a represent guilt. The jury rejects H_0 and judges the defendant to be guilty if it decides the evidence is sufficient to convict. A Type I error, rejecting a true H_0, occurs in convicting a defendant who is actually innocent. In a murder trial, suppose a convicted defendant gets the death penalty. Then, if a defendant is actually innocent, we would hope that the probability of conviction is much smaller than 0.05.

When we make a decision, we don't know whether we've made a Type I or Type II error, just as we don't know whether a particular confidence interval truly contains the parameter value. However, we can control the probability of an incorrect decision for either type of inference.

As *P* (Type I Error) Goes Down, *P* (Type II Error) Goes Up

In an ideal world, Type I or Type II errors would not occur. In practice, errors do happen. We've all read about defendants who were convicted but later determined to be innocent. When we make a decision, why don't we use an extremely small P(Type I error), such as $\alpha = 0.000001$? For instance, why don't we make it almost impossible to convict someone who is really innocent?

When we make α smaller in a significance test, we need a smaller P-value to reject H_0. It then becomes harder to reject H_0. But this means that it will also be harder even if H_0 is false. The stronger the evidence required to convict someone, the more likely we will fail to convict defendants who are actually guilty. In other words, the smaller we make P (Type I error), the larger P (Type II error) becomes; that is, the probability of failing to reject H_0 even though it is false.

If we tolerate only an extremely small P(Type I error), such as $\alpha = 0.000001$, the test may be unlikely to reject H_0 even if it is false—for instance, we may be unlikely to convict someone even if he or she is guilty. This reasoning reflects the fundamental relation:

- The smaller P(Type I error) is, the larger P(Type II error) is.

Section 6.6 shows that P(Type II error) depends on just how far the true parameter value falls from H_0. If the parameter is nearly equal to the value in H_0, P(Type II error) is relatively high. If it falls far from H_0, P(Type II error) is relatively low. The farther the parameter falls from the H_0 value, the less likely the sample is to result in a Type II error.

For a fixed P(Type I error), P(Type II error) depends also on the sample size n. The larger the sample size, the more likely we are to reject a false H_0. To keep both P(Type I error) and P(Type II error) at low levels, it may be necessary to use a very large sample size. The P(Type II error) may be quite large when the sample size is small, unless the parameter falls quite far from the H_0 value.

Except in Section 6.6, we shall not calculate P(Type II error), because such calculations are complex. In practice, making a decision requires setting only α, the P(Type I error).

Equivalence between Confidence Intervals and Test Decisions

We now elaborate on the equivalence between decisions from two-sided tests and conclusions from confidence intervals, first alluded to in Example 6.3 (page 151). Consider the large-sample test of

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_a : \mu \neq \mu_0.$$

When $P < 0.05$, H_0 is rejected at the $\alpha = 0.05$ level. This happens when the test statistic $t = (\bar{y} - \mu_0)/se$ is greater than about 1.96 in absolute value (when n is large), which means that $\bar{y}$ falls more than $1.96(se)$ from μ_0. But if this happens, then the 95% confidence interval for μ, namely, $\bar{y} \pm 1.96(se)$, does not contain the null hypothesis value μ_0. See Figure 6.8. These two inference procedures are consistent.

> In testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, suppose we reject H_0 at the 0.05 α-level. Then the 95% confidence interval for μ does not contain μ_0. The 95% confidence interval for μ consists of those μ_0 values for which we do not reject H_0: $\mu = \mu_0$ at the 0.05 α-level.

In Example 6.2, about mean political ideology, the P-value for testing H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ was $P = 0.50$. At the $\alpha = 0.05$ level, we do not reject H_0: $\mu = 4.0$.

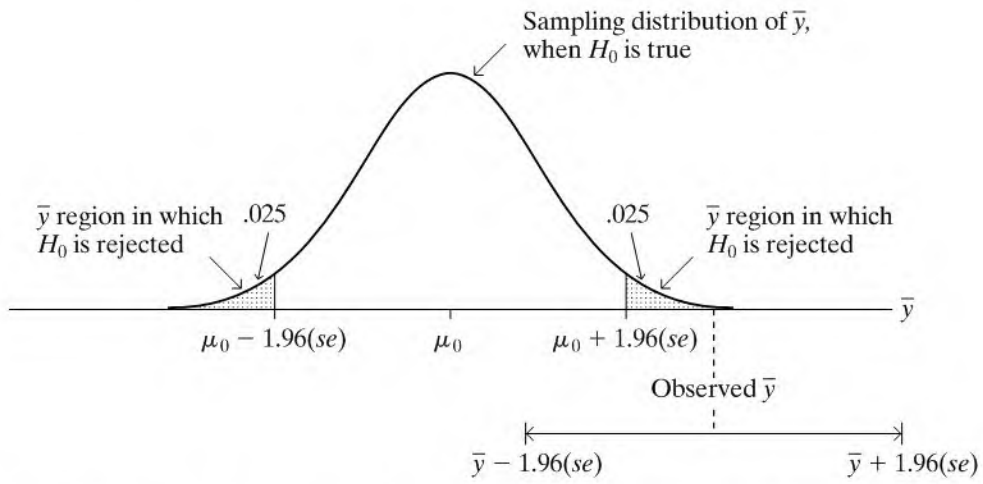

FIGURE 6.8: Relationship between Confidence Interval and Significance Test. The 95% confidence interval does not contain the H_0 value μ_0 when the sample mean falls more than 1.96 standard errors from μ_0, in which case the test statistic $|z| > 1.96$ and the P-value < 0.05.

It is believable that $\mu = 4.0$. Example 6.3 showed that a 95% confidence interval for μ is (3.9, 4.3), which contains $\mu_0 = 4.0$.

Rejecting H_0 at a particular α-level is equivalent to the confidence interval for μ with the same error probability not containing μ_0. For example, if a 99% confidence interval does not contain 0, then we would reject H_0: $\mu = 0$ in favor of H_a: $\mu \neq 0$ at the $\alpha = 0.01$ level with the test. The α-level is both P(Type I error) for the test and the probability that the confidence interval method does not contain the parameter.

Making Decisions versus Reporting the *P*-Value

The formal approach to hypothesis testing that this section has discussed was developed by the statisticians Jerzy Neyman and Egon Pearson in the late 1920s and early 1930s. In summary, this approach formulates null and alternative hypotheses, selects an α-level for the P(Type I error), determines the rejection region of test statistic values that provide enough evidence to reject H_0, and then makes a decision about whether to reject H_0 according to what is actually observed for the test statistic value. With this approach, it's not even necessary to find a P-value. The choice of α-level determines the rejection region, which together with the test statistic determines the decision.

The alternative approach of finding a P-value and using it to summarize evidence against a hypothesis is due to the great British statistician R. A. Fisher. He advocated merely reporting the P-value rather than using it to make a formal decision about H_0. Over time, this approach has gained favor, especially since software can now report precise P-values for a wide variety of significance tests.

This chapter has presented an amalgamation of the two approaches (the decision-based approach using an α-level and the P-value approach), so you can interpret a P-value yet also know how to use it to make a decision if that is needed. These days, most research articles merely report the P-value rather than a decision about whether

to reject H_0. From the P-value, readers can view the strength of evidence against H_0 and make their own decision, if they want to.

6.5 LIMITATIONS OF SIGNIFICANCE TESTS

A significance test makes an inference about whether a parameter differs from the H_0 value and about its direction from that value. In practice, we also want to know whether the parameter is sufficiently different from the H_0 value to be practically important. We'll see next that a test does not tell us as much as a confidence interval about practical importance.

Statistical Significance versus Practical Significance

It's important to distinguish between *statistical significance* and *practical significance*. A small P-value, such as $P = 0.001$, is highly statistically significant. It provides strong evidence against H_0. It does not, however, imply an *important* finding in any practical sense. The small P-value merely means that if H_0 were true, the observed data would be very unusual. It does not mean that the true parameter value is far from H_0 in practical terms.

EXAMPLE 6.7 Mean Political Ideology for All Americans

The mean political ideology of 4.08 in Example 6.2 refers to a sample of black Americans. The table also showed results for *white* and *other* categories. For a scoring of 1.0 through 7.0 for the ideology categories with 4.0 = moderate, the entire sample of 1331 observations has a mean of 4.12 and a standard deviation of 1.38. It appears that, on the average, conservatism was only slightly higher for the combined sample than for blacks alone (4.12 versus 4.08).

As in Example 6.2, we test H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ to analyze whether the population mean differs from the moderate ideology score of 4.0. Now, $se = s/\sqrt{n} = 1.38/\sqrt{1331} = 0.038$, and

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{4.12 - 4.0}{0.038} = 3.2.$$

The two-sided P-value is $P = 0.001$. There is *very* strong evidence that the true mean exceeds 4.0; that is, that the true mean falls on the conservative side of moderate. But on a scale of 1.0 to 7.0, 4.12 is close to the moderate score of 4.0. Although the difference of 0.12 between the sample mean of 4.12 and the H_0 mean of 4.0 is highly significant statistically, the magnitude of this difference is small in practical terms. The mean response on political ideology for all Americans is essentially a moderate one. ■

In Example 6.2, the sample mean ideology of 4.08 for $n = 186$ black Americans had $P = 0.50$, not much evidence against H_0. But if $\bar{y} = 4.08$ had been based on $n = 18{,}600$ (that is, 100 times as large as n was), again with $s = 1.51$, we would have instead found $z = 6.79$ and a two-sided P-value of $P = 0.00000000001$. This is highly statistically significant, but not practically significant. For practical purposes, a mean of 4.08 on a scale of 1.0 to 7.0 for political ideology does not differ from 4.00.

We've seen that, with large sample sizes, P-values can be small even when the point estimate falls near the H_0-value. The size of P merely summarizes the extent of evidence about H_0, not how far the parameter falls from H_0. Always inspect the difference between the estimate and the H_0-value to gauge the practical implications of a test result.

Significance Tests Are Less Useful than Confidence Intervals

Null hypotheses containing single values are rarely true. That is, rarely is the parameter *exactly* equal to the value listed in H_0. With sufficiently large samples, so that a Type II error is unlikely, these hypotheses will normally be rejected. What is more relevant is whether the parameter is sufficiently different from the H_0-value to be of practical importance.

Although significance tests can be useful, most statisticians believe they have been overemphasized in social science research. It is preferable to construct confidence intervals for parameters instead of performing only significance tests. A test merely indicates whether the particular value in H_0 is plausible. It does not tell us which other potential values are plausible. The confidence interval, by contrast, displays the entire set of believable values. It shows the extent to which H_0 may be false by showing whether the values in the interval are far from the H_0-value. Thus, it helps us to determine whether rejection of H_0 has practical importance.

To illustrate, for the political ideology data in the previous example, a 95% confidence interval for μ is $\bar{y} \pm 1.96(se) = 4.12 \pm 1.96(0.038)$, or (4.05, 4.20). This indicates that the difference between the population mean and the moderate score of 4.0 is small. Although the P-value of $P = 0.001$ provides very strong evidence against H_0: $\mu = 4.0$, in practical terms the confidence interval shows that H_0 is not wrong by much. By contrast, if $\bar{y}$ had been 6.125 (instead of 4.125), the 95% confidence interval would equal (6.05, 6.20). This indicates a substantial difference from 4.0, the mean response being near the conservative score rather than the moderate score.

When a P-value is not small but the confidence interval is quite wide, this forces us to realize that the parameter might well fall far from H_0 even though we cannot reject it. This also supports why it does not make sense to "accept H_0," as Section 6.3 discussed.

The remainder of the text presents significance tests for a variety of situations. It is important to become familiar with these tests, if for no other reason than their frequent use in social science research. However, we'll also introduce confidence intervals that describe how far reality is from the H_0-value.

Misinterpretations of Significance Tests and *P*-Values

We've seen it is improper to "accept H_0." We've also seen that statistical significance does not imply practical significance. Here's some other possible misinterpretations of significance tests:

- **It is misleading to report results only if they are statistically significant.** Some research journals have the policy of publishing results of a study only if the P-value ≤ 0.05. Here's a danger of this policy: Suppose there truly is no effect, but 20 researchers independently conduct studies. We would expect about $20(0.05) = 1$ of them to obtain significance at the 0.05 level merely by chance. (When H_0 is true, about 5% of the time we get a P-value below 0.05 anyway.) If that researcher then submits results to a journal but the other 19 researchers do not, the article published will be a Type I error. It will report an effect when there really is not one.

- **Some tests may be statistically significant just by chance.** You should never scan software output for results that are statistically significant and report only those. If you run 100 tests, even if all the null hypotheses are correct, you would expect to get P-values ≤ 0.05 about $100(0.05) = 5$ times. Be skeptical of reports of significance that might merely reflect ordinary random variability.
- **It is incorrect to interpret the P-value as the probability that H_0 is true.** The P-value is P(test statistic takes value like observed or even more extreme), presuming that H_0 is true. It is not $P(H_0$ true). Classical statistical methods calculate probabilities about variables and statistics (such as test statistics) that vary randomly from sample to sample, not about parameters. Statistics have sampling distributions; parameters do not. In reality, H_0 is not a matter of probability. It is either true or not true. We just don't know which is the case.
- **True effects may be smaller than reported estimates.** Even if a statistically significant result is a real effect, the true effect may be smaller than reported. For example, often several researchers perform similar studies, but the results that get attention are the most extreme ones. The researcher who decides to publicize the result may be the one who got the most impressive sample result, perhaps way out in the tail of the sampling distribution of all the possible results. See Figure 6.9.

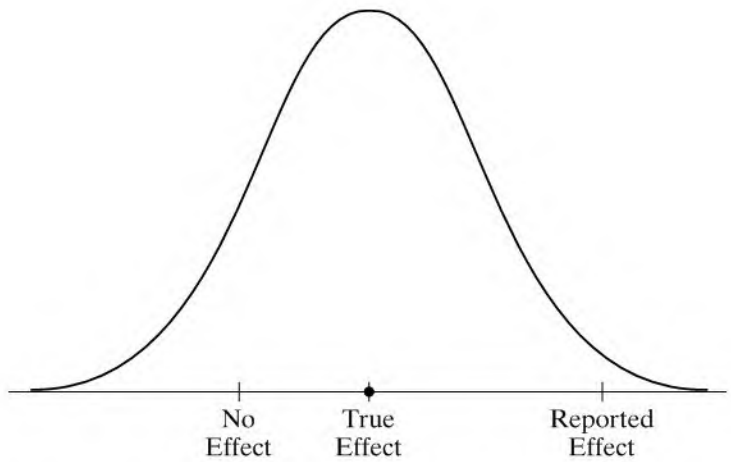

FIGURE 6.9: When Many Researchers Conduct Studies, the Statistically Significant Result Published Often Overestimates the True Effect

EXAMPLE 6.8 Are Many Medical "Discoveries" Actually Type I Errors?

In medical studies, suppose that a true effect exists only 10% of the time. Suppose also that when an effect truly exists, there's a 50% chance of making a Type II error and failing to detect it. These were the hypothetical percentages used in an article in a medical journal.[4] The authors noted that many medical studies have a high Type II error rate because they are not able to use a large sample size. Assuming these rates, could a substantial percentage of medical "discoveries" actually be Type I errors?

Figure 6.10 is a ***tree diagram*** showing what we'd would expect with 1000 medical studies that test various hypotheses. If a true effect exists only 10% of the time; this would be the case for 100 of the 1000 studies. We do not get a small enough P-value to

[4]J. Sterne, G. Smith, and D. R. Cox, *British Medical Journal*, vol. 322, 2001, pp. 226–231.

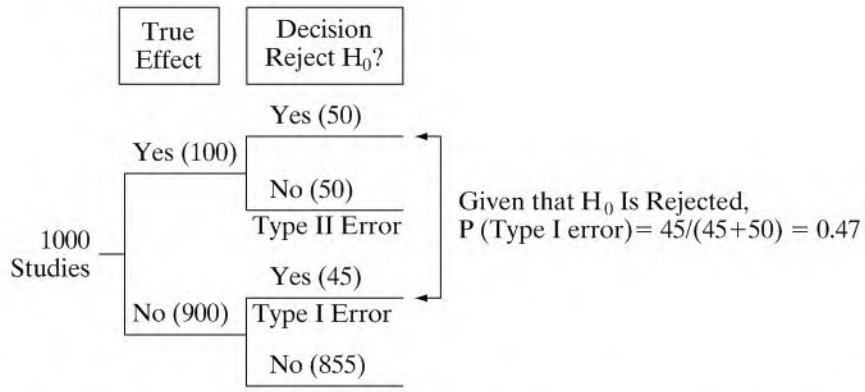

FIGURE 6.10: Tree Diagram of 1000 Hypothetical Medical Studies. This assumes a true effect exists 10% of the time and a 50% chance of a Type II error when an effect truly exists.

detect this true effect 50% of the time, that is, in 50 of these 100 studies. An effect will be reported for the other 50 of the 100 that do truly have an effect. For the 900 cases in which there truly is no effect, with the usual significance level of 0.05 we expect 5% of the 900 studies to incorrectly reject H_0. This happens for $(0.05)900 = 45$ studies. So, of the 1000 studies, we expect 50 to report an effect that is truly there, but we also expect 45 to report an effect that does not actually exist. So a proportion of $45/(45 + 50) = 0.47$ of medical studies that report effects are actually reporting Type I errors.

The moral is to be skeptical when you hear reports of new medical advances. The true effect may be weaker than reported, or there may actually be no effect at all. ■

6.6 CALCULATING *P* (TYPE II ERROR)*

We've seen that decisions in significance tests have two potential types of error. A Type I error results from rejecting H_0 when it is actually true. Given that H_0 is true, the probability of a Type I error is the α-level of the test; when $\alpha = 0.05$, the probability of rejecting H_0 equals 0.05.

When H_0 is false, a Type II error results from *not* rejecting it. This probability has more than one value, because H_a contains a range of possible values. Each value in H_a has its own P(Type II error). This section shows how to calculate P(Type II error) at a particular value.

EXAMPLE 6.9 Testing whether Astrology Really Works

One scientific test of the pseudo-science astrology used the following experiment[5]: For each of 116 adult subjects, an astrologer prepared a horoscope based on the positions of the planets and the moon at the moment of the person's birth. Each subject also filled out a California Personality Index survey. For each adult, his or her birth data and horoscope were shown to an astrologer with the results of the personality survey for that adult and for two other adults randomly selected from the experimental group. The astrologer was asked which personality chart of the three subjects was the correct one for that adult, based on his or her horoscope.

Let π denote the probability of a correct prediction by an astrologer. If the astrologers' predictions are like random guessing, then $\pi = 1/3$. To test this against the alternative that the guesses are better than random guessing, we can test H_0: $\pi =$

[5]S. Carlson, *Nature*, vol. 318, 1985, pp. 419–425.

$1/3$ against H_a: $\pi > 1/3$. The alternative hypothesis reflects the astrologers' belief that they can predict better than random guessing. In fact, the National Council for Geocosmic Research, which supplied the astrologers for the experiment, claimed π would be 0.50 or higher. So let's find P(Type II error) if actually $\pi = 0.50$, for an $\alpha = 0.05$-level test. That is, if actually $\pi = 0.50$, we'll find the probability we'd fail to reject H_0: $\pi = 1/3$.

To determine this, we'll first find the sample proportion values for which we would not reject H_0. For the test of H_0: $\pi = 1/3$, the sampling distribution of $\hat{\pi}$ is the curve shown on the left in Figure 6.11. With $n = 116$, this curve has standard error

$$se_0 = \sqrt{\pi_0(1 - \pi_0)/n} = \sqrt{[(1/3)(2/3)]/116} = 0.0438.$$

For H_a: $\pi > 1/3$, we get a P-value of 0.05 if the test statistic $z = 1.645$. That is, 1.645 is the z-score that has a right-tail probability of 0.05. So we *fail to reject* H_0, getting a P-value *above* 0.05, if $z < 1.645$. In other words, we fail to reject H_0: $\pi = 1/3$ if the sample proportion $\hat{\pi}$ falls less than 1.645 standard errors above 1/3, that is, if

$$\hat{\pi} < 1/3 + 1.645(se_0) = 1/3 + 1.645(0.0438) = 0.405.$$

So the right-tail probability above 0.405 is $\alpha = 0.05$ for the curve on the left in Figure 6.11.

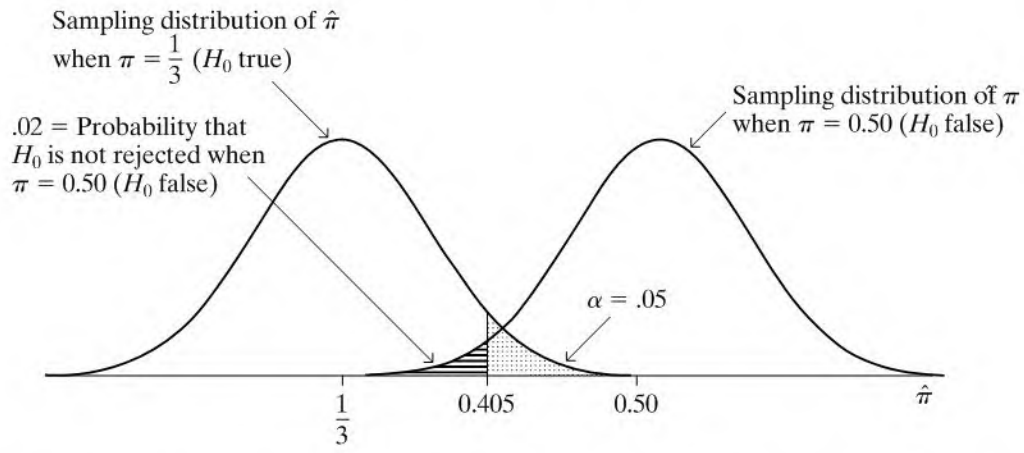

FIGURE 6.11: Calculation of P(Type II Error) for Testing H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$ at $\alpha = 0.05$ Level, when True Proportion is $\pi = 0.50$. A Type II error occurs if $\hat{\pi} < 0.405$, since then P-value >0.05 even though H_0 is false.

To find P(Type II error) if π actually equals 0.50, we must find $P(\hat{\pi} < 0.405)$ when $\pi = 0.50$. This is the left-tail probability *below* 0.405 for the curve on the right in Figure 6.11 (which is the curve that applies when $\pi = 0.50$). When $\pi = 0.50$, the standard error for a sample size of 116 is $\sqrt{[(0.50)(0.50)/116} = 0.0464$. (This differs a bit from se_0 for the test statistic, which uses 1/3 instead of 0.50 for π.) For the normal distribution with a mean of 0.50 and standard error of 0.0464, the $\hat{\pi}$ value of 0.405 has a z-score of

$$z = \frac{0.405 - 0.50}{0.0464} = -2.04.$$

The probability that $\hat{\pi} < 0.405$ is the probabilility that a standard normal variable falls below -2.04. From Table A, the left-tail probability below -2.04 equals 0.02. So, for a sample of size 116, the probability of not rejecting H_0: $\pi = 1/3$ is 0.02, if in fact $\pi = 0.50$.

In other words, if astrologers truly had the predictive power they claimed, the chance of failing to detect this with this experiment would have only been about 0.02. To see what actually happened in the experiment, see Exercise 6.17. ■

The probability of Type II error increases when the parameter value moves closer to H_0. To verify this, you can check that P(Type II error) = 0.54 at $\pi = 0.40$. So, if the parameter falls near the H_0 value, there may be a substantial chance of failing to reject H_0. Likewise, the farther the parameter falls from H_0, the less likely a Type II error. Figure 6.12 plots P(Type II error) for the various π values in H_a.

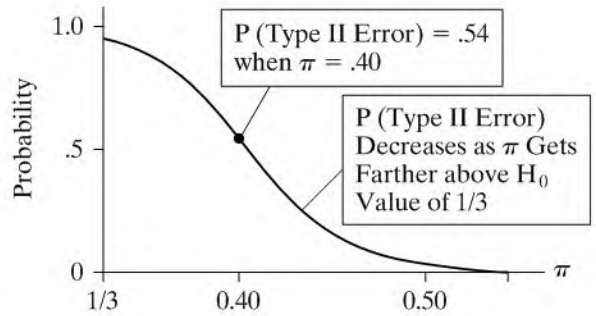

FIGURE 6.12: Probability of Type II Error for Testing H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$ at $\alpha = 0.05$ Level, Plotted for the Potential π Values in H_a

For a fixed α-level and alternative parameter value, P(Type II error) decreases when the sample size increases. If you can obtain more data, you will be less likely to make this sort of error.

Tests with Smaller α Have Greater P (Type II error)

As Section 6.4 discussed, the smaller $\alpha = P$(Type I error) is in a test, the larger P(Type II error) is. To illustrate, suppose Example 6.9 used $\alpha = 0.01$. Then you can verify that P(Type II error) = 0.08, compared to P(Type II error) = 0.02 when $\alpha = 0.05$.

The reason that extremely small values are not normally used for α, such as $\alpha = 0.0001$, is that P(Type II error) is too high. We may be unlikely to reject H_0 even if the parameter falls far from the null hypothesis. In summary, for fixed values of other factors,

- P(Type II error) decreases as
  - — the parameter value is farther from H_0.
  - — the sample size increases.
  - — P(Type I error) increases.

The Power of a Test

When H_0 is false, you want the probability of rejecting H_0 to be high. The probability of rejecting H_0 is called the ***power*** of the test. For a particular value of the parameter from within the H_a range,

$$\text{Power} = 1 - P\text{(Type II error)}.$$

In Example 6.9, for instance, the test of H_0: $\pi = 1/3$ has P(Type II error) = 0.02 at $\pi = 0.50$. Therefore, the power of the test at $\pi = 0.50$ equals $1 - 0.02 = 0.98$.

The power increases for values of the parameter falling farther from the H_0 value. Just as the curve for P(Type II error) in Figure 6.12 decreases as π gets farther above $\pi_0 = 1/3$, the curve for the power would increase.

In practice, studies should ideally have high power. Before granting financial support for a planned study, many research agencies expect principal investigators to show that reasonable power (usually, at least 0.80) exists at values of the parameter that are considered practically significant.

When you read that results of a study are insignificant, be skeptical if no information is given about the power. The power may be low, especially if n is small. For further details about calculating P(Type II error) or power, see Cohen (1988).

6.7 SMALL-SAMPLE TEST FOR A PROPORTION—THE BINOMIAL DISTRIBUTION*

For a population proportion π, Section 6.3 presented a significance test that is valid for large samples. The sampling distribution of the sample proportion $\hat{\pi}$ is then approximately normal, which justifies using a z test statistic.

For small n, the sampling distribution of $\hat{\pi}$ occurs at only a few points. If $n = 5$, for example, the only possible values for the sample proportion $\hat{\pi}$ are 0, 1/5, 2/5, 3/5, 4/5, and 1. A continuous approximation such as the normal distribution is inappropriate. In addition, we'll see that the closer the π is to 0 or 1 for a given sample size, the more skewed the actual sampling distribution becomes.

This section introduces a small-sample test for proportions. It uses the most important probability distribution for discrete variables, the *binomial*.

The Binomial Distribution

For categorical data, often the following conditions hold:

1. Each observation falls into one of two categories.
2. The probabilities for the two categories are the same for each observation. We denote the probabilities by π for category 1 and $(1 - \pi)$ for category 2.
3. The outcomes of successive observations are independent. That is, the outcome for one observation does not depend on the outcomes of other observations.

Flipping a coin repeatedly is a prototype for these conditions. For each flip, we observe whether the outcome is head (category 1) or tail (category 2). The probabilities of the outcomes are the same for each flip (0.50 for each if the coin is balanced). The outcome of a particular flip does not depend on the outcome of other flips.

Now, for n observations, let x denote the number that occur in category 1. For example, for $n = 5$ coin flips, x = number of heads could equal 0, 1, 2, 3, 4, or 5. When the observations satisfy the above conditions, the probability distribution of x is the ***binomial distribution***.

The binomial variable x is discrete, taking one of the integer values $0, 1, 2, \ldots, n$. The formula for the binomial probabilities follows:

Probabilities for a Binomial Distribution

Denote the probability of category 1, for each observation, by π. For n independent observations, the probability of x outcomes in category 1 is

$$P(x) = \frac{n!}{x!(n - x)!}\pi^x(1 - \pi)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

The symbol $n!$ is called ***n factorial***. It represents $n! = 1 \times 2 \times 3 \cdots \times n$. For example, $1! = 1$, $2! = 1 \times 2 = 2$, $3! = 1 \times 2 \times 3 = 6$, and so forth. Also, $0!$ is defined to be 1.

For particular values for π and n, substituting the possible values for x into the formula for $P(x)$ provides the probabilities of the possible outcomes. The sum of the probabilities equals 1.0.

EXAMPLE 6.10 Gender and Selection of Managerial Trainees

Example 6.1 (page 143) discussed a case involving potential bias against females in selection of management trainees for a large supermarket chain. The pool of employees is half female and half male. Ten trainees are supposedly selected at random from this pool. If they are truly selected at random, how many females would we expect?

The probability that any one person selected is a female is $\pi = 0.50$, the proportion of available trainees who are female. Similarly, the probability that any one person selected is male is $(1 - \pi) = 0.50$. Let x = number of females selected. This has the binomial distribution with $n = 10$ and $\pi = 0.50$. For each x between 0 and 10, the probability that x of the ten people selected are female equals

$$P(x) = \frac{10!}{x!(10 - x)!}(0.50)^x(0.50)^{10-x}, \quad x = 0, 1, 2, \ldots, 10.$$

For example, the probability that no females are chosen ($x = 0$) equals

$$P(0) = \frac{10!}{0!10!}(0.50)^0(0.50)^{10} = (0.50)^{10} = 0.001.$$

Recall that any number raised to the power of 0 equals 1. Also, $0! = 1$, and the 10! terms in the numerator and denominator cancel, leaving $P(0) = (0.50)^{10}$. The probability that exactly one female is chosen equals

$$P(1) = \frac{10!}{1!9!}(0.50)^1(0.50)^9 = 10(0.50)(0.50)^9 = 0.010.$$

This computation simplifies considerably by using $10!/9! = 10$, since 10! is just 9! multiplied by 10. Table 6.6 lists the entire binomial distribution for $n = 10$, $\pi = 0.50$.

In Table 6.6, the probability is about 0.98 that x falls between 2 and 8, inclusive. The least likely values for x are 0, 1, 9, and 10, which have a combined probability of only 0.022. If the sample were randomly selected, somewhere between about two and eight females would probably be selected. It is especially unlikely that none or ten would be selected.

The probabilities for females determine those for males. For instance, the probability that nine of the ten people selected are male equals the probability that one of the ten selected is female. ■

TABLE 6.6: The Binomial Distribution for $n = 10$, $\pi = 0.50$. The binomial variable x can take any value between 0 and 10.

x	$P(x)$	x	$P(x)$
0	0.001	6	0.205
1	0.010	7	0.117
2	0.044	8	0.044
3	0.117	9	0.010
4	0.205	10	0.001
5	0.246		

Properties of the Binomial Distribution

The binomial distribution is perfectly symmetric only when $\pi = 0.50$. In Example 6.10 with $n = 10$, for instance, since the population proportion of females equals 0.50, $x = 10$ has the same probability as $x = 0$.

The sample proportion $\hat{\pi}$ relates to x by

$$\hat{\pi} = x/n.$$

For example, for $x = 1$ female chosen out of $n = 10$, $\hat{\pi} = 1/10 = 0.10$. The sampling distribution of $\hat{\pi}$ is also symmetric when $\pi = 0.50$. When $\pi \neq 0.50$, the distributions are skewed, the degree of skew increasing as π gets closer to 0 or 1. Figure 6.13 illustrates for the sampling distribution of $\hat{\pi}$. For instance, when $\pi = 0.10$, the sample proportion $\hat{\pi}$ can't fall much below 0.10 since it must be positive, but it could fall considerably above 0.10.

The binomial distribution has mean and standard deviation

$$\mu = n\pi, \quad \sigma = \sqrt{n\pi(1 - \pi)}.$$

For example, suppose the chance of a female in any one selection for management training is 0.50, as the supermarket chain claims. Then, out of 10 trainees, we expect $\mu = n\pi = 10(0.50) = 5.0$ females.

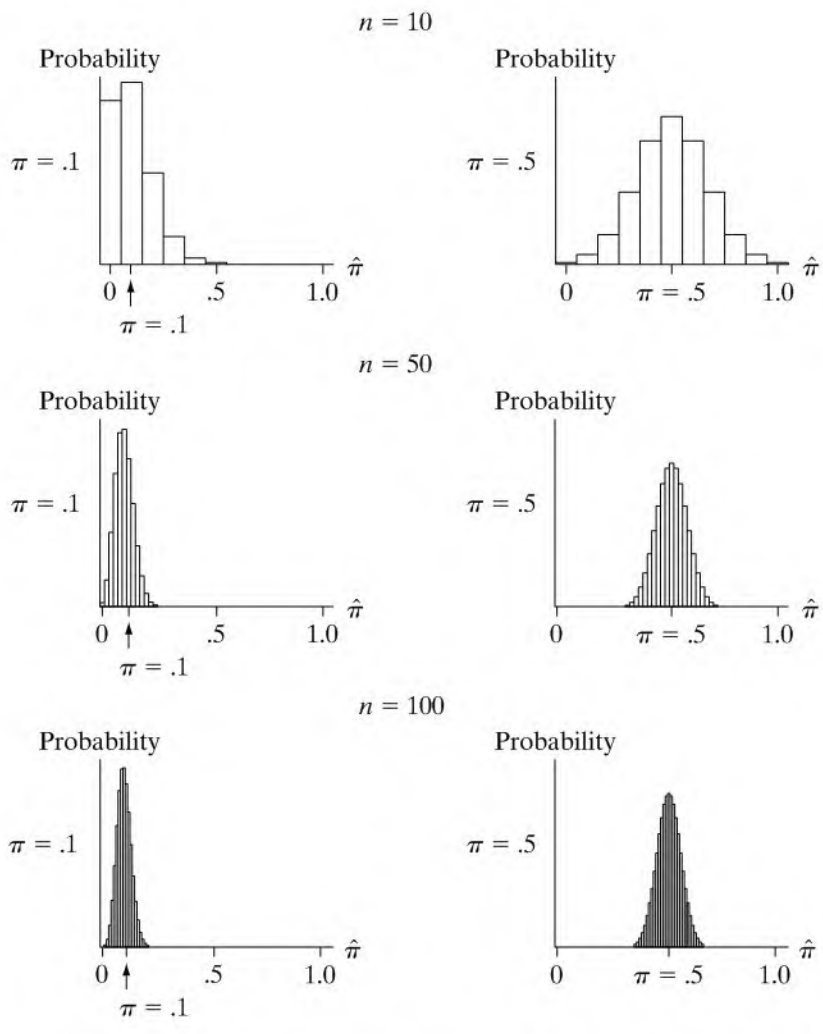

FIGURE 6.13: Sampling Distribution of $\hat{\pi}$ when $\pi = 0.10$ or 0.50, for $n = 10, 50, 100$

EXAMPLE 6.11 How Much Variability Can an Exit Poll Show?

Example 4.6 (page 85) discussed an exit poll of 2705 voters for the 2006 California gubernatorial election. Let x denote the number in the exit poll who voted for Arnold Schwarzenegger. In the population of nearly 7 million voters, 55.9% voted for him. If the exit poll was randomly selected, then the binomial distribution for x has $n = 2705$ and $\pi = 0.559$. The distribution is described by

$$\mu = 2705(0.559) = 1512, \quad \sigma = \sqrt{2705(0.559)(0.441)} = 26.$$

Almost certainly, x would fall within 3 standard deviations of the mean. This is the interval from 1434 to 1590. In fact, in that exit poll, 1528 reported voting for Schwarzenegger. ■

We've seen (Sections 5.2, 6.3) that the sampling distribution of the sample proportion $\hat{\pi}$ has mean π and standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$. To get these, we divide the binomial mean $\mu = n\pi$ and standard deviation $\sigma = \sqrt{n\pi(1 - \pi)}$ by n, since $\hat{\pi}$ divides x by n.

The binomial distribution and the sampling distribution of $\hat{\pi}$ are approximately normal for large n. This approximation is the basis of the large-sample test of Section 6.3. How large is "large"? A guideline is that the expected number of observations should be at least 10 for both categories. For example, if $\pi = 0.50$, we need at least about $n = 20$, because then we expect $20(0.50) = 10$ observations in one category and $20(1 - 0.50) = 10$ in the other category. For testing H_0: $\pi = 0.90$ or $H_0 : \pi = 0.10$, we need $n \geq 100$. The sample size requirement reflects the fact that a symmetric bell shape for the sampling distribution of $\hat{\pi}$ requires larger sample sizes when π is near 0 or 1 than when π is near 0.50.

The Binomial Test

If the sample size is not large enough to use the normal test, we can use the binomial distribution directly. Refer to Example 6.10 (page 170) about potential gender discrimination. For random sampling, the probability π that a person selected for management training is female equals 0.50. If there is bias against females, then $\pi < 0.50$. Thus, we can test the company's claim of random sampling by testing

$$H_0 : \pi = 0.50 \qquad \text{versus} \qquad H_a : \pi < 0.50.$$

Of the ten employees chosen for management training, let x denote the number of women. Under H_0, the sampling distribution of x is the binomial distribution with $n = 10$ and $\pi = 0.50$. Table 6.6 tabulated it. As in Example 6.1 (page 143), suppose $x = 1$. The P-value is then the left-tail probability of an outcome at least this extreme; that is, $x = 1$ or 0. From Table 6.6, the P-value equals

$$P = P(0) + P(1) = 0.001 + 0.010 = 0.011.$$

If the company selected trainees randomly, the probability of choosing one or fewer females is only 0.011. This result provides evidence against the null hypothesis of a random selection process. We can reject H_0 for $\alpha = 0.05$, though not for $\alpha = 0.010$.

Even if we suspect bias in a particular direction, the most even-handed way to perform a test uses a two-sided alternative. For H_a: $\pi \neq 0.50$, the P-value is $2(0.011) = 0.022$. This is a two-tail probability of the outcome that one or fewer *of either sex* is selected. Figure 6.14 shows the formation of this P-value.

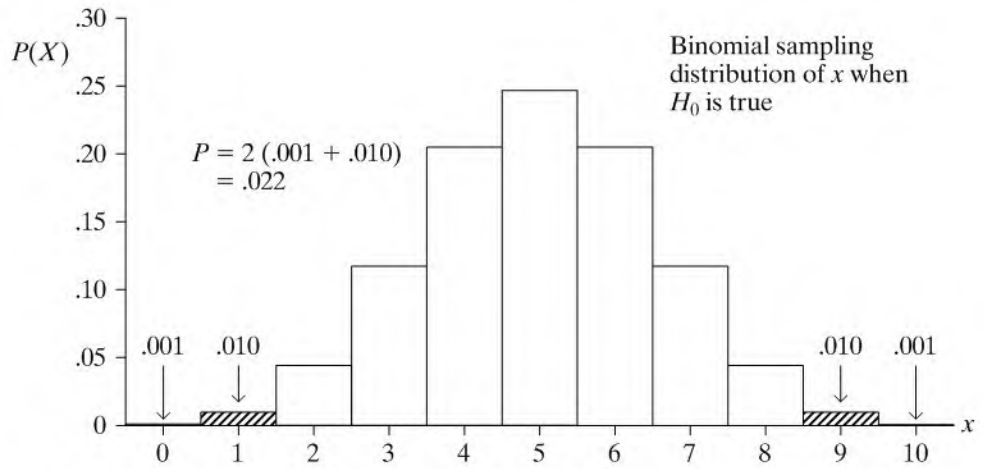

FIGURE 6.14: Calculation of *P*-Value in Testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$, when $n = 10$ and $x = 1$

The assumptions for the binomial test are the three conditions for the binomial distribution. Here, the conditions are satisfied. Each observation has only two possible outcomes, female or male. The probability of each outcome is the same for each selection, 0.50 for selecting a female and 0.50 for selecting a male (under H_0). For random sampling, the outcome of any one selection does not depend on any other one.

In the rare cases that the population size is small, the binomial conditions are not all satisfied. To illustrate, suppose a population contains only four persons, of whom two are female. If we randomly sample two separate individuals, the second observation has different probabilities than the first. For example, if the first person selected was female, then the probability that the second person selected is female equals 1/3, since 1 female remains among the 3 subjects. Thus, the probabilities are not the same for each selection, which the binomial requires. For the successive observations to have essentially the same probabilities and be independent, the population size must be much larger than the sample size. The sample size should be no more than about 10% as large as the minimum of the population numbers of subjects in the two categories. This is usually easily satisfied in practice.

6.8 CHAPTER SUMMARY

Statistical inference uses sample data to make predictions about population parameters. Chapters 5 and 6 have introduced two inference methods—***estimation*** and ***significance tests***. The estimation method called *confidence intervals* provides a range of the most plausible values for a parameter. A significance test judges whether a particular value for the parameter is plausible. Both methods utilize the sampling distribution of the estimator of the parameter.

Significance tests have five parts:

1. ***Assumptions:***
    - Tests for *means* apply with quantitative variables, whereas tests for *proportions* apply with categorical variables.
    - Tests assume *randomization*, such as a random sample.
    - Large-sample tests about proportions require no assumption about the population distribution, because the Central Limit Theorem implies approximate normality of the sampling distribution of the sample proportion. This

justifies using the z test statistic. Small-sample tests for proportions use the *binomial distribution*.

- Tests for means use the t distribution for the t test statistic. The test assumes the population distribution is normal. In practice, two-sided tests (like confidence intervals) are *robust* to violations of the normality assumption, especially for large samples because of the Central Limit Theorem.

2. ***Null and alternative hypotheses*** about the parameter: The null hypothesis has form H_0: $\mu = \mu_0$ for a mean and H_0: $\pi = \pi_0$ for a proportion. Here, μ_0 and π_0 denote values hypothesized for the parameters, such as 0.50 in H_0: $\pi = 0.50$. The most common alternative is *two sided*, such as H_a: $\pi \neq 0.50$. Hypotheses such as H_a: $\pi > 0.50$ and H_a: $\pi < 0.50$ are *one sided*, designed to detect departures from H_0 in a particular direction.
3. A ***test statistic*** describes how far the point estimate falls from the H_0 value. The z statistic for proportions and t statistic for means measure the number of standard errors that the point estimate ($\hat{\pi}$ or $\bar{y}$) falls from the H_0-value.
4. The ***P-value*** describes the weight of evidence the data provide about H_0.

   - The P-value is calculated by presuming that H_0 is true. It equals the probability that the test statistic equals the observed value or a value even more extreme.
   - The "more extreme" results are determined by the alternative hypothesis. For two-sided H_a, the P-value is a two-tail probability.
   - Small P-values result when the point estimate falls far from the H_0 value, so that the test statistic is large. When the P-value is small, it would be unusual to observe such data if H_0 were true. The smaller the P-value, the stronger the evidence against H_0.

5. A ***conclusion*** based on the sample evidence about H_0: We report and interpret the P-value. Sometimes it is necessary to make a decision. If the P-value is less than or equal to a fixed α-level (such as $\alpha = 0.05$), we reject H_0. Otherwise, we cannot reject it.

When we make a decision, two types of errors can occur.

- When H_0 is true, a Type I error results if we reject it.
- When H_0 is false, a Type II error results if we fail to reject it.

The choice of α, the cutoff point for the P-value in making a decision, equals P(Type I error). Normally, we choose small values such as $\alpha = 0.05$ or 0.01. For fixed α, P(Type II error) decreases as the distance increases between the parameter and the H_0 value or as the sample size increases.

Table 6.7 summarizes the five parts of the tests this chapter presented.

Sample size is a critical factor in both estimation and significance tests. With small sample sizes, confidence intervals are wide, making estimation imprecise. Small sample sizes also make it difficult to reject false null hypotheses unless the true parameter is far from the null hypothesis value. P(Type II error) may be high for parameter values of interest.

To introduce estimation and significance tests, Chapters 5 and 6 presented inference about a single parameter for a single variable. In practice, it is usually artificial to have a particular fixed number for the H_0 value of a parameter. One of the few times this happens is when the response score results from taking a difference of two

TABLE 6.7: Summary of Significance Tests for Means and Proportions

Parameter	Mean	Proportion
1. Assumptions	Random sample, quantitative variable normal population	Random sample, categorical variable null expected counts at least 10
2. Hypotheses	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$ H_a: $\mu > \mu_0$ H_a: $\mu < \mu_0$	H_0: $\pi = \pi_0$ H_a: $\pi \neq \pi_0$ H_a: $\pi > \pi_0$ H_a: $\pi < \pi_0$
3. Test statistic	$t = \frac{\bar{y} - \mu_0}{se}$ with $se = \frac{s}{\sqrt{n}}$, $df = n - 1$	$z = \frac{\hat{\pi} - \pi_0}{se_0}$ with $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$
4. *P*-value	Two-tail probability in sampling distribution for two-sided test (H_0: $\mu \neq \mu_0$ or H_a: $\pi \neq \pi_0$); One-tail probability for one-sided test	
5. Conclusion	Reject H_0 if *P*-value $\leq \alpha$-level such as 0.05	

values, such as the change in weight in Example 6.4 (page 151). In that case, $\mu_0 = 0$ is a natural baseline. Significance tests much more commonly refer to comparisons of means for two samples than to a fixed value of a parameter for a single sample. The next chapter shows how to compare means or proportions for two groups.

PROBLEMS

Practicing the Basics

6.1. For (a)–(c), is it a null hypothesis, or an alternative hypothesis?

(a) In Canada, the proportion of adults who favor legalized gambling equals 0.50.

(b) The proportion of all Canadian college students who are regular smokers now is less than 0.24 (the value it was ten years ago).

(c) The mean IQ of all students at Lake Wobegon High School is larger than 100.

(d) Introducing notation for a parameter, state the hypotheses in (a)–(c) in terms of the parameter values.

6.2. You want to know whether adults in your country think the ideal number of children is equal to 2, or higher or lower than that.

(a) Define notation and state the null and alternative hypotheses for studying this.

(b) For responses in a recent GSS to the question, "What do you think is the ideal number of children to have?", software shows results:

```
------------------------------------------------------------
Test of mu = 2.0 vs mu not = 2.0

Variable   N      Mean     StDev    SE Mean      T      P-value
Children  1302    2.490    0.850    0.0236     20.80     0.0000
------------------------------------------------------------
```

Report the test statistic value, and show how it was obtained from other values reported in the table.

(c) Explain what the *P*-value represents and interpret its value.

6.3. For a test of H_0: $\mu = 0$ against H_a: $\mu \neq 0$ with $n = 1000$, the t test statistic equals 1.04.

(a) Find the *P*-value and interpret it. (*Note*: You can use the standard normal to approximate the t distribution.)

(b) Suppose $t = -2.50$ rather than 1.04. Find the *P*-value. Does this provide stronger, or weaker, evidence against the null hypothesis? Explain.

(c) When $t = 1.04$, find the *P*-value for (i) H_a: $\mu > 0$, (ii) H_a: $\mu < 0$.

6.4. The *P*-value for a test about a mean with $n = 25$ is $P = 0.05$.

(a) Find the t test statistic value that has this *P*-value for (i) H_a: $\mu \neq 0$, (ii) H_a: $\mu > 0$, (iii) H_a: $\mu < 0$.

(b) Does this P-value provide stronger, or weaker, evidence against the null hypothesis than $P = 0.01$? Explain.

6.5. Find and interpret the P-value for testing H_0: $\mu = 100$ against $H_0 : \mu \neq 100$, if a sample has
(a) $n = 400, \bar{y} = 103, s = 40$.
(b) $n = 1600, \bar{y} = 103, s = 40$. Comment on the effect of n on the results of a significance test.

6.6. Example 6.4 (page 151) described a study about therapies for teenage girls suffering from anorexia. For the 17 girls who received the family therapy, the changes in weight were

$$11, 11, 6, 9, 14, -3, 0, 7, 22, -5, -4, 13, 13, 9, 4, 6, 11.$$

Part of an SPSS printout for the data shows:

```
------------------------------------------------
Lower     Upper     t-value     df     2-Tail Sig
3.60                                     .0007
------------------------------------------------
```

Fill in the missing results.

6.7. According to a union agreement, the mean income for all senior-level assembly-line workers in a large company equals \$500 per week. A representative of a women's group decides to analyze whether the mean income μ for female employees matches this norm. For a random sample of nine female employees, $\bar{y} = \$410$ and $s = \$90$.
(a) Test whether the mean income of female employees differs from \$500 per week. Include assumptions, hypotheses, test statistic, and P-value. Interpret the result.
(b) Report the P-value for H_a: $\mu < 500$. Interpret.
(c) Report and interpret the P-value for H_a: $\mu > 500$. (*Hint*: The P-values for the two possible one-sided tests must sum to 1.)

6.8. By law, an industrial plant can discharge no more than 500 gallons of waste water per hour, on the average, into a neighboring lake. Based on other infractions they have noticed, an environmental action group believes this limit is being exceeded. Monitoring the plant is expensive, and a random sample of four hours is selected over a period of a week. Software reports:

```
Variable  No. Cases  Mean     SD     SE of Mean
 WASTE        4     1000.0  400.0     200.0
```

(a) Test whether the mean discharge equals 500 gallons per hour against the alternative that the limit is being exceeded. Find the P-value and interpret.
(b) Explain why the test may be highly approximate or even invalid if the population distribution of discharge is far from normal.
(c) Explain how your one-sided analysis implicitly tests the broader null hypothesis that $\mu \leq 500$.

6.9. In response to the statement "A preschool child is likely to suffer if his or her mother works," the response categories (Strongly agree, Agree, Disagree, Strongly disagree) had counts (91, 385, 421, 99) for responses in a GSS. To treat this ordinal variable as quantitative, we assign scores to the categories. For the scores $(2, 1, -1, -2)$, which treat the distance between Agree and Disagree as twice the distance between Strongly agree and Agree or between Disagree and Strongly disagree, software reports:

```
-----------------------------------
  N       Mean     Std Dev    Std Err
 996     -0.052     1.253      0.0397
-----------------------------------
```

(a) Set up null and alternative hypotheses to test whether the population mean response differs from the neutral value, 0.
(b) Find the test statistic and P-value. Interpret, and make a decision about H_0, using $\alpha = 0.05$.
(c) Based on (b), can you "accept" H_0: $\mu = 0$? Why or why not?
(d) Construct a 95% confidence interval for μ. Show the correspondence between whether 0 falls in the interval and the decision about H_0.

6.10. In Example 6.2 on political ideology (page 149), suppose we use the scores $(-3, -2, -1, 0, 1, 2, 3)$ instead of the scores (1, 2, 3, 4, 5, 6, 7) used there. We then test H_0: $\mu = 0$. Explain the effect of the change in scores on **(a)** the sample mean and standard deviation, **(b)** the test statistic, **(c)** the P-value and interpretation.

6.11. Results of 99% confidence intervals for means are consistent with results of two-sided tests with which α-level? Explain the connection.

6.12. For a test of H_0: $\pi = 0.50$, the z test statistic equals 1.04.
(a) Find the P-value for H_a: $\pi > 0.50$.
(b) Find the P-value for H_a: $\pi \neq 0.50$.
(c) Find the P-value for H_a: $\pi < 0.50$.
(d) Do any of the P-values in (a), (b), or (c) give strong evidence against H_0? Explain.

6.13. For a test of H_0: $\pi = 0.50$, the sample proportion is 0.35 based on a sample size of 100.
(a) Show that the test statistic is $z = -3.0$.
(b) Find and interpret the P-value for H_a: $\pi < 0.50$.
(c) For a significance level of $\alpha = 0.05$, what decision do you make?
(d) If the decision in (c) was in error, what type of error was it? What could you do to reduce the chance of that type of error?

6.14. Same-sex marriage was legalized across Canada by the Civil Marriage Act enacted in 2005. Is this supported by a majority, or a minority, of the Canadian population? A poll conducted for the *Globe and Mail* newspaper in July 2005 of 1000 Canadians asked whether this bill should stand or be repealed. The responses were 55% should stand, 39% should repeal, 6% don't know. Let π denote the population proportion of Canadian adults who believe it should stand. For testing $H_0 : \pi = 0.50$ against $H_a : \pi \neq 0.50$:

(a) Find the standard error, and interpret.

(b) Find the test statistic, and interpret.

(c) Find the P-value, and interpret in context.

6.15. When a recent GSS asked, "Would you be willing to pay much higher taxes in order to protect the environment?" (variable GRNTAXES), 369 people answered yes and 483 answered no. Software shows the following results to analyze whether a majority or minority of Americans would answer yes:

```
-----------------------------------------------
Test of proportion = 0.5 vs not = 0.5

  N  Sample prop     95% CI      Z-Value P-Value
852    0.4331    (0.400, 0.466)  -3.91   0.000
-----------------------------------------------
```

(a) Specify the hypotheses that are tested.

(b) Report and interpret the test statistic value.

(c) Report and interpret the P-value as a probability.

(d) Explain an advantage of the confidence interval shown over the significance test.

6.16. A Pew Research Center poll (May 14, 2003) of 1201 adults asked, "All in all, do you think affirmative action programs designed to increase the number of black and minority students on college campuses are a good thing or a bad thing?" Sixty percent said good, 30% said bad, and 10% said don't know. Let π denote the population proportion who said it is good. Find the P-value for testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$. Interpret.

6.17. In the scientific test of astrology discussed in Example 6.9 (page 166), the astrologers were correct with 40 of their 116 predictions. Test H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$. Find the P-value, make a decision using $\alpha = 0.05$, and interpret.

6.18. The previous exercise analyzed whether astrologers could predict the correct personality chart for a given horoscope better than by random guessing. In the words of that study, what would be a

(a) Type I error,

(b) Type II error?

6.19. A mayoral election in Madison, Wisconsin, has two candidates. Exactly half the residents currently prefer each candidate.

(a) For a random sample of 400 voters, 230 voted for a particular candidate. Are you willing to predict the winner? Why?

(b) For a random sample of 40 voters, 23 voted for a particular candidate. Would you be willing to predict the winner? Why? (The sample proportion is the same in (a) and (b), but the sample sizes differ.)

6.20. The authorship of an old document is in doubt. A historian hypothesizes that the author was a journalist named Jacalyn Levine. Upon a thorough investigation of Levine's known works, it is observed that one unusual feature of her writing was that she consistently began 6% of her sentences with the word *whereas*. To test the historian's hypothesis, it is decided to count the number of sentences in the disputed document that begin with *whereas*. Out of the 300 sentences, none do. Let π denote the probability that any one sentence written by the unknown author of the document begins with *whereas*. Test H_0: $\pi = 0.06$ against H_a: $\pi \neq 0.06$. What assumptions are needed for your conclusion to be valid? (F. Mosteller and D. L. Wallace conducted this type of investigation to determine whether Alexander Hamilton or James Madison authored 12 of the *Federalist Papers*. See *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, 1964.)

6.21. A multiple-choice test question has four possible responses. The question is difficult, with none of the four responses being obviously wrong, yet with only one correct answer. It first occurs on an exam taken by 400 students. Test whether more people answer the question correctly than would be expected just due to chance (i.e., if everyone randomly guessed the correct answer).

(a) Set up the hypotheses for the test.

(b) Of the 400 students, 125 correctly answer the question. Find the P-value and interpret.

6.22. Example 6.4 (page 151) tested a therapy for anorexia, using H_0: $\mu = 0$ and H_a: $\mu > 0$ about the population mean weight change.

(a) In the words of that example, what would be a (i) Type I error, (ii) Type II error?

(b) The P-value was 0.017. If the decision for $\alpha = 0.05$ were in error, what type of error is it?

(c) Suppose instead $\alpha = 0.01$. What decision would you make? If it is in error, what type of error is it?

6.23. Jones and Smith separately conduct studies to test H_0: $\mu = 500$ against H_0: $\mu \neq 500$, each with $n = 1000$. Jones gets $\bar{y} = 519.5$, with $se = 10.0$. Smith gets $\bar{y} = 519.7$, with $se = 10.0$.

(a) Show that $t = 1.95$ and P-value $= 0.051$ for Jones. Show that $t = 1.97$ and P-value $= 0.049$ for Smith.

(b) Using $\alpha = 0.050$, for each study indicate whether the result is "statistically significant."

(c) Using this example, explain the misleading aspects of reporting the result of a test as "$P \leq 0.05$" versus "$P > 0.05$," or as "reject H_0" versus "Do not reject H_0," without reporting the actual P-value.

6.24. Jones and Smith separately conduct studies to test H_0: $\pi = 0.50$ against H_0: $\pi \neq 0.50$, each with $n = 400$. Jones gets $\hat{\pi} = 220/400 = 0.550$. Smith gets $\hat{\pi} = 219/400 = 0.5475$.

(a) Show that $z = 2.00$ and P-value $= 0.046$ for Jones. Show that $z = 1.90$ and P-value $= 0.057$ for Smith.

(b) Using $\alpha = 0.05$, indicate in each case whether the result is "statistically significant." Interpret.

(c) Use this example to explain why important information is lost by reporting the result of a test as "P-value ≤ 0.05" versus "P-value > 0.05," or as "reject H_0" versus "Do not reject H_0", without reporting the P-value.

(d) The 95% confidence interval for π is (0.501, 0.599) for Jones and (0.499, 0.596) for Smith. Explain how this method shows that, in practical terms, the two studies had very similar results.

6.25. A study considers whether the mean score μ on a college entrance exam for students in 2007 is any different from the mean of 500 for students in 1957. Test H_0: $\mu = 500$ against H_a: $\mu \neq 500$, if for a nationwide random sample of 10,000 students who took the exam in 2007, $\bar{y} = 497$ and $s = 100$. Show that the result is highly significant statistically, but not practically significant.

6.26. A report released on September 25, 2006 by the Collaborative on Academic Careers in Higher Education indicated that there is a notable gap between female and male academics in their confidence that tenure rules are clear, with men feeling more confident. The 4500 faculty members in the survey were asked to evaluate policies on a scale of 1 to 5 (very unclear to very clear). The mean response about the criteria for tenure was 3.51 for females and 3.55 for males, which was indicated to meet the test for statistical significance, with the mean for females being significantly less than the mean for males. Use this study to explain the distinction between statistical significance and practical significance.

6.27. Refer to Example 6.8 on "medical discoveries" (page 165). Using a tree diagram, approximate P(Type I error) under the assumption that a true effect exists 20% of the time and that P(Type II error) $= 0.30$.

6.28. A decision is planned in a test of H_0: $\mu = 0$ against H_a: $\mu > 0$, using $\alpha = 0.05$. If $\mu = 5$, P(Type II error) $= 0.17$.

(a) Explain the meaning of this last sentence.

(b) If the test used $\alpha = 0.01$, would P(Type II error) be less than, equal to, or greater than 0.17? Explain.

(c) If $\mu = 10$, would P(Type II error) be less than, equal to, or greater than 0.17? Explain.

6.29. Let π denote the proportion of schizophrenics who respond positively to treatment. A test is conducted of H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$, for a sample of size 25, using $\alpha = 0.05$.

(a) Find the region of sample proportion values for which H_0 is rejected.

(b) Suppose that $\pi = 0.60$. Find P(Type II error).

6.30. Studies have considered whether neonatal sex differences exist in behavioral and physiological reactions to stress. One study[6] evaluated changes in heart rate for a sample of infants placed in a stressful situation. The sample mean change in heart rate was small for males compared to females: -1.2 compared to 10.7, each with standard deviations of about 18. Suppose we are skeptical of the result for males and plan a larger experiment to test whether the mean heart rate increases when male infants undergo the stressful experience. Let μ denote the population mean of the difference in heart rates, after versus before the stress. We'll test H_0: $\mu = 0$ against H_a: $\mu > 0$, at the $\alpha = 0.05$ level using $n = 30$ infant males. Suppose the standard deviation is 18. Find P(Type II error) if $\mu = 10$ by showing (a) a test statistic of $t = 1.699$ has a P-value of 0.05, (b) we fail to reject H_0 if $\bar{y} < 5.6$, (c) this happens if $\bar{y}$ falls more than 1.33 standard errors below 10, (d) this happens with probability about 0.10.

6.31. Refer to the previous exercise.

(a) Find P(Type II error) if $\mu = 5$. How does P(Type II error) depend on the value of μ?

(b) Find P(Type II error) if $\mu = 10$ and $\alpha = 0.01$. How does P(Type II error) depend on α?

(c) How does P(Type II error) depend on n?

6.32. A jury list contains the names of all individuals who may be called for jury duty. The proportion of women on the list is 0.53. A jury of size 12 is selected at random from the list. None selected are women.

(a) Find the probability of selecting 0 women.

(b) Test the hypothesis that the selections are random against the alternative of bias against women. Report the P-value, and interpret.

6.33. A person claiming to possess extrasensory perception (ESP) says she can guess more often than not the outcome of a flip of a balanced coin in another room, not visible to her.

[6] M. Davis and E. Emory, *Child Development*, vol. 66, 1995, pp. 14–27.

(a) Introduce appropriate notation, and state hypotheses for testing her claim.

(b) Of 5 coin flips, she guesses the correct result 4 times. Find the P-value and interpret.

6.34. In a CNN exit poll of 1336 voters in the 2006 Senatorial election in New York State, let x = number in exit poll who voted for the Democratic candidate, Hillary Clinton.

(a) Explain why this scenario would seem to satisfy the three conditions needed to use the binomial distribution.

(b) If the population proportion voting for Clinton had been 0.50, find the mean and standard deviation of the probability distribution of x.

(c) For (b), using the normal distribution approximation, give an interval in which x would almost certainly fall.

(d) Actually, the exit poll had $x = 895$. Explain how you could make an inference about whether π is above or below 0.50.

6.35. In a given year, the probability that an American female dies in a motor vehicle accident equals 0.0001 (*Statistical Abstract of the United States*).

(a) In a city having 1 million females, find the mean and standard deviation of x = number of deaths from motor vehicle accidents. State the assumptions for these to be valid. (*Hint*: Find μ and σ for the binomial distribution.)

(b) Would it be surprising if $x = 0$? Explain. (*Hint*: How many standard deviations is 0 from the expected value?)

(c) Based on the normal approximation to the binomial, find an interval within which x has probability 0.95 of occurring.

(d) The probability for American males is 0.0002. Repeat (a) for males, and compare results to those for females.

Concepts and Applications

6.36. You can use an *applet* to repeatedly generate random samples and conduct significance tests, to illustrate their behavior when used for many samples. To try this, go to the *significance test for a proportion* applet at www.prenhall.com/???. Set the null hypothesis as H_0: $\pi = 1/3$ for a one-sided test ($\pi > 1/3$) with sample size 116, a case Example 6.9 (page 166) on the astrology experiment considered. At the menu, set the true proportion value to 0.33.

(a) Click *Simulate* and 100 samples of this size will be taken, with the P-value found for each sample. What percentage of the tests were significant at the 0.05 significance level?

(b) To get a feel for what happens "in the long run," do this simulation 50 times, so you will have a total of 5000 samples of size 116. What percentage of the samples resulted in a Type I error? What percentage would you expect to do so, resulting in a Type I error?

(c) Next, change π to 0.50, so H_0 is actually false. Simulate 5000 samples. What percentage of times did you make a Type II error? By Example 6.9 this should happen only about 2% of the time.

6.37. Refer to the "Student survey" data file (Exercise 1.11 on page 8).

(a) Test whether the population mean political ideology differs from 4.0. Report the P-value, and interpret.

(b) Test whether the proportion favoring legalized abortion equals, or differs from, 0.50. Report the P-value, and interpret.

6.38. Refer to the data file your class created in Exercise 1.12 (page 9). For variables chosen by your instructor, state a research question and conduct inferential statistical analyses. Also use graphical and numerical methods presented earlier in this text to describe the data and, if necessary, to check assumptions for your analyses. Prepare a report, summarizing and interpreting your findings.

6.39. A study considered the effects of a special class designed to improve children's verbal skills. Each child took a verbal skills test before and after attending the class for three weeks. Let y = second exam score − first exam score. The scores on y for a random sample of four children having learning problems were 3, 7, 3, 3. Conduct inferential statistical methods to determine whether the class has a positive effect. Summarize your analyses and interpretations in a short report. (*Note*: The scores could improve merely from the students feeling more comfortable with the testing process. A more appropriate design would also administer the exam twice to a control group that does not take the special class, comparing the changes for the experimental and control groups using methods of Chapter 7.)

6.40. The 49 students in a class at the University of Florida made blinded evaluations of pairs of cola drinks. For the 49 comparisons of Coke and Pepsi, Coke was preferred 29 times. In the population that this sample represents, is this strong evidence that a majority prefers one of the drinks? Refer to the following printout.

```
-------------------------------------------------
Test of parameter = 0.50 vs not = 0.50

  N  Sample prop     95.0% CI     Z-Value  P-Value
 49    0.5918     (0.454, 0.729)  1.286    0.1985
-------------------------------------------------
```

Explain how each result on this printout was obtained. Summarize results in a way that would be clear to someone who is not familiar with statistical inference.

6.41. In the 1990s, the U.S. Justice Department and other groups studied possible abuse by Philadelphia police officers in their treatment of minorities. One study, conducted by the American Civil Liberties Union, analyzed whether African-American drivers were more likely than others in the population to be targeted by police for traffic stops. Researchers studied the results of 262 police car stops during one week in 1997. Of those, 207 of the drivers were African-American, or 79% of the total. At that time, Philadelphia's population was 42.2% African-American. Does the number of African-Americans stopped give strong evidence of possible bias, being higher than you'd expect if we take into account ordinary random variation? Explain your reasoning in a report of at most 250 words.

6.42. An experiment with 26 students in an Israeli classroom consisted of giving everyone a lottery ticket, and then later asking if they would be willing to exchange their ticket for another one, plus a small monetary incentive. Only 7 students agreed to the exchange. In a separate experiment, 31 students were given a new pen and then later asked to exchange it for another pen and a small monetary incentive. All 31 agreed.[7] Conduct inferential statistical methods to analyze the data. Summarize your analyses and interpretations in a short report.

6.43. Ideally, results of a statistical analysis should not depend greatly on a single observation. To check this, it's a good idea to conduct a ***sensitivity study***: Redo the analysis after deleting an outlier from the data set or changing its value to a more typical value, and check whether results change much. For the anorexia data shown in Example 5.5 (page 120), the weight change of 20.9 pounds was a severe outlier. Suppose this observation was actually 2.9 pounds but was incorrectly recorded. Redo the one-sided test of Example 6.4 (page 151), and summarize the influence of that observation.

6.44. In making a decision in a test, a researcher worries about the possibility of rejecting H_0 when it is actually true. Explain how to control the probability of this type of error.

6.45. Consider the analogy between making a decision in a test and making a decision about the innocence or guilt of a defendant in a criminal trial.

(a) Explain what Type I and Type II errors are in the trial.

(b) Explain intuitively why decreasing P(Type I error) increases P(Type II error).

(c) Defendants are convicted if the jury finds them to be guilty "beyond a reasonable doubt." A jury interprets this to mean that if the defendant is innocent, the probability of being found guilty should be only 1 in a billion. Describe any problems this strategy has.

6.46. Medical tests for diagnosing conditions such as breast cancer are fallible, just like decisions in significance tests. Identify (H_0 true, H_0 false) with disease (absent, present), and (Reject H_0, Do not reject H_0) with diagnostic test (positive, negative), where a positive diagnosis means that the test predicts that the disease is present. Explain the difference between Type I and Type II errors in this context. Explain why decreasing P(Type I error) increases P(Type II error), in this context.

6.47. An article in a sociology journal that deals with changes in religious beliefs over time states, "For these subjects, the difference in their mean responses on the scale of religiosity between age 16 and the current survey was significant ($P < 0.05$)."

(a) Explain what it means for the result to be "significant."

(b) Explain why it would have been more informative if the authors provided the actual P-value rather than merely indicating that it is below 0.05. What other information might they have provided?

6.48. An article in a political science journal states that "no significant difference was found between men and women in their voting rates ($P = 0.63$)." Can we conclude that the population voting rates are identical for men and women? Explain.

6.49. You conduct a significance test using software. The output reports a P-value of 0.4173545. In summarizing your analyses in a research article, explain why it makes more sense to report $P = 0.42$ rather than $P = 0.4173545$.

6.50. A research study conducts 60 significance tests. Of these, 3 are significant at the 0.05 level. The authors write a report stressing only the three "significant" results, not mentioning the other 57 tests that were "not significant." Explain what is misleading about their report.

6.51. Some journals have a policy of publishing research results only if they achieve statistical significance at the 0.05 α-level.

(a) Explain the dangers of this.

(b) When medical stories in the mass media report supposed large dangers or benefits of certain agents (e.g., coffee drinking, fiber in cereal), later research often suggests that the effects

[7] M. Bar-Hillel and E. Neter, *J. Personality and Social Psych.*, vol. 70, 1996, pp. 17–27.

are smaller than first believed, or may not even exist. Explain why.

Select the correct response(s) in Exercises 6.52–6.56. (More than one may be correct.)

6.52. We analyze whether the true mean discharge of wastewater per hour from an industrial plant exceeds the company claim of 1000 gallons. For the decision in the one-sided test using $\alpha = 0.05$:

(a) If the plant is not exceeding the limit, but actually $\mu = 1000$, there is only a 5% chance that we will conclude that they are exceeding the limit.

(b) If the plant is exceeding the limit, there is only a 5% chance that we will conclude that they are not exceeding the limit.

(c) The probability that the sample mean equals exactly the observed value would equal 0.05 if H_0 were true.

(d) If we reject H_0, the probability that it is actually true is 0.05.

(e) All of the above.

6.53. The P-value for testing H_0: $\mu = 100$ against H_a: $\mu \neq 100$ is $P = 0.001$. This indicates that

(a) There is strong evidence that $\mu = 100$.

(b) There is strong evidence that $\mu \neq 100$.

(c) There is strong evidence that $\mu > 100$.

(d) There is strong evidence that $\mu < 100$.

(e) If μ were equal to 100, it would be unusual to obtain data such as those observed.

6.54. In the previous exercise, suppose the test statistic $t = 3.29$.

(a) There is strong evidence that $\mu = 100$.

(b) There is strong evidence that $\mu > 100$.

(c) There is strong evidence that $\mu < 100$.

6.55. A 95% confidence interval for μ is (96, 110). Which two statements about significance tests for the same data are correct?

(a) In testing H_0: $\mu = 100$ against H_a: $\mu \neq 100$, $P > 0.05$.

(b) In testing H_0: $\mu = 100$ against H_a: $\mu \neq 100$, $P < 0.05$.

(c) In testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, $P > 0.05$ if μ_0 is any of the numbers inside the confidence interval.

(d) In testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, $P > 0.05$ if μ_0 is any of the numbers outside the confidence interval.

6.56. Let β denote P(Type II error). For an $\alpha = 0.05$-level test of H_0: $\mu = 0$ against H_a: $\mu > 0$ with $n = 30$ observations, $\beta = 0.36$ at $\mu = 4$. Then

(a) At $\mu = 5$, $\beta > 0.36$.

(b) If $\alpha = 0.01$, then at $\mu = 4$, $\beta > 0.36$.

(c) If $n = 50$, then at $\mu = 4$, $\beta > 0.36$.

(d) The power of the test is 0.64 at $\mu = 4$.

(e) This must be false, because necessarily $\alpha + \beta = 1$.

6.57. Answer true or false for each of the following, and explain your answer:

(a) P(Type II error) = 1 − P(Type I error).

(b) If we reject H_0 using $\alpha = 0.01$, then we also reject it using $\alpha = 0.05$.

(c) The P-value is the probability that H_0 is true. (*Hint*: Do we find probabilities about variables and their statistics, or about parameters?)

(d) An article in an anthropology journal reports $P = 0.063$ for testing H_0: $\mu = 0$ against H_a: $\mu \neq 0$. If the authors had instead reported a 95% confidence interval for μ, then the interval would have contained 0, and readers could have better judged just which values are plausible for μ.

6.58. Explain the difference between one-sided and two-sided alternative hypotheses, and explain how this affects calculation of the P-value.

6.59. Explain why the terminology "do not reject H_0" is preferable to "accept H_0."

6.60. Your friend plans to survey students in your college to study whether a majority feel that the legal age for drinking alcohol should be reduced. He has never studied statistics. How would you explain to him the concepts of

(a) null and alternative hypotheses,

(b) P-value,

(c) α-level,

(d) Type II error?

6.61. A random sample of size 40 has $\bar{y} = 120$. The P-value for testing H_0: $\mu = 100$ against H_a: $\mu \neq 100$ is $P = 0.057$. Explain what is incorrect about each of the following interpretations of this P-value, and provide a proper interpretation.

(a) The probability that the null hypothesis is correct equals 0.057.

(b) The probability that $\bar{y} = 120$ if H_0 is true equals 0.057.

(c) If in fact $\mu \neq 100$, the probability equals 0.057 that the data would be at least as contradictory to H_0 as the observed data.

(d) The probability of Type I error equals 0.057.

(e) We can accept H_0 at the $\alpha = 0.05$ level.

(f) We can reject H_0 at the $\alpha = 0.05$ level.

***6.62.** Refer to the previous exercise and the P-value of 0.057.

(a) Explain why the P-value is the smallest α-level at which H_0 can be rejected; that is, P equals the smallest level at which the data are significant.

(b) Refer to the correspondence between results of confidence intervals and two-sided tests. When the P-value is 0.057, explain why the

94.3% confidence interval is the narrowest confidence interval for μ that contains $\mu_0 = 100$.

***6.63.** A researcher conducts a significance test every time she analyzes a new data set. Over time, she conducts 100 tests.

(a) Suppose H_0 is true in every case. What is the distribution of the number of times she rejects H_0 at the 0.05 level?

(b) Suppose she rejects H_0 in five of the tests. Is it plausible that H_0 is correct in every case? Explain.

***6.64.** Each year in Liverpool, New York, a public librarian estimates the mean number of times the books in that library have been checked out in the previous year. To do this, the librarian randomly samples computer records for 100 books and forms a 95% confidence interval for the mean. This has been done for 20 years.

(a) Find the probability that all the confidence intervals contain the true means. (*Hint*: Use the binomial distribution.)

(b) Find the probability that at least one confidence interval does not contain the true mean.

***6.65.** Suppose you wanted to test H_0: $\pi = 0.50$, but of the $n = 30$ observations, 0 were in the category of interest. If you found the z test statistic using the $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ for confidence intervals, show what happens to the test statistic. Explain why the $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$ is more appropriate for tests.

***6.66.** You test H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$, using $\alpha = 0.05$. In fact, H_a is true. Explain why P(Type II error) increases toward 0.95 as π moves down toward 0.50. (Assume n and α stay fixed.)

***6.67.** Refer to the ESP experiment in Exercise 6.33, with $n = 5$.

(a) For what value(s) of x = number of correct guesses can you reject H_0: $\pi = 0.50$ in favor of H_a: $\pi > 0.50$, using $\alpha = 0.05$?

(b) For what value(s) of x can you reject H_0 using $\alpha = 0.01$? (*Note*: For small samples, it may not be possible to achieve very small P-values.)

(c) Suppose you test H_0 using $\alpha = 0.05$. If $\pi = 0.50$, what is P(Type I error)? (*Note*: For discrete distributions, P(Type I error) may be less than intended. It is better to report the P-value.)

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.

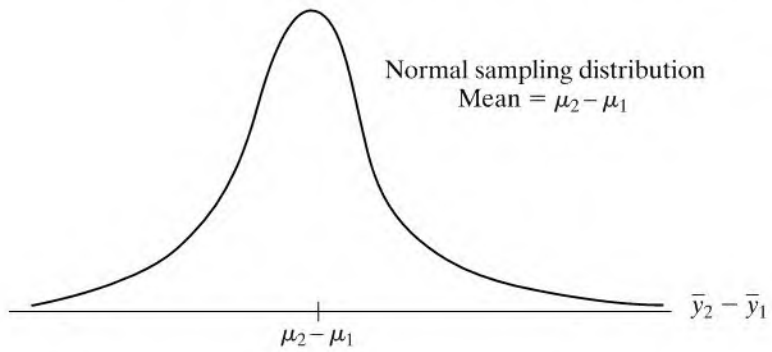

FIGURE 7.1: For Random Samples, the Sampling Distribution of the Difference between the Sample Means $\bar{y}_2 - \bar{y}_1$ Is Approximately Normal about $\mu_2 - \mu_1$

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.

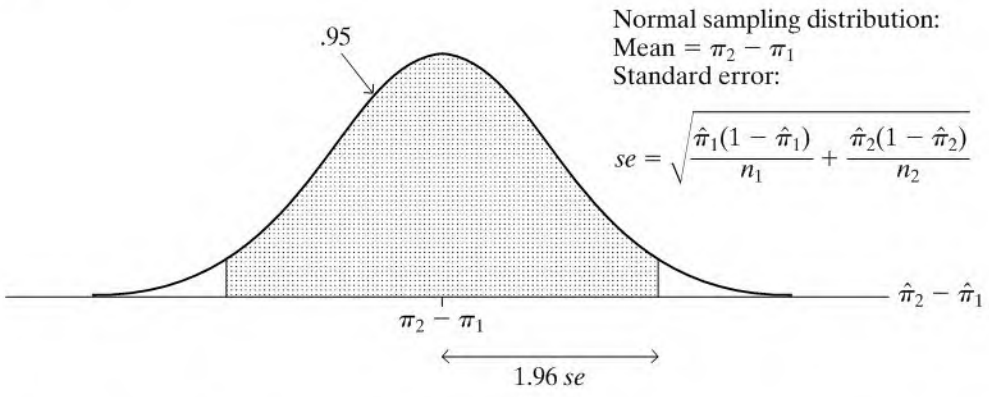

FIGURE 7.2: For Large Random Samples, the Sampling Distribution of the Estimator $\hat{\pi}_2 - \hat{\pi}_1$ of the Difference of Proportions Is Approximately Normal

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 expecially 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

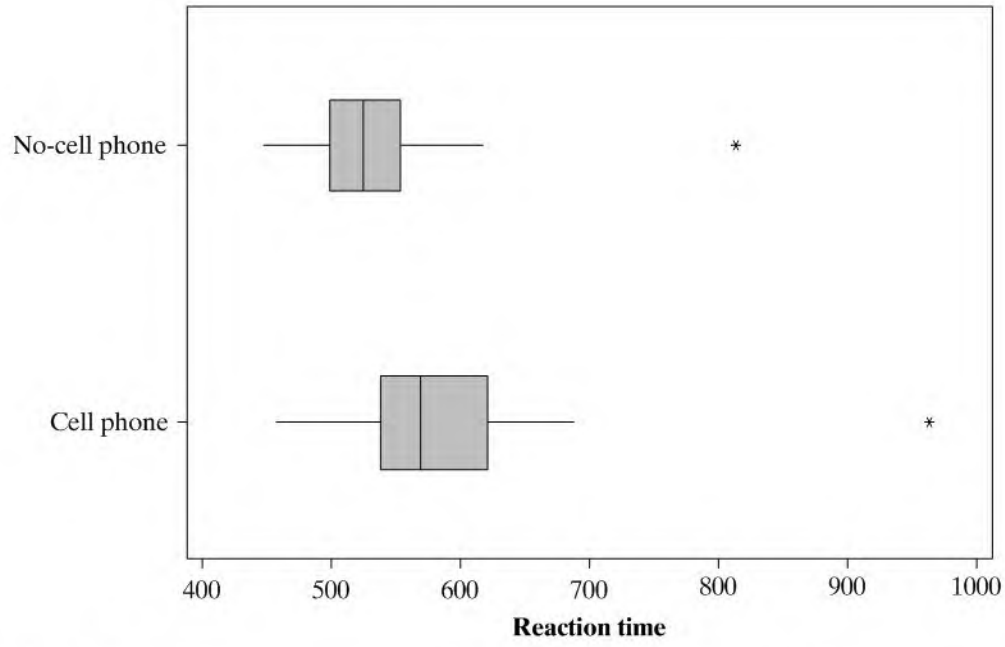

FIGURE 7.3: Box Plots of Observations for the Experiment on the Effects of Cell Phone Use on Reaction Times

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

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

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.