CHAPTER 14

Tests of Significance:
The Basics

Confidence intervals are one of the two most common types of statistical inference. Use a confidence interval when your goal is to estimate a population parameter. The second common type of inference, called tests of significance, has a different goal: to assess the evidence provided by data about some claim concerning a population. Here is the reasoning of statistical tests in a nutshell.

EXAMPLE 14.1  I’m a great free-throw shooter

I claim that I make 80% of my basketball free throws. To test my claim, you ask me to shoot 20 free throws. I make only 8 of the 20. “Aha!” you say. “Someone who makes 80% of his free throws would almost never make only 8 out of 20. So I don’t believe your claim.”

Your reasoning is based on asking what would happen if my claim were true and we repeated the sample of 20 free throws many times—I would almost never make as few as 8. This outcome is so unlikely that it gives strong evidence that my claim is not true.

You can say how strong the evidence against my claim is by giving the probability that I would make as few as 8 out of 20 free throws if I really make 80% in the long run. This probability is 0.0001. I would make as few as 8 of 20 only once in 10,000 tries in the long run if my claim to make 80% is true. The small probability convinces you that my claim is false.
Significance tests use an elaborate vocabulary, but the basic idea is simple: an outcome that would rarely happen if a claim were true is good evidence that the claim is not true.

The reasoning of statistical tests, like that of confidence intervals, is based on asking what would happen if we repeated the sample or experiment many times. We will again start with the simple conditions listed on page 321: an SRS from an exactly Normal population with standard deviation \( \sigma \) known to us. Here is an example we will explore.

**EXAMPLE 14.2 Sweetening colas**

Diet colas use artificial sweeteners to avoid sugar. These sweeteners gradually lose their sweetness over time. Manufacturers therefore test new colas for loss of sweetness before marketing them. Trained tasters sip the cola along with drinks of standard sweetness and score the cola on a “sweetness score” of 1 to 10. The cola is then stored for a month at high temperature to imitate the effect of four months’ storage at room temperature. Each taster scores the cola again after storage. This is a matched pairs experiment. Our data are the differences (score before storage minus score after storage) in the tasters’ scores. The bigger these differences, the bigger the loss of sweetness.

Suppose we know that for any cola, the sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation \( \sigma = 1 \). The mean \( \mu \) for all tasters measures loss of sweetness, and is different for different colas.

The following are the sweetness losses for a new cola, as measured by 10 trained tasters:

\[
2.0 \quad 0.4 \quad 0.7 \quad 2.0 \quad -0.4 \quad 2.2 \quad -1.3 \quad 1.2 \quad 1.1 \quad 2.3
\]

Most are positive. That is, most tasters found a loss of sweetness. But the losses are small, and two tasters (the negative scores) thought the cola gained sweetness. The average sweetness loss is given by the sample mean:

\[
\bar{x} = \frac{2.0 + 0.4 + \cdots + 2.3}{10} = 1.02
\]

Are these data good evidence that the cola lost sweetness in storage?

The reasoning is the same as in Example 14.1. We make a claim and ask if the data give evidence against it. We seek evidence that there is a sweetness loss, so the claim we test is that there is not a loss. In that case, the mean loss for the population of all trained testers would be \( \mu = 0 \).

- If the claim that \( \mu = 0 \) is true, the sampling distribution of \( \bar{x} \) from 10 tasters is Normal with mean \( \mu = 0 \) and standard deviation

\[
\sigma_{\bar{x}} = \frac{1}{\sqrt{10}} = 0.316
\]
Figure 14.1 If the cola does not lose sweetness in storage, the mean score $\bar{x}$ for 10 tasters will have this sampling distribution. The actual result for one cola was $\bar{x} = 0.3$. That could easily happen just by chance. Another cola had $\bar{x} = 1.02$. That’s so far out on the Normal curve that it is good evidence that this cola did lose sweetness.

Figure 14.1 shows this sampling distribution. We can judge whether any observed $\bar{x}$ is surprising by locating it on this distribution.

- Suppose that the 10 tasters had mean loss $\bar{x} = 0.3$. It is clear from Figure 14.1 that an $\bar{x}$ this large could easily occur just by chance when the population mean is $\mu = 0$. That 10 tasters find $\bar{x} = 0.3$ is not evidence of a sweetness loss.

- In fact, the taste test produced $\bar{x} = 1.02$. That’s way out on the Normal curve in Figure 14.1—so far out that an observed value this large would rarely occur just by chance if the true $\mu$ were 0. This observed value is good evidence that in fact the true $\mu$ is greater than 0, that is, that the cola lost sweetness. The manufacturer must reformulate the cola and try again.

**APPLY YOUR KNOWLEDGE**

14.1 Anemia. Hemoglobin is a protein in red blood cells that carries oxygen from the lungs to body tissues. People with less than 12 grams of hemoglobin per deciliter of blood (g/dl) are anemic. A public health official in Jordan suspects that the mean $\mu$ for all children in Jordan is less than 12. He measures a sample of 50 children. Suppose we know that hemoglobin level for all children this age follows a Normal distribution with standard deviation $\sigma = 1.6$ g/dl.

(a) We seek evidence against the claim that $\mu = 12$. What is the sampling distribution of $\bar{x}$ in many samples of size 50 if in fact $\mu = 12$? Make a sketch of the Normal curve for this distribution. (Sketch a Normal curve, then mark the axis using what you know about locating the mean and standard deviation on a Normal curve.)

(b) The sample mean was $\bar{x} = 11.3$. Mark this outcome on the sampling distribution. Also mark the outcome $\bar{x} = 11.8$ g/dl of a different study of 50 children. Explain carefully from your sketch
why one of these outcomes is good evidence that \( \mu \) is lower than 12, and also why the other outcome is not good evidence for this conclusion.

14.2 Student attitudes. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures students’ study habits and attitude toward school. Scores range from 0 to 200. The mean score for college students is about 115, and the standard deviation is about 30. A teacher suspects that the mean \( \mu \) for older students is higher than 115. She gives the SSHA to an SRS of 25 students who are at least 30 years old. Suppose we know that scores in the population of older students are Normally distributed with standard deviation \( \sigma = 30 \).

(a) We seek evidence against the claim that \( \mu = 115 \). What is the sampling distribution of the mean score \( \bar{x} \) of a sample of 25 students if the claim is true? Sketch the density curve of this distribution. (Sketch a Normal curve, then mark the axis using what you know about locating the mean and standard deviation on a Normal curve.)

(b) Suppose that the sample data give \( \bar{x} = 118.6 \). Mark this point on the axis of your sketch. In fact, the result was \( \bar{x} = 125.8 \). Mark this point on your sketch. Using your sketch, explain in simple language why one result is good evidence that the mean score of all older students is greater than 115 and why the other outcome is not.

Stating hypotheses

A statistical test starts with a careful statement of the claims we want to compare. In Example 14.2, we asked whether the taste test data are plausible if, in fact, there is no loss of sweetness. Because the reasoning of tests looks for evidence against a claim, we start with the claim we seek evidence against, such as “no loss of sweetness.”

**NULL HYPOTHESIS** \( H_0 \)

The statement being tested in a statistical test is called the null hypothesis. The test is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of “no effect” or “no difference.”

The claim about the population that we are trying to find evidence for is the alternative hypothesis, written \( H_1 \). In Example 14.2, we are seeking evidence of a loss in sweetness. The null hypothesis says “no loss” on the average in a large population of tasters. The alternative hypothesis says “there is a loss.” So the hypotheses are

\[
H_0: \mu = 0 \\
H_1: \mu > 0
\]
CHAPTER 14  •  Tests of Significance: The Basics

Hypotheses always refer to some population or model, not to a particular outcome. Be sure to state \( H_0 \) and \( H_a \) in terms of population parameters. Because \( H_a \) expresses the effect that we hope to find evidence for, it is often easier to begin by stating \( H_a \) and then set up \( H_0 \) as the statement that the hoped-for effect is not present. Stating \( H_0 \) is not always straightforward. It is not always clear, in particular, whether \( H_a \) should be one-sided or two-sided.

The alternative \( H_a: \mu > 0 \) is one-sided because we are interested only in whether the cola lost sweetness. Here is an example in which the alternative hypothesis is two-sided.

**EXAMPLE 14.3**  Studying job satisfaction

Does the job satisfaction of assembly workers differ when their work is machine-paced rather than self-paced? Assign workers either to an assembly line moving at a fixed pace or to a self-paced setting. All subjects work in both settings, in random order. This is a matched pairs design. After two weeks in a work setting, the workers take a test of job satisfaction. The response variable is the difference in satisfaction scores, self-paced minus machine-paced.

The parameter of interest is the mean \( \mu \) of the differences in scores in the population of all assembly workers. The null hypothesis says that there is no difference between self-paced and machine-paced work; that is,

\[
H_0: \mu = 0
\]

The authors of the study wanted to know if the two work conditions have different levels of job satisfaction. They did not specify the direction of the difference. The alternative hypothesis is therefore two-sided:

\[
H_a: \mu \neq 0
\]

The alternative hypothesis should express the hopes or suspicions we bring to the data. It is cheating to first look at the data and then frame \( H_a \) to fit what the data show. Thus, the fact that the workers in the study of Example 14.3 were more satisfied with self-paced work should not influence our choice of \( H_a \). If you do not have a specific direction firmly in mind in advance, use a two-sided alternative.

**APPLY YOUR KNOWLEDGE**

14.3 Anemia. State the null and alternative hypotheses for the anemia study described in Exercise 14.1.

14.4 Student attitudes. State the null and alternative hypotheses for the study of older students’ attitudes described in Exercise 14.2.

14.5 Gas mileage. Larry’s car averages 26 miles per gallon on the highway. He switches to a new motor oil that is advertised as increasing gas mileage. After driving 3000 highway miles with the new oil, he wants to determine if his average gas mileage has increased. What are the null and alternative hypotheses? Explain briefly what the parameter \( \mu \) in your hypotheses represents.
14.6 Diameter of a part. The diameter of a spindle in a small motor is supposed to be 5 millimeters. If the spindle is either too small or too large, the motor will not work properly. The manufacturer measures the diameter in a sample of motors to determine whether the mean diameter has moved away from the target. What are the null and alternative hypotheses? Explain briefly the distinction between the mean $\mu$ in your hypotheses and the mean $\bar{x}$ of the spindles the manufacturer measures.

Test statistics

A significance test uses data in the form of a test statistic. Here are some principles that apply to most tests:

- The test is based on a statistic that compares the value of the parameter stated by the null hypothesis with an estimate of the parameter from the sample data. The estimate is usually the same one used in a confidence interval for the parameter.
- Large values of the test statistic indicate that the estimate is far from the parameter value specified by $H_0$. These values give evidence against $H_0$.
- The alternative hypothesis determines which directions count against $H_0$.

**EXAMPLE 14.4 Sweetening colas: the test statistic**

In Example 14.2, the null hypothesis is $H_0: \mu = 0$ and the estimate of $\mu$ is $\bar{x} = 1.02$. The test statistic for hypotheses about the mean $\mu$ of a Normal distribution is the standardized version of $\bar{x}$:

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

The statistic $z$ says how far $\bar{x}$ is from $\mu$ in standard deviation units. For Example 14.2,

$$z = \frac{1.02 - 0}{1/\sqrt{10}} = 3.23$$

Because the sample result is more than 3 standard deviations above the hypothesized mean 0, it gives good evidence that the mean sweetness loss is not 0, but positive.

**APPLY YOUR KNOWLEDGE**

14.7 Sweetening colas. Figure 14.1 compares two possible results for the taste test of Example 14.2. Mean $\bar{x} = 1.02$ is far out on the Normal curve and so is good evidence against $H_0: \mu = 0$. Mean $\bar{x} = 0.3$ is not far enough out to convince us that the population mean is greater than 0. Example 14.4 shows that the test statistic is $z = 3.23$ for $\bar{x} = 1.02$. What is $z$ for $\bar{x} = 0.3$? The standard scale makes it easier to compare the two results.

14.8 Anemia. What are the values of the test statistic $z$ for the two outcomes in the anemia study of Exercise 14.1?
14.9 Student attitudes. What are the values of the test statistic \( z \) for the two outcomes for mean SSHA of older students in Exercise 14.2?

**P-values**

The null hypothesis \( H_0 \) states the claim we are seeking evidence against. The test statistic measures how far the sample data diverge from the null hypothesis. If the test statistic is large and is in the direction suggested by the alternative hypothesis \( H_a \), we have data that would be unlikely if \( H_0 \) were true. We make “unlikely” precise by calculating a probability.

**P-VALUE**

The probability, computed assuming that \( H_0 \) is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the \( P \)-value of the test. The smaller the \( P \)-value, the stronger the evidence against \( H_0 \) provided by the data.

Small \( P \)-values are evidence against \( H_0 \), because they say that the observed result is unlikely to occur when \( H_0 \) is true. Large \( P \)-values fail to give evidence against \( H_0 \).

**EXAMPLE 14.5 Sweetening colas: the \( P \)-value**

The 10 tasters in Example 14.2 found mean sweetness loss \( \bar{x} = 1.02 \). This is far from the value \( H_0: \mu = 0 \). The test statistic says just how far, in the standard scale,

\[
\frac{1.02 - 0}{\sqrt{10}} = 3.23
\]

The alternative \( H_a \) says that \( \mu > 0 \), so positive values of \( z \) favor \( H_a \) over \( H_0 \).

When \( H_0 \) is true, \( \mu = 0 \). The test statistic \( z \) is then the standardized version of the sample mean \( \bar{x} \). Because \( \bar{x} \) has a Normal distribution, \( z \) has the standard Normal distribution.

Figure 14.2 shows the \( P \)-value on the standard Normal curve that displays the distribution of \( z \). The \( P \)-value is the probability that a standard Normal variable is 3.23 or larger. Using Table A,

\[
P = P(Z > 3.23) = 1 - 0.9994 = 0.0006
\]

We would very rarely observe a sample sweetness loss as large as 1.02 if \( H_0 \) were true. The small \( P \)-value provides strong evidence against \( H_0 \) and in favor of the alternative \( H_a: \mu > 0 \).

The \( P \)-value in Example 14.5 is the probability of getting a \( z \) as large or larger than the observed \( z = 3.23 \). The alternative hypothesis sets the direction that counts as evidence against \( H_0 \). If the alternative is two-sided, both directions count. Here is an example of the \( P \)-value for a two-sided test.
The $P$-value for $z = 3.23$ is the tail area to the right of 3.23, $P = 0.0006$.

When $H_0$ is true, the test statistic $z$ has the standard Normal distribution.

**Figure 14.2** The $P$-value for the value $z = 3.23$ of the test statistic in Example 14.5. The $P$-value is the probability (when $H_0$ is true) that $z$ takes a value as large or larger than the actually observed value.

**EXAMPLE 14.6** Job satisfaction: the $P$-value

Suppose we know that differences in job satisfaction scores in Example 14.3 follow a Normal distribution with standard deviation $\sigma = 60$. If there is no difference in job satisfaction between the two work environments, the mean is $\mu = 0$. This is $H_0$.

The alternative hypothesis says simply "there is a difference," $H_a: \mu \neq 0$.

Data from 18 workers gave $x = 17$. That is, these workers preferred the self-paced environment on the average. The test statistic is

$$z = \frac{x - \mu}{\sigma/\sqrt{n}} = \frac{17 - 0}{60/\sqrt{18}} = 1.20$$

Because the alternative is two-sided, the $P$-value is the probability of getting a $z$ at least as far from 0 in either direction as the observed $z = 1.20$. As always, calculate the $P$-value taking $H_0$ to be true. When $H_0$ is true, $\mu = 0$ and $z$ has the standard Normal distribution. Figure 14.3 shows the $P$-value as an area under the standard Normal curve. It is

$$P = P(Z < -1.20 \text{ or } Z > 1.20) = 2P(Z < -1.20) = (2)(0.1151) = 0.2302$$

Values as far from 0 as $x = 17$ would happen $23\%$ of the time when the true population mean is $\mu = 0$. An outcome that would occur so often when $H_0$ is true is not good evidence against $H_0$.

The conclusion of Example 14.6 is not that $H_0$ is true. The study looked for evidence against $H_0: \mu = 0$ and failed to find strong evidence. That is all we
The two-sided $P$-value for $z = 1.20$ is the area at least 1.2 away from 0 in either direction, $P = 0.2302$.

![Figure 14.3 The $P$-value for the two-sided test in Example 14.6. The observed value of the test statistic is $z = 1.20$.](image)

can say. No doubt the mean $\mu$ for the population of all assembly workers is not exactly equal to 0. A large enough sample would give evidence of the difference, even if it is very small. Tests of significance assess the evidence against $H_0$. If the evidence is strong, we can confidently reject $H_0$ in favor of the alternative. Failing to find evidence against $H_0$ means only that the data are consistent with $H_0$, not that we have clear evidence that $H_0$ is true.

**APPLY YOUR KNOWLEDGE**

14.10 Sweetening colas. Figure 14.1 shows that the outcome $x = 0.3$ from the cola taste test is not good evidence that the mean sweetness loss is greater than 0. What is the $P$-value for this outcome? This $P$-value says, “A sample outcome this large or larger would often occur just by chance when the true mean is really 0.”

14.11 Anemia. What are the $P$-values for the two outcomes of the anemia study in Exercise 14.1? Explain briefly why these values tell us that one outcome is strong evidence against the null hypothesis and that the other outcome is not.

14.12 Student attitudes. What are the $P$-values for the two outcomes of the study of SSHA scores of older students in Exercise 14.2? Explain briefly why these values tell us that one outcome is strong evidence against the null hypothesis and that the other outcome is not.

14.13 Job satisfaction with a larger sample. Suppose that the job satisfaction study had produced exactly the same outcome $x = 17$ as in Example 14.6, but from a sample of 75 workers rather than just 18 workers. Find the test statistic $z$ and its two-sided $P$-value. Do the data give good evidence that the population mean is not zero?
Statistical significance

We sometimes take one final step to assess the evidence against $H_0$. We can compare the $P$-value with a fixed value that we regard as decisive. This amounts to announcing in advance how much evidence against $H_0$ we will insist on. The decisive value of $P$ is called the significance level. We write it as $\alpha$, the Greek significance level letter alpha. If we choose $\alpha = 0.05$, we are requiring that the data give evidence against $H_0$ so strong that it would happen no more than 5% of the time (1 time in 20 samples in the long run) when $H_0$ is true. If we choose $\alpha = 0.01$, we are insisting on stronger evidence against $H_0$, evidence so strong that it would appear only 1% of the time (1 time in 100 samples) if $H_0$ is in fact true.

“Significant” in the statistical sense does not mean “important.” It means simply “not likely to happen just by chance.” The significance level $\alpha$ makes “not likely” more exact. Significance at level 0.01 is often expressed by the statement “The results were significant ($P < 0.01$).” Here $P$ stands for the $P$-value. The actual $P$-value is more informative than a statement of significance, because it allows us to assess significance at any level we choose. For example, a result with $P = 0.03$ is significant at the $\alpha = 0.05$ level but is not significant at the $\alpha = 0.01$ level.

A P P L Y Y O U R K N O W L E D G E

14.14 Anemia. In Exercises 14.8 and 14.11, you found the $t$ test statistic and the $P$-value for the outcome $x = 11.8$ in the anemia study of Exercise 14.1. Is this outcome statistically significant at the $\alpha = 0.05$ level? At the $\alpha = 0.01$ level?  

14.15 Student attitudes. In Exercises 14.9 and 14.12, you found the $t$ test statistic and the $P$-value for the outcome $x = 125.8$ in the attitudes study of Exercise 14.2. Is this outcome statistically significant at the $\alpha = 0.05$ level? At the $\alpha = 0.01$ level?  

14.16 Protecting ultramarathon runners. Exercise 8.14 (page 212) describes an experiment designed to learn whether taking vitamin C reduces the incidence of respiratory infections among ultramarathon runners. The report of the study said: Sixty-eight percent of the runners in the placebo group reported the development of symptoms of upper respiratory tract infection after the race; this was significantly more ( $P < 0.01$) than that reported by the vitamin C-supplemented group (33%).
Tests for a population mean

There are four steps in carrying out a significance test:
1. State the hypotheses.
2. Calculate the test statistic.
3. Find the P-value.
4. State your conclusion in the context of your specific setting.

Once you have stated your hypotheses and identified the proper test, you or your computer can do Steps 2 and 3 by following a recipe. Here is the recipe for the test we have used in our examples.

\[ z \] TEST FOR A POPULATION MEAN

Draw an SRS of size \( n \) from a Normal population that has unknown mean \( \mu \) and known standard deviation \( \sigma \). To test the null hypothesis that \( \mu \) has a specified value, \( H_0: \mu = \mu_0 \), calculate the one-sample \( z \) statistic

\[ z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \]

In terms of a variable \( Z \) having the standard Normal distribution, the \( P \)-value for a test of \( H_0 \) against

\[ H_a: \mu > \mu_0 \text{ is } P(Z \geq z) \]

\[ H_a: \mu < \mu_0 \text{ is } P(Z \leq z) \]

\[ H_a: \mu \neq \mu_0 \text{ is } 2P(Z \geq |z|) \]
EXAMPLE 14.7  Executives’ blood pressures

The National Center for Health Statistics reports that the systolic blood pressure for males 35 to 44 years of age has mean 128 and standard deviation 15. The medical director of a large company looks at the medical records of 72 executives in this age group and finds that the mean systolic blood pressure in this sample is $\bar{x} = 126.07$. Is this evidence that the company’s executives have a different mean blood pressure from the general population?

Suppose we know that executives’ blood pressures follow a Normal distribution with standard deviation $\sigma = 15$.

Step 1. Hypotheses. The null hypothesis is “no difference” from the national mean $\mu_0 = 128$. The alternative is two-sided, because the medical director did not have a particular direction in mind before examining the data. So the hypotheses about the unknown mean $\mu$ of the executive population are

$H_0: \mu = 128$

$H_1: \mu \neq 128$

Step 2. Test statistic. The one-sample $z$ statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{126.07 - 128}{15/\sqrt{72}} = -1.09$$

Step 3. P-value. To help find a P-value, sketch the standard Normal curve and mark on it the observed value of $z$. Figure 14.4 shows that the $P$-value is the probability that a standard Normal variable $Z$ takes a value at least 1.09 away from zero. From Table A we find that this probability is

$$P = 2P(Z \geq 1.09) = 2(1 - 0.8621) = 0.2758$$

Figure 14.4  The $P$-value for the two-sided test in Example 14.7. The observed value of the test statistic is $z = -1.09$. 
Conclusion. More than 27% of the time, an SRS of size 72 from the
general male population would have a mean blood pressure at least as far
from 128 as that of the executive sample. The observed $x = 126.07$ is
therefore not good evidence that executives differ from other men.

The $z$ test requires that the 72 executives in the sample are an SRS from
the population of all middle-aged male executives in the company. We should
check this requirement by asking how the data were produced. If medical
records are available only for executives with recent medical problems, for ex-
ample, the data are of little value for our purpose. It turns out that all executives
are given a free annual medical exam, and that the medical director selected
72 exam results at random.

**EXAMPLE 14.8** Can you balance your checkbook?

In a discussion of the education level of the American workforce, someone says, "The
average young person can't even balance a checkbook." The National Assessment
of Educational Progress says that a score of 275 or higher on its quantitative test
(see Example 13.1 on page 321) reflects the skill needed to balance a checkbook.
The NAEP random sample of 840 young men had a mean score of $\bar{x} = 272$, a bit
below the checkbook-balancing level. Is this sample result good evidence that the
mean for all young men is less than 275? As in Example 13.1, suppose we know that
$\sigma = 60$.

**Step 1. Hypotheses.** The hypotheses are

$H_0: \mu = 275$

$H_a: \mu < 275$

**Step 2. Test statistic.** The $z$ statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{272 - 275}{60 / \sqrt{840}} = -1.45$$

**Step 3. P-value.** Because $H_a$ is one-sided on the low side, small values of $z$ count
against $H_0$. Figure 14.5 illustrates the $P$-value. Using Table A, we find that

$$P = P(Z \leq -1.45) = 0.0735$$

**Conclusion.** A mean score as low as 272 would occur about 7 times in
100 samples if the population mean were 275. This is modest evidence
that the mean NAEP score for all young men is less than 275. It is sig-
ificant at the $\alpha = 0.10$ level but not at the $\alpha = 0.05$ level.

**APPLY YOUR KNOWLEDGE**

14.17 Water quality. An environmentalist group collects a liter of water
from each of 45 random locations along a stream and measures the
amount of dissolved oxygen in each specimen. The mean is
Tests for a population mean

4.62 milligrams (mg). Is this strong evidence that the stream has a mean oxygen content of less than 5 mg per liter? (Suppose we know that dissolved oxygen varies among locations according to a Normal distribution with \( \sigma = 0.92 \) mg.)

14.18 Improving your SAT score. We suspect that on the average students will score higher on their second attempt at the SAT mathematics exam than on their first attempt. Suppose we know that the changes in score (second try minus first try) follow a Normal distribution with standard deviation \( \sigma = 50 \). Here are the results for 46 randomly chosen high school students:

<table>
<thead>
<tr>
<th>Change in Score (mg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-30</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>-47</td>
</tr>
<tr>
<td>70</td>
</tr>
<tr>
<td>-62</td>
</tr>
<tr>
<td>55</td>
</tr>
<tr>
<td>-41</td>
</tr>
<tr>
<td>-32</td>
</tr>
<tr>
<td>128</td>
</tr>
<tr>
<td>-11</td>
</tr>
<tr>
<td>-43</td>
</tr>
<tr>
<td>122</td>
</tr>
<tr>
<td>-10</td>
</tr>
<tr>
<td>56</td>
</tr>
<tr>
<td>-30</td>
</tr>
<tr>
<td>-28</td>
</tr>
<tr>
<td>-19</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>57</td>
</tr>
<tr>
<td>-14</td>
</tr>
<tr>
<td>-68</td>
</tr>
<tr>
<td>77</td>
</tr>
<tr>
<td>-63</td>
</tr>
<tr>
<td>51</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>-67</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>94</td>
</tr>
<tr>
<td>-11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>-53</td>
</tr>
<tr>
<td>-49</td>
</tr>
<tr>
<td>49</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>-24</td>
</tr>
<tr>
<td>96</td>
</tr>
<tr>
<td>120</td>
</tr>
<tr>
<td>-33</td>
</tr>
<tr>
<td>-2</td>
</tr>
<tr>
<td>-39</td>
</tr>
<tr>
<td>99</td>
</tr>
</tbody>
</table>

Do these data give good evidence that the mean change in the population is greater than zero? State hypotheses, calculate a test statistic and its \( P \)-value, and state your conclusion.

14.19 Engine crankshafts. Here are measurements (in millimeters) of a critical dimension on a sample of automobile engine crankshafts:

<table>
<thead>
<tr>
<th>Measurement (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>224.120</td>
</tr>
<tr>
<td>224.001</td>
</tr>
<tr>
<td>224.017</td>
</tr>
<tr>
<td>223.982</td>
</tr>
<tr>
<td>223.989</td>
</tr>
<tr>
<td>223.961</td>
</tr>
<tr>
<td>223.960</td>
</tr>
<tr>
<td>224.089</td>
</tr>
<tr>
<td>223.987</td>
</tr>
<tr>
<td>223.976</td>
</tr>
<tr>
<td>223.902</td>
</tr>
<tr>
<td>223.980</td>
</tr>
<tr>
<td>224.098</td>
</tr>
<tr>
<td>224.057</td>
</tr>
<tr>
<td>223.913</td>
</tr>
<tr>
<td>223.999</td>
</tr>
</tbody>
</table>
The manufacturing process is known to vary Normally with standard deviation \( \sigma = 0.060 \) mm. The process mean is supposed to be 224 mm. Do these data give evidence that the process mean is not equal to the target value 224 mm? State hypotheses and calculate a test statistic and its \( P \)-value. Are you convinced that the process mean is not 224 mm?

**P-values and significance levels**

Sometimes we demand a specific degree of evidence in order to reject the null hypothesis. A level of significance \( \alpha \) says how much evidence we require. In terms of the \( P \)-value, the outcome of a test is significant at level \( \alpha \) if \( P \leq \alpha \). Significance at any level is easy to assess once you have the \( P \)-value. When you do not use software, the \( P \)-value can be difficult to calculate. Fortunately, you can decide whether a result is statistically significant by using a table of critical values, the same table we use for confidence intervals. The table also allows you to approximate the \( P \)-value without calculation. Here is an example.

**EXAMPLE 14.9**  Is it significant?

In Example 14.8, we examined whether the mean NAEP quantitative score of young men is less than 275. The hypotheses are

\[
H_0: \mu = 275 \\
H_a: \mu < 275
\]

The \( z \)-statistic takes the value \( z = -1.45 \). How significant is the evidence against \( H_0 \)?

To determine significance, compare the observed \( z = -1.45 \) with the critical values \( z^* \) in the last row of Table C. The tail area for each \( z^* \) appears at the top of the table. The value \( z = -1.45 \) (ignoring its sign) falls between the critical values 1.282 and 1.645. Because \( z \) is farther from 0 than 1.282, the test is significant at level \( \alpha = 0.10 \). Because \( z = 1.45 \) is not farther from 0 than the critical value 1.645 for tail area 0.05, the test is not significant at level \( \alpha = 0.05 \).

Figure 14.6 locates \( z = -1.45 \) between the two tabled critical values, with minus signs added because the alternative is one-sided on the low side. The figure also shows how the critical value \( z^* = -1.645 \) separates values of \( z \) that are significant at the \( \alpha = 0.05 \) level from values that are not significant.

The \( P \)-value for \( z = -1.45 \) is the area under the curve in Figure 14.6 to the left of \(-1.45\). This area is greater than the area 0.05 to the left of \( z^* = -1.645 \) and less than the area 0.10 to the left of \( z^* = -1.282 \). We can say without any calculations that 0.05 < \( P \) < 0.10.

**EXAMPLE 14.10**  Is the concentration OK?

The analytical laboratory of Example 13.4 (page 330) is asked to evaluate the claim that the concentration of the active ingredient in a specimen is 0.86%. The lab
makes 3 repeated analyses of the specimen. The mean result is $\overline{x} = 0.8404$. The true concentration is the mean $\mu$ of the population of all analyses of the specimen. The standard deviation of the analysis process is known to be $\sigma = 0.0068$. Is there significant evidence at the 1% level that $\mu \neq 0.86$?

**Step 1. Hypotheses.** The hypotheses are

- $H_0: \mu = 0.86$
- $H_a: \mu \neq 0.86$

**Step 2. Test statistic.** The $z$ statistic is

$$ z = \frac{0.8404 - 0.86}{0.0068/\sqrt{3}} = -4.99 $$

**Step 3. Significance.** Because the alternative is two-sided, the $P$-value is the area under the standard Normal curve below $-4.99$ and above $4.99$. This area is double the area in either tail alone. For significance at level $\alpha = 0.01$, $z$ must be in the extreme $0.005 (\alpha/2)$ in either tail. Compare $z = -4.99$ (ignoring its sign) with the critical value for tail area $0.005$ from Table C. This critical value is $z^* = 2.576$. Figure 14.7 locates $z = -4.99$ and the critical values on the standard Normal curve. Because $z$ is farther from $0$ than the critical values, we have significant evidence ($P < 0.01$) that the concentration is not as claimed.

In fact, $z = -4.99$ lies beyond all the critical values in Table C. The largest critical value is 3.291, for tail area 0.0005. So we can say that the two-sided
CHAPTER 14 • Tests of Significance: The Basics

Figure 14.7 Deciding whether a $z$ statistic is significant at the $\alpha = 0.01$ level in the two-sided test of Example 14.10. The observed value $z = -4.99$ is significant because it is in the extreme 1% of the standard Normal distribution.

The test is significant at the 0.001 level, not just at the 0.01 level. Software gives the exact $P$-value as

$$P = 2P(Z \geq 4.99) = 0.0000006$$

No wonder Figure 14.7 places $z = -4.99$ so far out that the Normal curve is not visibly above its axis.

Because the practice of statistics almost always employs software that calculates $P$-values automatically, tables of critical values are becoming outdated. Tables of critical values such as Table C appear in this book for learning purposes and to rescue students without good computing facilities.

14.20 Significance. You are testing $H_0: \mu = 0$ against $H_a: \mu \neq 0$ based on an SRS of 20 observations from a Normal population. What values of the $z$ statistic are statistically significant at the $\alpha = 0.005$ level?

14.21 Significance. You are testing $H_0: \mu = 0$ against $H_a: \mu > 0$ based on an SRS of 20 observations from a Normal population. What values of the $z$ statistic are statistically significant at the $\alpha = 0.005$ level?

14.22 Testing a random number generator. A random number generator is supposed to produce random numbers that are uniformly distributed on the interval from 0 to 1. If this is true, the numbers generated come from a population with $\mu = 0.5$ and $\sigma = 0.2887$. A command to
Tests from confidence intervals

The calculation in Example 14.10 for a 1% significance test is very similar to that in Example 13.3 for a 99% confidence interval. In fact, a two-sided test at significance level \( \alpha \) can be carried out directly from a confidence interval with confidence level \( C = 1 - \alpha \).

**CONFIDENCE INTERVALS AND TWO-SIDED TESTS**

A level \( \alpha \) two-sided significance test rejects a hypothesis \( H_0: \mu = \mu_0 \) exactly when the value \( \mu_0 \) falls outside a level \( 1 - \alpha \) confidence interval for \( \mu \).

**EXAMPLE 14.11** Tests from a confidence interval

The 99% confidence interval for \( \mu \) in Example 13.3 is

\[
\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 0.8404 \pm 0.0101
\]

\[
= 0.8303 \text{ to } 0.8505
\]

The hypothesized value \( \mu_0 = 0.86 \) in Example 14.10 falls outside this confidence interval, so we reject

\[ H_0: \mu = 0.86 \]
CHAPTER 14 * Tests of Significance: The Basics

### Figure 14.8
Values of \( \mu \) falling outside a 99% confidence interval can be rejected at the 1% significance level. Values falling inside the interval cannot be rejected.

at the 1% significance level. On the other hand, we cannot reject \( H_0: \mu = 0.85 \) at the 1% level in favor of the two-sided alternative \( H_a: \mu \neq 0.85 \), because 0.85 lies inside the 99% confidence interval for \( \mu \). Figure 14.8 illustrates both cases.

#### APPLY YOUR KNOWLEDGE

14.24 Test and confidence interval. The \( P \)-value for a two-sided test of the null hypothesis \( H_0: \mu = 10 \) is 0.06.
(a) Does the 95% confidence interval include the value 10? Why?
(b) Does the 90% confidence interval include the value 10? Why?

14.25 Confidence interval and test. A 95% confidence interval for a population mean is 31.5 ± 3.5.
(a) Can you reject the null hypothesis that \( \mu = 34 \) at the 5% significance level? Why?
(b) Can you reject the null hypothesis that \( \mu = 36 \) at the 5% significance level? Why?

### Chapter 14 SUMMARY

A test of significance assesses the evidence provided by data against a null hypothesis \( H_0 \) in favor of an alternative hypothesis \( H_a \).

Hypotheses are stated in terms of population parameters. Usually, \( H_0 \) is a statement that no effect is present, and \( H_a \) says that a parameter differs from its null value in a specific direction (one-sided alternative) or in either direction (two-sided alternative).

The essential reasoning of a significance test is as follows. Suppose for the sake of argument that the null hypothesis is true. If we repeated our data production many times, would we often get data as inconsistent with \( H_0 \) as the data we actually have? If the data are unlikely when \( H_0 \) is true, they provide evidence against \( H_0 \).

A test is based on a test statistic. The \( P \)-value is the probability, computed supposing \( H_0 \) to be true, that the test statistic will take a value at least as
extreme as that actually observed. Small P-values indicate strong evidence against \( H_0 \). Calculating P-values requires knowledge of the sampling distribution of the test statistic when \( H_0 \) is true.

If the P-value is as small or smaller than a specified value \( \alpha \), the data are statistically significant at significance level \( \alpha \).

Significance tests for the hypothesis \( H_0: \mu = \mu_0 \) concerning the unknown mean \( \mu \) of a population are based on the one-sample \( z \) statistic

\[
\xi = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}
\]

The \( \xi \) test assumes an SRS of size \( n \), known population standard deviation \( \sigma \), and either a Normal population or a large sample. P-values are computed from the Normal distribution (Table A). Fixed \( \alpha \) tests use the table of standard Normal critical values (bottom row of Table C).

**Chapter 14 EXERCISES**

14.26 Student study times. A student group claims that first-year students at a university must study 2.5 hours per night during the school week. A skeptic suspects that they study less than that on the average. A class survey finds that the average study time claimed by 269 students is \( \bar{x} = 137 \) minutes. Regard these students as a random sample of all first-year students and suppose we know that study times follow a Normal distribution with standard deviation 65 minutes. Carry out a test of \( H_0: \mu = 150 \) against \( H_a: \mu < 150 \). What do you conclude?

14.27 IQ test scores. Exercise 13.6 (page 329) gives the IQ test scores of 31 seventh-grade girls in a Midwest school district. IQ scores follow a Normal distribution with standard deviation \( \sigma = 15 \). Treat these 31 girls as an SRS of all seventh-grade girls in this district. IQ scores in a broad population are supposed to have mean \( \mu = 100 \). Is there evidence that the mean in this district differs from 100? State hypotheses, find the test statistic and its P-value, and state your conclusion.

14.28 This wine stinks. Sulfur compounds cause “off-odors” in wine, so winemakers want to know the odor threshold, the lowest concentration of a compound that the human nose can detect. The odor threshold for dimethyl sulfide (DMS) in trained wine tasters is about 25 micrograms per liter of wine (\( \mu g/l \)). The untrained noses of consumers may be less sensitive, however. Here are the DMS odor thresholds for 10 untrained students:

| 31 | 31 | 43 | 36 | 23 | 34 | 32 | 30 | 20 | 24 |

Assume that the odor threshold for untrained noses is Normal with \( \sigma = 7 \) \( \mu g/l \). Is there evidence that the mean threshold for untrained tasters is greater than 25 \( \mu g/l \)?
14.29 Healing of skin wounds. Exercise 13.14 (page 334) gives data and information about the rate at which skin wounds heal in newts. A newt expert says that 25 micrometers per hour is the usual rate. Do the data give evidence against this claim?

14.30 $P$ without pain. You can approximate $P$ for any $z$ by comparing $z$ with the critical values in the bottom row of Table C. Between what values from Table C does the $P$-value for your $z$ in Exercise 14.26 lie?

14.31 $P$ without pain. Use Table C to approximate the $P$-value for the $z$ you obtained in Exercise 14.27 without a probability calculation. That is, give two values from the table between which $P$ must lie.

14.32 Tracking the placebo effect. The placebo effect (see Chapter 8, page 202) is particularly strong in patients with Parkinson’s disease. To understand the workings of the placebo effect, scientists made chemical measurements at a key point in the brain when patients received a placebo that they thought was an active drug and also when no treatment was given. They hoped to find that the placebo reduced the mean response. State $H_0$ and $H_a$ for the significance test.

14.33 Tracking the placebo effect, continued. The report of the study described in the previous exercise says that the placebo reduced the chemical response by an average of 17% and that $P < 0.005$. What can you conclude?

14.34 Fortified breakfast cereals. The Food and Drug Administration recommends that breakfast cereals be fortified with folic acid. In a matched pairs study, volunteers ate either fortified or unfortified cereal for some time, then switched to the other cereal. The response variable is the difference in blood folic acid, fortified minus unfortified. Does eating fortified cereal raise the level of folic acid in the blood? State $H_0$ and $H_a$ for a test to answer this question. State carefully what the parameter $\mu$ in your hypotheses is.

14.35 What’s the $P$-value? A test of the null hypothesis $H_0: \mu = 0$ gives test statistic $z = 1.8$.
(a) What is the $P$-value if the alternative is $H_a: \mu > 0$?
(b) What is the $P$-value if the alternative is $H_a: \mu < 0$?
(c) What is the $P$-value if the alternative is $H_a: \mu \neq 0$?

14.36 $P$ and significance. The $P$-value for a significance test is 0.078.
(a) Do you reject the null hypothesis at level $\alpha = 0.05$? Explain your answer.
(b) Do you reject the null hypothesis at level $\alpha = 0.01$? Explain your answer.

14.37 $P$ and significance. The $P$-value for a significance test is 0.033.
(a) Do you reject the null hypothesis at level $\alpha = 0.05$? Explain your answer.
(b) Do you reject the null hypothesis at level $\alpha = 0.01$? Explain your answer.

14.38 The Supreme Court speaks. Court cases in such areas as employment discrimination often involve statistical evidence. The Supreme Court has said that $z$-scores beyond $z^* = 2$ or 3 are generally convincing statistical evidence. For a two-sided test, what significance level $\alpha$ corresponds to $z^* = 2$? To $z^* = 3$?

14.39 Diet and diabetes. Does eating more fiber reduce the blood cholesterol level of patients with diabetes? A randomized clinical trial compared normal and high-fiber diets. Here is part of the researchers’ conclusion:

The high-fiber diet reduced plasma total cholesterol concentrations by 6.7 percent ($P = 0.02$), triglyceride concentrations by 10.2 percent ($P = 0.02$), and very-low-density lipoprotein cholesterol concentrations by 12.5 percent ($P = 0.01$).

A doctor who knows no statistics says that a drop of 6.7% in cholesterol isn’t a lot—maybe it’s just an accident due to the chance assignment of patients to the two diets. Explain in simple language how “$P = 0.02$” answers this objection.

14.40 Diet and bowel cancer. It has long been thought that eating a healthier diet reduces the risk of bowel cancer. A large study cast doubt on this advice. The subjects were 2079 people who had polyps removed from their bowels in the past six months. Such polyps may lead to cancer. The subjects were randomly assigned to a low-fat, high-fiber diet or to a control group in which subjects ate their usual diets. All subjects were checked for polyps over the next four years.

(a) Outline the design of this experiment.

(b) Surprisingly, the occurrence of new polyps “did not differ significantly between the two groups.” Explain clearly what this finding means.

14.41 How to show that you are rich. Every society has its own marks of wealth and prestige. In ancient China, it appears that owning pigs was such a mark. Evidence comes from examining burial sites. The skulls of sacrificed pigs tend to appear along with expensive ornaments, which suggests that the pigs, like the ornaments, signal the wealth and prestige of the person buried. A study of burials from around 3500 B.C. concluded that “there are striking differences in grave goods between burials with pig skulls and burials without them…. A test indicates that the two samples of total artifacts are significantly different at the 0.01 level.” Explain clearly why “significantly different at the 0.01 level” gives good reason to think that there really is a systematic difference between burials that contain pig skulls and those that lack them.
14.42 Forests and windstorms. Does the destruction of large trees in a windstorm change forests in any important way? Here is the conclusion of a study that found that the answer is no:

We found surprisingly little divergence between treefall areas and adjacent control areas in the richness of woody plants ($P = 0.62$), in total stem densities ($P = 0.98$), or in population size or structure for any individual shrub or tree species. The two $P$-values refer to null hypotheses that say "no change" in measurements between treefall and control areas. Explain clearly why these values provide no evidence of change.

14.43 Reporting $P$. The report of a study of seat belt use by drivers says, "Hispanic drivers were not significantly more likely than White/non-Hispanic drivers to overreport safety belt use (27.4 vs. 21.1%, respectively; $z = 1.33, P = 0.09$)." How do you know that the $P$-value given is incorrect? What is the correct one-sided $P$-value for test statistic $z = 1.33$?

14.44 5% versus 1%. Make a sketch that shows why a value of the $z$ test statistic that is significant at the 1% level must always be significant at the 5% level. If $z$ is significant at the 5% level, what can you say about its significance at the 1% level?

14.45 Is this what $P$ means? When asked to explain the meaning of "the $P$-value was $P = 0.03$," a student says, "This means there is only probability 0.03 that the null hypothesis is true." Is this an essentially correct explanation? Explain your answer.

14.46 Is this what significance means? Another student, when asked why statistical significance appears so often in research reports, says, "Because saying that results are significant tells us that they cannot easily be explained by chance variation alone." Do you think that this statement is essentially correct? Explain your answer.

14.47 Workers’ earnings. The Bureau of Labor Statistics generally uses 90% confidence in its reports. One report gives a 90% confidence interval for the mean hourly earnings of American workers in 2000 as $15.49 to $16.11. This result was calculated from the National Compensation Survey, a multistage probability sample of businesses.

(a) Would a 95% confidence interval be wider or narrower?
(b) Would the null hypothesis that the 2000 mean hourly earnings of all workers was $16 be rejected at the 10% significance level in favor of the two-sided alternative? What about the null hypothesis that the mean was $15$?

14.48 Pulling wood apart. In Exercise 13.15 (page 334), you found a 90% confidence interval for the mean load required to pull apart pieces of wood.
Chapter 14 Media Exercises

14.49 I’m a great free-throw shooter. The Test of Significance applet animates Example 14.1. That example asks if a basketball player’s actual performance gives evidence against the claim that he or she makes 80% of free throws. The parameter in question is the percent \( p \) of free throws that the player will make if he or she shoots free throws forever. The population is all free throws the player will ever shoot. The null hypothesis is always the same, that the player makes 80% of shots taken:

\[ H_0: p = 80\% \]

The applet does not do a formal statistical test. Instead, it allows you to ask the player to shoot until you are reasonably confident that the true percent of hits is or is not very close to 80%.

I claim that I make 80% of my free throws. To test my claim, we go to the gym and I shoot 20 free throws. Set the applet to take 20 shots. Check “Show null hypothesis” so that my claim is visible in the graph.

(a) Click “Shoot.” How many of the 20 shots did I make? Are you convinced that I really make less than 80%?

(b) If you are not convinced, click “Shoot” again for 20 more shots. Keep going until either you are convinced that I don’t make 80% of my shots or it appears that my true percent made is pretty close to 80%. How many shots did you watch me shoot? How many did I make? What did you conclude? Then click “Show true %” to reveal the truth. Was your conclusion correct?

Comment: You see why statistical tests say how strong the evidence is against some claim. If I make only 10 of 40 shots, you are pretty sure I can’t make 80% in the long run. But even if I make exactly 80 of 100, my true long-term percent might be 78% or 81% instead of 80%. It’s hard to be convinced that I make exactly 80%.

14.50 Significance at the 0.0125 level. The Normal Curve applet allows you to find critical values of the standard Normal distribution and to visualize the values of the \( z \) statistic that are significant at any level.
Max is interested in whether a one-sided $z$ test is statistically significant at the $\alpha = 0.0125$ level. Use the Normal Curve applet to tell Max what values of $z$ are significant. Sketch the standard Normal curve marked with the values that led to your result.

14.51 Significance at the 0.011 level. Jaran wants to know if a two-sided $z$ test statistic is significant at the $\alpha = 0.011$ level. Use the Normal Curve applet to say what values of $z$ are significant at this level. Sketch the standard Normal curve marked with the values that led to your result.