The $z$ test assumes an SRS of size $n$ from a Normal population with known population standard deviation $\sigma$. $P$-values can be obtained either with computations from the standard Normal distribution or by using technology (applet or software).

**LINK IT**

In this chapter we discuss tests of significance, the second type of statistical inference. The mathematics of probability, in particular the sampling distributions discussed in Chapter 11, provides the formal basis for a test of significance. The sampling distribution allows us to assess "probabilistically" the strength of evidence against a null hypothesis, either through a level of significance or a $P$-value. The goal of hypothesis testing, which is used to assess the evidence provided by data about some claim concerning a population, is different from the goal of confidence interval estimation, which is used to estimate a population parameter.

Although we apply the reasoning of tests of significance for the mean of a population that has a Normal distribution in a simple and artificial setting (we assume that we know the population standard deviation), we will use the same logic in future chapters to construct tests of significance for population parameters in more realistic settings.

**CHECK YOUR SKILLS**

15.19 You use software to carry out a test of significance. The program tells you that the $P$-value is $P = 0.031$. You conclude

(a) that 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 0.031.
(b) that the probability, computed assuming that $H_0$ is true, that the test statistic would take a value as extreme or less extreme than that actually observed is 0.031.
(c) that the probability, computed assuming that $H_0$ is false, that the test statistic would take a value as extreme or more extreme than that actually observed is 0.031.

15.20 You use software to carry out a test of significance. The program tells you that the $P$-value is $P = 0.031$. This result is

(a) not significant at either $\alpha = 0.05$ or $\alpha = 0.01$.
(b) significant at $\alpha = 0.05$ but not at $\alpha = 0.01$.
(c) significant at both $\alpha = 0.05$ and $\alpha = 0.01$.

15.21 The $z$ statistic for a one-sided test is $z = 2.433$. This test is

(a) not significant at either $\alpha = 0.05$ or $\alpha = 0.01$.
(b) significant at $\alpha = 0.05$ but not at $\alpha = 0.01$.
(c) significant at both $\alpha = 0.05$ and $\alpha = 0.01$.

15.22 The gas mileage for a particular model car is known to have a standard deviation of $\sigma = 1.0$ miles per gallon in repeated tests in a controlled laboratory environment at a fixed speed. For a fixed speed, gas mileages in repeated tests are Normally distributed. Tests on three cars of this model at 35 miles per hour give gas mileages of 29.3, 29.9, and 29.8 miles per gallon. The $z$ statistic for testing $H_0: \mu = 30$ miles per gallon based on these three measurements is

(a) $z = 0.286$.  (b) $z = 0.5$.  (c) $z = -0.286$.

15.23 Experiments on learning in animals sometimes measure how long it takes mice to find their way through a maze. The mean time is 18 seconds for one particular maze. A researcher thinks that a loud noise will cause the mice to complete the maze faster. She measures how long each of 10 mice takes with a noise as stimulus. The sample mean is $\bar{x} = 16.5$ seconds. The null hypothesis for the significance test is

(a) $H_0: \mu = 18$.  (b) $H_0: \mu = 16.5$.  (c) $H_0: \mu < 18$.  

[Image: Garry Gay/Getty Images]
15.24 The alternative hypothesis for the test in Exercise 15.23 is
(a) $H_a: \mu \neq 18.$  \hspace{1cm} (b) $H_a: \mu < 18.$  \hspace{1cm} (c) $H_a: \mu = 16.5.$

15.25 You read an article about an experiment in which the researcher conducted a test of significance. The article tells you that the $P$-value is $P = 0.19$. This means that
(a) the probability that the null hypothesis is true is 0.19.
(b) the value of the test statistic is not particularly large.
(c) neither of the above.

15.26 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?
(a) All values for which $z > 2.576$
(b) All values for which $z > 2.807$
(c) All values for which $|z| > 2.807$

15.27 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?
(a) All values for which $z > 2.576$
(b) All values for which $z > 2.807$
(c) All values for which $|z| > 2.807$

15.28 Student study times. Exercise 14.19 (page 365) describes a class survey in which students claimed to study an average of $\bar{x} = 118$ minutes on a typical weeknight. Regard these students as an SRS from the population of all first-year students at this university. Does the study give good evidence that students claim to study less than 2 hours per night on the average?
(a) State null and alternative hypotheses in terms of the mean study time in minutes for the population.
(b) What is the value of the test statistic $z$?
(c) What is the $P$-value of the test? Can you conclude that students do claim to study less than 2 hours per weeknight on the average?

15.29 I want more muscle. If young men thought that their own level of muscle was about what women prefer, the mean "muscle gap" in the study described in Exercise 14.20 (page 365) would be 0. We suspect (before seeing the data) that young men think women prefer more muscle than they themselves have.
(a) State null and alternative hypotheses for testing this suspicion.
(b) What is the value of the test statistic $z$?
(c) You can tell just from the value of $z$ that the evidence in favor of the alternative is very strong (that is, the $P$-value is very small). Explain why this is true.

15.30 Hotel managers’ personalities. Successful hotel managers must have personality characteristics often thought of as feminine (such as “compassionate”) as well as those often thought of as masculine (such as “forceful”). The Bern Sex-Role Inventory (BSRI) is a personality test that gives separate ratings for female and male stereotypes, both on a scale of 1 to 7. A sample of 148 male general managers of three-star and four-star hotels had mean BSRI femininity score $\bar{x} = 5.29$. The mean score for the general male population is $\mu = 5.19$. Do hotel managers, on the average, differ significantly in femininity score from men in general? Assume that the standard deviation of scores in the population of all male hotel managers is the same as the $\sigma = 0.78$ for the adult male population.
(a) State null and alternative hypotheses in terms of the mean femininity score $\mu$ for male hotel managers.
(b) Find the $z$ test statistic.
(c) What is the $P$-value for your $z$? What do you conclude about male hotel managers?

15.31 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.” Explain what $P = 0.03$ really means in a way that makes it clear that the student’s explanation is wrong.

15.32 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.

15.33 Alleviating test anxiety. Research suggests that pressure to perform well can reduce performance on exams. Are there effective strategies to deal with pressure? In an experiment, researchers had students take a test on mathematical skills. The same students were asked to take a second test on the same skills, but now each student was paired with a partner and only if both improved their scores would they receive a monetary reward for participating in the experiment. They were also told that their performance would be videotaped and watched by teachers and students. To help them cope with the pressure, ten minutes before the second exam they were asked to write as candidly as possible about their thoughts and feelings regarding the exam. "Students who expressed their thoughts before the high-pressure test showed a significant 5% math accuracy improvement from the pretest to posttest" (P < 0.03). A colleague who knows no statistics says that an increase of 5% isn’t a lot—maybe it’s just an accident due to natural variation among the students. Explain in simple language how "P < 0.03" answers this objection.

15.34 Treating Parkinson’s disease. A randomized comparative experiment compared the effects of two types of deep-brain stimulation (pallidal stimulation and subthalamic stimulation) on change in motor function, as blindly assessed on the Unified Parkinson’s Disease Rating Scale, part III (UPDRS-III). The abstract of the study said: "Mean changes in the primary outcome did not differ significantly between the two study groups (P = 0.50)." The P-value refers to a null hypothesis of "no change" in measurements between pallidal stimulation and subthalamic stimulation. Explain clearly why this value provides no evidence of change.

15.35 5% versus 1%. Sketch the standard Normal curve for the z test statistic and mark off areas under the curve to show why a value of z that is significant at the 1% level in a one-sided test is always significant at the 5% level. If z is significant at the 5% level, what can you say about its significance at the 1% level?

15.36 The wrong alternative. A graduate student is comparing final-exam test scores of male and female students in an introductory physics class. She starts with no expectations as to which sex will score more highly. After seeing that men did better than women on the first quiz, she tests a one-sided alternative about the mean final-exam scores,

\[ H_0: \mu_M = \mu_F \]
\[ H_1: \mu_M > \mu_F \]

She finds \( z = 1.9 \) with one-sided P-value \( P = 0.0287 \).
(a) Explain why she should have used the two-sided alternative hypothesis.
(b) What is the correct P-value for \( z = 1.9 \)?

15.37 The wrong 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 over-report safety belt use (27.4 vs. 21.1%, respectively; \( z = 1.33, P > 1.0 \))." How do you know that the P-value given is incorrect? What is the correct one-sided P-value for test statistic \( z = 1.33 \)?

Exercises 15.38 to 15.41 ask you to answer questions from data. Assume that the "simple conditions" hold in each case. The exercise statements give you the State step of the four-step process. In your work, follow the Plan, Solve, and Conclude steps, illustrated in Example 14.3 (page 359) for a confidence interval and in Example 15.6 (page 380) for a test of significance.

15.38 Pulling wood apart. How heavy a load (pounds) is needed to pull apart pieces of Douglas fir 4 inches long and 1.5 inches square? Here are data from students doing a laboratory exercise:

<table>
<thead>
<tr>
<th>Load (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>33,190</td>
</tr>
<tr>
<td>31,860</td>
</tr>
<tr>
<td>32,590</td>
</tr>
<tr>
<td>26,520</td>
</tr>
<tr>
<td>32,820</td>
</tr>
<tr>
<td>32,320</td>
</tr>
<tr>
<td>33,020</td>
</tr>
<tr>
<td>32,030</td>
</tr>
<tr>
<td>30,460</td>
</tr>
<tr>
<td>32,700</td>
</tr>
<tr>
<td>23,040</td>
</tr>
<tr>
<td>30,930</td>
</tr>
<tr>
<td>32,720</td>
</tr>
<tr>
<td>33,650</td>
</tr>
<tr>
<td>32,340</td>
</tr>
<tr>
<td>24,050</td>
</tr>
<tr>
<td>30,170</td>
</tr>
<tr>
<td>31,300</td>
</tr>
<tr>
<td>28,730</td>
</tr>
<tr>
<td>31,920</td>
</tr>
</tbody>
</table>

We are willing to regard the wood pieces prepared for the lab session as an SRS of all similar pieces of Douglas fir. Engineers also commonly assume that characteristics of materials vary Normally. Suppose that the strength of pieces of wood like these follows a Normal distribution with standard deviation 3000 pounds.
(a) Is there significant evidence at the \( \alpha = 0.10 \) level against the hypothesis that the mean is 32,500 pounds for the two-tailed alternative?
(b) Is there significant evidence at the \( \alpha = 0.10 \) level against the hypothesis that the mean is 31,500 pounds for the two-tailed alternative?

15.39 Bone loss by nursing mothers. As discussed in Exercise 14.26 (page 366), breast-feeding mothers secrete calcium into their milk. Some of the calcium
may come from their bones, so mothers may lose bone mineral. Researchers measured the percent change in mineral content of the spines of 47 mothers during three months of breast-feeding. Here are the data:

\[-4.7 -2.5 -4.9 -2.7 -0.8 -5.3 -8.3 -2.1 -6.8 -4.3
2.2 -7.7 -3.1 -1.0 -6.5 -1.8 -5.2 -5.7 -7.0 -2.2
-6.5 -1.6 -3.0 -3.6 -5.2 -2.0 -2.1 -5.6 -4.4 -3.3
-4.0 -4.9 -4.7 -3.8 -5.9 -2.5 -0.3 -6.2 -6.8 1.7
0.3 -2.3 0.4 -5.3 0.2 -2.2 -5.1\]

The researchers are willing to consider these 47 women as an SRS from the population of all nursing mothers. Suppose that the percent change in this population has a Normal distribution with standard deviation \( \sigma = 2.5\% \). Do these data give good evidence that, on the average, nursing mothers lose bone mineral?

15.40 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:

\[30 30 42 35 22 33 31 29 19 23\]

Assume that the odor threshold for untrained noses is Normally distributed with \( \sigma = 7 \mu g/l \). Is there evidence that the mean threshold for untrained tasters is greater than 25 \( \mu g/l \)?

15.41 Eye grease. Athletes performing in bright sunlight often smear black eye grease under their eyes to reduce glare. Does eye grease work? In one study, 16 student subjects took a test of sensitivity to contrast after 3 hours facing into bright sun, both with and without eye grease. This is a matched pairs design. Here are the differences in sensitivity, with eye grease minus without eye grease:

\[0.07 0.64 -0.12 -0.05 -0.18 0.14 -0.16 0.03
0.05 0.02 0.43 0.24 -0.11 0.28 0.05 0.29\]

We want to know whether eye grease increases sensitivity on the average.

(a) What are the null and alternative hypotheses? Say in words what mean \( \mu \) your hypotheses concern.
(b) Suppose that the subjects are an SRS of all young people with normal vision, that contrast differences follow a Normal distribution in this population, and that the standard deviation of differences is \( \sigma = 0.22 \). Carry out a test of significance.

15.42 Tests from confidence intervals. A confidence interval for the population mean \( \mu \) tells us which values of \( \mu \) are plausible (those inside the interval) and which values are not plausible (those outside the interval) at the chosen level of confidence. You can use this idea to carry out a test of any null hypothesis \( H_0: \mu = \mu_0 \) starting with a confidence interval: reject \( H_0 \) if \( \mu_0 \) is outside the interval and fail to reject if \( \mu_0 \) is inside the interval.

The alternative hypothesis is always two-sided, \( H_1: \mu \neq \mu_0 \) because the confidence interval extends in both directions from \( \bar{x} \). A 95% confidence interval leads to a test at the 5% significance level because the interval is wrong 5% of the time. In general, confidence level \( C \) leads to a test at significance level \( \alpha = 1 - C \).

(a) In Example 15.6 (page 380), a medical director found mean blood pressure \( \bar{x} = 126.07 \) for an SRS of 72 executives. The standard deviation of the blood pressures of all executives is \( \sigma = 15 \). Give a 90% confidence interval for the mean blood pressure \( \mu \) of all executives.
(b) The hypothesized value \( \mu_0 = 128 \) falls inside this confidence interval. Carry out the \( z \) test for \( H_0: \mu = 128 \) against the two-sided alternative. Show that the test is not significant at the 10% level.
(c) The hypothesized value \( \mu_0 = 129 \) falls outside this confidence interval. Carry out the \( z \) test for \( H_0: \mu = 129 \) against the two-sided alternative. Show that the test is significant at the 10% level.

15.43 Tests from confidence intervals. A 95% confidence interval for a population mean is 30.7 ± 3.2. Use the method described in the previous exercise to answer these questions.

(a) With a two-sided alternative, can you reject the null hypothesis that \( \mu = 33 \) at the 5% (\( \alpha = 0.05 \)) significance level? Why?
(b) With a two-sided alternative, can you reject the null hypothesis that \( \mu = 34 \) at the 5% significance level? Why?
15.44 Significance in journals. Choose a major journal in your field of study. Use a Web search engine to find its Web site—just search on the journal's name. Find a paper that uses a phrase like "significant (P < 0.01)" and summarize the findings in the paper.

15.45 A statistics glossary. An editorial was published in the Journal of the National Cancer Institute, Vol. 101, No. 23 (December 2, 2009) that announced some online resources for journalists, including a statistics glossary. The glossary can be found at www.oxfordjournals.org/our_journals/jnc/resource/statistics%20glossary.pdf. Read the definition of a P-value. Is this an accurate definition? Explain your answer.