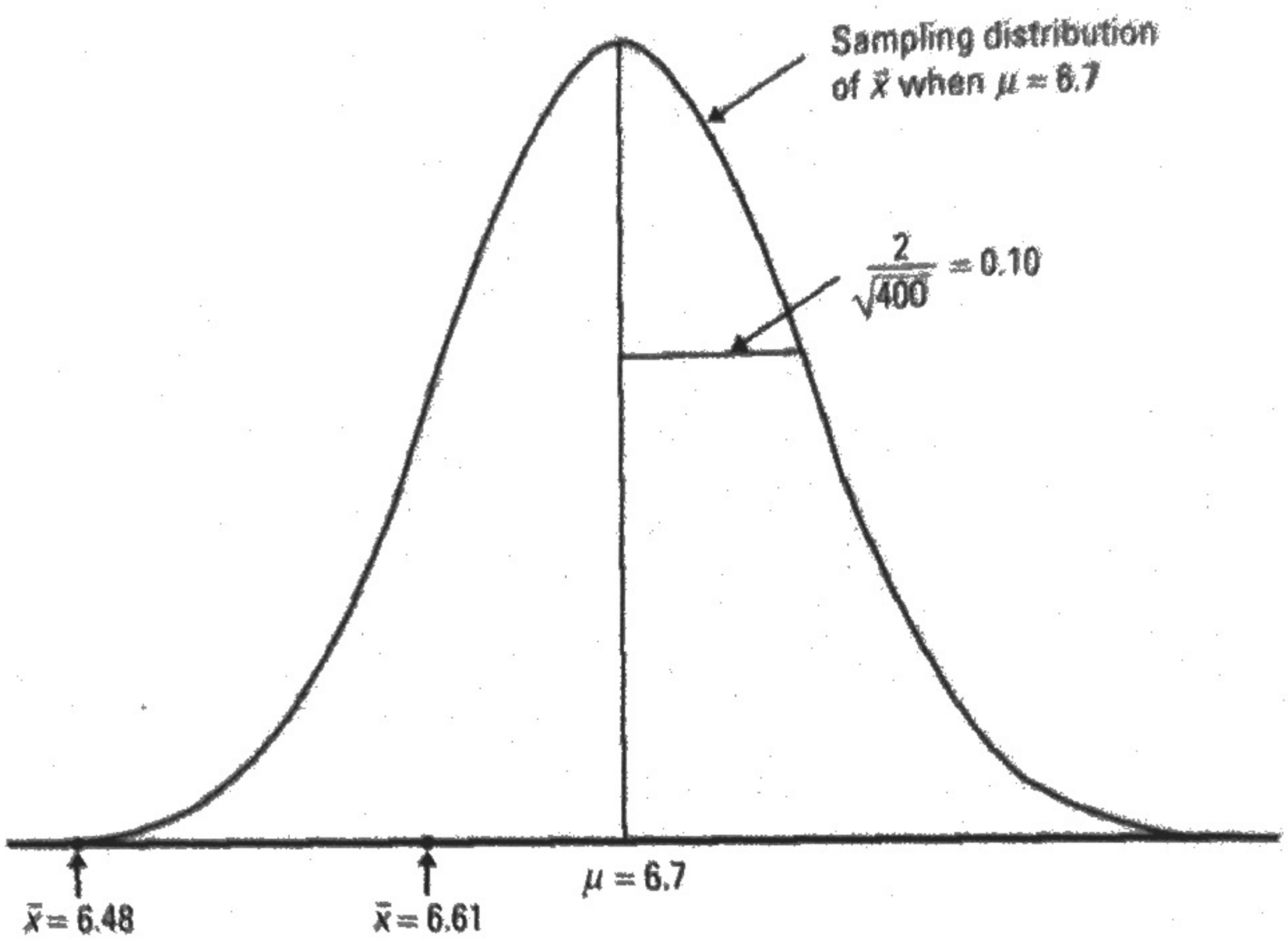
Figure 11.1 *If response times have not decreased, the mean response time $\bar{x}$ for 400 calls will have this sampling distribution. If the result had been $\bar{x} = 6.61$ minutes, that could easily happen just by chance. But the actual result was $\bar{x} = 6.48$ minutes. That's so far out on the Normal curve that it's good evidence of a decrease in paramedics' response times.*



Exercises

11.1 Student attitudes, I The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures students' attitudes toward school and study habits. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher suspects that older students have better attitudes toward school. She gives the SSHA to a random sample of 25 students at her college who are at least 30 years of age. Assume that scores in the population of older students are Normally distributed with standard deviation $\sigma = 30$.

(a) Carefully define the parameter μ in this setting.

(b) 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 older students if the null hypothesis is true? 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.)

(c) Suppose that the sample data give $\bar{x} = 118.6$. Mark this point on the axis of your sketch. In fact, the result was $\bar{x} = 125.7$. 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.

(d) Did we *need* to assume that the distribution of SSHA scores for older students is Normal in this problem? Justify your answer.

(e) Can we generalize our findings about SSHA scores to the population of all older students in U.S. colleges? Explain why or why not.

11.2 Anemia, I 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 μ for all children in Jordan is less than 12. He measures a sample of 50 children. Suppose we know that the hemoglobin level for all children of this age follows a Normal distribution with standard deviation $\sigma = 1.6$ g/dl.

(a) Carefully define the parameter μ in this case.

(b) We seek evidence *against* the claim that $\mu = 12$. What is the sampling distribution of $\bar{x}$ of a sample of 50 children if $\mu = 12$? 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.)

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

(d) Did we *need* to know that the distribution of hemoglobin level for children this age is Normal? Justify your answer.

(e) Can we generalize our findings about hemoglobin levels to the population of all children this age in Jordan? Explain why or why not.

Stating Hypotheses

A statistical test starts with a careful statement of the claims we want to compare. In Example 11.2, we asked whether the accident response time data are likely if, in fact, there is no decrease in paramedics' response times. Because the reasoning of tests looks for evidence *against* a claim, we start with the claim we seek evidence against, such as "no decrease in response time." This claim is our ***null hypothesis.***

Null and Alternative Hypotheses

The statement being tested in a significance test is called the **null hypothesis.** The significance test is designed to assess the strength of the evidence *against* the null hypothesis. Usually the null hypothesis is a statement of "no effect," "no difference," or no change from historical values.

The claim about the population that we are trying to find evidence *for* is the **alternative hypothesis.**

11.3 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.

Exercises

11.3 Student attitudes and anemia: stating hypotheses

(a) State appropriate null and alternative hypotheses for the study of older students' attitudes described in Exercise 11.1 (page 690).

(b) State appropriate null and alternative hypotheses for the anemia study described in Exercise 11.2 (page 691).

11.4 State your claims, I Each of the following situations calls for a significance test. State the appropriate null hypothesis H_0 and alternative hypothesis H_a in each case. Be sure to define your parameter each time.

(a) Larry's car averages 26 miles per gallon on the highway. He switches to a new brand of motor oil that is advertised to increase gas mileage. After driving 3000 highway miles with the new oil, he wants to determine if the average gas mileage has increased.

(b) A May 2005 Gallup Poll report on a national survey of 1028 teenagers revealed that 72% of teens said they rarely or never argue with their friends.[2] You wonder whether this national result would be true in your school. So you conduct your own survey of a random sample of students at your school.

11.5 State your claims, II Each of the following situations calls for a significance test. State the appropriate null hypothesis H_0 and alternative hypothesis H_a in each case. Be sure to define your parameter each time.

(a) In the setting of Example 11.2 (page 688), the city manager also noted that paramedics arrived within 8 minutes after 78% of all calls involving life-threatening injury last year. Based on this year's random sample of 400 calls, she wants to determine whether the paramedics have arrived within 8 minutes more frequently this year.

(b) A national study reports that households spend an average of 30% of their food expenditures in restaurants. A restaurant association in your area wonders if the national results apply locally. They interview a sample of households and ask about their total food budget and the amount spent in restaurants.

11.6 Wrong hypotheses Here are several situations where there is an incorrect application of the ideas presented in this section. Explain what is wrong in each situation and why it is wrong.

(a) A change is made that should improve student satisfaction with the parking situation at your school. The null hypothesis, that there is an improvement, is tested versus the alternative, that there is no change.

(b) A researcher tests the following null hypothesis: H_0: $\bar{x} = 10$.

(c) A climatologist wants to test the null hypothesis that it will rain tomorrow.

The *P-value of a Significance Test* applet at the book's Web site www.whfreeman.com/tps3e automates the work of finding *P*-values. The applet even displays *P*-values as areas under a Normal curve, just like Figures 11.2 and 11.3.

Exercises

11.7 Job satisfaction: the conditions Refer to Example 11.7.

(a) Verify that the three important conditions are satisfied.

(b) Was it necessary to know that differences in job satisfaction follow a Normal distribution? Why or why not?

11.8 Job satisfaction with a larger sample Suppose that the job satisfaction study had produced exactly the same outcome $\bar{x} = 17$ as in Example 11.7, but from a sample of 75 workers rather than just 18 workers.

(a) Now is it necessary to know that differences in job satisfaction follow a Normal distribution? Why or why not?

(b) Calculate the test statistic z and its two-sided *P*-value.

(c) Do the data give good evidence that the population mean is not zero? Justify your answer.

11.9 Student attitudes, II Return to Exercise 11.1 (page 690). Start with the picture you drew there, and then do the following.

(a) Shade the area under the curve that is the *P*-value for $\bar{x} = 118.6$. Then calculate the test statistic and the *P*-value.

(b) Shade differently the area under the curve that is the *P*-value for $\bar{x} = 125.7$. Then calculate the test statistic and the *P*-value.

(c) Explain what each of the *P*-values in parts (a) and (b) tells us about the evidence against the null hypothesis.

11.10 Anemia, II Return to Exercise 11.2 (page 691). Start with the picture you drew there, and then do the following:

(a) Shade the area under the curve that is the *P*-value for $\bar{x} = 11.3$. Then calculate the test statistic and the *P*-value.

(b) Shade differently the area under the curve that is the *P*-value for $\bar{x} = 11.8$. Then calculate the test statistic and the *P*-value.

(c) Explain what each of the *P*-values in parts (a) and (b) tells us about the evidence against the null hypothesis.

11.11 Coffee sales Weekly sales of regular ground coffee at a supermarket have in the recent past varied according to a Normal distribution with mean $\mu = 354$ units per week and standard deviation $\sigma = 33$ units. The store reduces the price by 5%. Sales in the next three

weeks are 405, 378, and 411 units. Is this good evidence that average sales are now higher? The hypotheses are

$$H_0: \mu = 354$$
$$H_a: \mu > 354$$

Assume that the standard deviation of the population of weekly sales remains $\sigma = 33$.

(a) Find the value of $\bar{x}$.

(b) Sketch the Normal curve for the sampling distribution of $\bar{x}$ when H_0 is true. Why is the sampling distribution Normal? Are the other conditions satisfied?

(c) Shade the area that represents the *P*-value for the observed outcome. Calculate the test statistic and the *P*-value.

(d) Do you think there is convincing evidence that mean sales are higher? Explain.

11.12 Finding *P*-values A test of the null hypothesis H_0: $\mu = 0$ gives test statistic $z = 1.6$.

(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: $\mu \neq 0$?

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 α, the Greek 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.

Statistical Significance

If the *P*-value is as small as or smaller than alpha, we say that the data are **statistically significant at level α.**

"Significant" in the statistical sense does not mean "important." It means simply "not likely to happen just by chance." The significance level α 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 *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.

Example 11.9 ***Paramedics and job satisfaction (continued)***
Deciding to reject or fail to reject H_0

In Example 11.6 (page 696), we calculated the *P*-value for the city manager's study of paramedic response times as $P = 0.0139$. If we were using an $\alpha = 0.05$ significance level, we would reject H_0: $\mu = 6.7$ minutes (conclusion) since our *P*-value, 0.0139, is less than $\alpha = 0.05$ (connection). It appears that the mean response time to all life-threatening calls this year is less than last year's average of 6.7 minutes (context).

For the job satisfaction study, the *P*-value was 0.2302. Using an $\alpha = 0.05$ significance level, we would fail to reject H_0: $\mu = 0$ since $0.2302 > \alpha = 0.05$. It is possible that the mean difference in job satisfaction scores for workers in a self-paced versus machine-paced environment is 0.

Warning: if you are going to draw a conclusion based on statistical significance, then the significance level α should be stated before the data are produced. Otherwise, a deceptive user of statistics might set an α level after the data have been analyzed in an obvious attempt to manipulate the conclusion. This is just as inappropriate as choosing an alternative hypothesis to be one-sided in a particular direction *after* looking at the data.

A *P*-value is more informative than a "reject" or "fail to reject" conclusion at a given significance level. For instance, a *P*-value of 0.0139 allows us to reject H_0 at the $\alpha = 0.05$ significance level. But the *P*-value, 0.0139, gives a better sense of how strong the evidence against H_0 is. *The P-value is the smallest α level at which the data are significant.* Knowing the *P*-value allows us to assess significance at any level. However, interpreting the *P*-value is more challenging than making a decision about H_0 based on statistical significance. A well-trained user of statistics should be able to handle either approach.

Exercises

11.13 Statistical significance Explain in plain language why a significance test that is significant at the 1% ($\alpha = 0.01$) level must always be significant at the 5% ($\alpha = 0.05$) level. If a test is significant at the 5% level, what can you say about its significance at the 1% level?

11.14 Nicotine in cigarettes To determine whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 milligrams, a health advocacy group tests H_0: $\mu = 1.4$ versus H_a: $\mu > 1.4$. The calculated value of the test statistic is $z = 2.42$.

(a) Is the result significant at the 5% level? Why or why not?

(b) Is the result significant at the 1% level? Why or why not?

(c) What decision would you make about H_0 in part (a)? Part (b)? Explain.

11.15 Significance tests You will perform a significance test of H_0: $\mu = 0$ versus H_a: $\mu > 0$.

(a) What values of z would lead you to reject H_0 at the 5% ($\alpha = 0.05$) significance level?

(b) If the alternative hypothesis was H_a: $\mu \neq 0$, what values of z would lead you to reject H_0 at the 5% significance level?

(c) Explain why your answers to parts (a) and (b) are different.

11.16 Testing a random number generator, I A certain 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 generate 100 random numbers gives outcomes with mean $\bar{x} = 0.4365$. Assume that the population σ remains fixed. We want to test H_0: $\mu = 0.5$ versus H_a: $\mu \neq 0.5$.

(a) Calculate the value of the z test statistic and the P-value.

(b) Is the result significant at the 5% level ($\alpha = 0.05$)? Why or why not?

(c) Is the result significant at the 1% level ($\alpha = 0.01$)? Why or why not?

(d) What decision would you make about H_0 in part (b)? Part (c)? Explain.

11.17 Testing a random number generator, II The `rand` function on the TI-83/84 (MATH / PRB / 1:rand) and the `rand83` function on the TI-89 (CATALOG/F3) generate a pseudo-random real number in the interval [0, 1)—that is, in the interval $0 \leq X < 1$. The command `rand(100)` (`tistat.rand83(100)` on the TI-89) generates 100 random real numbers in the interval [0, 1). Describe how you would use your calculator to carry out a simulation like the one described in the previous exercise. (*Hint:* Store the 100 values in a list.) As in Exercise 11.16, take $\sigma = 0.2887$. Then carry out your plan and answer the questions in Exercise 11.16.

11.18 *P*-values and statistical significance A research report described the results of two studies. The P-value for the first study was 0.049; for the second it was 0.00002. Write a few sentences comparing what these P-values tell you about statistical significance in each of the two studies.

Section 11.1 Summary

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

The 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 (the null hypothesis value) in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).

As with confidence intervals, you should verify that the three conditions—SRS, Normality, and independence—are met before you proceed to calculations.

The essential reasoning of a significance test is as follows. Suppose 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 as or smaller than a specified value α, the data are **statistically significant** at significance level α. In that case, we can **reject** H_0. If the *P*-value is larger than α, we **fail to reject** H_0. An alternative to making a decision about the null hypothesis in your conclusion is to interpret the *P*-value in context.

Section 11.1 Exercises

11.19 Research and hypotheses For each of the following research questions, (1) describe an appropriate method for producing data to answer the question, and (2) state an appropriate pair of hypotheses (null and alternative) to test.

(a) The mean area of the several thousand similar apartments in a new development is advertised to be 1250 square feet. A tenant group thinks the apartments are smaller than advertised.

(b) Last year, a company's service technicians took an average of 1.8 hours to respond to trouble calls from customers who had purchased service contracts. The company president wants to know whether response times have changed this year.

(c) Simon reads a newspaper report claiming that 12% of all adults in the U.S. are left-handed. He wonders if 12% of the students at his large public high school are left-handed.

11.20 Spending on housing The Census Bureau reports that households spend an average of 31% of their total spending on housing. A homebuilders association in Cleveland believes that this average is lower in their area. They interview a sample of 40 households in the Cleveland metropolitan area to learn what percent of their spending goes toward housing. Take μ to be the mean percent of spending devoted to housing among all Cleveland households. We want to test the hypotheses

$$H_0: \mu = 31\%$$

$$H_a: \mu < 31\%$$

Assume that the population standard deviation is $\sigma = 9.6\%$.

(a) What is the sampling distribution of the mean percent $\bar{x}$ that the sample spends on housing if the null hypothesis is true? Sketch the density curve of the sampling distribution. Be sure to locate μ and σ correctly.

(b) Suppose that the study finds $\bar{x} = 30.2\%$ for the 40 households in the sample. Mark this point on the axis in your sketch. Then suppose that the study result is $\bar{x} = 27.6\%$. Mark this point on your sketch. Referring to your sketch, explain in simple language why one result is good evidence that average Cleveland spending on housing is less than 31%, whereas the other result is not.

(c) Shade the area under the curve that gives the *P*-value for the result $\bar{x} = 30.2\%$. (Note that we are looking for evidence that spending is less than the null hypothesis states.) Now shade differently the area under the curve that represents the *P*-value for $\bar{x} = 27.6\%$. Find these two *P*-values.

(d) What conclusion would you draw for each of the two results at the $\alpha = 0.05$ level? At the $\alpha = 0.01$ level? Justify your answers.

11.21 Statistical significance: one-sided and two-sided tests You are performing a significance test of H_0: $\mu = 0$ based on an SRS of 20 observations from a Normal population.

(a) If the alternative hypothesis is H_a: $\mu \neq 0$, what values of the z statistic are significant at the $\alpha = 0.005$ level? Show your work.

(b) If the alternative hypothesis is H_a: $\mu > 0$, what values of the z statistic are significant at the $\alpha = 0.005$ level? Show your work.

11.22 Interpreting statistical significance Asked to explain the meaning of "statistically significant at the $\alpha = 0.05$ level," a student says: "This means that the probability that the null hypothesis is true is less than 0.05." Is this explanation correct? Why or why not?

11.23 Interpreting *P*-values Suppose you perform a significance test of H_0: $\mu = 15$ versus the two-sided alternative and obtain a P-value of 0.082.

(a) Explain to someone who knows no statistics what this P-value means.

(b) What decision would you make about the null hypothesis? Why?

(c) Suppose you decide to reject H_0. What is the probability that you are wrong (H_0 is true)?

11.24 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 considered convincing statistical evidence. For a two-sided test, what significance level α corresponds to $z^* = 2$? To $z^* = 3$?

11.25 Thinking about conditions Explain in your own words why each of the three conditions—SRS, Normality, and independence—is important when performing inference about a population mean μ.

11.26 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 concentrations by 12.5 percent ($P = 0.01$).[4]

A doctor who knows little 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 to the doctor in simple language how "$P = 0.02$" answers this objection.

ing Out Significance Tests

Although the reasoning of significance testing isn't simple, carrying out a test is. The process is very similar to the one we followed when constructing a confidence interval. With a few minor changes, the four-step Inference Toolbox will once again guide us through the inference procedure.

Step 4: Interpretation Since our P-value, 0.0170, is less than $\alpha = 0.05$, this result is statistically significant. We reject H_0 and conclude that the mean difference in blood pressure readings from before and after the campaign among this company's employees is negative. In other words, the data suggest that employees' blood pressure readings have decreased on average.

Our conclusion in Example 11.11 is cautious. We would like to conclude that the health campaign *caused* the drop in mean blood pressure. But there may be other possible explanations. Suppose the local television station runs a series on the risk of heart attacks and the value of better diet and exercise. Many employees may improve their health habits even without encouragement from the company. Only a randomized comparative experiment protects against such lurking variables. The medical director preferred to launch a company-wide campaign that appealed to all employees. This may be a good medical decision, but the lack of a control group weakens the statistical conclusion.

When you interpret your results in context (Step 4 of the Inference Toolbox), be sure to link your comments directly to your P-value or significance level. Do not simply say "reject H_0." Provide a basis for any decision that you make about the claim expressed in your hypotheses.

Exercises

11.27 Water quality An environmentalist group collects a liter of water from each of 45 randomly chosen locations along a stream and measures the amount of dissolved oxygen in each specimen. The mean is 4.62 milligrams per liter (mg/L). Is this strong evidence that the stream has a mean dissolved oxygen content of less than 5 mg per liter? (Suppose we know that dissolved oxygen varies among locations with $\sigma = 0.92$ mg/L.) Follow the Inference Toolbox.

11.28 Improving your SAT score We suspect that students will generally score higher the second time they take the SAT Mathematics exam than on their first attempt. Suppose we know that the changes in score (second try minus first try) have population standard deviation $\sigma = 50$. Here are the results for 46 randomly chosen high school students:

−30	24	47	70	−62	55	−41	−32	128	−11
−43	122	−10	56	32	−30	−28	−19	1	17
57	−14	−58	77	27	−33	51	17	−67	29
94	−11	2	12	−53	−49	49	8	−24	96
120	2	−33	−2	−39	99				

(a) Construct graphical displays and calculate numerical summaries for these data. Write a few sentences about the distribution of changes in SAT Math scores.

(b) Based on your work in part (a), do you believe that the population of differences in SAT Math score is Normally distributed? Why or why not?

(c) Do these data give good evidence that the mean change in the population is greater than 0? Follow the Inference Toolbox.

11.29 Pressing pills A drug manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. The hardness of a sample from each batch of tablets produced is measured in order to control the compression process. The target values for the hardness are $\mu = 11.5$ and $\sigma = 0.2$. The hardness data for a sample of 20 tablets are

11.627	11.613	11.493	11.602	11.360	11.374	11.592	11.458	11.552	11.463
11.383	11.715	11.485	11.509	11.429	11.477	11.570	11.623	11.472	11.531

(a) We are not told that the distribution of hardness measurements in the population is Normal, and the sample size is too small ($n = 20$) for the central limit theorem to help us. In cases like this, the best we can do is to examine the sample data. Make appropriate graphs and calculate numerical summaries for determining whether these data could have come from a Normal population. Write a few sentences explaining your conclusion.

(b) Is there significant evidence at the 5% level that the mean hardness of the tablets is different from the target value? Use the Inference Toolbox.

11.30 Filling cola bottles Bottles of a popular cola are supposed to contain 300 milliliters (ml) of cola. There is some variation from bottle to bottle because the filling machinery is not perfectly precise. The distribution of the contents is Normal with standard deviation $\sigma = 3$ ml. An inspector who suspects that the bottler is underfilling measures the contents of six randomly selected bottles from a single day's production. The results are

299.4	297.7	301.0	298.9	300.2	297.0

Is this convincing evidence that the mean amount of cola in all the bottles filled that day is less than the advertised 300 ml? Follow the Inference Toolbox.

Tests from Confidence Intervals

A 95% confidence interval captures the true value of μ in 95% of all samples. If we are 95% confident that the true μ lies in our interval, we are also confident that values of μ that fall outside our interval are incompatible with the data. That sounds like the conclusion of a significance test. In fact, there is an intimate connection between 95% confidence and significance at the 5% level. The same connection holds between 99% confidence intervals and significance at the 1% level, and so on.

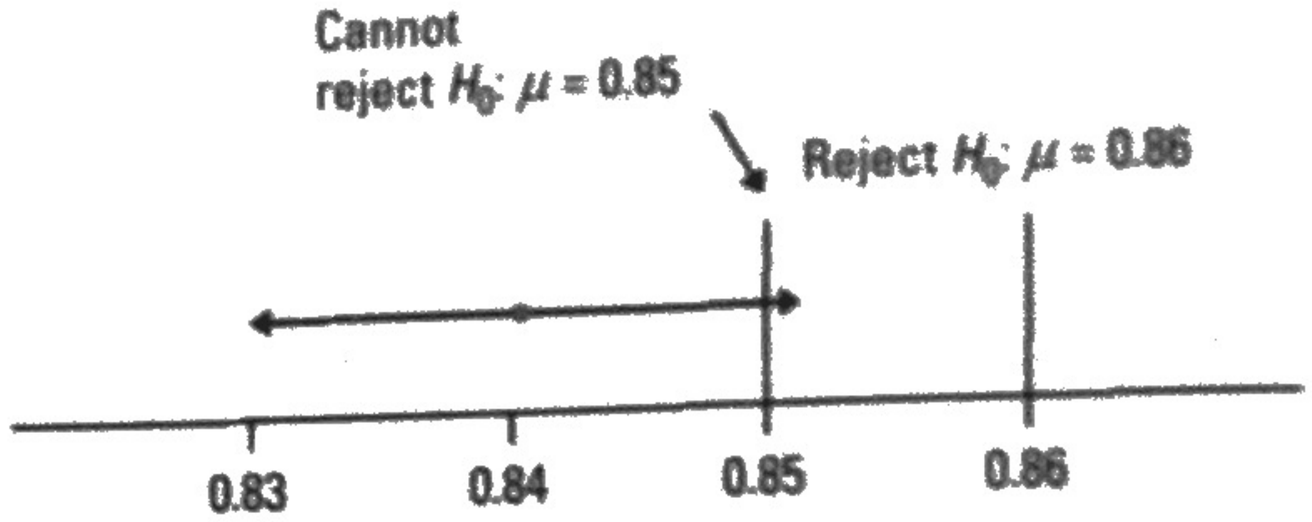
Confidence Intervals and Two-Sided Tests

A level α two-sided significance test rejects a hypothesis H_0: $\mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ.

The following example demonstrates the link between two-sided significance tests and confidence intervals (sometimes called ***duality***).

What if the null hypothesis in Example 11.12 had been H_0: $\mu = 0.85$? In that case, we would not be able to reject H_0: $\mu = 0.85$ in favor of the two-sided alternative H_a: $\mu \neq 0.85$, because 0.85 lies inside the 99% confidence interval for μ. Figure 11.7 shows both cases.

Figure 11.7 *Values of μ falling outside a 99% confidence interval can be rejected at the 1% significance level. Values falling inside the interval cannot be rejected.*



Exercises

11.31 Significance tests and confidence intervals 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 10? Why or why not?

(b) Does the 90% confidence interval include 10? Why or why not?

11.32 Confidence intervals and significance tests 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 or why not?

(b) Can you reject the null hypothesis that $\mu = 36$ at the 5% significance level? Why or why not?

11.33 Radon detectors Radon is a colorless, odorless gas that is naturally released by rocks and soils and may concentrate in tightly closed houses. Because radon is slightly radioactive, there is some concern that it may be a health hazard. Radon detectors are sold to homeowners worried about this risk, but the detectors may be inaccurate. University researchers placed a random sample of 12 detectors in a chamber where they were exposed to 105 picocuries per liter of radon over 3 days. Here are the readings given by the detectors:[5]

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

Assume that repeated readings using detectors of this type follow a Normal distribution with $\sigma = 9$.

11.36 Is this milk watered down? Cobra Cheese Company buys milk from several suppliers. Cobra suspects that some producers are adding water to their milk to increase their profits. Excess water can be detected by measuring the freezing point of the milk. The freezing temperature of natural milk varies Normally, with mean $\mu = -0.545°$ Celsius (C) and standard deviation $\sigma = 0.008°C$. Added water raises the freezing temperature toward 0°C, the freezing point of water. Cobra's laboratory manager measures the freezing temperature of a random sample of five containers of milk from one producer. The mean measurement is $\bar{x} = -0.538°C$. Is this good evidence that the producer is adding water to the milk? Provide appropriate statistical evidence to justify your answer.

11.37 Connecting confidence intervals and significance tests, I The P-value for a two-sided test of the null hypothesis H_0: $\mu = 30$ is 0.09.

(a) Does the 95% confidence interval include the value 30? Why?

(b) Does the 90% confidence interval include the value 30? Why?

11.38 Connecting confidence intervals and significance tests, II A 90% confidence interval for a population mean is (12, 15).

(a) Can you reject the null hypothesis that $\mu = 13$ against the two-sided alternative at the 10% significance level? Why?

(b) Can you reject the null hypothesis that $\mu = 13$ against a one-sided alternative at the 10% significance level? Why?

(c) Can you reject the null hypothesis that $\mu = 10$ against the two-sided alternative at the 10% significance level? Why?

(d) Can you reject the null hypothesis that $\mu = 10$ against a one-sided alternative at the 10% significance level? Why?

11.39 California SAT scores In a discussion of SAT scores, someone comments: "Because only a minority of high school students take the test, the scores overestimate the ability of typical high school seniors. The mean SAT Mathematics score is about 475, but I think that if all seniors took the test, the mean score would be no more than 450." You arrange to give the test to an SRS of 500 seniors from California. These students had a mean score of $\bar{x} = 461$. Assume that the population standard deviation is $\sigma = 100$. Is this good evidence against the claim that the mean for all California seniors is no more than 450? Give appropriate statistical evidence to justify your answer.

11.40 Cockroaches An understanding of cockroach biology may lead to an effective control strategy for these annoying insects. Researchers studying the absorption of sugar by insects feed cockroaches a diet containing measured amounts of a particular sugar. After 10 hours, the cockroaches are killed and the concentration of the sugar in various body parts is determined by a chemical analysis. The paper that reports the research states that a 95% confidence interval for the mean amount (in milligrams) of the sugar in the hindguts of the cockroaches is 4.2 ± 2.3.[6]

(a) Does this paper give evidence that the mean amount of sugar in the hindguts under these conditions is not equal to 7 mg? State H_0 and H_a and base a test on the confidence interval.

(a) Construct and interpret a 90% confidence interval for the mean reading μ for this type of detector.

(b) Is there significant evidence at the 10% level that the mean reading differs from the true value 105? State hypotheses and base a test on your confidence interval from (a).

11.34 One-sided tests and confidence intervals The P-value of a one-sided test of H_0: $\mu = 30$ is 0.04.

(a) Would the 95% confidence interval for μ include 30? Explain.

(b) Would the 90% confidence interval for μ include 30? Explain.

Section 11.2 Summary

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

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

The z test assumes an SRS of size n, known population standard deviation σ, independent observations, and either a Normal population or a large sample. P-values are computed from the standard Normal distribution (Table A). Use the Inference Toolbox as a guide when you perform a significance test.

Confidence intervals and two-sided significance tests are closely connected, provided that the significance level for the test and the confidence level for the interval add to 100%.

Section 11.2 Exercises

11.35 Calcium and pregnancy The level of calcium in the blood in healthy young adults varies with mean about 9.5 grams per deciliter and standard deviation about $\sigma = 0.4$. A clinic in rural Guatemala measures the blood calcium level of 160 healthy pregnant women at their first visit for prenatal care. The mean is $\bar{x} = 9.57$. Is this an indication that the mean calcium level in the population from which these women come differs from 9.5?

(a) Check the conditions for performing inference in this setting. Describe any concerns you may have about the data production.

(b) Carry out a significance test at the $\alpha = 0.05$ significance level. Report your conclusion.

(c) Construct and interpret a 95% confidence interval for the mean calcium level μ in the population. What additional information does this interval provide over the test in (b)?

Note: Based on our work in this problem, we are confident that μ lies quite close to 9.5. This illustrates the fact that **a test based on a large sample ($n = 160$ here) will often declare a small deviation from H_0 to be statistically significant.**

(b) Would the hypothesis that $\mu = 5$ mg be rejected at the 5% level in favor of a two-sided alternative?

11.41 *Statistical Significance* applet Go to the *Statistical Significance* applet on the book's Web site (www.whfreeman.com/tps3e). This applet illustrates statistical tests with a fixed significance level when sampling from a Normal population with known standard deviation.

(a) Open the applet and keep the default settings for the null ($\mu = 0$) and alternative ($\mu > 0$) hypotheses, the sample size ($n = 10$), the standard deviation ($\sigma = 1$), and the significance level ($\alpha = 0.05$). In the "I have data, and the observed $\bar{x}$ is $\bar{x} =$" box, enter the value 1. Is the difference between $\bar{x}$ and μ_0 significant at the 5% level?

(b) Repeat (a) for $\bar{x}$ equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Make a table giving $\bar{x}$ and the results of the significance tests. What do you conclude?

(c) Repeat parts (a) and (b) with significance level $\alpha = 0.01$. How does the choice of α change the statistical significance of the values of $\bar{x}$?

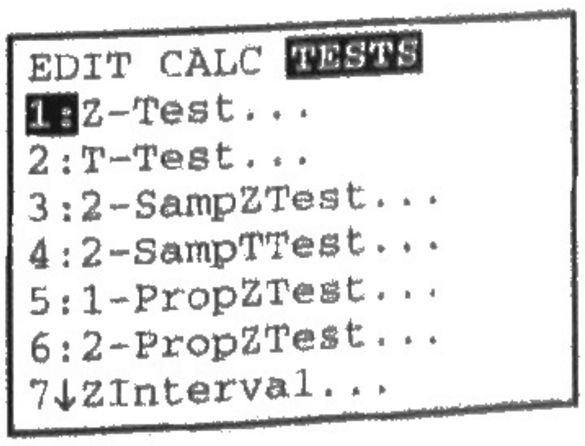
Technology Toolbox

Performing significance tests

The TI-83/84 and TI-89 can be used to conduct one-sample z tests, using either data stored in a list or summary statistics.

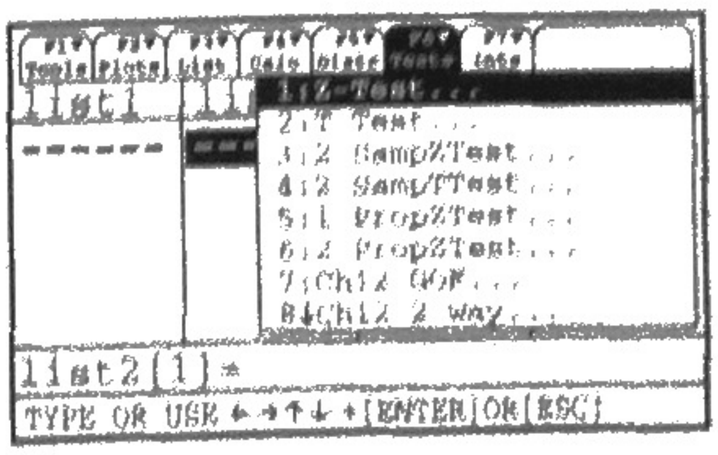
TI-83/84

- Press **STAT** and choose TESTS and 1:Z-Test to access the Z-Test screen, as shown.
- Choose "stats" as the input method.

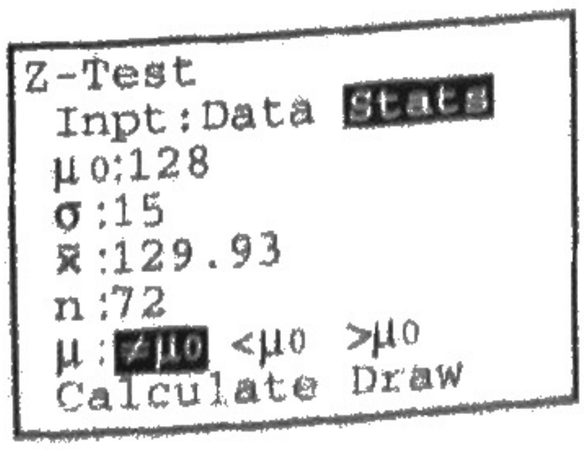


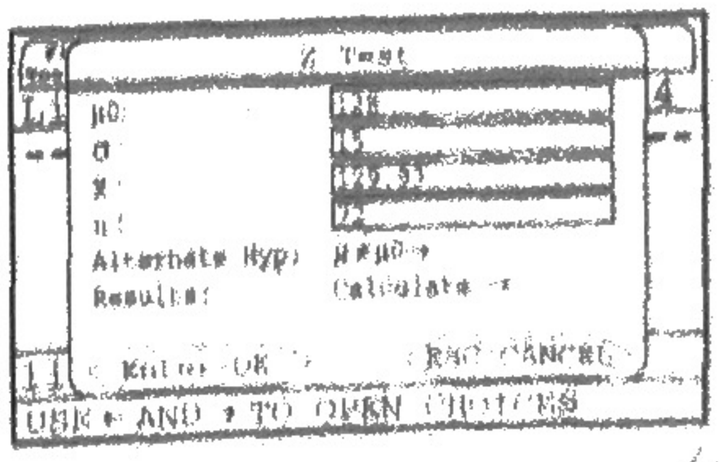
TI-89

- Press **2nd** **F1** ([F6]) and choose 1:Z-Test.
- Choose "Stats" as the input method.



For the executives' blood pressures of Example 11.10 (page 706), for example, you would enter 128 for the null hypothesized mean μ_0. Next enter 15 for σ, 129.93 for $\bar{x}$, and 72 for n. Select $\neq \mu_0$ for the alternative hypothesis, and choose "Calculate."





(continued)

the null hypothesis is true. Because 5% is 1/20, we expect about 1 of 20 tests to give a significant result just by chance. Running one test and reaching the $\alpha = 0.05$ level is reasonably good evidence that you have found something; running 20 tests and reaching that level only once is not.

Searching data for suggestive patterns is certainly legitimate. Exploratory data analysis is an important aspect of statistics. But the reasoning of formal inference does not apply when your search for a striking effect in the data is successful. The remedy is clear. Once you have a hypothesis, design a study to search specifically for the effect you now think is there. If the result of this study is statistically significant, you have real evidence.

Section 11.3 Summary

P-values are more informative than the reject-or-not result of a fixed level α test. Beware of placing too much weight on traditional values of α, such as $\alpha = 0.05$.

Very small effects can be highly significant (small *P*), especially when a test is based on a large sample. A statistically significant effect need not be practically important. Plot the data to display the effect you are seeking, and use confidence intervals to estimate the actual value of parameters.

On the other hand, lack of significance does not imply that H_0 is true, especially when the test is based on just a few observations.

Significance tests are not always valid. Faulty data collection, outliers in the data, and testing a hypothesis on the same data that suggested the hypothesis can invalidate a test. Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

Section 11.3 Exercises

11.43 Is it significant? In the absence of special preparation, SAT Math scores in recent years have varied Normally with mean $\mu = 518$ and $\sigma = 114$. One hundred students go through a rigorous training program designed to raise their SAT Math scores by improving their mathematics skills. Either by hand or using the *P-value of a Significance Test* applet, carry out a test of

$$H_0: \mu = 518$$
$$H_a: \mu > 518$$

(with $\sigma = 114$) in each of the following situations:

(a) The students' average score is $\bar{x} = 536.7$. Is this result significant at the 5% level?

(b) The average score is $\bar{x} = 536.8$. Is this result significant at the 5% level?

The difference between the two outcomes in (a) and (b) is of no importance. Beware attempts to treat $\alpha = 0.05$ as sacred.

11.44 Coaching and the SAT, I Suppose that SAT Math scores in the absence of coaching vary Normally with mean $\mu = 518$ and $\sigma = 100$. Suppose that coaching may change μ but does not change σ. A coaching service finds that a sample of students it has coached have mean score $\bar{x} = 522$. An increase in the SAT Math score from 518 to 522 is of no importance in seeking admission to college. But this service may still be able to advertise that its customers "score significantly higher." To see this, calculate the P-value for a test of

$$H_0: \mu = 518$$
$$H_a: \mu > 518$$

by hand or using the *P-value of a Significance Test* applet in each of the following situations.

(a) The service coaches 100 students. Their SAT Math scores average $\bar{x} = 522$.

(b) By the next year, the service has coached 1000 students. For these students, $\bar{x} = 522$.

(c) An advertising campaign brings the number of students coached to 10,000. Their average score is still $\bar{x} = 522$.

11.45 Coaching and the SAT, II Give a 99% confidence interval for the mean SAT Math score μ after coaching in each part of the previous exercise. For large samples, the confidence interval tells us, "Yes, the mean score is higher than 518 after coaching, but only by a small amount."

11.46 Ages of presidents Joe is writing a report on the backgrounds of American presidents. He looks up the ages of all the presidents when they entered office. Because Joe took a statistics course, he uses these numbers to perform a significance test about the mean age of all U.S. presidents. This makes no sense. Why not?

11.47 Do you have ESP? A researcher looking for evidence of extrasensory perception (ESP) tests 500 subjects. Four of these subjects do significantly better ($P \leq 0.01$) than random guessing.

(a) Is it proper to conclude that these four people have ESP? Explain your answer.

(b) What should the researcher now do to test whether any of these four subjects have ESP?

11.48 What is significance good for? Which of the following questions does a test of significance answer? Justify your answer.

(a) Is the sample or experiment properly designed?

(b) Is the observed effect due to chance?

(c) Is the observed effect important?

Inference to Make Decisions

In Section 11.1, we presented significance tests as methods for assessing the strength of evidence against the null hypothesis. Most users of statistics think of tests this way. But signs of another way of thinking were present in our discus-

The probability of a Type II error (sometimes called β) for the particular alternative $\mu = 6.4$ minutes in Example 11.21 is the probability that the test will fail to reject H_0 when μ has this alternative value. This is the probability that the sample mean $\bar{x}$ falls to the right of the critical value of $\bar{x}$ in Figure 11.9, calculated assuming $\mu = 6.4$. This probability is *not* $1 - 0.05$, because the probability 0.05 was found assuming that $\mu = 6.7$. Calculating Type II error by hand is possible but unpleasant. It's better to let technology do the work for you.

Awful accidents (one last time) Calculating Type II error probability

The TI-83/84 screen shot in the margin reproduces the two sampling distributions of Figure 11.9. This time, only the area corresponding to the Type II error probability is shaded. We can see that the probability β of making a Type II error (failing to reject H_0: $\mu = 6.7$ when actually $\mu = 6.4$) is about 8.8%. The city manager must decide whether this is an acceptable chance of failing to detect that the paramedics' mean response time has dropped to 6.4 minutes.

Where did the 6.53551 value for "low" in the screen shot come from? This is the critical value of $\bar{x}$ from Example 11.21. If $\bar{x}$ for our sample is less than 6.53551, we would reject H_0. If $\bar{x} \geq 6.53551$, we would fail to reject H_0. Since H_0 *is false* in this setting, rejecting H_0 is the correct decision. Failing to reject H_0 is a Type II error, which will occur in about 8.8% of all possible samples of 400 response times.

Exercises

11.48 Awful accidents Another way for the city manager of Example 11.20 to measure response times is to look at the *proportion* of calls for which paramedics arrived within 8 minutes. Last year, paramedics arrived on scene 75% of the time within 8 minutes. The city manager wants to determine whether they have done significantly better this year.

(a) State appropriate null and alternative hypotheses for the city manager to test.

(b) Describe a Type I and a Type II error in this setting.

(c) Explain the consequences of each type of error.

(d) Which is more serious: a Type I error or a Type II error? Justify your answer.

(e) If you sustain a life-threatening injury due to a vehicle accident, you want to receive medical treatment as quickly as possible. Which of the two significance tests—H_0: $\mu = 6.7$ versus H_a: $\mu < 6.7$ or the one from part (a) of this exercise—would you be more interested in? Justify your answer.

11.50 Blood pressure screening Your company markets a computerized device for detecting high blood pressure. The device measures an individual's blood pressure once per hour at a randomly selected time throughout a 12-hour period. Then it calculates the mean systolic (top number) pressure for the sample of measurements. Based on the

sample results, the device determines whether there is significant evidence that the individual's actual mean systolic pressure is greater than 130. If so, it recommends that the person seek medical attention.

(a) State appropriate null and alternative hypotheses in this setting.

(b) Describe a Type I and a Type II error.

(c) The program can be adjusted to decrease one error probability at the cost of an increase in the other error probability. Which error probability would you choose to make smaller, and why?

11.51 Catalog sales You want to see if the redesign of the cover of a mail-order catalog will increase sales. A very large number of customers will receive the original catalog, and a random sample of customers will receive the one with the new cover. Based on past sales, you are willing to assume that the sales from the new catalog will follow a Normal distribution with $\sigma = 60$ dollars and that the mean for the original catalog will be $\mu = 40$ dollars. You decide to use a sample size of $n = 1000$. You wish to test

$$H_0: \mu = 40$$

$$H_a: \mu > 40$$

at the 1% significance level ($\alpha = 0.01$).

(a) Describe a Type I error in this setting and the consequences of making a Type I error.

(b) Describe a Type II error in this setting and the consequences of making a Type II error.

(c) Which type of error is more serious in this case? Explain.

(d) Using the information provided, including the TI-83/84 calculator screen shot in the margin, determine the probability of each of the two types of error. (Refer to Example 11.21.)

(e) Try to determine where the value 44.4139 came from.

11.52 Beetles in the wood An outbreak of the mountain pine beetle has affected several types of pine trees in British Columbia. The beetle leaves behind a fungus that produces blue-colored stains in the wood. Some consumers might worry that lumber obtained from the blue-stained trees is weaker as a result of the effects of the fungus. A Canadian company performed a test on the breaking strength of blue-stained wood.[12] They measured the mean breaking strength of a sample of 100 pieces of blue-stained pine. The target breaking strength of lumber made from healthy pine trees is 10,000 pounds per square inch (psi).

(a) State appropriate null and alternative hypotheses for the company to test.

(b) Describe a Type I and a Type II error, and give the consequences of each.

(c) Which type of error is more serious? Why?

Power

A significance test makes a Type II error when it fails to reject a null hypothesis that really is false. A high probability of a Type II error for a particular alternative means

- Consider a particular alternative that is farther away from μ_0. Values of μ that are in H_a but lie close to the hypothesized value μ_0 are harder to detect (lower power) than values of μ that are far from μ_0.
- Increase the sample size. More data will provide more information about $\bar{x}$, so we will have a better chance of distinguishing values of μ.
- Decrease σ. This has the same effect as increasing the sample size: more information about $\bar{x}$. Improving the measurement process and restricting attention to a subpopulation are two common ways to decrease σ.

Power calculations are important in planning studies. Using a significance test with low power makes it unlikely that you will find a significant effect even if the truth is far from the null hypothesis. A null hypothesis that is in fact false can become widely believed if repeated attempts to find evidence against it fail because of low power. Our best advice for maximizing the power of a test is to choose as high an α level (Type I error probability) as you are willing to risk *and* as large a sample size as you can afford.

Exercises

11.53 Opening a restaurant You are thinking about opening a restaurant and are searching for a good location. From research you have done, you know that the mean income of those living near the restaurant must be over $85,000 to support the type of upscale restaurant you wish to open. You decide to take a simple random sample of 50 people living near one potential location. Based on the mean income of this sample, you will decide whether to open a restaurant there. A number of similar studies have shown that $\sigma = \$5000$.[13]

(a) State appropriate null and alternative hypotheses. Be sure to define your parameter.

(b) Describe the two types of errors that you might make. Identify which is a Type I error and which is a Type II error.

(c) Which of the two types of error is more serious? Explain.

(d) If you had to choose one of the "standard" significance levels for your significance test, would you choose $\alpha = 0.01$, 0.05, or 0.10? Justify your choice.

(e) Based on your choice in part (d), if the mean income in a certain area is $87,000, how likely are you to open a restaurant in that area? What is β? (Use the TYPE2 program provided by your teacher.)

11.54 Salty potato chips The mean salt content of a certain type of potato chip is supposed to be 2.0 milligrams (mg). The salt content of these chips varies Normally with standard deviation $\sigma = 0.1$ mg. From each batch produced, an inspector takes a sample of 50 chips and measures the salt content of each chip. The inspector rejects the entire batch if the sample mean salt content is significantly different from 2 mg at the 5% significance level.

Section 11.4 Summary

When we use statistical tests to make decisions, we view H_0 and H_a as two competing hypotheses that we must decide between. Our decision is based on comparing the P-value with a fixed significance level α. If the P-value is less than α, we reject H_0. Otherwise, we fail to reject H_0.

A **Type I error** occurs if we reject H_0 when it is in fact true. A **Type II error** occurs if we fail to reject H_0 when it is actually false.

The **power** of a significance test measures its ability to detect an alternative hypothesis. The power against a specific alternative is the probability that the test will reject H_0 when the alternative is true.

In a fixed level α significance test, the significance level α is the probability of a Type I error, and the power against a specific alternative is 1 minus the probability of a Type II error for that alternative.

Increasing the size of the sample increases the power (reduces the probability of a Type II error) when the significance level remains fixed. We can also increase the power of a test by using a higher significance level (say, $\alpha = 0.10$ instead of $\alpha = 0.05$).

Section 11.4 Exercises

11.59 Power and sample size Two studies are identical in all respects except for the sample sizes. Will the study with the larger sample size have more or less power than the one with the smaller sample size? Explain your answer in terms that could be understood by someone with very little knowledge of statistics.

11.60 Strong pipes A large high school needs to replace all of its water pipes as part of a major construction project. Such water pipes frequently experience water pressures of up to 100 pounds per square inch (psi). To be safe, the construction company requires that any pipes it will use have a mean breaking strength of greater than 120 pounds per square inch (psi). A water pipe supplier wants to provide the 100 pipes that will be needed for the project. They claim that their pipes are strong enough to meet the 120 psi standard. The supplier has 1500 pipes currently available for use. Before deciding whether to use this supplier's water pipes, the construction manager wants to test the breaking strength of a random sample of pipes from among the 1500 that are currently available. From previous experience with this company's pipes, we can assume that $\sigma = 8$ psi.

(a) State appropriate null and alternative hypotheses to test the supplier's claim.

(b) The supplier offers to provide 25 pipes for testing. Will this give the test in part (a) enough power to detect pipes that have a mean breaking strength of 125 pounds at the 5% significance level? Use technology to calculate the power.

(c) After some discussion, the supplier agrees to provide 40 pipes for testing. Describe a method for choosing a random sample of 40 pipes from the 1500 that are available using Table B of random digits. Use line 140 to carry out your method.

(a) What null and alternative hypotheses is the inspector testing? Be sure to define your parameter.

(b) Explain what a Type I error would mean in this setting. What's the probability of making a Type I error?

(c) Explain what a Type II error would mean in this setting. Use the TYPE2 program provided by your teacher to find β if $\mu = 2.05$.

(d) What is the power of the test to detect $\mu = 2.05$?

(e) What is the power of the test to detect $\mu = 1.95$? Why does this make sense?

(f) If the inspector used a 10% significance level instead of a 5% significance level, how would this affect the probability of a Type I error? A Type II error? The power of the test?

(g) Would you recommend a 1% significance level, a 5% significance level, or a 10% significance level to the company? Justify your answer.

11.55 More power to you! You are reviewing a research proposal that includes a section on sample size justification. A careful reading of this section indicates that the power is 20% for detecting an effect that most people would consider important. Write a short explanation of what this means and make a recommendation on whether the study should be run.

11.56 Power and the alternative A one-sided test of the null hypothesis $\mu = 50$ versus the alternative $\mu = 70$ has power equal to 0.5. Will the power for the alternative $\mu = 80$ be higher or lower than 0.5? Justify your answer.

11.57 Choose the right distribution You must decide which of two probability distributions a discrete random variable X has. We will call the probability distributions p_0 and p_1. Here are the probabilities they assign to the possible values of X:

X:	0	1	2	3	4	5	6
p_0:	0.1	0.1	0.1	0.1	0.2	0.1	0.3
p_1:	0.3	0.2	0.1	0.1	0.1	0.1	0.1

(a) Verify that both p_0 and p_1 are legitimate probability distributions.

You make a single observation on X and use it to test

$$H_0: p_0 \text{ is correct}$$

$$H_a: p_1 \text{ is correct}$$

One possible decision rule is to reject H_0 only if $X = 0$ or $X = 1$.

(b) Find the probability of making a Type I error, that is, the probability that you reject H_0 when p_0 is the correct distribution.

(c) Find the probability of making a Type II error. Show your work.

11.58 *Power* applet Go to the book's Web site and launch the *Power* applet. Redo Exercises 11.53 and 11.54 using the applet. In a few sentences, describe what the calculator program TYPE2 does that the applet does not do.

(d) Describe a Type I and a Type II error in this situation, and describe the consequences of each.

(e) Would you recommend a significance level of 0.01, 0.05, or 0.10 for this test? Justify your choice.

(f) If you were the construction company's manager, would you prefer the test about the mean breaking strength of the supplier's pipes or a test about the proportion of the supplier's pipes that have a breaking strength less than or equal to 120 psi? Explain.

11.61 Heavy bags Captain Ben flies small passenger jets. These jets carry 50 passengers, plus their luggage. On a full flight, these jets will perform properly as long as the total weight of passengers' checked baggage does not exceed 5000 pounds. Ben is concerned that passengers on a particular flight have brought unusually heavy bags. He selects a random sample of 10 passengers and weighs their checked baggage. Based on the results from this sample, he must decide whether it is safe to take off.

(a) Ben wants to perform a test to determine whether the mean weight μ of passengers' luggage on this flight is too heavy. State appropriate null and alternative hypotheses.

(b) Describe a Type I and a Type II error in this setting, and give the consequences of each.

(c) If you had to choose one of the "standard" significance levels for your significance test, would you choose $\alpha = 0.01$, 0.05, or 0.10? Justify your choice.

(d) Discuss any concerns you have about how the data were produced.

11.62 Power on either side The power for a two-sided test of the null hypothesis $\mu = 0$ versus the specific alternative $\mu = 10$ is 0.82.

(a) What is the probability of a Type II error in this setting?

(b) What is the power of the test versus the specific alternative $\mu = -10$? Explain your answer.

(c) Would the power of the test versus $\mu = -6$ be larger or smaller than the power you calculated in part (b)? Explain.

11.63 Choosing alpha You are the statistical expert on a team that is planning a study. After you have made a careful presentation of the mechanics of significance testing, one of the team members suggests using $\alpha = 0.50$ for this study because you would be more likely to obtain statistically significant results with this choice. Explain in simple terms why this would not be a good use of statistical methods.

11.64 Filling cola bottles: power Exercise 11.30 (page 710) concerns a test about the mean contents of cola bottles. The hypotheses are

$$H_0: \mu = 300 \text{ milliliters}$$

$$H_a: \mu < 300 \text{ milliliters}$$

The sample size is $n = 6$, and the population is assumed to have a Normal distribution with $\sigma = 3$. We'll use a 5% significance level. Power calculations help us see how large a shortfall in the bottle contents the test can be expected to detect.

Chapter Review Exercises

11.65 Stating hypotheses State the appropriate null and alternative hypotheses in each of the following cases.

(a) Census Bureau data show that the mean household income in the area served by a shopping mall is $72,500 per year. A market research firm questions shoppers at the mall to find out whether the mean household income of mall shoppers is higher than that of the general population.

(b) Mr. Starnes believes that fewer than 75% of the students at his school completed their math homework last night. The math teachers inspect the homework assignments from a random sample of students at the school to help Mr. Starnes test his claim.

(c) Experiments on learning in animals sometimes measure how long it takes a mouse to find its way through a maze. The mean time is 20 seconds for one particular maze. A researcher thinks that playing rap music will cause the mice to complete the maze faster. She measures how long each of 12 mice takes with the rap music as a stimulus.

11.66 Corn yield The mean yield of corn in the United States is about 120 bushels per acre. A survey of 40 farmers this year gives a sample mean yield of $\bar{x} = 123.8$ bushels per acre. We want to know whether this is good evidence that the national mean this year is not 120 bushels per acre. Assume that the farmers surveyed are an SRS from the population of all commercial corn growers and that the standard deviation of the yield in this population is $\sigma = 10$ bushels per acre. Are you convinced that the population mean is not 120 bushels per acre? Is your conclusion correct if the distribution of corn yields is somewhat non-Normal? Why?

11.67 Significance tests and confidence intervals The P-value of a one-sided test of the null hypothesis H_0: $\mu = 15$ is 0.02.

(a) Does the 99% confidence interval for μ include 15? Why or why not?

(b) Does the 95% confidence interval for μ include 15? Why or why not?

11.68 Analyzing study results A study with 12 subjects reported a result that failed to achieve statistical significance at the 5% level. The P-value was 0.052. Write a short summary of how you would interpret these findings.

11.69 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. What conclusion would you draw? Give appropriate evidence to support your conclusion.

11.70 CEO pay A study of the pay of corporate chief executive officers (CEOs) examined the increase in cash compensation of the CEOs of 104 companies, adjusted for inflation, in a recent year. The mean increase in real compensation was $\bar{x} = 6.9\%$, and the standard deviation of the increases was $s = 55\%$. Is this good evidence that the mean real

compensation μ of all CEOs increased that year? (Because the sample size is large, the sample s is close to the population σ, so take $\sigma = 55\%$.)

11.71 Why are large samples better? Statisticians prefer large samples. Describe briefly the effect of increasing the size of a sample (or the number of subjects in an experiment) on each of the following:

(a) The margin of error of a 95% confidence interval.

(b) The P-value of a test, when H_0 is false and all facts about the population remain unchanged as n increases.

(c) The power of a fixed level α test, when α, the alternative hypothesis, and all facts about the population remain unchanged.

11.72 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? Explain.

(b) Would the null hypothesis that the year 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? Justify your answers.

11.73 Strong chairs? A company that manufactures classroom chairs for high school students claims that the mean breaking strength of the chairs they make is 300 pounds. From years of production, they have seen that $\sigma = 15$ pounds. One of the chairs collapsed beneath a 220-pound student last week. You wonder whether the manufacturer is exaggerating the breaking strength of their chairs.

(a) State null and alternative hypotheses in words and symbols.

(b) Describe a Type I error and a Type II error in this situation. Which is more serious?

(c) There are 30 chairs in your classroom. You decide to determine the breaking strength of each chair and then to find the mean of those values. What values of $\bar{x}$ would cause you to reject H_0 at the 5% significance level?

(d) If the truth is that $\mu = 290$ pounds, use technology to find the probability that you will make a Type II error.

(e) Explain two ways that you could improve the power of this test.

11.74 Significance and sample size A study with 5000 subjects reported a result that was statistically significant at the 5% level. Explain why this result might not be particularly large or important.