(c) The margin of error is slightly less than a half hour. To reduce this down to 15 minutes, Gary says the sample size needs to be doubled to 200. What is wrong with this statement?

6.15 Importance of recreational sports. The National Intramural-Recreational Sports Association (NIRSA) performed a study to look at the value of recreational sports on college campuses.[7] A total of 2673 students were asked to indicate how important (on a 10-point scale) each of 21 factors was in terms of their college satisfaction and success. The factor "recreational sports and activities" resulted in a mean score of 7.5. Assume a standard deviation of 3.9.

(a) Give the margin of error and find the 95% confidence interval for this sample.

(b) Repeat these calculations for a 99% confidence interval. How do the results compare with those in part (a)?

6.16 Inference based on integer values. Refer to Exercise 6.15. The data for this study are integer values between 1 and 10. Explain why the confidence interval based on the Normal distribution should be a good approximation.

6.17 Mean serum TRAP in young women. For many important processes that occur in the body, direct measurement of characteristics of the process is not possible. In many cases, however, we can measure a *biomarker*, a biochemical substance that is relatively easy to measure and is associated with the process of interest. Bone turnover is the net effect of two processes: the breaking down of old bone, called resorption, and the building of new bone, called formation. One biochemical measure of bone resorption is tartrate resistant acid phosphatase (TRAP), which can be measured in blood. In a study of bone turnover in young women, serum TRAP was measured in 31 subjects.[8] The units are units per liter (U/l). The mean was 13.2 U/l. Assume that the standard deviation is known to be 6.5 U/l. Give the margin of error and find a 95% confidence interval for the mean for young women represented by this sample.

6.18 Mean OC in young women. Refer to the previous exercise. A biomarker for bone formation measured in the same study was osteocalcin (OC), measured in the blood. The units are nanograms per milliliter (ng/ml). For the 31 subjects in the study the mean was 33.4 ng/ml. Assume that the standard deviation is known to be 19.6 ng/ml. Report the 95% confidence interval.

6.19 Populations sampled and margins of error. Consider the following two scenarios. (A) Take a simple random sample of 100 sophomore students at your college or university. (B) Take a simple random sample of 100 sophomore students in your major at your college or university. For each of these samples you will record the amount spent on textbooks used for classes during the fall semester. Which sample should have the smaller margin of error? Explain your answer.

6.20 Average starting salary. The National Association of Colleges and Employers (NACE) Fall Salary Survey report shows that the current class of college graduates earned an average starting salary offer of \$48,633.[9] Your institution collected an SRS ($n = 2500$) of its recent graduates and obtained a 95% confidence interval of (\$45,330, \$46,156). This interval is below the NACE average. Using this information, compute the 99% confidence interval of the *difference* between the average starting salary from recent graduates at your institution and the overall NACE mean.

6.21 Consumption of sugar-sweetened beverages. A recent study estimated the U.S. per capita consumption of sugar-sweetened beverages among adults 20 to 44 years of age to be 289 kcal/day with a standard deviation of the mean equal to 7 kcal/day.[10]

(a) The 68–95–99.7 rule says that the probability is about 0.95 that $\bar{x}$ is within ______ kcal/day of the population mean μ. Fill in the blank.

(b) About 95% of all samples will capture the true mean of kcals consumed per day in the interval $\bar{x}$ plus or minus ______ kcal/day. Fill in the blank.

6.22 Apartment rental rates. You want to rent an unfurnished one-bedroom apartment in Dallas next year. The mean monthly rent for a random sample of 10 apartments advertised in the local newspaper is \$980. Assume the monthly rents in Dallas follow a Normal distribution with a standard deviation of \$290. Find a 95% confidence interval for the mean monthly rent for unfurnished one-bedroom apartments available for rent in this community.

6.23 More on apartment rental rates. Refer to the previous exercise. Will the 95% confidence interval include approximately 95% of the rents of all unfurnished one-bedroom apartments in this area? Explain why or why not.

6.24 Inference based on skewed data. The mean OC for the 31 subjects in Exercise 6.18 was 33.4 ng/ml. In our calculations, we assumed that the standard deviation was known to be 19.6 ng/ml. Use the 68–95–99.7 rule from Chapter 1 (page 56) to find the approximate bounds on the values of OC that would include these percents of the population. If the assumed standard deviation is correct, it would appear that this distribution may be highly skewed. Why? (*Hint:* The measured values for a variable such as this are all positive.) Do you think that this skewness will invalidate the use of the Normal confidence interval in this case? Explain your answer.

6.25 Average hours per week on the Internet. The *Student Monitor* surveys 1200 undergraduates from 100 colleges semiannually to understand trends among college students.[11] Recently, the *Student Monitor* reported that the average amount of time spent per week on the Internet was 19.0 hours. Assume that the standard deviation is 5.5 hours.

(a) Give a 95% confidence interval for the mean time spent per week on the Internet.

(b) Is it true that 95% of the students surveyed reported weekly times that lie in the interval you found in part (a)? Explain your answer.

6.26 Average minutes per week on the Internet. Refer to the previous exercise.

(a) Give the mean and standard deviation in minutes.

(b) Calculate the 95% confidence interval in minutes from your answer to part (a).

(c) Explain how you could have directly calculated this interval from the 95% interval that you calculated in the previous exercise.

6.27 Satisfied with your job? A Gallup Poll asked working adults about their job satisfaction.[12] One question was "How satisfied or dissatisfied are you with your job?" The possible answers were "completely satisfied," "somewhat satisfied," "somewhat dissatisfied," and "completely dissatisfied." Ninety percent responded that they were somewhat or completely satisfied. Material provided with the results of the poll noted:

Results are based on telephone interviews with 1,009 national adults, aged 18 and older, conducted Aug. 7–10, 2008. For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is ±3 percentage points. For results based on the sample of 557 adults employed full or part-time, the maximum margin of sampling error is ±5 percentage points.

The Gallup Poll uses a complex multistage sample design, but the sample percent has approximately a Normal sampling distribution.

(a) The announced poll result was 90% ± 5%. Can we be certain that the true population percent falls in this interval? Explain your answer.

(b) Explain to someone who knows no statistics what the announced result 90% ± 5% means.

(c) This confidence interval has the same form we have met earlier:

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

What is the standard deviation σ_{estimate} of the estimated percent?

(d) Does the announced margin of error include errors due to practical problems such as undercoverage and nonresponse?

6.28 Fuel efficiency. Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the mpg were recorded each time the gas tank was filled, and the computer was then reset.[13] Here are the mpg values for a random sample of 20 of these records: GASMILEAGE

41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3

Suppose that the standard deviation is known to be $\sigma = 3.5$ mpg.

(a) What is $\sigma_{\bar{x}}$, the standard deviation of $\bar{x}$?

(b) Examine the data for skewness and other signs of non-Normality. Show your plots and numerical summaries. Do you think it is reasonable to construct a confidence interval based on the Normal distribution? Explain your answer.

(c) Give a 95% confidence interval for μ, the mean mpg for this vehicle.

6.29 Fuel efficiency in metric units. In the previous exercise you found an estimate with a margin of error for the average miles per gallon. Convert your estimate and margin of error to the metric units kilometers per liter (kpl). To change mpg to kpl, use the fact that 1 mile = 1.609 kilometers and 1 gallon = 3.785 liters.

6.30 Percent coverage of 95% confidence interval. The *Confidence Interval* applet lets you simulate large numbers of confidence intervals quickly. Select 95% confidence and then sample 50 intervals. Record the number of intervals that cover the true value (this appears in the "Hit" box in the applet). Press the reset button and repeat 30 times. Make a stemplot of the results and find the mean. Describe the results. If you repeated this experiment very many times, what would you expect the average number of hits to be?

6.31 Required sample size for specifed margin of error. A new bone study is being planned that will measure the biomarker TRAP described in Exercise 6.17. Using the value of σ given there, 6.5 U/l, find the sample size required to provide an estimate of the mean TRAP with a margin of error of 1.5 U/l for 95% confidence.

6.32 Adjusting required sample size for drop out. Refer to the previous exercise. In similar previous studies, about 20% of the subjects drop out before the study is completed. Adjust your sample size requirement to have enough subjects at the end of the study to meet the margin of error criterion.

6.33 Radio poll. A college radio station invites listeners to enter a dispute about a proposed "pay as you throw" waste collection program. The station asks listeners to call in and state how much each bag of trash should cost. A total of 633 listeners call in. The station calculates the 95% confidence interval for the average fee desired by city residents to be \$0.83 to \$1.28. Is this result trustworthy? Explain your answer.

6.34 Accuracy of a laboratory scale. To assess the accuracy of a laboratory scale, a standard weight known to weigh 10 grams is weighed repeatedly. The scale readings are Normally distributed with unknown mean (this mean is 10 grams if the scale has no bias). The standard deviation of the scale readings is known to be 0.0002 gram.

(a) The weight is measured five times. The mean result is 10.0023 grams. Give a 98% confidence interval for the mean of repeated measurements of the weight.

(b) How many measurements must be averaged to get a margin of error of ±0.0001 with 98% confidence?

6.35 More than one confidence interval. As we prepare to take a sample and compute a 95% confidence interval, we know that the probability that the interval we compute will cover the parameter is 0.95. That's the meaning of 95% confidence. If we use several such intervals, however, our confidence that *all* of them give correct results is less than 95%. Suppose we take independent samples each month for five months and report a 95% confidence interval for each set of data.

(a) What is the probability that all five intervals cover the true means? This probability (expressed as a percent) is our overall confidence level for the five simultaneous statements.

(b) What is the probability that at least four of the five intervals cover the true means?

6.2 Tests of Significance

The confidence interval is appropriate when our goal is to estimate population parameters. The second common type of inference is directed at a quite different goal: to assess the evidence provided by the data in favor of some claim about the population parameters.

The reasoning of significance tests

A significance test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess. The hypothesis is a statement about the population parameters. The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree. We use the following examples to illustrate these concepts.

EXAMPLE

6.8 Credit card debt by U.S. region. One purpose of Sallie Mae's credit card usage studies described in Example 6.4 (page 349) is to compare the debt of different subgroups of student applicants. For example, the average debt of student applicants in the Midwest is \$3260, while those student applicants in the West had an average debt of \$3817. The difference of \$557 is fairly large, but we know that these numbers are estimates of the true means. If we took different samples, we would get different estimates. Can we conclude from

these data that the average credit card debts of student applicants in these two regions of the United States are different?

One way to answer this question is to compute the probability of obtaining a difference as large or larger than the observed \$557 assuming that, in fact, there is no difference in the true means. This probability is 0.14. Because this probability is not particularly small, we conclude that observing a difference of \$557 is not very surprising when the true means are equal. The data do not provide evidence for us to conclude that the credit card debts for student applicants in the Midwest and West are different.

Here is an example with a different conclusion.

EXAMPLE

6.9 Credit card debt by undergraduate grade level. This study also reports that the average debt among senior applicants is \$4138, while it is \$2912 among juniors. Is the average debt among senior applicants higher than the average debt among junior applicants? The observed difference is \$1226 but as we learned in the previous example, an observed difference in means is not necessarily sufficient for us to conclude that the true means are different. Do the data provide evidence that there is a difference in debt? Again, we answer this question with a probability calculated under the assumption that there is *no difference in the true means*. The probability is 0.0000004 of observing a difference of mean debt that is \$1226 or more when there really is no difference. Because this probability is so small, we have sufficient evidence in the data to conclude that senior applicants have a higher average credit card debt.

What are the key steps in these examples?

- We started each with a question about the difference between two mean debts. In Example 6.8, we compare student applicants in the West and Midwest. In Example 6.9, we compare junior applicants with senior applicants. In both cases, we ask whether or not the data are compatible with no difference, that is, a difference of \$0.
- Next we compared the data, \$557 in the first case and \$1226 in the second, with the value that comes from the question, \$0.
- The results of the comparisons are probabilities, 0.14 in the first case and 0.0000004 in the second.

The 0.14 probability is not particularly small, so we have limited evidence to question the possibility that the true difference is zero. In the second case, however, the probability is very small. Something that happens with probability 0.0000004 occurs only about 4 times out of 10,000,000. In this case we have two possible explanations:

1. We have observed something that is very unusual, or
2. The assumption that underlies the calculation, no difference in mean debt, is not true.

The z test assumes an SRS of size n, known population standard deviation σ, and either a Normal population or a large sample. P-values are computed from the Normal distribution (Table A). Fixed α tests use the table of **standard Normal critical values** (Table D).

SECTION 6.2 Exercises

For Exercises 6.36 and 6.37, see page 363; for Exercises 6.38 and 6.39, see pages 366–367; for Exercises 6.40 to 6.42, see page 369; for Exercises 6.43 and 6.44, see page 373; for Exercises 6.45 and 6.46, see page 375; and for Exercises 6.47 to 6.49, see page 377.

6.50 What's wrong? Here are several situations where there is an incorrect application of the ideas presented in this section. Write a short paragraph explaining what is wrong in each situation and why it is wrong.

(a) A researcher tests the following null hypothesis: H_0: $\bar{x} = 23$.

(b) A random sample of size 30 is taken from a population that is assumed to have a standard deviation of 5. The standard deviation of the sample mean is 5/30.

(c) A study with $\bar{x} = 45$ reports statistical significance for $H_a : \mu > 50$.

(d) A researcher tests the hypothesis H_0: $\mu = 350$ and concludes that the population mean is equal to 350.

6.51 What's wrong? Here are several situations where there is an incorrect application of the ideas presented in this section. Write a short paragraph explaining what is wrong in each situation and why it is wrong.

(a) A significance test rejected the null hypothesis that the sample mean is equal to 500.

(b) A test preparation company wants to test that the average score of their students on the ACT is better than the national average score of 21.2. They state their null hypothesis to be $H_0 : \mu > 21.2$.

(c) A study summary says that the results are statistically significant and the P-value is 0.98.

(d) The z statistic is equal to 0.018. Because this is less than $\alpha = 0.05$, the null hypothesis was rejected.

6.52 Determining hypotheses. State the appropriate null hypothesis H_0 and alternative hypothesis H_a in each of the following cases.

(a) A 2008 study reported that 88% of students owned a cell phone. You plan to take an SRS of students to see if the percentage has increased.

(b) The examinations in a large freshman chemistry class are scaled after grading so that the mean score is 75. The professor thinks that students who attend early morning recitation sections will have a higher mean score than the class as a whole. Her students this semester can be considered a sample from the population of all students she might teach, so she compares their mean score with 75.

(c) The student newspaper at your college recently changed the format of their opinion page. You take a random sample of students and select those who regularly read the newspaper. They are asked to indicate their opinions on the changes using a five-point scale: −2 if the new format is much worse than the old, −1 if the new format is somewhat worse than the old, 0 if the new format is the same as the old, +1 if the new format is somewhat better than the old, and +2 if the new format is much better than the old.

6.53 More on determining hypotheses. State the null hypothesis H_0 and the alternative hypothesis H_a in each case. Be sure to identify the parameters that you use to state the hypotheses.

(a) A university gives credit in first-year calculus to students who pass a placement test. The mathematics department wants to know if students who get credit in this way differ in their success with second-year calculus. Scores in second-year calculus are scaled so the average each year is equivalent to a 77. This year 21 students who took second-year calculus passed the placement test.

(b) 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 slower. She measures how long each of 12 mice takes with the rap music as a stimulus.

(c) The average square footage of one-bedroom apartments in a new student-housing development is advertised to be 880 square feet. A student group thinks that the apartments are smaller than advertised. They hire an engineer to measure a sample of apartments to test their suspicion.

6.54 Even more on determining hypotheses. In each of the following situations, state an appropriate null hypothesis H_0 and alternative hypothesis H_a. Be sure to identify the parameters that you use to state the hypotheses. (We have not yet learned how to test these hypotheses.)

(a) A sociologist asks a large sample of high school students which television channel they like best. She suspects that a higher percent of males than of females will name MTV as their favorite channel.

(b) An education researcher randomly divides sixth-grade students into two groups for physical education class. He teaches both groups basketball skills, using the same methods of instruction in both classes. He encourages Group A with compliments and other positive behavior but acts cool and neutral toward Group B. He hopes to show that positive teacher attitudes result in a higher mean score on a test of basketball skills than do neutral attitudes.

(c) An education researcher believes that among college students there is a negative correlation between time spent at social network sites and self-esteem. To test this, she gathers social networking information and self-esteem data from a sample of students at your college.

6.55 Translating research questions into hypotheses. Translate each of the following research questions into appropriate H_0 and H_a.

(a) Census Bureau data show that the mean household income in the area served by a shopping mall is $42,800 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) Last year, your online registration technicians took an average of 0.4 hours to respond to trouble calls from students trying to register. Do this year's data show a different average response time?

6.56 Computing the P-value. A test of the null hypothesis H_0: $\mu = \mu_0$ gives test statistic $z = 1.63$.

(a) What is the P-value if the alternative is H_a: $\mu > \mu_0$?

(b) What is the P-value if the alternative is H_a: $\mu < \mu_0$?

(c) What is the P-value if the alternative is H_a: $\mu \neq \mu_0$?

6.57 More on computing the P-value. A test of the null hypothesis H_0: $\mu = \mu_0$ gives test statistic $z = -1.82$.

(a) What is the P-value if the alternative is H_a: $\mu > \mu_0$?

(b) What is the P-value if the alternative is H_a: $\mu < \mu_0$?

(c) What is the P-value if the alternative is H_a: $\mu \neq \mu_0$?

6.58 A two-sided test and the confidence interval. The P-value for a two-sided test of the null hypothesis H_0: $\mu = 30$ is 0.032.

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

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

6.59 More on a two-sided test and the confidence interval. A 90% confidence interval for a population mean is (25, 32).

(a) Can you reject the null hypothesis that $\mu = 24$ against the two-sided alternative at the 10% significance level? Why?

(b) Can you reject the null hypothesis that $\mu = 30$ against the two-sided alternative at the 10% significance level? Why?

6.60 Use of bed nets. A study found that the use of bed nets was associated with a lower prevalence of malarial infections in the Gambia.[17] A report of the study states that the significance is $P < 0.001$. Explain what this means in a way that could be understood by someone who has not studied statistics.

6.61 Peer pressure and choice of major. A recent study followed a cohort of students entering a business/economics program.[18] All students followed a common track during the first three semesters and then chose to specialize in either business or economics. Through a series of surveys, the researchers were able to classify roughly 50% of the students as either peer driven (ignored abilities and chose major to follow peers) or ability driven (ignored peers and chose major based on ability). When looking at entry wages after graduation, the researchers conclude that a peer-driven student can expect an average wage that is 13% less than that of an ability-driven student. The report states that the significance level is $P = 0.09$. Can you be confident of this statement regarding the wage decrease? Discuss.

6.62 Symbol of wealth in ancient China? Every society has its own symbols of wealth and prestige. In ancient China, it appears that owning pigs was such a symbol. Evidence comes from examining burial sites. If the skulls of sacrificed pigs tend to appear along with expensive ornaments, that 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."[19] 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.

6.63 Alcohol awareness among college students. A study of alcohol awareness among college students reported a higher awareness for students enrolled in a health and safety class than for those enrolled in a statistics class.[20] The difference is described as being statistically significant. Explain what this means in simple terms and offer an explanation for why the health and safety students had a higher mean score.

6.64 Change in eighth-grade average mathematics score. A report based on the 2009 National Assessment of Educational Progress (NAEP)[21] states that the average score on their mathematics test for eighth-grade students is significantly higher than in 2007. They also state that the average score in Minnesota was not significantly different from the average score in 2007. A footnote states that comparisons are determined by statistical tests with 0.05 as the level of significance. Explain what this means in language understandable to someone who knows no statistics. Do not use the word "significance" in your answer.

6.65 Sleep quality and elevated blood pressure. A recent study looked at $n = 238$ adolescents, all free of severe illness.[22] Subjects wore a wrist actigraph, which allowed the researchers to estimate sleep patterns. For those subjects classified as having low sleep efficiency, they are reported as having an average systolic blood pressure that is 4 mm Hg higher than other children with a standard deviation of the mean equal to 1.2 mm Hg. Based on these results, test whether this difference is significant at the 0.01 level.

6.66 Are the pine trees randomly distributed north to south? In Example 6.1 we looked at the distribution of longleaf pine trees in the Wade Tract. One way to formulate hypotheses about whether or not the trees are randomly distributed in the tract is to examine the average location in the north-south direction. The values range from 0 to 200, so if the trees are uniformly distributed in this direction, any difference from the middle value (100) should be due to chance variation. The sample mean for the 584 trees in the tract is 99.74. A theoretical calculation based on the assumption that the trees are uniformly distributed gives a standard deviation of 58. Carefully state the null and alternative hypotheses in terms of this variable. Note that this requires that you translate the research question about the random distribution of the trees into specific statements about the mean of a probability distribution. Test your hypotheses, report your results, and write a short summary of what you have found.

6.67 Are the pine trees randomly distributed east to west? Answer the questions in the previous exercise for the east-west direction, where the sample mean is 113.8.

6.68 Who is the author? Statistics can help decide the authorship of literary works. Sonnets by a certain Elizabethan poet are known to contain an average of $\mu = 8.9$ new words (words not used in the poet's other works). The standard deviation of the number of new words is $\sigma = 2.5$. Now a manuscript with six new sonnets has come to light, and scholars are debating whether it is the poet's work. The new sonnets contain an average of $\bar{x} = 10.2$ words not used in the poet's known works. We expect poems by another author to contain more new words, so to see if we have evidence that the new sonnets are not by our poet we test

$$H_0: \mu = 8.9$$
$$H_a: \mu > 8.9$$

Give the z test statistic and its P-value. What do you conclude about the authorship of the new poems?

6.69 Attitudes toward school. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. 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 who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is $\bar{x} = 127.8$.

(a) Assuming that $\sigma = 30$ for the population of older students, carry out a test of

$$H_0: \mu = 115$$
$$H_a: \mu > 115$$

Report the P-value of your test, and state your conclusion clearly.

(b) Your test in part (a) required two important assumptions in addition to the assumption that the value of σ is known. What are they? Which of these assumptions is most important to the validity of your conclusion in part (a)?

6.70 Nutritional intake in Canadian high performance athletes. Since previous studies have reported that elite athletes are often deficient in their nutritional intake (e.g., total calories, carbohydrates, protein), a group of researchers decided to evaluate Canadian high performance athletes.[23] A total of $n = 324$ athletes from eight Canadian sports centers participated in the study. One reported finding was that the average caloric intake among the $n = 201$ women was 2403.7 kcal/day. The recommended amount is 2811.5 kcal/day. Is there evidence that female Canadian athletes are deficient in the caloric intake?

(a) State the appropriate H_0 and H_a to test this.

(b) Assuming a standard deviation of 880 kcal/day, carry out the test. Give the P-value, and then interpret the result in plain language.

6.71 Are the mpg measurements similar? Refer to Exercise 6.28 (page 359). In addition to the computer computing mpg, the driver also recorded the mpg by dividing the miles driven by the number of gallons at each fill-up. The following data are the differences between the computer's and the driver's calculations for that random sample of 20 records. The driver wants to determine if these calculations are different. Assume the standard deviation of a difference to be $\sigma = 3.0$. MPGCOMPARISON

5.0	6.5	−0.6	1.7	3.7	4.5	8.0	2.2	4.9	3.0
4.4	0.1	3.0	1.1	1.1	5.0	2.1	3.7	−0.6	−4.2

(a) State the appropriate H_0 and H_a to test this suspicion.

(b) Carry out the test. Give the P-value, and then interpret the result in plain language.

6.72 Adjusting for the cost of living. In Example 6.8 (page 360), we compared the average credit card debt between undergraduate applicants in the Midwest and West. In computing the difference, we did not adjust for differing costs in the two regions. Assuming that \$1 in the West is worth about \$1.06 in the Midwest, redo the test based on Midwest dollars. For simplicity, assume the standard deviation is unchanged.

6.73 Level of nicotine in cigarettes. According to data from the Tobacco Institute Testing Laboratory, Camel Lights king size cigarettes contain an average of 0.9 milligrams of nicotine. An advocacy group commissions an independent test to see if the mean nicotine content is higher than the industry laboratory claims.

(a) What are H_0 and H_a?

(b) Suppose that the test statistic is $z = 1.83$. Is this result significant at the 5% level?

(c) Is the result significant at the 1% level?

6.74 Changes of $\bar{x}$ on significance. The *Statistical Significance* applet illustrates statistical tests with a fixed level of significance for Normally distributed data with known standard deviation. Open the applet and keep the default settings for the null ($\mu = 0$) and the 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? Repeat for $\bar{x}$ equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving $\bar{x}$ and the results of the significance tests. What do you conclude?

6.75 Changes of α on significance. Repeat the previous exercise with significance level $\alpha = 0.01$. How does the choice of α affect the values of $\bar{x}$ that are far enough away from μ_0 to be statistically significant?

6.76 Changes of $\bar{x}$ on the P-value. The *P-Value of a Test of Significance* applet illustrates P-values of significance tests for Normally distributed data with known standard deviation. Open the applet and keep the default settings for the null ($\mu = 0$) and the 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. What is the P-value? Repeat for $\bar{x}$ equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving $\bar{x}$ and P-values. How does the P-value change as $\bar{x}$ moves farther away from μ_0?

6.77 Understanding levels of significance. Explain in plain language why a significance test that is significant at the 5% level must always be significant at the 10% level.

6.78 More on understanding levels of significance. You are told that a significance test is significant at the 5% level. From this information can you determine whether or not it is significant at the 1% level? Explain your answer.

6.79 Test statistic and levels of significance. Consider a significance test for a null hypothesis versus a two-sided alternative. Give a value of z that will give a result significant at the 1% level but not at the 0.5% level.

6.80 Using Table D to find a P-value. You have performed a two-sided test of significance and obtained a value of $z = 2.52$. Use Table D to find the approximate P-value for this test.

6.81 More on using Table D to find a P-value. You have performed a one-sided test of significance and obtained a value of $z = 0.63$. Use Table D to find the approximate P-value for this test.

6.82 Using Table A and Table D to find a P-value. Consider a significance test for a null hypothesis versus a two-sided alternative. Between what values from Table D does the P-value for an outcome $z = 1.92$ lie? Calculate the P-value using Table A, and verify that it lies between the values you found from Table D.

6.83 More on using Table A and Table D to find a P-value. Refer to the previous exercise. Find the P-value for $z = -1.92$.

USE YOUR KNOWLEDGE

6.85 Home security systems. A recent TV advertisement for home security systems said that homes without an alarm system are three times more likely to be broken into. Suppose this conclusion was obtained by examining an SRS of police records of break-ins and determining whether the percent of homes with alarm systems was significantly smaller than 50%. Explain why the significance of this study is suspect and propose an alternative study that would help clarify the importance of an alarm system.

Beware of searching for significance

Statistical significance is an outcome much desired by researchers. It means (or ought to mean) that you have found an effect that you were looking for. *The reasoning behind statistical significance works well if you decide what effect you are seeking, design an experiment or sample to search for it, and use a test of significance to weigh the evidence you get.* But because a successful search for a new scientific phenomenon often ends with statistical significance, it is all too tempting to make significance itself the object of the search. There are several ways to do this, none of them acceptable in polite scientific society.

EXAMPLE

6.28 Genomics studies. In genomic experiments, it is common to assess the differences in expression for tens of thousands of genes. If each of these genes was examined separately and statistical significance declared for all that had P-values that pass the 0.05 standard, we would have quite a mess. In the absence of any real biological effects, we would expect that, by chance alone, approximately 5% of these tests will show statistical significance. Much research in genomics is directed toward appropriate ways to deal with this situation.[28]

We do not mean that searching data for suggestive patterns is not proper scientific work. It certainly is. Many important discoveries have been made by accident rather than by design. Exploratory analysis of data is an essential part of statistics. We do mean that the usual reasoning of statistical inference does not apply when the search for a pattern is successful. *You cannot legitimately test a hypothesis on the same data that first suggested that hypothesis.* 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 6.3 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 values of parameters.

On the other hand, lack of significance does not imply that H_0 is true, especially when the test has low power.

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 6.3 Exercises

For Exercise 6.84, see page 383; and for Exercise 6.85, see page 386.

6.86 A role as a statistical consultant. You are the statistical expert for a graduate student planning her PhD research. After you carefully present the mechanics of significance testing, she suggests using $\alpha = 0.20$ for the study because she would be more likely to obtain statistically significant results and she *really* needs significant results to graduate. Explain in simple terms why this would not be a good use of statistical methods.

6.87 What do you know? A research report described two results that both achieved statistical significance at the 5% level. The P-value for the first is 0.048; for the second it is 0.0002. Do the P-values add any useful information beyond that conveyed by the statement that both results are statistically significant? Write a short paragraph explaining your views on this question.

6.88 Selective publication based on results. In addition to statistical significance, selective publication can also be due to the observed outcome. A recent review of 74 FDA-registered studies of antidepressant agents found 38 studies with positive results and 36 studies with negative or questionable results. All but 1 of the 38 positive studies were published. Of the remaining 36, 22 were not published with another 11 published in such a way as to convey a positive outcome.[29] Describe how this selective reporting can have adverse consequences on health care.

6.89 What a test of significance can answer. Explain whether a test of significance can answer each of the following questions.

(a) Is the sample or experiment properly designed?

(b) Is the observed effect compatible with the null hypothesis?

(c) Is the observed effect important?

6.90 Vitamin C and colds. In a study to investigate whether vitamin C will prevent colds, 400 subjects are assigned at random to one of two groups. The experimental group takes a vitamin C tablet daily, while the control group takes a placebo. At the end of the experiment, the researchers calculate the difference between the percents of subjects in the two groups who were free of colds. This difference is statistically significant ($P = 0.03$) in favor of the vitamin C group. Can we conclude that vitamin C has a strong effect in preventing colds? Explain your answer.

6.91 How far do rich parents take us? How much education children get is strongly associated with the wealth and social status of their parents, termed "socioeconomic status," or SES. The SES of parents, however, has little influence on whether children who have graduated from college continue their education. One study looked at whether college graduates took the graduate admissions tests for business, law, and other graduate programs. The effects of the parents' SES on taking the LSAT test for law school were "both statistically insignificant and small."

(a) What does "statistically insignificant" mean?

(b) Why is it important that the effects were small in size as well as insignificant?

6.92 Do you agree? State whether or not you agree with each of the following statements and provide a short summary of the reasons for your answers.

(a) If the P-value is larger than 0.05, the null hypothesis is true.

(b) Practical significance is not the same as statistical significance.

(c) We can perform a statistical analysis using any set of data.

(d) If you find an interesting pattern in a set of data, it is appropriate to then use a significance test to determine its significance.

6.93 Practical significance and sample size. Every user of statistics should understand the distinction between statistical significance and practical importance. A sufficiently large sample will declare very small effects statistically significant. Consider the study of elite female Canadian athletes in Exercise 6.70 (page 380). Female athletes were consuming an average of 2403.7 kcal/day with a standard deviation of 880 kcal/day. Suppose a nutritionist is brought in to implement a new health program for these athletes. This program should increase mean caloric intake but not change the standard deviation. Given the standard deviation and how coloric deficient these athletes are, a change in the mean of 50 kcal/day to 2453.7 is of little importance. However, with a large enough sample, this change can be significant. To see this, calculate the P-value for the test of

$$H_0: \mu = 2403.7$$
$$H_a: \mu > 2403.7$$

in each of the following situations:

(a) A sample of 100 athletes; their average coloric intake is $\bar{x} = 2453.7$.

(b) A sample of 500 athletes; their average coloric intake is $\bar{x} = 2453.7$.

(c) A sample of 2500 athletes; their average coloric intake is $\bar{x} = 2453.7$.

6.94 Statistical versus practical significance. A study with 7500 subjects reported a result that was statistically significant at the 5% level. Explain why this result might not be particularly important.

6.95 More on statistical versus practical significance. A study with 14 subjects reported a result that failed to achieve statistical significance at the 5% level. The P-value was 0.051. Write a short summary of how you would interpret these findings.

6.96 Find journal articles. Find two journal articles that report results with statistical analyses. For each article, summarize how the results are reported and write a critique of the presentation. Be sure to include details regarding use of significance testing at a particular level of significance, P-values, and confidence intervals.

6.97 Create example of your own. For each case, provide an example and an explanation as to why it is appropriate.

(a) A set of data or experiment for which statistical inference is not valid.

(b) A set of data or experiment for which statistical inference is valid.

6.98 Predicting success of trainees. What distinguishes managerial trainees who eventually become executives from those who, after expensive training, don't succeed and leave the company? We have abundant data on past trainees—data on their personalities and goals, their college preparation and performance, even their family backgrounds and their hobbies. Statistical software makes it easy to perform dozens of significance tests on these dozens of variables to see which ones best predict later success. We find that future executives are significantly more likely than washouts to have an urban or suburban upbringing and an undergraduate degree in a technical field.

Explain clearly why using these "significant" variables to select future trainees is not wise. Then suggest a follow-up study using this year's trainees as subjects that should clarify the importance of the variables identified by the first study.

6.99 Searching for significance. Give an example of a situation where searching for significance would lead to misleading conclusions.

6.100 More on searching for significance. You perform 1000 significance tests using $\alpha = 0.05$. Assuming that all null hypotheses are true, about how many of the test results would you expect to be statistically significant? Explain how you obtained your answer.

6.101 Interpreting a very small P-value. Assume that you are performing a large number of significance tests. Let n be the number of these tests. How large would n need to be for you to expect about one P-value to be 0.00001 or smaller? Use this information to write an explanation of how to interpret a result that has $P = 0.00001$ in this setting.

6.102 An adjustment for multiple tests. One way to deal with the problem of misleading P-values when performing more than one significance test is to adjust the criterion you use for statistical significance. The **Bonferroni procedure** does this in a simple way. If you perform two tests and want to use the $\alpha = 5\%$ significance level, you would require a P-value of $0.05/2 = 0.025$ to declare either one of the tests significant. In general, if you perform k tests and want protection at level α, use α/k as your cutoff for statistical significance. You perform six tests and obtain individual P-values 0.083, 0.032, 0.246, 0.003, 0.010, and < 0.001. Which of these are statistically significant using the Bonferroni procedure with $\alpha = 0.05$?

6.103 Significance using the Bonferroni procedure. Refer to the previous problem. A researcher has performed 12 tests of significance and wants to apply the Bonferroni procedure with $\alpha = 0.05$. The calculated P-values are 0.041, 0.569, 0.050, 0.416, 0.002, 0.006, 0.286, 0.021, 0.888, 0.010, < 0.002, and 0.533. Which of these tests reject their null hypotheses with this procedure?

6.4 Power and Inference as a Decision*

Although we prefer to use P-values rather than the reject-or-not view of the fixed α significance test, the latter view is very important for planning studies and for understanding statistical decision theory. We will discuss these two topics in this section.

Power

Fixed level α significance tests are closely related to confidence intervals—in fact, we saw that a two-sided test can be carried out directly from a confidence interval. The significance level, like the confidence level, says how reliable the method is in repeated use. If we use 5% significance tests repeatedly when H_0 is in fact true, we will be wrong (the test will reject H_0) 5% of the time and right (the test will fail to reject H_0) 95% of the time.

The ability of a test to detect that H_0 is false is measured by the probability that the test will reject H_0 when an alternative is true. The higher this probability is, the more sensitive the test is.

POWER

The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test to detect that alternative.

EXAMPLE

6.29 Power of TBBMC significance test. Can a six-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. Based on the results of a previous study, they are willing to assume that $\sigma = 2$ for the percent change in TBBMC over the six-month period. They also feel a change in TBBMC of 1% is important, so they would like to have a reasonable chance of detecting a change this large or larger. Is 25 subjects a large enough sample for this project?

We will answer this question by calculating the power of the significance test that will be used to evaluate the data to be collected. The calculation consists of three steps:

1. State H_0, H_a, the particular alternative we want to detect, and the significance level α.
2. Find the values of $\bar{x}$ that will lead us to reject H_0.
3. Calculate the probability of observing these values of $\bar{x}$ when the alternative is true.

*Although the topics in this section are important in planning and interpreting significance tests, they can be omitted without loss of continuity.