Most statistics packages compute the test statistic for you and report a P-value corresponding to that statistic. And, of course, statistics packages make it easy to examine the boxplots and histograms of the two groups, so you have no excuse for skipping this important check.

Some statistics software automatically tries to test whether the variances of the two groups are equal. Some automatically offer both the two-sample-t and pooled-t results. Ignore the test for the variances; it has little power in any situation in which its results could matter. If the pooled and two-sample methods differ in any important way, you should stick with the two-sample method. Most likely, the Equal Variance Assumption needed for the pooled method has failed.

The degrees of freedom approximation usually gives a fractional value. Most packages seem to round the approximate value down to the next smallest integer (although they may actually compute the P-value with the fractional value, gaining a tiny amount of power).

TWO-SAMPLE METHODS ON THE COMPUTER

Here's some typical computer package output with comments:

May just say “difference of means”
Test Statistic

Some programs will draw a conclusion about the test. Others just give the P-value and let you decide for yourself.

df found from approximation formula and rounded down. The unrounded value may be given, or may be used to find the P-value.

Many programs give far too many digits. Ignore the excess digits.

2-Sample t-Test of \( \mu_1 - \mu_2 = 0 \) vs \( \neq 0 \)
Difference Between Means = 0.99145299 t-Statistic = 1.540
w/196 df
Fail to reject Ho at Alpha = 0.05
P = 0.1251

EXERCISES

1. **Dogs and calories.** In July 2007, Consumer Reports examined the calorie content of two kinds of hot dogs: meat (usually a mixture of pork, turkey, and chicken) and all beef. The researchers purchased samples of several different brands. The meat hot dogs averaged 111.7 calories, compared to 135.4 for the beef hot dogs. A test of the null hypothesis that there’s no difference in mean calorie content yields a P-value of 0.124. Would a 95% confidence interval for \( \mu_{\text{Meat}} - \mu_{\text{Beef}} \) include 0? Explain.

2. **Dogs and sodium.** The Consumer Reports article described in Exercise 1 also listed the sodium content (in mg) for the various hot dogs tested. A test of the null hypothesis that beef hot dogs and meat hot dogs don’t differ in the mean amounts of sodium yields a P-value of 0.11. Would a 95% confidence interval for \( \mu_{\text{Meat}} - \mu_{\text{Beef}} \) include 0? Explain.

3. **Dogs and fat.** The Consumer Reports article described in Exercise 1 also listed the fat content (in grams) for samples of beef and meat hot dogs. The resulting 90% confidence interval for \( \mu_{\text{Meat}} - \mu_{\text{Beef}} \) is \((-6.5, -1.4)\).
   a) The endpoints of this confidence interval are negative numbers. What does that indicate?
   b) What does the fact that the confidence interval does not contain 0 indicate?
   c) If we use this confidence interval to test the hypothesis that \( \mu_{\text{Meat}} - \mu_{\text{Beef}} = 0 \), what’s the corresponding alpha level?

4. **Washers.** In June 2007, Consumer Reports examined top-loading and front-loading washing machines, testing samples of several different brands of each type. One of the variables the article reported was “cycle time”, the number of minutes it took each machine to wash a load of clothes. Among the machines rated good to excellent, the 98% confidence interval for the difference in mean cycle time (\( \mu_{\text{Top}} - \mu_{\text{Front}} \)) is \((-40, -22)\).
   a) The endpoints of this confidence interval are negative numbers. What does that indicate?
b) What does the fact that the confidence interval does not contain 0 indicate?
c) If we use this confidence interval to test the hypothesis that \( \mu_{\text{Top}} - \mu_{\text{Front}} = 0 \), what’s the corresponding alpha level?

5. Dogs and fat, second helping. In Exercise 3, we saw a 90% confidence interval of \((-6.5, -1.4)\) grams for \( \mu_{\text{Meat}} - \mu_{\text{Beef}} \), the difference in mean fat content for meat vs. all-beef hot dogs. Explain why you think each of the following statements is true or false:

a) If I eat a meat hot dog instead of a beef dog, there’s a 90% chance I’ll consume less fat.
b) 90% of meat hot dogs have between 1.4 and 6.5 grams less fat than a beef hot dog.
c) I’m 90% confident that meat hot dogs average 1.4–6.5 grams less fat than the beef hot dogs.
d) If I were to get more samples of both kinds of hot dogs, 90% of the time the meat hot dogs would average 1.4–6.5 grams less fat than the beef hot dogs.
e) If I tested many samples, I’d expect about 90% of the resulting confidence intervals to include the true difference in mean fat content between the two kinds of hot dogs.

6. Second load of wash. In Exercise 4, we saw a 98% confidence interval of \((-40, -22)\) minutes for \( \mu_{\text{Top}} - \mu_{\text{Front}} \), the difference in time it takes top-loading and front-loading washers to do a load of clothes. Explain why you think each of the following statements is true or false:

a) 98% of top loaders are 22 to 40 minutes faster than front loaders.
b) If I choose the laundromat’s top loader, there’s a 98% chance that my clothes will be done faster than if I had chosen the front loader.
c) If I tried more samples of both kinds of washing machines, in about 98% of these samples I’d expect the top loaders to be an average of 22 to 40 minutes faster.
d) If I tried more samples, I’d expect about 98% of the resulting confidence intervals to include the true difference in mean cycle time for the two types of washing machines.
e) I’m 98% confident that top loaders wash clothes an average of 22 to 40 minutes faster than front-loaders.

7. Learning math. The Core Plus Mathematics Project (CPMP) is an innovative approach to teaching Mathematics that engages students in group investigations and mathematical modeling. After field tests in 36 high schools over a three-year period, researchers compared the performances of CPMP students with those taught using a traditional curriculum. In one test, students had to solve applied Algebra problems using calculators. Scores for 320 CPMP students were compared to those of a control group of 273 students in a traditional Math program. Computer software was used to create a confidence interval for the difference in mean scores. (Journal for Research in Mathematics Education, 31, no. 3[2000])

Conf level: 95% Variable: Mu(CPMP) - Mu(Control)
Inter val: (5.573, 11.427)

a) What’s the margin of error for this confidence interval?
b) If we had created a 98% CI, would the margin of error be larger or smaller?
c) Explain what the calculated interval means in context.
d) Does this result suggest that students who learn Mathematics with CPMP will have significantly higher mean scores in Algebra than those in traditional programs? Explain.

8. Stereograms. Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can “fuse” into a three-dimensional (3D) image by staring at the dots while defocusing the eyes. An experiment was performed to determine whether knowledge of the form of the embedded image affected the time required for subjects to fuse the images. One group of subjects (group NV) received no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (specifically, a drawing of the object). The experimenters measured how many seconds it took for the subject to report that he or she saw the 3D image.

\[
\begin{align*}
2-\text{Sample t-Interval for } & \mu_{\text{Top}} - \mu_{\text{Front}} \\
\text{Conf level} = 90% & \\
\text{df} = 70 & \\
\mu_{\text{NV}} - \mu_{\text{VV}} & = (0.55, 5.47)
\end{align*}
\]

a) Interpret your interval in context.
b) Does it appear that viewing a picture of the image helps people “see” the 3D image in a stereogram?
c) What’s the margin of error for this interval?
d) Explain what the 90% confidence level means.
e) Would you expect a 99% confidence level to be wider or narrower? Explain.
f) Might that change your conclusion in part b? Explain.

9. CPMP, again. During the study described in Exercise 7, students in both CPMP and traditional classes took another Algebra test that did not allow them to use calculators. The table below shows the results. Are the mean scores of the two groups significantly different?

<table>
<thead>
<tr>
<th>Math Program</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPMP</td>
<td>312</td>
<td>29.0</td>
<td>18.8</td>
</tr>
<tr>
<td>Traditional</td>
<td>265</td>
<td>38.4</td>
<td>16.2</td>
</tr>
</tbody>
</table>

Performance on Algebraic Symbolic Manipulation Without Use of Calculators

a) Write an appropriate hypothesis.
b) Do you think the assumptions for inference are satisfied? Explain.
c) Here is computer output for this hypothesis test. Explain what the P-value means in this context.

\[
\begin{align*}
\text{2-Sample t-Test of } & \mu_1 - \mu_2 \\
\text{t-Statistic} & = -6.451 \quad w/574.8761 \quad \text{df} \\
P & < 0.0001
\end{align*}
\]

d) State a conclusion about the CPMP program.

10. CPMP and word problems. The study of the new CPMP Mathematics methodology described in Exercise 7 also tested students’ abilities to solve word problems. This
table shows how the CPMP and traditional groups performed. What do you conclude?

<table>
<thead>
<tr>
<th>Math Program</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPMP</td>
<td>320</td>
<td>57.4</td>
<td>32.1</td>
</tr>
<tr>
<td>Traditional</td>
<td>273</td>
<td>53.9</td>
<td>28.5</td>
</tr>
</tbody>
</table>

11. **Commuting.** A man who moves to a new city sees that there are two routes he could take to work. A neighbor who has lived there a long time tells him Route A will average 5 minutes faster than Route B. The man decides to experiment. Each day he flips a coin to determine which way to go, driving each route 20 days. He finds that Route A takes an average of 40 minutes, with standard deviation 3 minutes, and Route B takes an average of 43 minutes, with standard deviation 2 minutes. Histograms of travel times for the routes are roughly symmetric and show no outliers.

a) Find a 95% confidence interval for the difference in average commuting time for the two routes.

b) Should the man believe the old-timer’s claim that he can save an average of 5 minutes a day by always driving Route A? Explain.

12. **Pulse rates.** A researcher wanted to see whether there is a significant difference in resting pulse rates for men and women. The data she collected are displayed in the boxplots and summarized below.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Count</th>
<th>Mean</th>
<th>Median</th>
<th>StdDev</th>
<th>Range</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>28</td>
<td>72.75</td>
<td>73</td>
<td>5.37225</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>Female</td>
<td>24</td>
<td>72.625</td>
<td>73</td>
<td>7.69987</td>
<td>29</td>
<td>12.5</td>
</tr>
</tbody>
</table>

a) What do the boxplots suggest about differences between male and female pulse rates?

b) Is it appropriate to analyze these data using the methods of inference discussed in this chapter? Explain.

c) Create a 90% confidence interval for the difference in mean pulse rates.

d) Does the confidence interval confirm your answer to part a? Explain.

13. **Cereal.** The data below show the sugar content (as a percentage of weight) of several national brands of children’s and adults’ cereals. Create and interpret a 95% confidence interval for the difference in mean sugar content. Be sure to check the necessary assumptions and conditions.

14. **Egyptians.** Some archaeologists theorize that ancient Egyptians interbred with several different immigrant populations over thousands of years. To see if there is any indication of changes in body structure that might have resulted, they measured 30 skulls of male Egyptians dated from 4000 B.C.E and 30 others dated from 200 B.C.E. (A. Thomson and R. Randall-Maciver, *Ancient Races of the Thebaid*, Oxford: Oxford University Press, 1905)

a) Are these data appropriate for inference? Explain.

b) Create a 95% confidence interval for the difference in mean skull breadth between these two eras.

c) Do these data provide evidence that the mean breadth of males’ skulls changed over this period? Explain.

15. **Reading.** An educator believes that new reading activities for elementary school children will improve reading comprehension scores. She randomly assigns third graders to an eight-week program in which some will use these activities and others will experience traditional teaching methods. At the end of the experiment, both groups take a reading comprehension exam. Their scores are shown in the back-to-back stem-and-leaf display. Do these results suggest that the new activities are better? Test an appropriate hypothesis and state your conclusion.

10. **Pulse rates.** A researcher wanted to see whether there is a significant difference in resting pulse rates for men and women. The data she collected are displayed in the boxplots and summarized below.
16. Streams. Researchers collected samples of water from streams in the Adirondack Mountains to investigate the effects of acid rain. They measured the pH (acidity) of the water and classified the streams with respect to the kind of substrate (type of rock over which they flow). A lower pH means the water is more acidic. Here is a plot of the pH of the streams by substrate (limestone, mixed, or shale):

Here are selected parts of a software analysis comparing the pH of streams with limestone and shale substrates:

\[ t-Test: \mu_1 - \mu_2 = 0.735 \]
\[ t = 16.30 \text{ with 133 degrees of freedom} \]
\[ p < 0.0001 \]

a) State the null and alternative hypotheses for this test.

b) From the information you have, do the assumptions and conditions appear to be met?

c) What conclusion would you draw?

17. Baseball 2006. American League baseball teams play their games with the designated hitter rule, meaning that pitchers do not bat. The league believes that replacing the pitcher, traditionally a weak hitter, with another player in the batting order produces more runs and generates more interest among fans. Below are the average numbers of runs scored in American League and National League stadiums for the 2006 season.

<table>
<thead>
<tr>
<th></th>
<th>American</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11.4</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>10.5</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>10.4</td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td>10.3</td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td>10.2</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td>10.0</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td>9.9</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>8.8</td>
<td>9.5</td>
</tr>
</tbody>
</table>

a) Create an appropriate display of these data. What do you see?

b) With a 95% confidence interval, estimate the mean number of runs scored in American League games.

c) Coors Field, in Denver, stands a mile above sea level, an altitude far greater than that of any other National League ball park. Some believe that the thinner air makes it harder for pitchers to throw curve balls and easier for batters to hit the ball a long way. Do you think the 10.5 runs scored per game at Coors is unusual? Explain.

d) Explain why you should not use two separate confidence intervals to decide whether the two leagues differ in average number of runs scored.

18. Handy. A factory hiring people to work on an assembly line gives job applicants a test of manual agility. This test counts how many strangely shaped pegs the applicant can fit into matching holes in a one-minute period. The table below summarizes the data by sex of the job applicant. Assume that all conditions necessary for inference are met.

<table>
<thead>
<tr>
<th>Pegs placed:</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Mean</td>
<td>19.39</td>
<td>17.91</td>
</tr>
<tr>
<td>SD</td>
<td>2.52</td>
<td>3.39</td>
</tr>
</tbody>
</table>

a) Find 95% confidence intervals for the average number of pegs that males and females can each place.

b) Those intervals overlap. What does this suggest about any sex-based difference in manual agility?

c) Find a 95% confidence interval for the difference in the mean number of pegs that could be placed by men and women.

d) What does this interval suggest about any difference in manual agility between men and women?

e) The two results seem contradictory. Which method is correct: doing two-sample inference or doing one-sample inference twice?

f) Why don’t the results agree?

19. Double header 2006. Do the data in Exercise 17 suggest that the American League’s designated hitter rule may lead to more runs?

a) Using a 95% confidence interval, estimate the difference between the mean number of runs scored in American and National League games.

b) Interpret your interval.

c) Does that interval suggest that the two leagues may differ in average number of runs scored per game?

20. Hard water. In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, ppm) in the drinking water. The data set also notes, for each town, whether it was south or north of Derby. Is there a significant difference in mortality rates in the two regions? Here are the summary statistics.

<table>
<thead>
<tr>
<th></th>
<th>Derby</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>34</td>
</tr>
<tr>
<td>Mean</td>
<td>1631.59</td>
</tr>
<tr>
<td>Median</td>
<td>1631</td>
</tr>
<tr>
<td>StdDev</td>
<td>138.470</td>
</tr>
<tr>
<td>Count</td>
<td>27</td>
</tr>
<tr>
<td>Mean</td>
<td>1388.85</td>
</tr>
<tr>
<td>Median</td>
<td>1369</td>
</tr>
<tr>
<td>StdDev</td>
<td>151.114</td>
</tr>
</tbody>
</table>

a) Test appropriate hypotheses and state your conclusion.

b) On the next page, the boxplots of the two distributions show an outlier among the data north of Derby. What effect might that have had on your test?
21. **Job satisfaction.** A company institutes an exercise break for its workers to see if this will improve job satisfaction, as measured by a questionnaire that assesses workers’ satisfaction. Scores for 10 randomly selected workers before and after implementation of the exercise program are shown. The company wants to assess the effectiveness of the exercise program. Explain why you can’t use the methods discussed in this chapter to do that. (Don’t worry, we’ll give you another chance to do this the right way.)

<table>
<thead>
<tr>
<th>Worker Number</th>
<th>Job Satisfaction Index Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>41</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>41</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>27</td>
<td>37</td>
</tr>
</tbody>
</table>

22. **Summer school.** Having done poorly on their math final exams in June, six students repeat the course in summer school, then take another exam in August. If we consider these students representative of all students who might attend this summer school in other years, do these results provide evidence that the program is worthwhile?

<table>
<thead>
<tr>
<th></th>
<th>June</th>
<th>Aug</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of subjects</td>
<td>54</td>
<td>50</td>
</tr>
<tr>
<td>Mean</td>
<td>49.5</td>
<td>66.5</td>
</tr>
<tr>
<td>SD</td>
<td>6.5</td>
<td>4.7</td>
</tr>
</tbody>
</table>

23. **Sex and violence.** In June 2002, the *Journal of Applied Psychology* reported on a study that examined whether the content of TV shows influenced the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned volunteers to watch one of three programs, each containing the same nine commercials. One of the programs had violent content, another sexual content, and the third neutral content. After the shows ended, the subjects were asked to recall the brands of products that were advertised. Here are summaries of the results:

<table>
<thead>
<tr>
<th>Program Type</th>
<th>Violent</th>
<th>Sexual</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of subjects</td>
<td>108</td>
<td>108</td>
<td>108</td>
</tr>
<tr>
<td>Brands recalled</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.08</td>
<td>1.71</td>
<td>3.17</td>
</tr>
<tr>
<td>SD</td>
<td>1.87</td>
<td>1.76</td>
<td>1.77</td>
</tr>
</tbody>
</table>

a) Do these results indicate that viewer memory for ads may differ depending on program content? A test of the hypothesis that there is no difference in ad memory between programs with sexual content and those with violent content has a P-value of 0.136. State your conclusion.

b) Is there evidence that viewer memory for ads may differ between programs with sexual content and those with neutral content? Test an appropriate hypothesis and state your conclusion.

24. **Ad campaign.** You are a consultant to the marketing department of a business preparing to launch an ad campaign for a new product. The company can afford to run ads during one TV show, and has decided not to sponsor a show with sexual content. You read the study described in Exercise 23, then use a computer to create a confidence interval for the difference in mean number of brand names remembered between the groups watching violent shows and those watching neutral shows.

\[ \text{TWO-SAMPLE T} \]

95\% CI FOR \( \mu_{\text{viol}} - \mu_{\text{neut}} \): (–1.578, –0.602)

a) At the meeting of the marketing staff, you have to explain what this output means. What will you say?

b) What advice would you give the company about the upcoming ad campaign?

25. **Sex and violence II.** In the study described in Exercise 23, the researchers also contacted the subjects again, 24 hours later, and asked them to recall the brands advertised. Results are summarized below.

<table>
<thead>
<tr>
<th>Program Type</th>
<th>Violent</th>
<th>Sexual</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of subjects</td>
<td>101</td>
<td>106</td>
<td>103</td>
</tr>
<tr>
<td>Brands recalled</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3.02</td>
<td>2.72</td>
<td>4.65</td>
</tr>
<tr>
<td>SD</td>
<td>1.61</td>
<td>1.85</td>
<td>1.62</td>
</tr>
</tbody>
</table>

a) Is there a significant difference in viewers’ abilities to remember brands advertised in shows with violent vs. neutral content?

b) Find a 95\% confidence interval for the difference in mean number of brand names remembered between the groups watching shows with sexual content and those watching neutral shows. Interpret your interval in this context.

26. **Ad recall.** In Exercises 23 and 25, we see the number of advertised brand names people recalled immediately after watching TV shows and 24 hours later. Strangely...
enough, it appears that they remembered more about the ads the next day. Should we conclude this is true in general about people’s memory of TV ads?

a) Suppose one analyst conducts a two-sample hypothesis test to see if memory of brands advertised during violent TV shows is higher 24 hours later. If his P-value is 0.00013, what might he conclude?
b) Explain why his procedure was inappropriate. Which of the assumptions for inference was violated?
c) How might the design of this experiment have tainted the results?
d) Suggest a design that could compare immediate brand-name recall with recall one day later.

27. Hungry? Researchers investigated how the size of a bowl affects how much ice cream people tend to scoop when serving themselves. At an “ice cream social,” people were randomly given either a 17 oz or a 34 oz bowl (both large enough that they would not be filled to capacity). They were then invited to scoop as much ice cream as they liked. Did the bowl size change the selected portion size? Here are the summaries:

<table>
<thead>
<tr>
<th>Small Bowl</th>
<th>Large Bowl</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>26</td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>5.07 oz</td>
</tr>
<tr>
<td>s</td>
<td>1.84 oz</td>
</tr>
</tbody>
</table>

Test an appropriate hypothesis and state your conclusions. Assume any assumptions and conditions that you cannot test are sufficiently satisfied to proceed.

28. Thirsty? Researchers randomly assigned participants either a tall, thin “highball” glass or a short, wide “tumbler,” each of which held 355 ml. Participants were asked to pour a shot (1.5 oz = 44.3 ml) into their glass. Did the shape of the glass make a difference in how much liquid they poured?

Here are the summaries:

<table>
<thead>
<tr>
<th>highball</th>
<th>tumbler</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>99</td>
</tr>
<tr>
<td>$\bar{y}$</td>
<td>42.2 ml</td>
</tr>
<tr>
<td>s</td>
<td>16.2 ml</td>
</tr>
</tbody>
</table>

Test an appropriate hypothesis and state your conclusions. Assume any assumptions and conditions that you cannot test are sufficiently satisfied to proceed.

29. Lower scores? Newspaper headlines recently announced a decline in science scores among high school seniors. In 2000, a total of 15,109 seniors tested by The National Assessment in Education Program (NAEP) scored a mean of 147 points. Four years earlier, 7537 seniors had averaged 150 points. The standard error of the difference in the mean scores for the two groups was 1.22.

a) Have the science scores declined significantly? Cite appropriate statistical evidence to support your conclusion.
b) The sample size in 2000 was almost double that in 1996. Does this make the results more convincing or less? Explain.

30. The Internet. The NAEP report described in Exercise 29 compared science scores for students who had home Internet access to the scores of those who did not, as shown in the graph. They report that the differences are statistically significant.

a) Explain what “statistically significant” means in this context.
b) If their conclusion is incorrect, which type of error did the researchers commit?
c) Does this prove that using the Internet at home can improve a student’s performance in science?

31. Running heats. In Olympic running events, preliminary heats are determined by random draw, so we should expect that the abilities of runners in the various heats to be about the same, on average. Here are the times (in seconds) for the 400-m women’s run in the 2004 Olympics in Athens for preliminary heats 2 and 5. Is there any evidence that the mean time to finish is different for randomized heats? Explain. Be sure to include a discussion of assumptions and conditions for your analysis.

32. Swimming heats. In Exercise 31 we looked at the times in two different heats for the 400-m women’s run from the 2004 Olympics. Unlike track events, swimming heats are not determined at random. Instead, swimmers...
Exercises 585

are seeded so that better swimmers are placed in later heats. Here are the times (in seconds) for the women’s 400-m freestyle from heats 2 and 5. Do these results suggest that the mean times of seeded heats are not equal? Explain. Include a discussion of assumptions and conditions for your analysis.

<table>
<thead>
<tr>
<th>Country</th>
<th>Name</th>
<th>Heat</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>BIAGIOLI Cecilia Elizabeth</td>
<td>2</td>
<td>256.42</td>
</tr>
<tr>
<td>SLO</td>
<td>CARMAN Anja</td>
<td>2</td>
<td>257.79</td>
</tr>
<tr>
<td>CHI</td>
<td>KOBRICH Kristel</td>
<td>2</td>
<td>258.68</td>
</tr>
<tr>
<td>MKD</td>
<td>STOJANOVSKA Vesna</td>
<td>2</td>
<td>259.39</td>
</tr>
<tr>
<td>JAM</td>
<td>ATKINSON Janelle</td>
<td>2</td>
<td>260.00</td>
</tr>
<tr>
<td>NZL</td>
<td>LINTON Rebecca</td>
<td>2</td>
<td>261.58</td>
</tr>
<tr>
<td>KOR</td>
<td>HA Eun-Ju</td>
<td>2</td>
<td>261.65</td>
</tr>
<tr>
<td>UKR</td>
<td>BERESNYEVA Olga</td>
<td>2</td>
<td>266.30</td>
</tr>
<tr>
<td>FRA</td>
<td>MANAOUOU Laure</td>
<td>5</td>
<td>246.76</td>
</tr>
<tr>
<td>JPN</td>
<td>YAMADA Sachiko</td>
<td>5</td>
<td>249.10</td>
</tr>
<tr>
<td>ROM</td>
<td>PADARARU Simona</td>
<td>5</td>
<td>250.39</td>
</tr>
<tr>
<td>GER</td>
<td>STOCKBAUER Hannah</td>
<td>5</td>
<td>250.46</td>
</tr>
<tr>
<td>AUS</td>
<td>GRAHAM Elka</td>
<td>5</td>
<td>251.67</td>
</tr>
<tr>
<td>CHN</td>
<td>PANG Jiaying</td>
<td>5</td>
<td>251.81</td>
</tr>
<tr>
<td>CAN</td>
<td>REIMER Brittany</td>
<td>5</td>
<td>252.33</td>
</tr>
<tr>
<td>BRA</td>
<td>FERREIRA Monique</td>
<td>5</td>
<td>253.75</td>
</tr>
</tbody>
</table>

33. **Tees.** Does it matter what kind of tee a golfer places the ball on? The company that manufactures “Stinger” tees claims that the thinner shaft and smaller head will lessen drag, reducing spin and allowing the ball to travel farther. In August 2003, Golf Laboratories, Inc., compared the distance traveled by golf balls hit off regular wooden tees to those hit off Stinger tees. All the balls were struck by the same golf club using a robotic device set to swing the club head at approximately 95 miles per hour. Summary statistics from the test are shown in the table. Assume that 6 balls were hit off each tee and that the data were suitable for inference.

<table>
<thead>
<tr>
<th>Country</th>
<th>Name</th>
<th>Heat</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>BIAGIOLI Cecilia Elizabeth</td>
<td>2</td>
<td>256.42</td>
</tr>
<tr>
<td>SLO</td>
<td>CARMAN Anja</td>
<td>2</td>
<td>257.79</td>
</tr>
<tr>
<td>CHI</td>
<td>KOBRICH Kristel</td>
<td>2</td>
<td>258.68</td>
</tr>
<tr>
<td>MKD</td>
<td>STOJANOVSKA Vesna</td>
<td>2</td>
<td>259.39</td>
</tr>
<tr>
<td>JAM</td>
<td>ATKINSON Janelle</td>
<td>2</td>
<td>260.00</td>
</tr>
<tr>
<td>NZL</td>
<td>LINTON Rebecca</td>
<td>2</td>
<td>261.58</td>
</tr>
<tr>
<td>KOR</td>
<td>HA Eun-Ju</td>
<td>2</td>
<td>261.65</td>
</tr>
<tr>
<td>UKR</td>
<td>BERESNYEVA Olga</td>
<td>2</td>
<td>266.30</td>
</tr>
<tr>
<td>FRA</td>
<td>MANAOUOU Laure</td>
<td>5</td>
<td>246.76</td>
</tr>
<tr>
<td>JPN</td>
<td>YAMADA Sachiko</td>
<td>5</td>
<td>249.10</td>
</tr>
<tr>
<td>ROM</td>
<td>PADARARU Simona</td>
<td>5</td>
<td>250.39</td>
</tr>
<tr>
<td>GER</td>
<td>STOCKBAUER Hannah</td>
<td>5</td>
<td>250.46</td>
</tr>
<tr>
<td>AUS</td>
<td>GRAHAM Elka</td>
<td>5</td>
<td>251.67</td>
</tr>
<tr>
<td>CHN</td>
<td>PANG Jiaying</td>
<td>5</td>
<td>251.81</td>
</tr>
<tr>
<td>CAN</td>
<td>REIMER Brittany</td>
<td>5</td>
<td>252.33</td>
</tr>
<tr>
<td>BRA</td>
<td>FERREIRA Monique</td>
<td>5</td>
<td>253.75</td>
</tr>
</tbody>
</table>

34. **Golf again.** Given the test results on golf tees described in Exercise 33, is there evidence that balls hit off Stinger tees would travel farther? Again, assume that 6 balls were hit off each tee and that the data were suitable for inference.

<table>
<thead>
<tr>
<th>Country</th>
<th>Name</th>
<th>Heat</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>BIAGIOLI Cecilia Elizabeth</td>
<td>2</td>
<td>256.42</td>
</tr>
<tr>
<td>SLO</td>
<td>CARMAN Anja</td>
<td>2</td>
<td>257.79</td>
</tr>
<tr>
<td>CHI</td>
<td>KOBRICH Kristel</td>
<td>2</td>
<td>258.68</td>
</tr>
<tr>
<td>MKD</td>
<td>STOJANOVSKA Vesna</td>
<td>2</td>
<td>259.39</td>
</tr>
<tr>
<td>JAM</td>
<td>ATKINSON Janelle</td>
<td>2</td>
<td>260.00</td>
</tr>
<tr>
<td>NZL</td>
<td>LINTON Rebecca</td>
<td>2</td>
<td>261.58</td>
</tr>
<tr>
<td>KOR</td>
<td>HA Eun-Ju</td>
<td>2</td>
<td>261.65</td>
</tr>
<tr>
<td>UKR</td>
<td>BERESNYEVA Olga</td>
<td>2</td>
<td>266.30</td>
</tr>
<tr>
<td>FRA</td>
<td>MANAOUOU Laure</td>
<td>5</td>
<td>246.76</td>
</tr>
<tr>
<td>JPN</td>
<td>YAMADA Sachiko</td>
<td>5</td>
<td>249.10</td>
</tr>
<tr>
<td>ROM</td>
<td>PADARARU Simona</td>
<td>5</td>
<td>250.39</td>
</tr>
<tr>
<td>GER</td>
<td>STOCKBAUER Hannah</td>
<td>5</td>
<td>250.46</td>
</tr>
<tr>
<td>AUS</td>
<td>GRAHAM Elka</td>
<td>5</td>
<td>251.67</td>
</tr>
<tr>
<td>CHN</td>
<td>PANG Jiaying</td>
<td>5</td>
<td>251.81</td>
</tr>
<tr>
<td>CAN</td>
<td>REIMER Brittany</td>
<td>5</td>
<td>252.33</td>
</tr>
<tr>
<td>BRA</td>
<td>FERREIRA Monique</td>
<td>5</td>
<td>253.75</td>
</tr>
</tbody>
</table>

35. **Crossing Ontario.** Between 1954 and 2003, swimmers have crossed Lake Ontario 43 times. Both women and men have made the crossing. Here are some plots (we’ve omitted a crossing by Vikki Keith, who swam a round trip—North to South to North—in 3390 minutes):

The summary statistics are:

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
<th>Mean</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>22</td>
<td>1271.59</td>
<td>261.111</td>
</tr>
<tr>
<td>M</td>
<td>20</td>
<td>1196.75</td>
<td>304.369</td>
</tr>
</tbody>
</table>

How much difference is there between the mean amount of time (in minutes) it would take female and male swimmers to swim the lake?

a) Construct and interpret a 95% confidence interval for the difference between female and male times.

b) Comment on the assumptions and conditions.

36. **Music and memory.** Is it a good idea to listen to music when studying for a big test? In a study conducted by some Statistics students, 62 people were randomly assigned to listen to rap music, music by Mozart, or no music while attempting to memorize objects pictured on a page. They were then asked to list all the objects they could remember. Here are summary statistics:

<table>
<thead>
<tr>
<th>Music</th>
<th>Count</th>
<th>Mean</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rap</td>
<td>29</td>
<td>10.72</td>
<td>3.99</td>
</tr>
<tr>
<td>Mozart</td>
<td>20</td>
<td>10.00</td>
<td>3.19</td>
</tr>
<tr>
<td>No Music</td>
<td>13</td>
<td>12.77</td>
<td>4.73</td>
</tr>
</tbody>
</table>

a) Does it appear that it is better to study while listening to Mozart than to rap music? Test an appropriate hypothesis and state your conclusion.

b) Create a 90% confidence interval for the mean difference in memory score between students who study to Mozart and those who listen to no music at all. Interpret your interval.

37. **Rap.** Using the results of the experiment described in Exercise 36, does it matter whether one listens to rap music while studying, or is it better to study without music at all?

a) Test an appropriate hypothesis and state your conclusion.

b) If you concluded there is a difference, estimate the size of that difference with a confidence interval and explain what your interval means.
38. Cuckoos. Cuckoos lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. But the potential host birds lay eggs of different sizes. Does the cuckoo change the size of her eggs for different foster species? The numbers in the table are lengths (in mm) of cuckoo eggs found in nests of three different species of other birds. The data are drawn from the work of O.M. Latter in 1902 and were used in a fundamental textbook on statistical quality control by L.H.C. Tippett (1902–1985), one of the pioneers in that field.

<table>
<thead>
<tr>
<th>CUCKOO EGG LENGTH (MM)</th>
<th>Foster Parent Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sparrow</td>
</tr>
<tr>
<td>20.85</td>
<td>21.05</td>
</tr>
<tr>
<td>21.65</td>
<td>21.85</td>
</tr>
<tr>
<td>22.05</td>
<td>22.05</td>
</tr>
<tr>
<td>22.85</td>
<td>22.05</td>
</tr>
<tr>
<td>23.05</td>
<td>22.05</td>
</tr>
<tr>
<td>23.45</td>
<td>22.45</td>
</tr>
<tr>
<td>23.05</td>
<td>22.45</td>
</tr>
<tr>
<td>23.05</td>
<td>22.65</td>
</tr>
<tr>
<td>23.05</td>
<td>23.05</td>
</tr>
<tr>
<td>23.05</td>
<td>23.05</td>
</tr>
<tr>
<td>23.05</td>
<td>23.05</td>
</tr>
<tr>
<td>24.05</td>
<td>23.05</td>
</tr>
<tr>
<td>25.05</td>
<td>23.05</td>
</tr>
<tr>
<td>23.25</td>
<td>24.85</td>
</tr>
</tbody>
</table>

Investigate the question of whether the mean length of cuckoo eggs is the same for different species, and state your conclusion.