CHAPTER 18 • Inference about a Population Mean

CAUTION: Students should be warned that in practice it is imperative to consider these conditions for inference about a mean prior to any analysis. They should be encouraged to check these conditions.

CONDITIONS FOR INference ABOUT A MEAN

- We can regard our data as a simple random sample (SRS) from the population. This condition is very important.
- Observations from the population have a Normal distribution with mean \( \mu \) and standard deviation \( \sigma \). In practice, it is enough that the distribution be symmetric and single-peaked unless the sample is very small. Both \( \mu \) and \( \sigma \) are unknown parameters.

There is another condition that applies to all of the inference methods in this book: the population must be much larger than the sample, say at least 20 times as large. All of our examples and exercises satisfy this condition. Practical settings in which the sample is a large part of the population are rather special, and we will not discuss them.

When the conditions for inference are satisfied, the sample mean \( \bar{x} \) has the Normal distribution with mean \( \mu \) and standard deviation \( \sigma / \sqrt{n} \). Because we don’t know \( \sigma \), we estimate it by the sample standard deviation \( s \). We then estimate the standard deviation of \( \bar{x} \) by \( s / \sqrt{n} \). This quantity is called the standard error of the sample mean \( \bar{x} \).

STANDARD ERROR

When the standard deviation of a statistic is estimated from data, the result is called the standard error of the statistic. The standard error of the sample mean \( \bar{x} \) is \( s / \sqrt{n} \).

APPLY YOUR KNOWLEDGE

18.1 Travel time to work. A study of commuting times reports the travel times to work of a random sample of 20 employed adults in New York State. The mean is \( \bar{x} = 31.25 \) minutes and the standard deviation is \( s = 21.88 \) minutes. What is the standard error of the mean?

18.2 Rats eating oat bran. In a study of the effect of diet on cholesterol, rats were fed several different diets. One diet had 5% added fiber from oat bran. The study report gives results in the form “mean plus or minus the standard error of the mean.” This form is very common in scientific publications. For the 6 rats fed this diet, blood cholesterol levels (in milligrams per deciliter of blood) were 89.01 ± 5.36. What are \( \bar{x} \) and \( s \) for these 6 rats?
The t distributions

If we knew the value of \( \sigma \), we would base confidence intervals and tests for \( \mu \) on the one-sample \( z \) statistic

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

This \( z \) statistic has the standard Normal distribution \( N(0, 1) \). In practice, we don’t know \( \sigma \), so we substitute the standard error \( s/\sqrt{n} \) of \( \bar{x} \) for its standard deviation \( \sigma/\sqrt{n} \). The statistic that results does not have a Normal distribution. It has a distribution that is new to us, called a \( t \) distribution.

THE ONE-SAMPLE \( t \) STATISTIC AND THE \( t \) DISTRIBUTIONS

Draw an SRS of size \( n \) from a large population that has the Normal distribution with mean \( \mu \) and standard deviation \( \sigma \). The one-sample \( t \) statistic

\[
t = \frac{\bar{x} - \mu}{s/\sqrt{n}}
\]

has the \( t \) distribution with \( n - 1 \) degrees of freedom.

The \( t \) statistic has the same interpretation as any standardized statistic: it says how far \( \bar{x} \) is from its mean \( \mu \) in standard deviation units. There is a different \( t \) distribution for each sample size. We specify a particular \( t \) distribution by giving its degrees of freedom. The degrees of freedom for the one-sample \( t \) statistic come from the sample standard deviation \( s \) in the denominator of \( t \). We saw in Chapter 2 (page 49) that \( s \) has \( n - 1 \) degrees of freedom. There are other \( t \) statistics with different degrees of freedom, some of which we will meet later. We will write the \( t \) distribution with \( n - 1 \) degrees of freedom as \( t(n - 1) \) for short.

Figure 18.1 compares the density curves of the standard Normal distribution and the \( t \) distributions with 2 and 9 degrees of freedom. The figure illustrates these facts about the \( t \) distributions:

- The density curves of the \( t \) distributions are similar in shape to the standard Normal curve. They are symmetric about 0, single-peaked, and bell-shaped.
- The spread of the \( t \) distributions is a bit greater than that of the standard Normal distribution. The \( t \) distributions in Figure 18.1 have more probability in the tails and less in the center than does the standard Normal. This is true because substituting the estimate \( s \) for the fixed parameter \( \sigma \) introduces more variation into the statistic.
- As the degrees of freedom increase, the \( t \) density curve approaches the \( N(0, 1) \) curve ever more closely. This happens because \( s \) estimates \( \sigma \) more accurately as the sample size increases. So using \( s \) in place of \( \sigma \) causes little extra variation when the sample is large.
The $F$ test for $H_0$: $\sigma_1 = \sigma_2$ and other procedures for inference on the spread of one or more Normal distributions are so strongly affected by lack of Normality that we do not recommend them for regular use.

**CHECK YOUR SKILLS**

19.22 In 2003, the National Assessment of Educational Progress gave a mathematics test to 4398 eighth-graders in Texas. The mean score was 277 out of 500, with standard error 1.1. To give a confidence interval for the mean score of all Texas eighth-graders, you would use
(a) the one-sample $t$ interval.
(b) the matched pairs $t$ interval.
(c) the two-sample $t$ interval.

19.23 In 2003, the National Assessment of Educational Progress gave a mathematics test to 4398 eighth-graders in Texas. To test whether there is a difference in the mean scores of all female and male eighth-grade students in Texas, you would use
(a) the one-sample $t$ test.
(b) the matched pairs $t$ test.
(c) the two-sample $t$ test.

19.24 There are two common methods for measuring the concentration of a pollutant in fish tissue. Do the two methods differ on the average? You apply both methods to a sample of 18 carp and use
(a) the one-sample $t$ test.
(b) the matched pairs $t$ test.
(c) the two-sample $t$ test.

19.25 A study of the effects of exercise used rats bred to have high or low capacity for exercise. There were 8 high-capacity and 8 low-capacity rats. To compare the mean blood pressure of the two types of rats using the conservative Option 2 $t$ procedures, the correct degrees of freedom is
(a) 7.  (b) 14.  (c) 15.

19.26 The 8 high-capacity rats had mean blood pressure 89 with standard deviation 9; the 8 low-capacity rats had mean blood pressure 105 with standard deviation 13. (Blood pressure is measured in millimeters of mercury.) The two-sample $t$ statistic for comparing the population means has value
(a) 0.5.  (b) 2.86.  (c) 9.65.

19.27 A study of road rage asked samples of 596 men and 523 women about their behavior while driving. Based on their answers, each subject was assigned a road rage score on a scale of 0 to 20. The subjects were chosen by random digit dialing of telephone numbers. Are the conditions for two-sample $t$ inference satisfied?
(a) Maybe: the SRS condition is OK but we need to look at the data to check Normality.
(b) No: scores in a range between 0 and 20 can't be Normal.
(c) Yes: the SRS condition is OK and large sample sizes make the Normality condition unnecessary.

19.28 We suspect that men are more prone to road rage than women. To see if this is true, test these hypotheses for the mean road rage scores of all male and female drivers:
(a) $H_0: \mu_M = \mu_F$ versus $H_a: \mu_M > \mu_F$.
(b) $H_0: \mu_M = \mu_F$ versus $H_a: \mu_M \neq \mu_F$.
(c) $H_0: \mu_M = \mu_F$ versus $H_a: \mu_M < \mu_F$.

19.29 The two-sample $t$ statistic for the road rage study (male mean minus female mean) is $t = 3.18$. The $P$-value for testing the hypotheses from the previous exercise satisfies
(a) $0.001 < P < 0.005$. (b) $0.0005 < P < 0.001$. (c) $0.001 < P < 0.002$.

CHAPTER 19 EXERCISES

In exercises that call for two-sample $t$ procedures, use Option 1 if you have technology that implements that method. Otherwise, use Option 2 (degrees of freedom the smaller of $n_1 - 1$ and $n_2 - 1$). Many of these exercises ask you to think about issues of statistical practice as well as to carry out procedures.

19.30 Active versus passive learning. A study of computer-assisted learning examined the learning of “Blissymbols” by children. Blissymbols are pictographs (think of Egyptian hieroglyphs) that are sometimes used to help learning-impaired children communicate. The researcher designed two computer lessons that taught the same content using the same examples. One lesson required the children to interact with the material, while in the other the children controlled only the pace of the lesson. Call these two styles “Active” and “Passive.” Children were assigned at random to Active and Passive groups. After the lesson, the computer presented a quiz that asked the children to identify 56 Blissymbols. Here are the numbers of correct identifications by the 24 children in the Active group:

| 29 | 28 | 24 | 31 | 15 | 24 | 27 | 23 | 20 | 22 | 23 | 21 |
| 24 | 35 | 21 | 24 | 44 | 28 | 17 | 21 | 21 | 20 | 28 | 16 |

The 24 children in the Passive group had these counts of correct identifications:

| 16 | 14 | 17 | 15 | 26 | 17 | 12 | 25 | 21 | 20 | 18 | 21 |
| 20 | 16 | 18 | 15 | 26 | 15 | 13 | 17 | 21 | 19 | 15 | 12 |

Is there good evidence that active learning is superior to passive learning? Follow the four-step process as illustrated in Examples 19.2 and 19.3. That is, state hypotheses, make graphs to examine the data, discuss the conditions for inference, carry out a test, and state your conclusion.

19.31 IQ scores for boys and girls. Here are the IQ test scores of 31 seventh-grade girls in a Midwest school district:

| 114 | 100 | 104 | 89 | 102 | 91 | 114 | 114 | 103 | 105 |
| 108 | 130 | 120 | 132 | 111 | 128 | 118 | 119 | 86 | 72 |
| 111 | 103 | 74 | 112 | 107 | 103 | 98 | 96 | 112 | 112 | 93 |

\[ \bar{x}_a = 114.167 \quad s_a = 6.31022 \]

\[ \bar{x}_b = 107.750 \quad s_b = 4.02506 \]
The IQ test scores of 47 seventh-grade boys in the same district are

<table>
<thead>
<tr>
<th>111</th>
<th>107</th>
<th>100</th>
<th>107</th>
<th>115</th>
<th>111</th>
<th>97</th>
<th>112</th>
<th>104</th>
<th>106</th>
<th>113</th>
</tr>
</thead>
<tbody>
<tr>
<td>109</td>
<td>113</td>
<td>128</td>
<td>128</td>
<td>118</td>
<td>113</td>
<td>124</td>
<td>127</td>
<td>136</td>
<td>106</td>
<td>123</td>
</tr>
<tr>
<td>124</td>
<td>126</td>
<td>116</td>
<td>127</td>
<td>119</td>
<td>97</td>
<td>102</td>
<td>110</td>
<td>120</td>
<td>103</td>
<td>115</td>
</tr>
<tr>
<td>93</td>
<td>123</td>
<td>79</td>
<td>119</td>
<td>110</td>
<td>110</td>
<td>107</td>
<td>105</td>
<td>105</td>
<td>110</td>
<td>77</td>
</tr>
<tr>
<td>90</td>
<td>114</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Make stemplots or histograms of both sets of data. Because the distributions are reasonably symmetric with no extreme outliers, the t procedures will work well.

(b) Treat these data as SRSs from all seventh-grade students in the district. Is there good evidence that girls and boys differ in their mean IQ scores?

19.32 Active versus passive learning, continued.

(a) Use the data in Exercise 19.30 to give a 90% confidence interval for the difference in mean number of Blisymbols identified correctly by children after active and passive lessons.

(b) Give a 90% confidence interval for the mean number of Blisymbols identified correctly by children after the active lesson.

19.33 IQ scores for boys and girls, continued. Use the data in Exercise 19.31 to give a 95% confidence interval for the difference between the mean IQ scores of all boys and girls in the district.

19.34 Students’ attitudes. 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. A selective private college gives the SSHA to an SRS of both male and female first-year students. The data for the women are as follows:

<table>
<thead>
<tr>
<th>154</th>
<th>109</th>
<th>137</th>
<th>115</th>
<th>152</th>
<th>140</th>
<th>154</th>
<th>178</th>
<th>101</th>
</tr>
</thead>
<tbody>
<tr>
<td>103</td>
<td>126</td>
<td>126</td>
<td>137</td>
<td>165</td>
<td>165</td>
<td>129</td>
<td>200</td>
<td>148</td>
</tr>
</tbody>
</table>

Here are the scores of the men:

<table>
<thead>
<tr>
<th>108</th>
<th>140</th>
<th>114</th>
<th>91</th>
<th>180</th>
<th>115</th>
<th>126</th>
<th>92</th>
<th>169</th>
<th>146</th>
</tr>
</thead>
<tbody>
<tr>
<td>109</td>
<td>132</td>
<td>75</td>
<td>88</td>
<td>113</td>
<td>151</td>
<td>70</td>
<td>115</td>
<td>187</td>
<td>104</td>
</tr>
</tbody>
</table>

Most studies have found that the mean SSHA score for men is lower than the mean score in a comparable group of women. Is this true for first-year students at this college? Follow the four-step process as illustrated in Examples 19.2 and 19.3. That is, state hypotheses, make graphs to examine the data, discuss the conditions for inference, carry out a test, and state your conclusion.

19.35 Fungus in the air. The air in poultry-processing plants often contains fungus spores. Inadequate ventilation can affect the health of the workers. The problem is most serious during the summer and least serious during the winter. To measure the presence of spores, air samples are pumped to an agar plate and “colony forming units (CFUs)” are counted after an incubation period. Here are data from the “kill room” of a plant that processes 37,000 turkeys per day, taken on four separate days in the summer and in the winter. The units are CFUs per cubic meter of air.14
<table>
<thead>
<tr>
<th></th>
<th>Summer</th>
<th>Winter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3175</td>
<td>2526</td>
</tr>
<tr>
<td></td>
<td>1763</td>
<td>1090</td>
</tr>
<tr>
<td></td>
<td>384</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>251</td>
<td>97</td>
</tr>
</tbody>
</table>

The counts are clearly much higher in the summer. Give a 90% confidence interval to estimate how much higher the mean count is during the summer. Follow the four-step process as illustrated in Example 19.4.

**Do birds learn to time their breeding?** Blue tits like lots of caterpillars around when they have young to feed, but they breed earlier than peak caterpillar season. Do the birds learn from one year’s experience when they time breeding the next year? Researchers randomly assigned 7 pairs of birds to have the natural caterpillar supply supplemented while feeding their young and another 6 pairs to serve as a control group relying on natural food supply. The next year, they measured how many days after the caterpillar peak the birds produced their nestlings. Exercises 19.36 to 19.38 are based on this experiment.

19.36 **Did the randomization produce similar groups?** First, compare the two groups in the first year. The only difference should be the chance effect of the random assignment. The study report says: "In the experimental year, the degree of synchronization did not differ between food-supplemented and control females." For this comparison, the report gives $t = -1.05$. What type of $t$ statistic (paired or two-sample) is this? Show that this $t$ leads to the quoted conclusion.

19.37 **Did the treatment have an effect?** The investigators expected the control group to adjust their breeding date the next year, whereas the well-fed supplemented group had no reason to change. The report continues: "but in the following year food-supplemented females were more out of synchrony with the caterpillar peak than the controls." Here are the data (days behind the caterpillar peak):

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Supplemented</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.6</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>2.3</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>7.7</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>6.0</td>
<td>16.5</td>
</tr>
<tr>
<td></td>
<td>4.6</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>-1.2</td>
<td>11.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.7</td>
</tr>
</tbody>
</table>

Carry out a $t$ test and show that it leads to the quoted conclusion.

19.38 **Year-to-year comparison.** Rather than comparing the two groups in each year, we could compare the behavior of each group in the first and second years. The study report says: "Our main prediction was that females receiving additional food in the nestling period should not change laying date the next year, whereas controls, which (in our area) breed too late in their first year, were expected to advance their laying date in the second year."

Comparing days behind the caterpillar peak in Years 1 and 2 gave $t = 0.63$ for the control group and $t = -2.63$ for the supplemented group. Are these paired or two-sample $t$ statistics? What are the degrees of freedom for each $t$? Show that these $t$-values do not agree with the prediction.

Exercises 19.39 to 19.44 are based on summary statistics rather than raw data. This information is typically all that is presented in published reports. Inference procedures can be calculated by hand from the summaries. You must trust that the authors understood the conditions for inference and verified that they apply. This isn’t always true.
19.39 *Eating potato chips.* Healthy women aged 18 to 40 participated in a study of eating habits. Subjects were given bags of potato chips and bottled water and invited to snack freely. Interviews showed that some women were trying to restrain their diet out of concern about their weight. How much effect did these good intentions have on their eating habits? Here are the data on grams of potato chips consumed (note that the study report gave the standard error of the mean rather than the standard deviation):\(^{16}\)

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>(\bar{x})</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrestrained</td>
<td>9</td>
<td>59</td>
<td>7</td>
</tr>
<tr>
<td>Restrained</td>
<td>11</td>
<td>32</td>
<td>10</td>
</tr>
</tbody>
</table>

Give a 90% confidence interval that describes the effect of restraint. Based on this interval, is there a significant difference between the two groups? At what significance level does the interval allow this conclusion?

19.40 **Hispanic customers and Anglo customers.** As the presence of Hispanics in the United States grows, businesses are trying to understand what Hispanics like. One study sampled customers leaving a bank. Customers were classified as Hispanic if they preferred to be interviewed in Spanish and as Anglo if they preferred English. Each customer rated the importance of several aspects of bank service on a 10-point scale.\(^{17}\) Here are summary results for the importance of “reliability” (the accuracy of account records and so on):

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>(\bar{x})</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anglo</td>
<td>92</td>
<td>6.37</td>
<td>0.60</td>
</tr>
<tr>
<td>Hispanic</td>
<td>86</td>
<td>5.91</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Another aspect of service quality is “empathy,” the relationship that bank employees have with customers. The summary data are

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>(\bar{x})</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anglo</td>
<td>92</td>
<td>6.00</td>
<td>0.89</td>
</tr>
<tr>
<td>Hispanic</td>
<td>86</td>
<td>6.43</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Do Hispanic and Anglo bank customers differ in the importance they place on either or both of these qualities? Write a one-sentence description of the differences in what the two groups find important.

19.41 **Extraterrestrial handedness?** Molecules often have “left-handed” and “right-handed” versions. Some classes of molecules found in life on earth are almost entirely left-handed. Did this left-handedness precede the origin of life? To find out, scientists analyzed meteorites from space. To correct for bias in the delicate analysis, they also analyzed standard compounds known to have equal proportions of left-handed and right-handed forms. Here are the results for the percents of left-handed forms of one molecule in two analyses:\(^{18}\)
The researchers used the $t$ test to see if the meteorite had a significantly higher percent than the standard. Carry out the tests for both analyses and report the results. The researchers concluded: "The observations suggest that organic matter of extraterrestrial origin could have played an essential role in the origin of terrestrial life."

### Coaching and SAT scores

Coaching companies claim that their courses can raise the SAT scores of high school students. Of course, students who retake the SAT without paying for coaching generally raise their scores. A random sample of students who took the SAT twice found 427 who were coached and 2733 who were uncoached. Starting with their verbal scores on the first and second tries, we have these summary statistics:

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Meteorite</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>$\bar{x}$</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>52.6</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>51.7</td>
</tr>
</tbody>
</table>

Let's first ask if students who are coached significantly increased their scores.

(a) You could use the information given to carry out either a two-sample $t$ test comparing Try 1 with Try 2 for coached students or a matched pairs $t$ test using Gain. Which is the correct test? Why?

(b) Carry out the proper test. What do you conclude?

(c) Give a 99% confidence interval for the mean gain of all students who are coached.

### Coaching and SAT scores, continued

What we really want to know is whether coached students improve more than uncoached students, and whether any advantage is large enough to be worth paying for. Use the information in the previous exercise to answer these questions:

(a) Is there good evidence that coached students gained more on the average than uncoached students?

(b) How much more do coached students gain on the average? Give a 99% confidence interval.

(c) Based on your work, what is your opinion: do you think coaching courses are worth paying for?

### Coaching and SAT scores: critique

The data you used in the previous two problems came from a random sample of students who took the SAT twice. The response rate was 63%, which is pretty good for non-government surveys, so let's accept that the respondents do represent all students who took the exam twice.

19.42 (a) Matched pairs test—a student's score on Try 1 is correlated with his/her score on Try 2. (b) $H_0: \mu = 0; H_1: \mu > 0$; $t = 10.16$ (df = 426); $P < 0.0005$. (c) 21.50 to 36.50 points (software: 21.61 to 36.39).

19.44 This was an observational study, not an experiment. Students who choose coaching might have other motivating factors that help them do better the second time.
probably shorten their lives. Harris announces a margin of error of ±3 percentage points for all samples of about this size. Opinion polls announce the margin of error for 95% confidence.

(a) What is the actual margin of error (in percent) for the large-sample confidence interval from this sample?

(b) The margin of error is largest when \( \hat{p} = 0.5 \). What would the margin of error (in percent) be if the sample had resulted in \( \hat{p} = 0.5 \)?

(c) Why do you think that Harris announces a ±3% margin of error for all samples of about this size?

19.28 Do college students pray? Social scientists asked 127 undergraduate students “from courses in psychology and communications” about prayer and found that 107 prayed at least a few times a year.\(^{17}\)

(a) To use any inference procedure, we must be willing to regard these 127 students, as far as their religious behavior goes, as an SRS from the population of all undergraduate students. Do you think it is reasonable to do this? Why or why not?

(b) If we act as if the sample is an SRS, what is the large-sample 99% confidence interval for the proportion \( p \) of all students who pray?

(c) Give the plus four 99% confidence interval for \( p \). If you express the two intervals in percents and round to the nearest tenth of a percent, how do they differ? (As always, the plus four method pulls results away from 0% or 100%, whichever is closer. Although the condition for the large-sample interval is met, the plus four interval is more trustworthy.)

19.29 I don’t like my life. The Pew Research Center asked a random sample of 1128 adult women, “How satisfied are you with your life overall?” Of these women, 56 said either “Mostly dissatisfied” or “Very dissatisfied.”\(^{18}\)

(a) Pew dialed residential telephone numbers at random in the continental United States in an attempt to contact a random sample of adults. Based on what you know about national sample surveys, what is likely to be the biggest weakness in the survey?

(b) Act as if the sample is an SRS. Give a large-sample 90% confidence interval for the proportion \( p \) of all adult women who are mostly or very dissatisfied with their lives.

(c) Give the plus four confidence interval for \( p \). If you express the two confidence intervals in percents and round to the nearest tenth of a percent, how do they differ? (As always, the plus four method pulls results away from 0% or 100%, whichever is closer. Although the condition for the large-sample interval is met, we can place more trust in the plus four interval.)

19.30 Which font? Plain type fonts such as Times New Roman are easier to read than fancy fonts such as Gigi. A group of 25 volunteer subjects read the same text in both fonts. (This is a matched pairs design. One-sample procedures for proportions, like those for means, are used to analyze data from matched pairs designs.) Of the 25 subjects, 17 said that they preferred Times New Roman for Web use. But 20 said that Gigi was more attractive.\(^{19}\)
(a) Because the subjects were volunteers, conclusions from this sample can be challenged. Show that the sample size condition for the large-sample confidence interval is not met, but that the condition for the plus four interval is met.

(b) Give a 95% confidence interval for the proportion of all adults who prefer Times New Roman for Web use. Give a 90% confidence interval for the proportion of all adults who think Gigi is more attractive.

19.31 Detecting genetically modified soybeans. Most soybeans grown in the United States are genetically modified to, for example, resist pests and so reduce use of pesticides. Because some nations do not accept genetically modified (GM) foods, grain-handling facilities routinely test soybean shipments for the presence of GM beans. In a study of the accuracy of these tests, researchers submitted shipments of soybeans containing 1% of GM beans to 23 randomly selected facilities. Eighteen detected the GM beans.20

(a) Show that the conditions for the large-sample confidence interval are not met. Show that the conditions for the plus four interval are met.

(b) Use the plus four method to give a 90% confidence interval for the percent of all grain-handling facilities that will correctly detect 1% of GM beans in a shipment.

19.32 Running red lights. A random digit dialing telephone survey of 880 drivers asked, “Recalling the last ten traffic lights you drove through, how many of them were red when you entered the intersections?” Of the 880 respondents, 171 admitted that at least one light had been red.21

(a) Give a 95% confidence interval for the proportion of all drivers who ran one or more of the last ten red lights they met.

(b) Nonresponse is a practical problem for this survey—only 21.6% of calls that reached a live person were completed. Another practical problem is that people may not give truthful answers. What is the likely direction of the bias do you think more or fewer than 171 of the 880 respondents really ran a red light? Why?

19.33 The IRS plans an SRS. The Internal Revenue Service plans to examine an SRS of individual federal income tax returns from each state. One variable of interest is the proportion of returns claiming itemized deductions. The total number of tax returns in a state varies from more than 15 million in California to fewer than 250,000 in Wyoming.

(a) Will the margin of error for estimating the population proportion change from state to state if an SRS of 2000 tax returns is selected in each state? Explain your answer.

(b) Will the margin of error change from state to state if an SRS of 1% of all tax returns is selected in each state? Explain your answer.

19.34 Customer satisfaction. An automobile manufacturer would like to know what proportion of its customers are not satisfied with the service provided by the local dealer. The customer relations department will survey a random sample of customers and compute a 99% confidence interval for the proportion who are not satisfied.
where the standard error of $\hat{p}_1 - \hat{p}_2$ is

$$SE = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

and $z^*$ is a standard Normal critical value.

- The true confidence level of the large-sample interval can be substantially less than the planned level $C$. Use this interval only if the counts of successes and failures in both samples are 10 or greater.
- To get a more accurate confidence interval, add four imaginary observations, one success and one failure in each sample. Then use the same formula for the confidence interval. This is the plus four confidence interval. You can use it whenever both samples have 5 or more observations.
- Significance tests for $H_0: p_1 = p_2$ use the pooled sample proportion

$$\hat{p} = \frac{\text{number of successes in both samples combined}}{\text{number of individuals in both samples combined}}$$

and the $z$ statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$P$-values come from the standard Normal distribution. Use this test when there are 5 or more successes and 5 or more failures in both samples.

**Check Your Skills**

A sample survey interviews SRSs of 500 female college students and 550 male college students. Each student is asked if he or she worked for pay last summer. In all, 410 of the women and 484 of the men say “Yes.” Exercises 20.9 to 20.13 are based on this survey.

**20.9** Take $p_M$ and $p_F$ to be the proportions of all college males and females who worked last summer. We conjectured before seeing the data that men are more likely to work. The hypotheses to be tested are

(a) $H_0: p_M = p_F$ versus $H_a: p_M \neq p_F$.
(b) $H_0: p_M = p_F$ versus $H_a: p_M > p_F$.
(c) $H_0: p_M = p_F$ versus $H_a: p_M < p_F$.

**20.10** The sample proportions of college males and females who worked last summer are about

(a) $\hat{p}_M = 0.88$ and $\hat{p}_F = 0.82$.
(b) $\hat{p}_M = 0.82$ and $\hat{p}_F = 0.88$.
(c) $\hat{p}_M = 0.75$ and $\hat{p}_F = 0.97$. 
20.11 The pooled sample proportion who worked last summer is about
(a) $\hat{p} = 1.70$. (b) $\hat{p} = 0.89$. (c) $\hat{p} = 0.85$.

20.12 The $z$ statistic for a test comparing the proportions of college men and women who worked last summer is about
(a) $z = 2.66$. (b) $z = 2.72$. (c) $z = 3.10$.

20.13 The 95% large-sample confidence interval for the difference $p_M - p_F$ in the proportions of college men and women who worked last summer is about
(a) $0.06 \pm 0.00095$. (b) $0.06 \pm 0.043$. (c) $0.06 \pm 0.036$.

20.14 In an experiment to learn if substance M can help restore memory, the brains of 20 rats were treated to damage their memories. The rats were trained to run a maze. After a day, 10 rats were given M and 7 of them succeeded in the maze; only 2 of the 10 control rats were successful. The $z$ test for "no difference" against "a higher proportion of the M group succeeds" has
(a) $z = 2.25$, $P < 0.02$.
(b) $z = 2.60$, $P < 0.005$.
(c) $z = 2.25$, $P < 0.04$ but not $< 0.02$.

20.15 The $z$ test in the previous exercise
(a) may be inaccurate because the populations are too small.
(b) may be inaccurate because some counts of successes and failures are too small.
(c) is reasonably accurate because the conditions for inference are met.

20.16 The plus four 90% confidence interval for the difference between the proportion of rats that succeed when given M and the proportion that succeed without it is
(a) $0.455 \pm 0.312$. (b) $0.417 \pm 0.304$. (c) $0.417 \pm 0.185$.

CHAPTER 20 EXERCISES

We recommend using the plus four method for all confidence intervals for proportions. However, the large-sample method is acceptable when the guidelines for its use are met.

20.17 Truthfulness in online profiles. Many teens have posted profiles on sites such as MySpace. A sample survey asked random samples of teens with online profiles if they included false information in their profiles. Of 170 younger teens (ages 12 to 14), 117 said "Yes." Of 317 older teens (ages 15 to 17), 152 said "Yes."14
(a) Do these samples satisfy the guidelines for the large-sample confidence interval?
(b) Give a 95% confidence interval for the difference between the proportions of younger and older teens who include false information in their online profiles.

20.18 Drug testing in schools. In 2002 the Supreme Court ruled that schools could require random drug tests of students participating in competitive after-school activities such as athletics. Does drug testing reduce use of illegal drugs? A study compared two similar high schools in Oregon. Wahtonka High School tested athletes