In this chapter we cover...

Comparing several means
The analysis of variance
\( F \) test
Using technology
The idea of analysis of variance
Conditions for ANOVA
\( F \) distributions and degrees of freedom
Some details of ANOVA: the two-sample case
Some details of ANOVA

CHAPTER

25

One-Way Analysis of Variance: Comparing Several Means

The two-sample t procedures of Chapter 19 compare the means of two populations or the mean responses to two treatments in an experiment. Of course, studies don’t always compare just two groups. We need a method for comparing any number of means.

EXAMPLE 25.1 Comparing tropical flowers

STATE: Ethan Temeles of Amherst College, with his colleague W. John Kress, studied the relationship between varieties of the tropical flower Heliconia on the island of Dominica and the different species of hummingbirds that fertilize the flowers.\(^1\) Over time, the researchers believe, the lengths of the flowers and the form of the hummingbirds’ beaks have evolved to match each other. If that is true, flower varieties fertilized by different hummingbird species should have distinct distributions of length.

Table 25.1 gives length measurements (in millimeters) for samples of three varieties of Heliconia, each fertilized by a different species of hummingbird. Do the three varieties display distinct distributions of length? In particular, are the average lengths of their flowers different?

FORMULATE: Use graphs and numerical descriptions to describe and compare the three distributions of flower length. Finally, ask whether the differences among the mean lengths of the three varieties are statistically significant.
TABLE 25.1  Flower lengths (millimeters) for three *Heliconia* varieties

<table>
<thead>
<tr>
<th>Variety</th>
<th>Sample size</th>
<th>Mean length</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. bihai</em></td>
<td>16</td>
<td>47.60</td>
<td>1.213</td>
</tr>
<tr>
<td><em>H. caribaea red</em></td>
<td>23</td>
<td>39.71</td>
<td>1.799</td>
</tr>
<tr>
<td><em>H. caribaea yellow</em></td>
<td>15</td>
<td>36.18</td>
<td>0.975</td>
</tr>
</tbody>
</table>

SOLVE (first steps): Perhaps these data seem familiar. We first met them in Chapter 2 (page 54), where we compared the distributions. Figure 25.1 repeats a stemplot display from Chapter 2. The lengths have been rounded to the nearest tenth of a millimeter. Here are the summary measures we will use in further analysis:

<table>
<thead>
<tr>
<th>Sample</th>
<th>Variety</th>
<th>Sample size</th>
<th>Mean length</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>bhai</em></td>
<td>16</td>
<td>47.60</td>
<td>1.213</td>
</tr>
<tr>
<td>2</td>
<td>red</td>
<td>23</td>
<td>39.71</td>
<td>1.799</td>
</tr>
<tr>
<td>3</td>
<td>yellow</td>
<td>15</td>
<td>36.18</td>
<td>0.975</td>
</tr>
</tbody>
</table>

\[ \text{FIGURE 25.1} \] Side-by-side stemplots comparing the lengths in millimeters of samples of flowers from three varieties of *Heliconia*, from Table 25.1.
CONCLUDE (first steps): The three varieties differ so much in flower length that there is little overlap among them. In particular, the flowers of *bihai* are longer than either red or yellow. The mean lengths are 47.6 mm for *H. bihai*, 39.7 mm for *H. caribaea* red, and 36.2 mm for *H. caribaea* yellow. Are these observed differences in sample means statistically significant? We must develop a test for comparing more than two population means.

Comparing several means

Call the mean lengths for the three populations of flowers $\mu_1$ for *bihai*, $\mu_2$ for red, and $\mu_3$ for yellow. The subscript reminds us which group a parameter or statistic describes. To compare these three population means, we might use the two-sample $t$ test several times:

- Test $H_0: \mu_1 = \mu_2$ to see if the mean length for *bihai* differs from the mean for red.
- Test $H_0: \mu_1 = \mu_3$ to see if *bihai* differs from yellow.
- Test $H_0: \mu_2 = \mu_3$ to see if red differs from yellow.

The weakness of doing three tests is that we get three $P$-values, one for each test alone. That doesn’t tell us how likely it is that three sample means are spread apart as far as these are. It may be that $\bar{x}_1 = 47.60$ and $\bar{x}_3 = 36.18$ are significantly different if we look at just two groups but not significantly different if we know that they are the largest and the smallest means in three groups. As we look at more groups, we expect the gap between the largest and smallest sample mean to get larger. (Think of comparing the tallest and shortest person in larger and larger groups of people.) We can’t safely compare many parameters by doing tests or confidence intervals for two parameters at a time.

The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions is common in statistics. This is the problem of **multiple comparisons**. Statistical methods for dealing with multiple comparisons usually have two steps:

1. An overall test to see if there is good evidence of any differences among the parameters that we want to compare.
2. A detailed follow-up analysis to decide which of the parameters differ and to estimate how large the differences are.

The overall test, though more complex than the tests we met earlier, is often reasonably straightforward. The follow-up analysis can be quite elaborate. In our basic introduction to statistical practice, we will concentrate on the overall test, along with data analysis that points to the nature of the differences.
The analysis of variance \( F \) test

We want to test the null hypothesis that there are no differences among the mean lengths for the three populations of flowers:

\[ \text{H}_0: \mu_1 = \mu_2 = \mu_3 \]

The alternative hypothesis is that there is some difference. That is, not all three population means are equal:

\[ \text{H}_a: \text{not all of } \mu_1, \mu_2, \text{ and } \mu_3 \text{ are equal} \]

The alternative hypothesis is no longer one-sided or two-sided. It is "many-sided," because it allows any relationship other than "all three equal." For example, \( \text{H}_a \) includes the case in which \( \mu_2 = \mu_3 \) but \( \mu_1 \) has a different value. The test of \( \text{H}_0 \) against \( \text{H}_a \) is called the analysis of variance \( F \) test. Analysis of variance is usually abbreviated as ANOVA. The ANOVA \( F \) test is almost always carried out by software that reports the test statistic and its \( P \)-value.

**EXAMPLE 25.2** Comparing tropical flowers: ANOVA

**SOLVE (inference):** Software tells us that for the flower length data in Table 25.1, the test statistic is \( F = 259.12 \) with \( P \)-value \( P < 0.0001 \). There is very strong evidence that the three varieties of flowers do not all have the same mean length.

The \( F \) test does not say which of the three means are significantly different. It appears from our preliminary data analysis that bihai flowers are distinctly longer than either red or yellow. Red and yellow are closer together, but the red flowers tend to be longer.

**CONCLUDE:** There is strong evidence (\( P < 0.0001 \)) that the population means are not all equal. The most important difference among the means is that the bihai variety has longer flowers than the red and yellow varieties.

Example 25.2 illustrates our approach to comparing means. The ANOVA \( F \) test (done by software) assesses the evidence for some difference among the population means. In most cases, we expect the \( F \) test to be significant. We would not undertake a study if we did not expect to find some effect. The formal test is nonetheless important to guard against being misled by chance variation. We will not do the formal follow-up analysis that is often the most useful part of an ANOVA study. Follow-up analysis would allow us to say which means differ and by how much, with (say) 95% confidence that all our conclusions are correct. We rely instead on examination of the data to show what differences are present and whether they are large enough to be interesting.

**APPLY YOUR KNOWLEDGE**

25.1 **Do fruit flies sleep?** Mammals and birds sleep. Insects such as fruit flies rest, but is this rest sleep? Biologists now think that insects do sleep. One experiment gave caffeine to fruit flies to see if it affected their rest. We know that caffeine reduces sleep in mammals, so if it reduces rest in fruit flies that’s another hint that the rest
is really sleep. The paper reporting the study contains a graph similar to Figure 25.2 and states, “Flies given caffeine obtained less rest during the dark period in a dose-dependent fashion (n = 36 per group, P < 0.0001).”

(a) The explanatory variable is amount of caffeine, in milligrams per milliliter of blood. The response variable is minutes of rest (measured by an infrared motion sensor) during a 12-hour dark period. Outline the design of this experiment.

(b) The P-value in the report comes from the ANOVA F test. What means does this test compare? State in words the null and alternative hypotheses for the test in this setting. What do the graph and the statistical test together lead you to conclude?

25.2 Road rage. “The phenomenon of road rage has been frequently discussed but infrequently examined.” So begins a report based on interviews with 1382 randomly selected drivers. The respondents’ answers to interview questions produced scores on an “angry/threatening driving scale” with values between 0 and 19. What driver characteristics go with road rage? There were no significant differences among races or levels of education. What about the effect of the driver’s age? Here are the mean responses for three age groups:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 30 yr</td>
<td>2.22</td>
</tr>
<tr>
<td>30-55 yr</td>
<td>1.33</td>
</tr>
<tr>
<td>&gt; 55 yr</td>
<td>0.66</td>
</tr>
</tbody>
</table>
The report says that $F = 34.96$, with $P < 0.01$.

(a) What are the null and alternative hypotheses for the ANOVA $F$ test? Be sure to explain what means the test compares.

(b) Based on the sample means and the $F$ test, what do you conclude?

**Using technology**

Any technology used for statistics should perform analysis of variance. Figure 25.3 displays ANOVA output for the data of Table 25.1 from a graphing calculator, two statistical programs, and a spreadsheet program.

**TI-83**

![TI-83 ANOVA output](image)

**CrunchIt!**

![CrunchIt! ANOVA output](image)

**Figure 25.3** ANOVA for the flower length data: output from a graphing calculator, two statistical programs, and a spreadsheet program (continued).
One-Way Analysis of Variance: Comparing Several Means

Minitab

![Minitab Output](image)

Excel

![Excel Output](image)

FIGURE 25.3 (continued)

The three software outputs give the sizes of the three samples and their means. These agree with those in Example 25.1. Minitab also gives the standard deviations. You should be able to recover the standard deviations from either the variances (Excel) or the standard errors of the means (CrunchIt!). The most important part of all four outputs reports the $F$ test statistic, $F = 259.12$, and its $P$-value. CrunchIt! and Minitab sensibly report the $P$-value as 0 to three decimal places, which is all we need to know in practice. Excel and the TI-83 compute a
more specific value. The E – 27 in these displays means move the decimal point 27 digits to the left. There is very strong evidence that the three varieties of flowers do not all have the same mean length.

All four outputs also report degrees of freedom (df), sums of squares (SS), and mean squares (MS). We don’t need this information now.

Minitab also gives confidence intervals for all three means that help us see which means differ and by how much. None of the intervals overlap, and bihai is much above the other two. These are 95% confidence intervals for each mean separately. We are not 95% confident that all three intervals cover the three means. This is another example of the peril of multiple comparisons.

APPLY YOUR KNOWLEDGE

25.3 Logging in the rain forest. How does logging in a tropical rain forest affect the forest several years later? Researchers compared forest plots in Borneo that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). Although the study was not an experiment, the authors explain why we can consider the plots to be randomly selected. The data appear in Table 25.2. The variable Trees is the count of trees in a plot; Species is the count of tree species in a plot. The variable Richness is the number of species divided by the number of individual trees, Species/Trees.

(a) Make side-by-side stemplots of Trees for the three groups. Use stems 0, 1, 2, and 3 and split the stems (see page 21). What effects of logging are visible?

(b) Figure 25.4 shows Excel ANOVA output for Trees. What do the group means show about the effects of logging?

(c) What are the values of the ANOVA F statistic and its P-value? What hypotheses does F test? What conclusions about the effects of logging on number of trees do the data lead to?

![Figure 25.4](image-url) Excel output for analysis of variance on the number of trees in forest plots, for Exercise 25.3.
### Table 25.2: Data from a study of logging in Borneo

<table>
<thead>
<tr>
<th>Observation</th>
<th>Group</th>
<th>Trees</th>
<th>Species</th>
<th>Richness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>27</td>
<td>22</td>
<td>0.81481</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>22</td>
<td>18</td>
<td>0.81818</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>29</td>
<td>22</td>
<td>0.75862</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>21</td>
<td>20</td>
<td>0.95238</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>19</td>
<td>15</td>
<td>0.78947</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>33</td>
<td>21</td>
<td>0.86336</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>16</td>
<td>13</td>
<td>0.81250</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>20</td>
<td>13</td>
<td>0.65000</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>24</td>
<td>19</td>
<td>0.79167</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>27</td>
<td>13</td>
<td>0.48148</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>28</td>
<td>19</td>
<td>0.67857</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>19</td>
<td>15</td>
<td>0.78947</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>12</td>
<td>11</td>
<td>0.91667</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>12</td>
<td>11</td>
<td>0.91667</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>15</td>
<td>14</td>
<td>0.93333</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>9</td>
<td>7</td>
<td>0.77778</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>20</td>
<td>18</td>
<td>0.90000</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>18</td>
<td>15</td>
<td>0.83333</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>17</td>
<td>15</td>
<td>0.88235</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>14</td>
<td>12</td>
<td>0.85714</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>14</td>
<td>13</td>
<td>0.92857</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1.00000</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>17</td>
<td>15</td>
<td>0.88235</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>19</td>
<td>8</td>
<td>0.42105</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>18</td>
<td>17</td>
<td>0.94444</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1.00000</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>22</td>
<td>18</td>
<td>0.81818</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>15</td>
<td>14</td>
<td>0.93333</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>18</td>
<td>18</td>
<td>1.00000</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>19</td>
<td>15</td>
<td>0.78947</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>22</td>
<td>15</td>
<td>0.68182</td>
</tr>
<tr>
<td>32</td>
<td>3</td>
<td>12</td>
<td>10</td>
<td>0.83333</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>12</td>
<td>12</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

### 25.4 Dogs, friends, and stress

If you are a dog lover, perhaps having your dog along reduces the effect of stress. To examine the effect of pets in stressful situations, researchers recruited 45 women who said they were dog lovers. The EESEE story “Stress among Pets and Friends” describes the results. Fifteen of the subjects were randomly assigned to each of three groups to do a stressful task alone (the control group), with a good friend present, or with their dog present. The subject’s mean heart rate during the task is one measure of the effect of stress. Table 25.3 contains the data.
TABLE 25.3  Mean heart rates during stress with a pet (P), with a friend (F), and for the control group (C)

<table>
<thead>
<tr>
<th>Group</th>
<th>Rate</th>
<th>Group</th>
<th>Rate</th>
<th>Group</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>69.169</td>
<td>P</td>
<td>68.862</td>
<td>C</td>
<td>84.738</td>
</tr>
<tr>
<td>F</td>
<td>99.692</td>
<td>C</td>
<td>87.231</td>
<td>C</td>
<td>84.877</td>
</tr>
<tr>
<td>P</td>
<td>70.169</td>
<td>P</td>
<td>64.169</td>
<td>P</td>
<td>58.692</td>
</tr>
<tr>
<td>C</td>
<td>80.369</td>
<td>C</td>
<td>91.754</td>
<td>P</td>
<td>79.662</td>
</tr>
<tr>
<td>C</td>
<td>87.446</td>
<td>C</td>
<td>87.785</td>
<td>P</td>
<td>69.231</td>
</tr>
<tr>
<td>P</td>
<td>75.985</td>
<td>F</td>
<td>91.354</td>
<td>C</td>
<td>73.277</td>
</tr>
<tr>
<td>F</td>
<td>83.400</td>
<td>F</td>
<td>100.877</td>
<td>C</td>
<td>84.523</td>
</tr>
<tr>
<td>F</td>
<td>102.154</td>
<td>C</td>
<td>77.800</td>
<td>C</td>
<td>70.877</td>
</tr>
<tr>
<td>P</td>
<td>86.446</td>
<td>P</td>
<td>97.538</td>
<td>F</td>
<td>89.815</td>
</tr>
<tr>
<td>F</td>
<td>80.277</td>
<td>P</td>
<td>85.000</td>
<td>F</td>
<td>98.200</td>
</tr>
<tr>
<td>C</td>
<td>90.015</td>
<td>F</td>
<td>101.062</td>
<td>F</td>
<td>76.908</td>
</tr>
<tr>
<td>C</td>
<td>99.046</td>
<td>F</td>
<td>97.046</td>
<td>P</td>
<td>69.538</td>
</tr>
<tr>
<td>C</td>
<td>75.477</td>
<td>C</td>
<td>62.646</td>
<td>P</td>
<td>70.077</td>
</tr>
<tr>
<td>F</td>
<td>88.015</td>
<td>F</td>
<td>81.600</td>
<td>F</td>
<td>86.985</td>
</tr>
<tr>
<td>F</td>
<td>92.492</td>
<td>P</td>
<td>72.262</td>
<td>P</td>
<td>65.446</td>
</tr>
</tbody>
</table>

(a) Make stemplots of the heart rates for the three groups (round to the nearest whole number of beats). Do any of the groups show outliers or extreme skewness?

(b) Figure 25.5 gives the Minitab ANOVA output for these data. Do the mean heart rates for the groups appear to show that the presence of a pet or a friend reduces heart rate during a stressful task?

![Minitab output](image)

**FIGURE 25.5** Minitab output for the data in Table 25.3 on heart rates (beats per minute) during stress, for Exercise 25.4. The “Control” group worked alone, the “Friend” group had a friend present, and the “Pet” group had a pet dog present.
(c) What are the values of the ANOVA $F$ statistic and its $P$-value? What hypotheses does $F$ test? Briefly describe the conclusions you draw from these data. Did you find anything surprising?

**The idea of analysis of variance**

The details of ANOVA are a bit daunting (they appear in an optional section at the end of this chapter). The main idea of ANOVA is both more accessible and much more important. Here it is: when we ask if a set of sample means gives evidence for differences among the population means, what matters is not how far apart the sample means are but how far apart they are relative to the variability of individual observations.

Look at the two sets of boxplots in Figure 25.6. For simplicity, these distributions are all symmetric, so that the mean and median are the same. The centerline in each boxplot is therefore the sample mean. Both sets of boxplots compare three samples with the same three means. Could differences this large easily arise just due to chance, or are they statistically significant?

- The boxplots in Figure 25.6(a) have tall boxes, which show lots of variation among the individuals in each group. With this much variation among individuals, we would not be surprised if another set of samples gave quite different sample means. The observed differences among the sample means could easily happen just by chance.
- The boxplots in Figure 25.6(b) have the same centers as those in Figure 25.6(a), but the boxes are much shorter. That is, there is much less variation among the individuals in each group. It is unlikely that any sample from the first group would have a mean as small as the mean of the second group.

![Figure 25.6](image-url) **Figure 25.6** Boxplots for two sets of three samples each. The sample means are the same in (a) and (b). Analysis of variance will find a more significant difference among the means in (b) because there is less variation among the individuals within those samples.
Because means as far apart as those observed would rarely arise just by chance in repeated sampling, they are good evidence of real differences among the means of the three populations we are sampling from.

You can use the One-Way ANOVA applet to demonstrate the analysis of variance idea for yourself. The applet allows you to change both the group means and the spread within groups. You can watch the ANOVA F statistic and its P-value change as you work.

This comparison of the two parts of Figure 25.6 is too simple in one way. It ignores the effect of the sample sizes, an effect that boxplots do not show. Small differences among sample means can be significant if the samples are large. Large differences among sample means can fail to be significant if the samples are small. All we can be sure of is that for the same sample size, Figure 25.6(b) will give a much smaller P-value than Figure 25.6(a). Despite this qualification, the big idea remains: if sample means are far apart relative to the variation among individuals in the same groups, that's evidence that something other than chance is at work.

**THE ANALYSIS OF VARIANCE IDEA**

Analysis of variance compares the variation due to specific sources with the variation among individuals who should be similar. In particular, ANOVA tests whether several populations have the same mean by comparing how far apart the sample means are with how much variation there is within the samples.

It is one of the oddities of statistical language that methods for comparing means are named after the variance. The reason is that the test works by comparing two kinds of variation. Analysis of variance is a general method for studying sources of variation in responses. Comparing several means is the simplest form of ANOVA, called one-way ANOVA. One-way ANOVA is the only form of ANOVA that we will study.

**THE ANOVA F STATISTIC**

The analysis of variance F statistic for testing the equality of several means has this form:

\[ F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}} \]
If you want more detail, read the optional section at the end of this chapter. The $F$ statistic can take only values that are zero or positive. It is zero only when all the sample means are identical and gets larger as they move farther apart. Large values of $F$ are evidence against the null hypothesis $H_0$ that all population means are the same. Although the alternative hypothesis $H_a$ is many-sided, the ANOVA $F$ test is one-sided because any violation of $H_0$ tends to produce a large value of $F$.

**APPLY YOUR KNOWLEDGE**

25.5 ANOVA compares several means. The One-Way ANOVA applet displays the observations in three groups, with the group means highlighted by black dots. When you open or reset the applet, the scale at the bottom of the display shows that for these groups the ANOVA $F$ statistic is $F = 31.74$, with $P < 0.001$. (The $P$-value is marked by a red dot that moves along the scale.)

(a) The middle group has larger mean than the other two. Grab its mean point with the mouse. How small can you make $F$? What did you do to the mean to make $F$ small? Roughly how significant is your small $F$?

(b) Starting with the three means aligned from your configuration at the end of (a), drag any one of the group means either up or down. What happens to $F$? What happens to the $P$-value? Convince yourself that the same thing happens if you move any one of the means, or if you move one slightly and then another slightly in the opposite direction.

25.6 ANOVA uses within-group variation. Reset the One-Way ANOVA applet to its original state. As in Figure 25.6(b), the differences among the three means are highly significant (large $F$, small $P$-value) because the observations in each group cluster tightly about the group mean.

(a) Use the mouse to slide the Pooled Standard Error at the top of the display to the right. You see that the group means do not change, but the spread of the observations in each group increases. What happens to $F$ and $P$ as the spread among the observations in each group increases? What are the values of $F$ and $P$ when the slider is all the way to the right? This is similar to Figure 25.6(a): variation within groups hides the differences among the group means.

(b) Leave the Pooled Standard Error slider at the extreme right of its scale, so that spread within groups stays fixed. Use the mouse to move the group means apart. What happens to $F$ and $P$ as you do this?

**Conditions for ANOVA**

Like all inference procedures, ANOVA is valid only in some circumstances. Here are the conditions under which we can use ANOVA to compare population means.
CONDITIONS FOR APPLYING ANOVA

- We have $I$ independent SRSs, one from each of $I$ populations.
- The $i$th population has a Normal distribution with unknown mean $\mu_i$. The means may be different in the different populations. The ANOVA $F$ statistic tests the null hypothesis that all of the populations have the same mean:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

$$H_a: \text{not all of the } \mu_i \text{ are equal}$$

- All of the populations have the same standard deviation $\sigma$, whose value is unknown.

There are $I + 1$ population parameters that we must estimate from the data: the $I$ population means and the standard deviation $\sigma$.

The first two requirements are familiar from our study of the two-sample $t$ procedures for comparing two means. As usual, the design of the data production is the most important condition for inference. Biased sampling or confounding can make any inference meaningless. If we do not actually draw separate SRSs from each population or carry out a randomized comparative experiment, it may be unclear to what population the conclusions of inference apply. ANOVA, like other inference procedures, is often used when random samples are not available. You must judge each use on its merits, a judgment that usually requires some knowledge of the subject of the study in addition to some knowledge of statistics.

Because no real population has exactly a Normal distribution, the usefulness of inference procedures that assume Normality depends on how sensitive they are to departures from Normality. Fortunately, procedures for comparing means are not very sensitive to lack of Normality. The ANOVA $F$ test, like the $t$ procedures, is robust. What matters is Normality of the sample means, so ANOVA becomes safer as the sample sizes get larger, because of the central limit theorem effect. Remember to check for outliers that change the value of sample means and for extreme skewness. When there are no outliers and the distributions are roughly symmetric, you can safely use ANOVA for sample sizes as small as 4 or 5. (Don’t confuse the ANOVA $F$, which compares several means, with the $F$ statistic discussed in Chapter 19, which compares two standard deviations and is not robust against non-Normality.)

The third condition is annoying: ANOVA assumes that the variability of observations, measured by the standard deviation, is the same in all populations. You may recall from Chapter 19 (page 476) that there is a special version of the two-sample $t$ test that assumes equal standard deviations in both populations. The ANOVA $F$ for comparing two means is exactly the square of this special
One-Way Analysis of Variance: Comparing Several Means

We prefer the $t$ test that does not assume equal standard deviations, but for comparing more than two means there is no general alternative to the ANOVA $F$. It is not easy to check the condition that the populations have equal standard deviations. Statistical tests for equality of standard deviations are very sensitive to lack of Normality, so much so that they are of little practical value. You must either seek expert advice or rely on the robustness of ANOVA.

How serious are unequal standard deviations? ANOVA is not too sensitive to violations of the condition, especially when all samples have the same or similar sizes and no sample is very small. When designing a study, try to take samples of about the same size from all the groups you want to compare. The sample standard deviations estimate the population standard deviations, so check before doing ANOVA that the sample standard deviations are similar to each other. We expect some variation among them due to chance. Here is a rule of thumb that is safe in almost all situations.

**CHECKING STANDARD DEVIATIONS IN ANOVA**

The results of the ANOVA $F$ test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation.

**EXAMPLE 25.3 Comparing tropical flowers: conditions for ANOVA**

The study of *Heliconia* blossoms is based on three independent samples that the researchers consider to be random samples from all flowers of these varieties in Dominica. The stemplots in Figure 25.1 show that the *bihai* and red varieties have slightly skewed distributions, but the sample means of samples of sizes 16 and 23 will have distributions that are close to Normal. The sample standard deviations for the three varieties are

$$s_1 = 1.213 \quad s_2 = 1.799 \quad s_3 = 0.975$$

These standard deviations satisfy our rule of thumb:

$$\frac{\text{largest } s}{\text{smallest } s} = \frac{1.799}{0.975} = 1.85$$

We can safely use ANOVA to compare the mean lengths for the three populations.

**EXAMPLE 25.4 Which color attracts beetles best?**

**STATE:** To detect the presence of harmful insects in farm fields, we can put up boards covered with a sticky material and examine the insects trapped on the boards. Which colors attract insects best? Experimenters placed six boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped. Here are the data:

[Data provided]
FORMULATE: Examine the data to determine the effect of board color on beetles trapped and check that we can safely use ANOVA. If the data allow ANOVA, assess the significance of the observed differences in mean counts of beetles trapped.

SOLVE: Because the samples are small, we plot the data in side-by-side stemplots in Figure 25.7. CrunchIt! output for ANOVA appears in Figure 25.8. The yellow boards attract by far the most beetles ($\bar{x}_4 = 47.2$), with green next ($\bar{x}_3 = 31.2$) and blue and white far behind.

Check that we can safely use ANOVA to test equality of the four means. Because the standard error of $\bar{x}$ is $s/\sqrt{n}$, each sample standard deviation is $\sqrt{6}$ times the standard error given by CrunchIt!. The largest of the four standard deviations is 6.795 and the smallest is 3.764. The ratio

$$\frac{\text{largest } s}{\text{smallest } s} = \frac{6.795}{3.764} = 1.8$$

is less than 2, so these data satisfy our rule of thumb. The shapes of the four distributions are irregular, as we expect with only 6 observations in each group, but there are no outliers. The ANOVA results will be approximately correct. The $F$ statistic is $F = 42.84$, a large $F$ with $P < 0.0001$.

CONCLUDE: Despite the small samples, the experiment gives very strong evidence of differences among the colors. Yellow boards appear best at attracting leaf beetles.

### APPLY YOUR KNOWLEDGE

25.7 Checking standard deviations. Verify that the sample standard deviations for these sets of data do allow use of ANOVA to compare the population means.
CHAPTER 25 • One-Way Analysis of Variance: Comparing Several Means

(Figure 25.8) CrunchIt! ANOVA output for comparing the four board colors in Example 25.4.

(a) The counts of trees in Exercise 25.3 and Figure 25.4.
(b) The heart rates of Exercise 25.4 and Figure 25.5.

25.8 Species richness after logging. Table 25.2 gives data on the species richness in rain forest plots, defined as the number of tree species in a plot divided by the number of trees in the plot. ANOVA may not be trustworthy for the richness data. Do data analysis: make side-by-side stemplots to examine the distributions of the response variable in the three groups, and also compare the standard deviations. What characteristic of the data makes ANOVA risky?

25.9 Compressing soil. Farmers know that driving heavy equipment on wet soil compresses the soil and injures future crops. Table 2.3 (page 61) gives data on the “penetrability” of the same soil at three levels of compression. Penetrability is a measure of how much resistance plant roots will meet when they try to grow through the soil.
(a) Do the sample means suggest that penetrability decreases as soil is more compressed? We would like to use ANOVA to assess the significance of the differences among the means.
(b) Examine each of the three samples. Do the standard deviations satisfy our rule of thumb? What are the overall shapes of the distributions? Are there outliers? Can we safely use ANOVA on these data?
(c) Suppose that the outliers you found in part (b) were absent. Could we safely use ANOVA on the remaining data? Explain your answer.

**F distributions and degrees of freedom**

To find the *P*-value for the ANOVA *F* statistic, we must know the sampling distribution of *F* when the null hypothesis (all population means equal) is true. This sampling distribution is an *F* distribution. The *F* distributions are described on page 478 in Chapter 19. A specific *F* distribution is specified by two parameters: a numerator degrees of freedom and a denominator degrees of freedom. Table D in the back of the book contains critical values for *F* distributions with various degrees of freedom. You will rarely need Table D because software gives *P*-values directly.

**Example 25.5 Comparing tropical flowers: the *F* distribution**

Look again at the software output for the flower length data in Figure 25.3. All four outputs give the degrees of freedom for the *F* test, labeled “df” or “DF.” There are 2 degrees of freedom in the numerator and 51 in the denominator. *P*-values for the *F* test therefore come from the *F* distribution with 2 and 51 degrees of freedom. Figure 25.9 shows the density curve of this distribution. The 5% critical value is 3.179 and the 1% critical value is 5.047. The observed value \( F = 259.12 \) of the ANOVA *F* statistic lies far to the right of these values, so the *P*-value is extremely small.

The degrees of freedom of the ANOVA *F* statistic depend on the number of means we are comparing and the number of observations in each sample. That
is, the $F$ test takes into account the number of observations. Here are the details.

**degrees of freedom for the $F$ test**

We want to compare the means of $I$ populations. We have an SRS of size $n_i$ from the $i$th population, so that the total number of observations in all samples combined is

$$N = n_1 + n_2 + \cdots + n_I$$

If the null hypothesis that all population means are equal is true, the ANOVA $F$ statistic has the $F$ distribution with $I - 1$ degrees of freedom in the numerator and $N - I$ degrees of freedom in the denominator.

**Example 25.6 Degrees of freedom for $F$**

In Examples 25.1 and 25.2, we compared the mean lengths for three varieties of flowers, so $I = 3$. The three sample sizes are

$$n_1 = 16 \quad n_2 = 23 \quad n_3 = 15$$

The total number of observations is therefore

$$N = 16 + 23 + 15 = 54$$

The ANOVA $F$ test has numerator degrees of freedom

$$I - 1 = 3 - 1 = 2$$

and denominator degrees of freedom

$$N - I = 54 - 3 = 51$$

These are the degrees of freedom given in the outputs in Figure 25.3.

**Apply Your Knowledge**

25.10 Logging in the rain forest, continued. Exercise 25.3 compares the number of tree species in rain forest plots that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3).

(a) What are $I$, the $n_i$, and $N$ for these data? Identify these quantities in words and give their numerical values.

(b) Find the degrees of freedom for the ANOVA $F$ statistic. Check your work against the Excel output in Figure 25.4.

(c) For these data, $F = 11.43$. What does Table D tell you about the $P$-value of this statistic?

25.11 What music will you play? People often match their behavior to their social environment. One study of this idea first established that the type of music most preferred by black college students is R&B and that whites' most preferred music is rock. Will students hosting a small group of other students choose music that matches the makeup of the people attending? Assign 90 black business students at
random to three equal-sized groups. Do the same for 96 white students. Each student sees a picture of the people he or she will host. Group 1 sees 6 blacks, Group 2 sees 3 whites and 3 blacks, and Group 3 sees 6 whites. Ask how likely the host is to play the type of music preferred by the other race. Use ANOVA to compare the three groups to see whether the racial mix of the gathering affects the choice of music.7

(a) For the white subjects, \( F = 16.48 \). What are the degrees of freedom? Use Table D to give the approximate \( P \)-value. What do you conclude?
(b) For the black subjects, \( F = 2.47 \). What are the degrees of freedom? Use Table D to give the approximate \( P \)-value. What do you conclude?

---

**Some details of ANOVA: the two-sample case**

One-way ANOVA tests the hypotheses that all of \( I \) populations have the same mean,

\[
\begin{align*}
H_0: \mu_1 &= \mu_2 = \cdots = \mu_I \\
H_a: \text{not all of the } \mu_i \text{ are equal}
\end{align*}
\]

When there are just two populations, the hypotheses become

\[
\begin{align*}
H_0: \mu_1 &= \mu_2 \\
H_a: \mu_1 &\neq \mu_2
\end{align*}
\]

If the data suggest that the population distributions are at least roughly Normal (see the conditions in Chapter 19, page 462), we know what to do: use the two-sample \( t \) test. The idea of the \( t \) statistic is to standardize \( \bar{x}_1 - \bar{x}_2 \), that is,

\[
t = \frac{\bar{x}_1 - \bar{x}_2}{SE}
\]

The standard error \( SE \) estimates the standard deviation of \( \bar{x}_1 - \bar{x}_2 \), which is

\[
\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

The conditions for ANOVA say that two population standard deviations \( \sigma_1 \) and \( \sigma_2 \) have a common value (call it just \( \sigma \)). If we know that this is true, we should use this knowledge in forming the standard error. The key idea is to combine (the statistical term is “pool”) information from both samples to get a single estimate \( s_p \) of the single standard deviation \( \sigma \). When we do so, the standard error becomes

\[
SE = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}
\]

\[
= s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\]

*This more advanced section is optional if you are using software to find the \( F \) statistic.*
so that the pooled two-sample $t$ statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

What should we use for the estimate $s_p$ of $\sigma$? We should give more weight to the larger sample because larger samples contain more information than smaller samples. Theory that we won’t go into tells us the right way to do this: take a weighted average of the two sample variances $s_1^2$ and $s_2^2$, using their degrees of freedom as the weights. The pooled estimate of the variance $\sigma^2$ is then

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The advantage of the pooled $t$ statistic is that when the null hypothesis is true it has exactly a $t$ distribution, with $n_1 + n_2 - 2$ degrees of freedom. The larger degrees of freedom give slightly greater power than the conservative Option 2 method discussed in Chapter 19 (page 464). The disadvantage of pooled $t$ is that the condition that the population standard deviations are the same is hard to verify because tests for equal standard deviations are extremely sensitive to lack of Normality. The Option 1 method, with degrees of freedom provided by software, avoids this disadvantage. Extensive studies show that Option 1 $t$ performs essentially as well as pooled $t$ when the standard deviations really are equal and provides notably more accurate $P$-values when they are not.

When we want to compare more than two means, however, there is no simple way to avoid the equal standard deviations condition. The ANOVA $F$ statistic extends the idea behind any of the $t$ statistics. You can think of any $t$ statistic as comparing an average overall effect ($\bar{x}_1 - \bar{x}_2$ in the two-sample setting) with a measure of variation among individual observations ($s_p$, for example). This is the ANOVA idea, though the details are messier when we must compare more than two populations.

Because ANOVA requires equal population standard deviations, it is most closely related to the pooled two-sample $t$ statistic. In fact, in the two-sample case the $F$ statistic is exactly the square of the pooled $t$,

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals within samples}}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_p^2} \times \frac{n_1 n_2}{n_1 + n_2}$$

(The first term in this expression displays the comparison-of variation idea. The second term plays the role of the various constants in the $t$ statistics, arranging that we get exactly an $F$ distribution when the null hypothesis is true.)
The general ANOVA $F$ for more than two samples extends this idea in a straightforward way. In fact, the denominator of $F$ is exactly the weighted average of all the sample variances with degrees of freedom as weights. The numerator must compare $I$ sample means rather than just 2, so it becomes a bit more complicated.

Some details of ANOVA

Now we will give the actual formula for the ANOVA $F$ statistic. We have SRSs from each of $I$ populations. Subscripts from 1 to $I$ tell us which sample a statistic refers to:

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample size</th>
<th>Sample mean</th>
<th>Sample std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_1$</td>
<td>$\bar{x}_1$</td>
<td>$s_1$</td>
</tr>
<tr>
<td>2</td>
<td>$n_2$</td>
<td>$\bar{x}_2$</td>
<td>$s_2$</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>$I$</td>
<td>$n_I$</td>
<td>$\bar{x}_I$</td>
<td>$s_I$</td>
</tr>
</tbody>
</table>

You can find the $F$ statistic from just the sample sizes $n_i$, the sample means $\bar{x}_i$, and the sample standard deviations $s_i$. You don't need to go back to the individual observations.

The ANOVA $F$ statistic has the form

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

The measures of variation in the numerator and denominator of $F$ are called mean squares. A mean square is a more general form of a sample variance. An ordinary sample variance $s^2$ is an average (or mean) of the squared deviations of observations from their mean, so it qualifies as a "mean square."

The numerator of $F$ is a mean square that measures variation among the $I$ sample means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_I$. Call the overall mean response (the mean of all $N$ observations together) $\bar{x}$. You can find $\bar{x}$ from the $I$ sample means by

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \cdots + n_I \bar{x}_I}{N}$$

The sum of each mean multiplied by the number of observations it represents is the sum of all the individual observations. Dividing this sum by $N$, the total number of observations, gives the overall mean $\bar{x}$. The numerator mean square in $F$ is an

---

*This more advanced section is optional if you are using software to find the $F$ statistic.*
average of the $I$ squared deviations of the means of the samples from $\bar{x}$. We call it the mean square for groups, abbreviated as MSG.

$$MSG = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_I(\bar{x}_I - \bar{x})^2}{I - 1}$$

Each squared deviation is weighted by $n_i$, the number of observations it represents.

The mean square in the denominator of $F$ measures variation among individual observations in the same sample. For any one sample, the sample variance $s_i^2$ does this job. For all $I$ samples together, we use an average of the individual sample variances. It is again a weighted average in which each $s_i^2$ is weighted by one fewer than the number of observations it represents, $n_i - 1$. Another way to put this is that each $s_i^2$ is weighted by its degrees of freedom $n_i - 1$. The resulting mean square is called the mean square for error, MSE.

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{N - I}$$

“Error” doesn’t mean a mistake has been made. It’s a traditional term for chance variation. Here is a summary of the ANOVA test.

### THE ANOVA F TEST

Draw an independent SRS from each of $I$ populations. The $i$th population has the $N(\mu_i, \sigma)$ distribution, where $\sigma$ is the common standard deviation in all the populations. The $i$th sample has size $n_i$, sample mean $\bar{x}_i$, and sample standard deviation $s_i$.

The ANOVA F statistic tests the null hypothesis that all $I$ populations have the same mean:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

$$H_a: \text{not all of the } \mu_i \text{ are equal}$$

The statistic is

$$F = \frac{MSG}{MSE}$$

The numerator of $F$ is the mean square for groups

$$MSG = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_I(\bar{x}_I - \bar{x})^2}{I - 1}$$

The denominator of $F$ is the mean square for error

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{N - I}$$

When $H_0$ is true, $F$ has the $F$ distribution with $I - 1$ and $N - I$ degrees of freedom.
The denominators in the formulas for MSG and MSE are the two degrees of freedom $I - 1$ and $N - I$ of the $F$ test. The numerators are called **sums of squares**, from their algebraic form. It is usual to present the results of ANOVA in an **ANOVA table**. Output from software usually includes an ANOVA table.

**EXAMPLE 25.7 ANOVA calculations: software**

Look again at the four outputs in Figure 25.3. The three software outputs give the ANOVA table. The TI-83, with its small screen, gives the degrees of freedom, sums of squares, and mean squares separately. Each output uses slightly different language to identify the two sources of variation. The basic ANOVA table is

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$F$ statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation among samples</td>
<td>2</td>
<td>1082.87</td>
<td>MSG = 541.44</td>
<td>259.12</td>
</tr>
<tr>
<td>Variation within samples</td>
<td>51</td>
<td>106.57</td>
<td>MSE = 2.09</td>
<td></td>
</tr>
</tbody>
</table>

You can check that each mean square MS is the corresponding sum of squares SS divided by its degrees of freedom df. The $F$ statistic is MSG divided by MSE.

Because MSE is an average of the individual sample variances, it is also called the **pooled sample variance**, written as $s_p^2$. When all $I$ populations have the same population variance $\sigma^2$, as ANOVA assumes that they do, $s_p^2$ estimates the common variance $\sigma^2$. The square root of MSE is the **pooled standard deviation** $s_p$. It estimates the common standard deviation $\sigma$ of observations in each group. The Minitab and TI-83 outputs in Figure 25.3 give the value $s_p = 1.446$.

The pooled standard deviation $s_p$ is a better estimator of the common $\sigma$ than any individual sample standard deviation $s_i$ because it combines (pools) the information in all $I$ samples. We can get a confidence interval for any of the means $\mu_i$ from the usual form

$$\text{estimate} \pm t^* \text{SE}_{\text{estimate}}$$

using $s_p$ to estimate $\sigma$. The confidence interval for $\mu_i$ is

$$\overline{x}_i \pm t^* \frac{s_p}{\sqrt{n_i}}$$

Use the critical value $t^*$ from the $t$ distribution with $N - I$ degrees of freedom because $s_p$ has $N - I$ degrees of freedom. These are the confidence intervals that appear in Minitab ANOVA output.

**EXAMPLE 25.8 ANOVA calculations: without software**

We can do the ANOVA test comparing the mean lengths of bihai, red, and yellow flower varieties using only the sample sizes, sample means, and sample standard deviations. These appear in Example 25.1, but it is easy to find them with a calculator. There are $I = 3$ groups with a total of $N = 54$ flowers.
The overall mean of the 54 lengths in Table 25.1 is
\[ \overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + n_3\overline{x}_3}{N} \]
\[ = \frac{(16)(47.598) + (23)(39.711) + (15)(36.180)}{54} \]
\[ = \frac{2217.621}{54} = 41.067 \]

The mean square for groups is
\[ MSG = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + n_3(\overline{x}_3 - \overline{x})^2}{I - 1} \]
\[ = \frac{1}{3 - 1} \left[ (16)(47.598 - 41.067)^2 + (23)(39.711 - 41.067)^2 + (15)(36.180 - 41.067)^2 \right] \]
\[ = \frac{1082.996}{2} = 541.50 \]

The mean square for error is
\[ MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{N - I} \]
\[ = \frac{(15)(1.213^2) + (22)(1.799^2) + (14)(0.975^2)}{51} \]
\[ = \frac{106.580}{51} = 2.09 \]

Finally, the ANOVA test statistic is
\[ F = \frac{MSG}{MSE} = 541.50 \div 2.09 = 259.09 \]

Our work differs slightly from the output in Figure 25.3 because of roundoff error. We don't recommend doing these calculations, because tedium and roundoff errors cause frequent mistakes.

**APPLY YOUR KNOWLEDGE**

The calculations of ANOVA use only the sample sizes \( n_i \), the sample means \( \overline{x}_i \), and the sample standard deviations \( s_i \). You can therefore re-create the ANOVA calculations when a report gives these summaries but does not give the actual data. These optional exercises ask you to do the ANOVA calculations starting with the summary statistics.

**25.12 Road rage.** Exercise 25.2 describes a study of road rage. Here are the means and standard deviations for a measure of “angry/threatening driving” for random samples of drivers in three age groups:

<table>
<thead>
<tr>
<th>Age group</th>
<th>n</th>
<th>( \overline{x} )</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 30 years</td>
<td>244</td>
<td>2.22</td>
<td>3.11</td>
</tr>
<tr>
<td>30 to 55 years</td>
<td>734</td>
<td>1.33</td>
<td>2.21</td>
</tr>
<tr>
<td>Over 55 years</td>
<td>364</td>
<td>0.66</td>
<td>1.60</td>
</tr>
</tbody>
</table>
Some details of ANOVA

(a) The distributions of responses are somewhat right-skewed. ANOVA is nonetheless safe for these data. Why?
(b) Check that the standard deviations satisfy the guideline for ANOVA inference.
(c) Calculate the overall mean response $\bar{x}$, the mean squares MSG and MSE, and the ANOVA $F$ statistic.
(d) What are the degrees of freedom for $F$? How significant are the differences among the three mean responses?

25.13 Exercise and weight loss. What conditions help overweight people exercise regularly? Subjects were randomly assigned to three treatments: a single long exercise period 5 days per week; several 10-minute exercise periods 5 days per week; and several 10-minute periods 5 days per week on a home treadmill that was provided to the subjects. The study report contains the following information about weight loss (in kilograms) after six months of treatment:8

<table>
<thead>
<tr>
<th>Treatment</th>
<th>n</th>
<th>$\bar{x}$</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long exercise periods</td>
<td>37</td>
<td>10.2</td>
<td>4.2</td>
</tr>
<tr>
<td>Short exercise periods</td>
<td>36</td>
<td>9.3</td>
<td>4.5</td>
</tr>
<tr>
<td>Short periods with equipment</td>
<td>42</td>
<td>10.2</td>
<td>5.2</td>
</tr>
</tbody>
</table>

(a) Do the standard deviations satisfy the rule of thumb for safe use of ANOVA?  
(b) Calculate the overall mean response $\bar{x}$ and the mean square for groups MSG.  
(c) Calculate the mean square for error MSE.  
(d) Find the ANOVA $F$ statistic and its approximate $P$-value. Is there evidence that the mean weight losses of people who follow the three exercise programs differ?

25.14 Attitudes toward math. Do high school students from different racial/ethnic groups have different attitudes toward mathematics? Measure the level of interest in mathematics on a 5-point scale for a national random sample of students. Here are summaries for students who were taking math at the time of the survey:9

<table>
<thead>
<tr>
<th>Racial/ethnic group</th>
<th>n</th>
<th>$\bar{x}$</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>809</td>
<td>2.57</td>
<td>1.40</td>
</tr>
<tr>
<td>White</td>
<td>1860</td>
<td>2.32</td>
<td>1.36</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>654</td>
<td>2.63</td>
<td>1.32</td>
</tr>
<tr>
<td>Hispanic</td>
<td>883</td>
<td>2.51</td>
<td>1.31</td>
</tr>
<tr>
<td>Native American</td>
<td>207</td>
<td>2.51</td>
<td>1.28</td>
</tr>
</tbody>
</table>

Calculate the ANOVA table and the $F$ statistic. Show that there are significant differences among the mean attitudes of the five groups. What explains the small $P$-value? Do you think the differences are large enough to be important?
CHAPTER 25 SUMMARY

One-way analysis of variance (ANOVA) compares the means of several populations. The ANOVA $F$ test tests the overall $H_0$ that all the populations have the same mean. If the $F$ test shows significant differences, examine the data to see where the differences lie and whether they are large enough to be important.

The conditions for ANOVA state that we have an independent SRS from each population; that each population has a Normal distribution; and that all populations have the same standard deviation.

In practice, ANOVA inference is relatively robust when the populations are non-Normal, especially when the samples are large. Before doing the $F$ test, check the observations in each sample for outliers or strong skewness. Also verify that the largest sample standard deviation is no more than twice as large as the smallest standard deviation.

When the null hypothesis is true, the ANOVA $F$ statistic for comparing $I$ means from a total of $N$ observations in all samples combined has the $F$ distribution with $I - 1$ and $N - I$ degrees of freedom.

ANOVA calculations are reported in an ANOVA table that gives sums of squares, mean squares, and degrees of freedom for variation among groups and for variation within groups. In practice, we use software to do the calculations.

STATISTICS IN SUMMARY

Here are the most important skills you should have acquired from reading this chapter.

A. RECOGNITION

1. Recognize when testing the equality of several means is helpful in understanding data.

2. Recognize that the statistical significance of differences among sample means depends on the sizes of the samples and on how much variation there is within the samples.

3. Recognize when you can safely use ANOVA to compare means. Check the data production, the presence of outliers, and the sample standard deviations for the groups you want to compare.

B. INTERPRETING ANOVA

1. Explain what null hypothesis $F$ tests in a specific setting.

2. Locate the $F$ statistic and its $P$-value on the output of analysis of variance software.

3. Find the degrees of freedom for the $F$ statistic from the number and sizes of the samples. Use Table D of the $F$ distributions to approximate the $P$-value when software does not give it.

4. If the test is significant, use graphs and descriptive statistics to see what differences among the means are most important.
Check Your Skills

25.15 The purpose of analysis of variance is to compare
(a) the variances of several populations.
(b) the proportions of successes in several populations.
(c) the means of several populations.

25.16 A study of the effects of smoking classifies subjects as nonsmokers, moderate smokers, or heavy smokers. The investigators interview a sample of 200 people in each group. Among the questions is “How many hours do you sleep on a typical night?” The degrees of freedom for the ANOVA $F$ statistic comparing mean hours of sleep are
(a) 2 and 197.  (b) 2 and 597.  (c) 3 and 597.

25.17 The alternative hypothesis for the ANOVA $F$ test in the previous exercise is
(a) the mean hours of sleep in the groups are all the same.
(b) the mean hours of sleep in the groups are all different.
(c) the mean hours of sleep in the groups are not all the same.

The air in poultry processing plants often contains fungus spores. Large spore concentrations can affect the health of the workers. 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 slaughters 37,000 turkeys per day, taken at four seasons of the year. Each observation was made on a different day. The units are CFUs per cubic meter of air.\(^\text{10}\)

<table>
<thead>
<tr>
<th></th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
<th>Summer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1231</td>
<td>384</td>
<td>2105</td>
<td>3175</td>
<td></td>
</tr>
<tr>
<td>1254</td>
<td>104</td>
<td>701</td>
<td>2526</td>
<td></td>
</tr>
<tr>
<td>752</td>
<td>251</td>
<td>2947</td>
<td>1763</td>
<td></td>
</tr>
<tr>
<td>1088</td>
<td>97</td>
<td>842</td>
<td>1090</td>
<td></td>
</tr>
</tbody>
</table>

Here is Minitab output for ANOVA to compare mean CFUs in the four seasons:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Season</td>
<td>3</td>
<td>8236359</td>
<td>2745453</td>
<td>5.38</td>
<td>0.014</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>6124211</td>
<td>510351</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>14360570</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs For Mean Based on Pooled StDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>4</td>
<td>1081.3</td>
<td>231.5</td>
<td>(--<em>--</em>)</td>
<td>(--<em>--</em>)</td>
</tr>
<tr>
<td>Spring</td>
<td>4</td>
<td>1648.8</td>
<td>1071.2</td>
<td>(--<em>--</em>)</td>
<td>(--<em>--</em>)</td>
</tr>
<tr>
<td>Summer</td>
<td>4</td>
<td>2138.5</td>
<td>906.4</td>
<td>(--<em>--</em>)</td>
<td>(--<em>--</em>)</td>
</tr>
<tr>
<td>Winter</td>
<td>4</td>
<td>209.0</td>
<td>136.6</td>
<td>(--<em>--</em>)</td>
<td>(--<em>--</em>)</td>
</tr>
</tbody>
</table>

Exercises 25.18 to 25.23 are based on this study.
The most striking conclusion from the numerical summaries for the turkey-processing plant is that
(a) there appears to be little difference among the seasons.
(b) on the average, CFUs are much lower in winter than in other seasons.
(c) the air in the plant is clearly unhealthy.

We might use the two-sample $t$ procedures to compare summer and winter. The conservative 90% confidence interval for the difference in the two population means is
(a) $1929.5 \pm 458.3$.
(b) $1929.5 \pm 1078.4$.
(c) $1929.5 \pm 1458.4$.

In all, we would have to give 6 two-sample confidence intervals to compare all pairs of seasons. The weakness of doing this is that
(a) we don’t know how confident we can be that all 6 intervals cover the true differences in means.
(b) 90% confidence is OK for one comparison, but it isn’t high enough for 6 comparisons done at once.
(c) the conditions for two-sample $t$ inference are not met for all 6 pairs of seasons.

The conclusion of the ANOVA test is that
(a) there is quite strong evidence ($P = 0.014$) that the mean CFUs are not the same in all four seasons.
(b) there is quite strong evidence ($P = 0.014$) that the mean CFUs are much lower in winter than in any other season.
(c) the data give no evidence ($P = 0.014$) to suggest that mean CFUs differ from season to season.

Without software, we would compare $F = 5.38$ with critical values from Table D. This comparison shows that
(a) the $P$-value is greater than 0.1.
(b) the $P$-value is between 0.01 and 0.025.
(c) the $P$-value is between 0.001 and 0.01.

The $P$-value 0.014 in the output may not be accurate because the conditions for ANOVA are not satisfied. The most serious violation of the conditions is that
(a) the sample standard deviations are too different.
(b) there is an extreme outlier in the data.
(c) the data can’t be regarded as random samples.

**CHAPTER 25 EXERCISES**

Exercises 25.24 to 25.27 describe situations in which we want to compare the mean responses in several populations. For each setting, identify the populations and the response variable. Then give $I$, the $n_i$, and $N$. Finally, give the degrees of freedom of the ANOVA $F$ test.

25.24 How much do students borrow? A sample survey of students who had just received their bachelor’s degree found that about half had borrowed money to pay
for college. Does the type of college attended influence the amount borrowed? Here is information about the students in the sample who did borrow:

<table>
<thead>
<tr>
<th></th>
<th>Public nondoctrate</th>
<th>Public doctrate</th>
<th>Private nondoctrate</th>
<th>Private doctrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of students</td>
<td>137</td>
<td>313</td>
<td>148</td>
<td>95</td>
</tr>
<tr>
<td>Mean borrowed</td>
<td>$15,000</td>
<td>$17,500</td>
<td>$20,900</td>
<td>$28,000</td>
</tr>
</tbody>
</table>

**25.25 Morning or evening?** Are you a morning person, an evening person, or neither? Does this personality trait affect how well you perform? A sample of 100 students took a psychological test that found 16 morning people, 30 evening people, and 54 who were neither. All the students then took a test of their ability to memorize at 8 A.M. and again at 9 P.M. Analyze the score at 8 A.M. minus the score at 9 P.M.

**25.26 Writing essays.** Do strategies such as preparing a written outline help students write better essays? College students were divided at random into four groups of 20 students each, then asked to write an essay on an assigned topic. Group A (the control group) received no additional instruction. Group B was required to prepare a written outline. Group C was given 15 ideas that might be relevant to the essay topic. Group D was given the ideas and also required to prepare an outline. An expert scored the quality of the essays on a scale of 1 to 7.

**25.27 A medical study.** The Quebec (Canada) Cardiovascular Study recruited men aged 34 to 64 at random from towns in the Quebec City metropolitan area. Of these, 1824 met the criteria (no diabetes, free of heart disease, and so on) for a study of the relationship between being overweight and medical risks. The 719 normal-weight men had mean triglyceride level 1.5 millimoles per liter (mmol/l); the 885 overweight men had mean 1.7 mmol/l; and the 220 obese men had mean 1.9 mmol/l.

**25.28 Whose little bird is that?** The males of many animal species signal their fitness to females by displays of various kinds. For male barn swallows, the signal is the color of their plumage. Does better color really lead to more offspring? A randomized comparative experiment assigned barn swallow pairs who had already produced a clutch of eggs to three groups: 13 males had their color “enhanced” by the investigators, 9 were handled but their color was left alone, and 8 were not touched. DNA testing showed whether the eggs were sired by the male of the pair rather than by another male. The investigators then destroyed the eggs. Barn swallows usually produce a second brood in such circumstances. Sure enough, the males with enhanced color fathered more young in the second brood, and males in the other groups fathered fewer. The investigators used ANOVA to compare the three groups.

(a) What are the degrees of freedom for the $F$ statistic that compares the numbers of the male’s own young among the eggs in the first brood? The statistic was $F = 0.07$. What can you say about the $P$-value? What do you conclude?

(b) Of the 30 original pairs, 27 produced a second brood. What are the degrees of freedom for $F$ to compare differences in the number of the male’s own young
between the first and second broods produced by these 27 pairs? The statistic was $F = 5.45$. How significant is this $F$?

25.29 Plants defend themselves. When some plants are attacked by leaf-eating insects, they release chemical compounds that attract other insects that prey on the leaf-eaters. A study carried out on plants growing naturally in the Utah desert demonstrated both the release of the compounds and that they not only repel the leaf-eaters but attract predators that act as the plants’ bodyguards. The investigators chose 8 plants attacked by each of three leaf-eaters and 8 more that were undamaged, 32 plants of the same species in all. They then measured emissions of several compounds during seven hours. Here are data (mean ± standard error of the mean for eight plants) for one compound. The emission rate is measured in nanograms (ng) per hour.

<table>
<thead>
<tr>
<th>Group</th>
<th>Emission rate (ng/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>9.22 ± 5.93</td>
</tr>
<tr>
<td>Hornworm</td>
<td>31.03 ± 8.75</td>
</tr>
<tr>
<td>Leaf bug</td>
<td>18.97 ± 6.64</td>
</tr>
<tr>
<td>Flea beetle</td>
<td>27.12 ± 8.62</td>
</tr>
</tbody>
</table>

(a) Make a graph that compares the mean emission rates for the four groups. Does it appear that emissions increase when a plant is attacked?
(b) What hypotheses does ANOVA test in this setting?
(c) We do not have the full data. What would you look for in deciding whether you can safely use ANOVA?
(d) What is the relationship between the standard error of the mean (SEM) and the standard deviation for a sample? Do the standard deviations satisfy our rule of thumb for safe use of ANOVA?

25.30 Can you hear these words? To test whether a hearing aid is right for a patient, audiologists play a tape on which words are pronounced at low volume. The patient tries to repeat the words. There are several different lists of words that are supposed to be equally difficult. Are the lists equally difficult when there is background noise? To find out, an experimenter had subjects with normal hearing listen to four lists with a noisy background. The response variable was the percent of the 50 words in a list that the subject repeated correctly. The data set contains 96 responses. Here are two study designs that could produce these data:

**Design A.** The experimenter assigns 96 subjects to 4 groups at random. Each group of 24 subjects listens to one of the lists. All individuals listen and respond separately.

**Design B.** The experimenter has 24 subjects. Each subject listens to all four lists in random order. All individuals listen and respond separately.

Does Design A allow use of one-way ANOVA to compare the lists? Does Design B allow use of one-way ANOVA to compare the lists? Briefly explain your answers.

25.31 Nematodes and tomato plants. How do nematodes (microscopic worms) affect plant growth? A botanist prepares 16 identical planting pots and then introduces
different numbers of nematodes into the pots. He transplants a tomato seedling into each pot. Here are data on the increase in height of the seedlings (in centimeters) 16 days after planting.\textsuperscript{15}

<table>
<thead>
<tr>
<th>Nematodes</th>
<th>Seedling growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10.8 9.1 13.5 9.2</td>
</tr>
<tr>
<td>1,000</td>
<td>11.1 11.1 8.2 11.3</td>
</tr>
<tr>
<td>5,000</td>
<td>5.4 4.6 7.4 5.0</td>
</tr>
<tr>
<td>10,000</td>
<td>5.8 5.3 3.2 7.5</td>
</tr>
</tbody>
</table>

Figure 25.10 shows output from CrunchIt! for ANOVA comparing the four groups of pots.

(a) Make a table of the means and standard deviations in the groups. Make side-by-side stemplots to compare the treatments. What do the data appear to show about the effect of nematodes on growth? Is use of ANOVA justified?

(b) State $H_0$ and $H_a$ for the ANOVA test for these data, and explain in words what ANOVA tests in this setting.

(c) Report your overall conclusions about the effect of nematodes on plant growth.

\textbf{Figure 25.10} CrunchIt! ANOVA output for comparing the growth of tomato seedlings with different concentrations of nematodes in the soil, for Exercise 25.31.
25.32 Can you hear these words? Figure 25.11 displays the Minitab output for one-way ANOVA applied to the hearing data described in Exercise 25.30. The response variable is “Percent,” and “List” identifies the four lists of words. Based on this analysis, is there good reason to think that the four lists are not all equally difficult? Write a brief summary of the study findings.

25.33 Which blue is most blue? The color of a fabric depends on the dye used and also on how the dye is applied. This matters to clothing manufacturers, who want the color of the fabric to be just right. Dye fabric made of ramie with the same “procion blue” die applied in four different ways. Then use a colorimeter to measure the lightness of the color on a scale in which black is 0 and white is 100. Here are the data for 8 pieces of fabric dyed in each way:

- Method A: 41.72, 41.83, 42.05, 41.44, 41.27, 42.27, 41.12, 41.49
- Method B: 40.98, 40.88, 41.30, 41.28, 41.66, 41.50, 41.39, 41.27
- Method C: 42.30, 42.20, 42.65, 42.43, 42.50, 42.28, 43.13, 42.45
- Method D: 41.68, 41.65, 42.30, 42.04, 42.25, 41.99, 41.72, 41.97

(a) This is a randomized comparative experiment. Outline the design.
(b) A clothing manufacturer wants to know which method gives the darkest color. Follow the four-step process in answering this question.

25.34 Does nature heal best? Our bodies have a natural electrical field that helps wounds heal. Might higher or lower levels speed healing? An experiment with newts investigated this question. Newts were randomly assigned to five groups. In four of the groups, an electrode applied to one hind limb (chosen at random) changed the natural field, while the other hind limb was not manipulated. Both limbs in the fifth (control) group remained in their natural state.
**Table 25.4** Effect of electrical field on healing rate in newts

<table>
<thead>
<tr>
<th>Group</th>
<th>Diff</th>
<th>Group</th>
<th>Diff</th>
<th>Group</th>
<th>Diff</th>
<th>Group</th>
<th>Diff</th>
<th>Group</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-10</td>
<td>0.5</td>
<td>-1</td>
<td>1</td>
<td>-7</td>
<td>1.25</td>
<td>1</td>
<td>1.5</td>
<td>-13</td>
</tr>
<tr>
<td>0</td>
<td>-12</td>
<td>0.5</td>
<td>10</td>
<td>1</td>
<td>15</td>
<td>1.25</td>
<td>8</td>
<td>1.5</td>
<td>-49</td>
</tr>
<tr>
<td>0</td>
<td>-9</td>
<td>0.5</td>
<td>3</td>
<td>1</td>
<td>-4</td>
<td>1.25</td>
<td>-15</td>
<td>1.5</td>
<td>-16</td>
</tr>
<tr>
<td>0</td>
<td>-11</td>
<td>0.5</td>
<td>-3</td>
<td>1</td>
<td>-16</td>
<td>1.25</td>
<td>14</td>
<td>1.5</td>
<td>-8</td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>0.5</td>
<td>-31</td>
<td>1</td>
<td>-2</td>
<td>1.25</td>
<td>-7</td>
<td>1.5</td>
<td>-2</td>
</tr>
<tr>
<td>0</td>
<td>-6</td>
<td>0.5</td>
<td>4</td>
<td>1</td>
<td>-13</td>
<td>1.25</td>
<td>-1</td>
<td>1.5</td>
<td>-35</td>
</tr>
<tr>
<td>0</td>
<td>-31</td>
<td>0.5</td>
<td>-12</td>
<td>1</td>
<td>5</td>
<td>1.25</td>
<td>11</td>
<td>1.5</td>
<td>-11</td>
</tr>
<tr>
<td>0</td>
<td>-5</td>
<td>0.5</td>
<td>-3</td>
<td>1</td>
<td>-4</td>
<td>1.25</td>
<td>8</td>
<td>1.5</td>
<td>-46</td>
</tr>
<tr>
<td>0</td>
<td>13</td>
<td>0.5</td>
<td>-7</td>
<td>1</td>
<td>-2</td>
<td>1.25</td>
<td>11</td>
<td>1.5</td>
<td>-22</td>
</tr>
<tr>
<td>0</td>
<td>-2</td>
<td>0.5</td>
<td>-10</td>
<td>1</td>
<td>-14</td>
<td>1.25</td>
<td>-4</td>
<td>1.5</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>-7</td>
<td>0.5</td>
<td>-22</td>
<td>1</td>
<td>5</td>
<td>1.25</td>
<td>7</td>
<td>1.5</td>
<td>10</td>
</tr>
<tr>
<td>0</td>
<td>-8</td>
<td>0.5</td>
<td>-4</td>
<td>1</td>
<td>11</td>
<td>1.25</td>
<td>-14</td>
<td>1.5</td>
<td>-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>-1</td>
<td>1</td>
<td>10</td>
<td>1.25</td>
<td>0</td>
<td>1.5</td>
<td>-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>-3</td>
<td>1</td>
<td>3</td>
<td>1.25</td>
<td>5</td>
<td>1.5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>-1</td>
<td>6</td>
<td>1.25</td>
<td>-2</td>
<td>1.5</td>
<td>-5</td>
<td></td>
</tr>
</tbody>
</table>

Table 25.4 gives data from this experiment. The “Group” variable shows the field applied as a multiple of the natural field for each newt. For example, “0.5” is half the natural field, “1” is the natural level (the control group), and “1.5” indicates a field 1.5 times natural. “Diff” is the response variable, the difference in the healing rate (in micrometers per hour) of cuts made in the experimental and control limbs of that newt. Negative values mean that the experimental limb healed more slowly. The investigators conjectured that nature heals best, so that changing the field from the natural state (the “1” group) will slow healing.

Do a complete analysis to see whether the groups differ in the effect of the electrical field level on healing. Follow the four-step process in your work.

**25.35 Does polyester decay?** How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay; lower strength means the fabric has decayed.

Part of the study buried 20 polyester strips in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after each of 2 weeks, 4 weeks, and 8 weeks. Here are the breaking strengths in pounds:

<table>
<thead>
<tr>
<th>Time</th>
<th>18</th>
<th>22</th>
<th>126</th>
<th>120</th>
<th>129</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 weeks</td>
<td>118</td>
<td>120</td>
<td>126</td>
<td>120</td>
<td>129</td>
</tr>
<tr>
<td>4 weeks</td>
<td>130</td>
<td>120</td>
<td>114</td>
<td>126</td>
<td>128</td>
</tr>
<tr>
<td>8 weeks</td>
<td>122</td>
<td>136</td>
<td>128</td>
<td>146</td>
<td>140</td>
</tr>
</tbody>
</table>
The investigator conjectured that buried polyester loses strength over time. Do the data support this conjecture? Follow the four-step process in data analysis and ANOVA. Be sure to check the conditions for ANOVA.

**25.36 Durable press fabrics are weaker.** “Durable press” cotton fabrics are treated to improve their recovery from wrinkles after washing. Unfortunately, the treatment also reduces the strength of the fabric. A study compared the breaking strength of untreated fabric with that of fabrics treated by three commercial durable press processes. Five specimens of the same fabric were assigned at random to each group. Here are the data, in pounds of pull needed to tear the fabric:

<table>
<thead>
<tr>
<th>Process</th>
<th>Untreated</th>
<th>Permafresh 55</th>
<th>Permafresh 48</th>
<th>Hylite LF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60.1</td>
<td>29.9</td>
<td>24.8</td>
<td>28.8</td>
</tr>
<tr>
<td></td>
<td>56.7</td>
<td>30.7</td>
<td>24.6</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>61.5</td>
<td>30.0</td>
<td>27.3</td>
<td>27.0</td>
</tr>
<tr>
<td></td>
<td>55.1</td>
<td>29.5</td>
<td>28.1</td>
<td>22.1</td>
</tr>
<tr>
<td></td>
<td>59.4</td>
<td>27.6</td>
<td>30.3</td>
<td>24.2</td>
</tr>
</tbody>
</table>

The untreated fabric is clearly much stronger than any of the treated fabrics. We want to know if there is a significant difference in breaking strength among the three durable press treatments. Analyze the data for the three processes and write a clear summary of your findings. Which process do you recommend if breaking strength is a main concern? Use the four-step process to guide your discussion. (Although the standard deviations do not quite satisfy our rule of thumb, that rule is conservative and many statisticians would use ANOVA for these data.)

**25.37 Durable press fabrics wrinkle less.** The data in Exercise 25.36 show that durable press treatment greatly reduces the breaking strength of cotton fabric. Of course, durable press treatment also reduces wrinkling. How much? “Wrinkle recovery angle” measures how well a fabric recovers from wrinkles. Higher is better. Here are data on the wrinkle recovery angle (in degrees) for the same fabric specimens discussed in the previous exercise:

<table>
<thead>
<tr>
<th>Process</th>
<th>Untreated</th>
<th>Permafresh 55</th>
<th>Permafresh 48</th>
<th>Hylite LF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>79</td>
<td>136</td>
<td>125</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>135</td>
<td>131</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>132</td>
<td>125</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>137</td>
<td>145</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>134</td>
<td>145</td>
<td></td>
</tr>
</tbody>
</table>

The untreated fabric once again stands out, this time as inferior to the treated fabrics in wrinkle resistance. Examine the data for the three durable press processes and summarize your findings. How does the ranking of the three processes by wrinkle resistance compare with their ranking by breaking strength in Exercise 25.36? Explain why we can’t trust the ANOVA $F$ test.

**25.38 Logging in the rain forest: species counts.** Table 25.2 gives data on the number of trees per forest plot, the number of species per plot, and species richness. Exercise 25.3 analyzed the effect of logging on number of trees. Exercise 25.8 concludes that it would be risky to use ANOVA to analyze richness. Use software to analyze the effect of logging on the number of species.

(a) Make a table of the group means and standard deviations. Do the standard deviations satisfy our rule of thumb for safe use of ANOVA? What do the means suggest about the effect of logging on the number of species?
(b) Carry out the ANOVA. Report the $F$ statistic and its $P$-value and state your conclusion.

25.39 Plant defenses (optional). The calculations of ANOVA use only the sample sizes $n_i$, the sample means $\overline{x}_i$, and the sample standard deviations $s_i$. You can therefore re-create the ANOVA calculations when a report gives these summaries but does not give the actual data. Use the information in Exercise 25.29 to calculate the ANOVA table (sums of squares, degrees of freedom, mean squares, and the $F$ statistic). Note that the report gives the standard error of the mean (SEM) rather than the standard deviation. Are there significant differences among the mean emission rates for the four populations of plants?

25.40 $F$ versus $t$ (optional). We have two methods to compare the means of two groups: the two-sample $t$ test of Chapter 19 and the ANOVA $F$ test with $I = 2$. We prefer the $t$ test because it allows one-sided alternatives and does not assume that both populations have the same standard deviation. Let us apply both tests to the same data.

There are two types of life insurance companies. "Stock" companies have shareholders, and "mutual" companies are owned by their policyholders. Take an SRS of each type of company from those listed in a directory of the industry. Then ask the annual cost per $1000 of insurance for a $50,000 policy insuring the life of a 35-year-old man who does not smoke. Here are the data summaries:

<table>
<thead>
<tr>
<th></th>
<th>Stock companies</th>
<th>Mutual companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_i$</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>$\overline{x}_i$</td>
<td>$2.31$</td>
<td>$2.37$</td>
</tr>
<tr>
<td>$s_i$</td>
<td>$0.38$</td>
<td>$0.58$</td>
</tr>
</tbody>
</table>

(a) Calculate the two-sample $t$ statistic for testing $H_0: \mu_1 = \mu_2$ against the two-sided alternative. Use the conservative method to find the $P$-value.

(b) Calculate MSG, MSE, and the ANOVA $F$ statistic for the same hypotheses. What is the $P$-value of $F$?

(c) How close are the two $P$-values?

25.41 ANOVA or chi-square? (optional). Exercise 25.13 describes a randomized, comparative experiment that assigned subjects to three types of exercise programs intended to help them lose weight. Some of the results of this study were analyzed using the chi-square test for two-way tables, and some others were analyzed using one-way ANOVA. For each of the following excerpts from the study report, say which analysis is appropriate and explain how you made your choice.

(a) "Overall, 115 subjects (78% of 148 subjects randomized) completed 18 months of treatment, with no significant difference in attrition rates between the groups ($P = .12$)."

(b) "In analyses using only the 115 subjects who completed 18 months of treatment, there were no significant differences in weight loss at 6 months among the groups."
(c) “The duration of exercise for weeks 1 through 4 was significantly greater in the SB compared with both LB and SBEQ groups ($P < .05$). . . . However, exercise duration was greater in SBEQ compared with both LB and SB groups for months 13 through 18 ($P < .05$).”

**EESEE CASE STUDIES**

*The Electronic Encyclopedia of Statistical Examples and Exercises (EESEE) is available on the text CD and Web site. These more elaborate stories, with data, provide settings for longer case studies. Here are some suggestions for EESEE stories that apply ANOVA.*

25.42 Read the EESEE story “Blinded Knee Doctors.” Write a report that answers all questions for this case study.

25.43 Read the EESEE story “Stress among Pets and Friends.” Write a report that answers all questions for this case study.

25.44 Read the EESEE story “Nutrition and Breakfast Cereals.” Write a report that answers all questions for this case study.