Review of Statistics

We have found in teaching research methods that even students who have had a prior course in statistics find a review of some basic concepts very helpful. We believe that this results from the technical and abstract nature of statistical thinking. The purpose of this appendix is to review some concepts from introductory statistics in a nontechnical way to provide additional practice in understanding them. Because there are so many different statistical tests, and they differ among themselves, you will need to refer to a book on statistics to select the appropriate tests and to find the formulas and steps for conducting them.

SOME BASIC TERMS

empirical data: facts derived from experience

First, let us review a few basic terms. Statistics is—this is not a misprint; statistics as a field is a singular word—an area of study having to do with interpretation of empirical data. This term needs some definition. The word data refers to facts. The word empirical means based on experience. So, empirical data are facts that are obtained by observation or experiment. Now, these empirical data ordinarily exist in numerical form: The mean IQ of the population is 100, the average reaction time was 483 milliseconds [ms], and so forth. But it is important to note that not all numbers are empirical data. Mathematicians often talk about numbers in the abstract: Two plus two is four. These numbers are not empirical data because they do not refer to a specific observation or experiment; they are true by definition.

(It is worth noting that data is a plural word. The singular form is datum. A datum is a fact, and data are a collection of facts. Scientists say, “The data are such and such” rather than “The data is such and such.” In common usage, it is often considered pedantic [picky] to treat the word data as plural, but scientific usage still favors treating data as a plural word.)
A population is the entire collection of individuals being considered: all people who live in the United States (the US population), all students at State University, all college sophomores, all possible tosses of a pair of dice, and the like. Note from the last example that a population does not necessarily contain people or even animate objects. Nor is a population necessarily finite: You could toss dice forever and still be able to toss them some more. A sample, by contrast, is a subset of the population: 100 randomly selected people who live in the United States, every 100th student at State University, 100 tosses of a pair of dice, and so on. Statistics deal sometimes with populations and sometimes with samples.

A statistic is a quantity computed from a sample: The mean number of hours worked per week is a statistic if it is based on a sample of the students at State University. (This differs from the common usage in which a statistic is any sort of empirical datum: “Drive carefully or you will become a statistic.”) A parameter is a quantity computed from the population. The mean number of hours worked per week by students at State University would be a parameter if we had obtained the data from every student at the university. Note that the term parameter is used in statistics in a somewhat different sense than it is when one is talking about a function. A mean, then, can be either a statistic or a parameter, depending on whether it is based on a sample of the population or the entire population. This distinction is signified in statistical notation. A statistic is identified by a Roman letter (the ones we use every day), whereas a parameter is identified by a Greek letter. Thus, the mean of a sample is usually indicated by \( \bar{X} \), called “X bar,” whereas the mean of a population is indicated by \( \mu \), the lowercase Greek letter mu (pronounced “mew”).

There are two main uses of statistics: to describe a particular set of data and to use data to draw conclusions about a population. These two uses correspond to the distinction between descriptive statistics and inferential statistics. We use descriptive statistics to summarize what was found in a set of empirical data. For example, we might have found that the mean reaction time was 483 ms when subjects had to respond to the presence of a single target and 621 ms when they had to respond differentially to one of two possible targets. If we try to draw a conclusion about whether simple reaction time is shorter than choice reaction time in general, we need to use inferential statistics. Here we are trying to decide whether the mean of the population of simple reaction times is shorter than that of choice reaction times.

**DESCRIPTIVE STATISTICS**

The most common descriptive statistics are those that concern the average and the variability of a set of data, and those that describe the degree of relationship between two variables.

**Measures of Central Tendency**

A measure of central tendency is a single number that is used to represent the average score in the distribution. Three common measures of central tendency are the mode, the median, and the mean. All three are actually kinds of averages although people commonly use the term average to refer to the mean, one of the kinds of averages. Some of the meanings commonly associated with the
term *average* are these: a number that is typical of all the scores, a number that is in the middle of the scores, and a number that represents all scores. Although all three meanings are true of each measure of central tendency, each measure best captures one of them, as we see in the following sections.

**Mode**

The *mode* is the easiest measure of central tendency to define. It is the most common score in a frequency distribution. The mode has the advantage of representing the most typical score. It also has the practical advantage of being the easiest to compute because it literally sticks out in a frequency distribution. We can see in data shown in Table 6.1 on page 143 that the mode is 17, having been earned by three students.

In a large distribution, the mode will be fairly stable, but in a small data set, such as the one we are discussing from Chapter 6, the mode can bounce around considerably. For example, notice that if one student had earned a 14 instead of a 17, the mode would have changed by 3. The mode, therefore, is not very useful for small data sets. In large data sets, however, the mode is a good representation of the typical case. Another disadvantage of the mode is that it does not enter into any further statistical calculations. It is sort of a statistical orphan.

**Median**

The *median* is the middlemost score in a distribution. Computing the median requires one to rank-order the scores in a distribution from highest to lowest and find the middle one. The following equation tells you which score is the middle one:

\[
\text{middle score} = \frac{\text{number of scores} + 1}{2}
\]

For example, if there are nine scores, the middle score is:

\[
(9 + 1)/2 = 5\text{th score}
\]

If there is an even number of scores, then there are two middle scores. In that case, the median is halfway between the two middle scores. For example, if there are 10 scores, the middle one is:

\[
(10 + 1)/2 = 5.5
\]

You average the fifth and sixth scores to obtain the median.

For example, go back to Table 6.1 and find the median. We first count the number of scores and find that there are 10 scores. From the last equation we know that the median is halfway between the fifth and sixth scores. Counting from the bottom, we find that the fifth score is 15 and the sixth score is 16. The average of these two, and thus the median score, is 15.5.

The median has the advantage that half the scores in the distribution fall above it, and half fall below it. Thus, it is the middlemost score. It is not affected by how far other scores are from the median, only by how many scores fall above or below. Disadvantages of the median are that it requires ranking all the scores and counting to find the middle. This can be quite a chore in a large data set. Another disadvantage of the median is that its use in further statistical computations is somewhat limited, as we will see shortly.
Mean

The mean is the ordinary average that you learned to compute in grade school. As you may recall, the mean is computed by adding all the scores and dividing by the number of scores:

\[ \text{mean} = \frac{\Sigma X}{N} \]

or

\[ \text{mean} = \frac{\text{sum of scores}}{\text{number of scores}} \]

The main advantage of the mean is that it uses all the information in the distribution. In other words, the mean is influenced by the value of every score in the distribution. (Note that all the scores are summed and therefore enter into the value of the mean.) Thus, of the three measures of central tendency, the mean best captures the idea of the average as the quantity that represents all the scores in the distribution. The mode, by contrast, is not influenced at all by the other scores in the distribution; the median is influenced only by how many scores fall above and below it. A second advantage is that the mean is the basis of the most common and most powerful of the further statistical computations that we will discuss later.

Third, means of subgroups may be combined to obtain the mean of the entire group. Recall Professor Carlton from Chapter 6. If he has two sections of the same course, he can average the means of the two sections to get the mean of the whole course. There is no way that medians of subgroups can be combined to get the median of the entire group.

The mean does have disadvantages, however. Precisely because it uses every score in the distribution, it is sensitive to the value of extreme scores, as we will see shortly.

Behavior of the Mean, Median, and Mode with Various Shaped Distributions

When data are distributed symmetrically, the three measures of central tendency will be the same. When the data are skewed, however, they will be differentially affected. Refer back to Professor Carlton’s test, as shown in Figure 6.2. As is common with tests in college courses, the data are skewed to the left. Most students tend to do quite well, but a few fall at the low end of the scale. The mode is 17, the median is 15.5, and the mean is 15. This is what would be expected with a distribution skewed to the left: mode > median > mean. (Skewness to the right would produce the opposite order.)

---

1Provided each subgroup has the same number of cases, or the subgroup means are weighted by the number of cases.

\[ \text{Grand mean} = \frac{(\text{mean}_1)N_1 + (\text{mean}_2)N_2}{N_1 + N_2} \]

For example, if Group 1 has a mean of 16 and 4 cases, and Group 2 has a mean of 6 and 6 cases, the mean of all cases will be 10:

\[ \frac{(16 \times 4) + (6 \times 6)}{10} = 10 \]
### TABLE A.1

<table>
<thead>
<tr>
<th>Score</th>
<th>Upper Real Limit</th>
<th>Tally</th>
<th>Frequency</th>
<th>Cum Freq</th>
<th>%</th>
<th>Cum%</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20.5</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>19.5</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>18.5</td>
<td>I</td>
<td>11</td>
<td>7.1</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17.5</td>
<td>II</td>
<td>10</td>
<td>27.3</td>
<td>90.9</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16.5</td>
<td>I</td>
<td>1</td>
<td>7</td>
<td>9.1</td>
<td>63.6</td>
</tr>
<tr>
<td>15</td>
<td>15.5</td>
<td>I</td>
<td>1</td>
<td>6</td>
<td>9.1</td>
<td>55.5</td>
</tr>
<tr>
<td>14</td>
<td>14.5</td>
<td>II</td>
<td>2</td>
<td>5</td>
<td>18.2</td>
<td>45.5</td>
</tr>
<tr>
<td>13</td>
<td>13.5</td>
<td></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>27.3</td>
</tr>
<tr>
<td>12</td>
<td>12.5</td>
<td>I</td>
<td>1</td>
<td>3</td>
<td>9.1</td>
<td>27.3</td>
</tr>
<tr>
<td>11</td>
<td>11.5</td>
<td>O</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>18.2</td>
</tr>
<tr>
<td>10</td>
<td>10.5</td>
<td>I</td>
<td>2</td>
<td>9.1</td>
<td>18.2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9.5</td>
<td>O</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9.1</td>
</tr>
<tr>
<td>8</td>
<td>8.5</td>
<td>O</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9.1</td>
</tr>
<tr>
<td>7</td>
<td>7.5</td>
<td>O</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9.1</td>
</tr>
<tr>
<td>6</td>
<td>6.5</td>
<td>O</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9.1</td>
</tr>
<tr>
<td>5</td>
<td>5.5</td>
<td>O</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9.1</td>
</tr>
<tr>
<td>4</td>
<td>4.5</td>
<td>I</td>
<td>1</td>
<td>9.1</td>
<td>9.1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**N = 11**  
**Σ = 154**  
**X̄ = 14**  
**Median = 15**  
**Mode = 17**

The mode is not affected at all by the skewness of the data because, as we have seen, it is not affected by any other scores. The median is affected by the number of scores above and below it, so the skewness will pull it down somewhat. The mean, however, is lowest because it is affected by the distance of the low scores from the middle, as well as by the number of scores.

The effect of skewness on the various measures of central tendency can be seen by looking at Table A.1, which shows the same data that we have been talking about with the addition of a single outlier. Suppose that another student took the test late and got a very low score of 4. The mode is not changed by the addition of the score of 4. The median, however, has gone from 15.5 to 15, and the mean has gone from 15 to 14. These changes illustrate the differential effects of skew on the three measures of central tendency.
Professor Carlton will probably take the mode into account when assigning grades. He might decide to make the A cutoff at 18 because 17 was the most common score, and a lower cutoff would mean that an A would be the grade for the most common score. He will probably also take the median into account when trying to decide what he should consider the middle of the C range. Because of the skew of the data and the outlier, he may decide that the median is the better choice than the mean for the center of the distribution. (Then again, Professor Carlton may not grade on the curve and may pay no attention to these considerations! We are using this distribution simply to illustrate the behavior of kinds of averages, not professors.)

**Measures of Variability**

The second type of descriptive statistic is a measure of the variability of the data. Besides knowing the typical score, we generally want to know how much the data vary. Suppose you are a non-swimmer who wants to wade across a river that is two feet deep on the average. It makes a great deal of difference to you whether it is two feet deep all the way across, or whether it is 10 feet deep in places. Three general types of measures of variability are those based on the range, those based on percentiles, and those based on the mean.

**Range**

The range is the simplest measure of variability. It is simply the difference between the highest and lowest score in a distribution. In Professor Carlton’s class, the range was 18-10 = 8 when the first 10 students were considered. If a later student took the test and scored a 4, the range would become 18 – 4 = 14. Thus, although it is simple to compute, the range depends completely on the two extreme scores. For this reason it is highly unstable, as we can see. The mode and range can be thought of together, then, as descriptive statistics that depend completely on a few individual scores rather than on the entire distribution of scores.

**Percentile-Based Measures**

Although we defined the median as the middle score, we could have also defined it as the 50th percentile. A percentile is a score in a distribution below which a certain percentage of the cases fall. The 50th percentile is the score below which 50% of the cases fall. That score is by definition the middle score in the distribution, or the median.

We can use any arbitrary set of percentiles to describe the variability of scores around the median, but the interquartile range is generally used. This statistic is defined as the 75th percentile minus the 25th percentile. The 25th percentile is called the first quartile ($Q_1$), the 50th percentile is called the second quartile ($Q_2$), and so on.

$$\text{interquartile range} = 75\text{th percentile} - 25\text{th percentile} = Q_3 - Q_1$$

The interquartile range includes half the cases in a distribution. It has the advantage over the range that it is not affected by outliers.

A closely related measure of variability is the semi-interquartile range, which is simply half the interquartile range:

$$\text{semi-interquartile range} = (Q_3 - Q_1)/2$$
The advantage of percentile-based measures of variability is that they better represent skewed distributions than do other measures. They are more cumbersome to compute, however, and they do not enter into further statistical calculations.

**Variance and Standard Deviation**

The most commonly used measures of variability are the variance and the standard deviation. These measures are based on the mean. The variance, $\sigma^2$, is defined as the average of the squared deviations from the mean.

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{N}$$

Table A.2 shows how one would compute the variance. The first column contains all the individual scores. Below that column we see that the sum of all the scores is 150 and the mean is 15. The second column indicates how much each score deviates from the mean. Notice that the sum of these deviations is

<table>
<thead>
<tr>
<th>$X$</th>
<th>$(X - \bar{X})$</th>
<th>$(X - \bar{X})^2$</th>
<th>$X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>-1</td>
<td>1</td>
<td>256</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>9</td>
<td>144</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>196</td>
</tr>
<tr>
<td>17</td>
<td>-2</td>
<td>4</td>
<td>289</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>17</td>
<td>-2</td>
<td>4</td>
<td>289</td>
</tr>
<tr>
<td>18</td>
<td>-3</td>
<td>9</td>
<td>324</td>
</tr>
<tr>
<td>17</td>
<td>-2</td>
<td>4</td>
<td>289</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>169</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>225</td>
</tr>
</tbody>
</table>

$\Sigma X = 150 \quad \Sigma (X - \bar{X}) = 0 \quad \text{sum of squares} = 58 \quad 2308$

$$\bar{X} = \frac{150}{10} = 15$$

$$\sigma^2 = \frac{\Sigma X^2 - (\Sigma X)^2}{N} \quad \frac{2308 - 22500}{10} = \frac{22500}{10} \quad = 2308 - 2250 = 58 \quad \frac{22500}{10} = 10 \quad 58 \quad 10 \quad = 5.8$$
zero, as required by the definition of the mean. This tells you why we do not use the average deviation from the mean as a measure of variability: If the sum of the deviations from the mean is zero, the average deviation must also be zero.

The third column indicates the square of the deviation of each score from the mean. The square of a negative number is a positive number, so the sum of the squared deviations from the mean is greater than zero, and thus can be used as a measure of variability. The sum of the squared deviations from the mean is often called the sum of squares. The sum of squares is used to compute not only the variance but many other statistics as well.

The formula we have given for the variance is simple to define and is useful for conveying the idea of what the variance is, but it is somewhat cumbersome to compute because it requires one first to find the mean, then to subtract each score from the mean, and finally to square the deviation. Another formula, known as the computational formula, is more complicated looking but actually easier to use:

$$\sigma^2 = \frac{\sum X^2 - (\sum X)^2/N}{N}$$

The computational formula requires each score to be squared and then summed, as in the fourth column of Table A.2. The only other quantities required to compute the variance are the sum of the squares of all the scores, and the number of scores.

The variance is useful mainly because it enters into other statistical calculations, such as the analysis of variance (ANOVA). It has the disadvantage that it is not scaled in the same units as the original scores because it is expressed in terms of squared deviations from the mean. So the variance is analogous to a square foot, which cannot be used as a measure of distance. Fortunately, it is a simple matter to convert the variance into a measure of distance by taking the square root of the variance. The square root of the variance is known as the standard deviation.

The standard deviation (SD, or $\sigma$) has the advantage, as already mentioned, that it is related to other commonly used statistical procedures. The SD is widely used for this reason alone. Beyond that, the SD has the same advantages and disadvantages as the mean, on which it is based: SD represents all scores, but it is also affected by outliers.

**Choice of Measure of Variability**

The range is not very useful as a measure of variability because it depends completely on the two extreme scores. The interquartile range and the semi-interquartile range are useful in describing data when the median has been used as the measure of central tendency and when the data are skewed. The variance and standard deviation are the most widely used measures because they relate to the mean and other common statistics. Even when the data are skewed, the variance and standard deviation can be used after certain data transformations that reduce data skew.

**Correlation and Regression**

We said in Chapter 1 that one of the fundamental tasks of science is to establish that two variables are associated; for example, we may want to determine
whether grades on a test are related to time spent studying. Statisticians have developed techniques to measure the strength of a relationship between variables.

**Correlation**

The most commonly used measure of relationship between variables is the Pearson correlation coefficient, usually referred to simply as the correlation, or $r$. The correlation is expressed as a number that can take any value between +1.0 and −1.0. Figure A.1 shows five scattergrams, or scatterplots, depicting different sorts of correlations between two variables, $x$ and $y$.

In Figure A.1(a), there is a perfect correlation between $x$ and $y$. The value of $r$ is +1.00. Variable $y$ increases with increasing values of $x$, and for any given value of $x$, there is only one value of $y$. In other words, there is a perfect straight-line relationship between $x$ and $y$. This is an example of a linear function. An example of a correlation of 1.00 between two variables would be the relationship between weight in pounds and weight in kilograms. If you know someone’s weight in pounds, you can predict perfectly his weight in kilograms because one is a simple linear transformation of the other.

Figure A.1(b) shows another perfect correlation. There is a straight-line relationship between $x$ and $y$, but this time high values of $x$ are associated with low values of $y$. The correlation here is −1.00. Examples of perfect negative correlations are not common, but a trivial example would be the height of two ends of a seesaw.

Figure A.1(c) shows a more usual situation in which there is a correlation between $x$ and $y$, but the correlation is not perfect. High values of $x$ tend to be associated with high values of $y$, but the data are scattered instead of

![Scattergrams of paired values of $x$ and $y$.](image)

**FIGURE A.1**

Scattergram of paired values of $x$ and $y$: (a) $r = +1.00$, (b) $r = -1.00$, (c) $r = 0.50$, (d) $r = 0$, and (e) $r = 0$. 
falling exactly on a straight line. The correlation in this panel is about 0.5. There are many examples of such a correlation: height and weight of people, grades and time spent studying, education and income, and so forth.

Figure A.1(d) shows a situation in which there is no correlation between $x$ and $y$. A value of $x$ can be associated with any value of $y$, and vice versa. The correlation between the two variables is 0.00. An example of a zero correlation is eye color and income. People with each eye color are equally likely to be rich, poor, or in between.

It is common to describe various correlation coefficients by adjectives to indicate the strength of the relationship they represent. Correlations less than 0.2 are considered weak; 0.2 to 0.4, moderately weak; 0.4 to 0.6, moderate; 0.6 to 0.8, moderately strong; and 0.8 to 1.0, strong.

It is important to note that the Pearson correlation coefficient is a measure of a linear (straight-line) function. It is entirely possible that there may be a close relation between two variables but the relation does not fit a straight line. If the data actually fit a curved line very closely, a straight line cannot make a good fit. As you can see, the data in Figure A.1(c) fit a curvilinear function closely. The correlation here is 0.00 because both low and high values of $x$ are associated with the same values of $y$. Thus, if the data actually fit a curvilinear function, the correlation coefficient, $r$, will underestimate the amount of the relationship between the variables. There are other measures of correlation appropriate for curvilinear data.

**Regression**

An important property of the correlation coefficient is that it measures how well you can predict the value of one variable when you know the value of the other. When the correlation is 1.00, prediction is perfect: If you know the temperature outside in degrees Celsius, you can predict the temperature in degrees Fahrenheit perfectly. When the correlation is 0.00, prediction is impossible: If you know a person’s eye color, your prediction of her income will be no better than if you did not have that information.

**Regression** is the technical term for the process of predicting the value of one variable from another. When we predict $y$ from $x$, we use the familiar equation for a straight line:

$$y' = mx + b$$

The $y$ has an apostrophe next to it and is read “$y$ prime” to indicate that we are predicting $y$ from $x$. The $m$ is the slope of the line relating $y'$ to $x$. The value of $m$ depends on two things. The first is the correlation coefficient, $r$. Recall that if the correlation is perfect, the value of $r$ is 1.00, and if there is no correlation, $r$ is 0.00. Under certain conditions, $r$ is the slope of the regression line predicting $y'$ from $x$, and hence $r = m$. Return to Figure A.1 for a moment. Notice that the value of $r$ is the same as the slope of the line drawn predicting $y'$ from $x$ in each case.

Now, what else goes into the slope of the regression line besides the value of $r$? Recall that we said there is a perfect correlation between temperature in degrees Celsius and temperature in degrees Fahrenheit because one is a simple linear function of the other:

$$F = 9/5(C) + 32$$
But the regression equation between the two, shown here, has a slope of 9/5 rather than a slope of 1.00. The difference in slope is the result of the differing scales of measurement of the two variables. In fact, scale of measurement is all that differs between temperature in Fahrenheit and temperature in Celsius; otherwise the two are identical. So the slope of the regression line is a quantity that reflects both the correlation between the two variables and the scale of measurement of the two variables. The scale on which the two variables are measured is their variability, or standard deviation.

There is a mathematical procedure for determining the regression line, with which we will not concern ourselves here. (It essentially involves computing the correlation coefficient.) The important thing to note is that when the correlation is perfect, there will be a slope of 1.00 between the two variables (when scaled by their variability), and when there is no correlation, there will be a slope of 0.00. Values of $r$ between 0.00 and 1.00 indicate differing slopes of the line predicting $y$ from $x$.

The situation in which the correlation is 0.00 is very instructive. When there is no correlation between two variables, knowing the value of one does not help you to predict the other. If you know a person’s eye color and want to predict his income, your best bet is to guess the mean income of the population. Your best guess would be the same number whether the person’s eyes are blue, brown, green, or hazel. This simply puts into words the significance of a slope of 0.00 in the line predicting $y$ from $x$: Always predict the same value for $y$ no matter what the value of $x$.

Most of the time you will not see scattergrams in which the two variables are scaled by their variability; they are usually plotted by the units in which they are measured: IQ, GPA, cm, and so forth. Then the slope of the line predicting $y$ and $x$ will not be equal to the correlation coefficient.

**Example of Correlation and Regression**

Consider a group of 20 students who have taken the SAT and then attended college. Table A.3 shows their scores on the SAT and their first-year grade-point averages (GPAs). Figure A.2 shows the scattergram of their SAT scores plotted against their GPAs.

We find a correlation of 0.74 between SAT score and GPA. This is a fairly high correlation, and it suggests that we are justified in using SAT scores to select students for college because the SAT score allows us to predict the first-year GPA. The regression line predicting GPA from SAT is

$$\text{GPA}' = 0.0022(\text{SAT}) + 0.61$$

This equation predicts that the GPA will increase by 0.0022 for every point increase in SAT score. (The difference between the slope of 0.0022 and the correlation coefficient of 0.74 is accounted for by the difference in the range of GPAs and SAT scores: GPAs range from 1.9 to 3.8, whereas SAT scores range from 750 to 1350.)

According to this equation, a student who has an SAT score of 1400 would be predicted to achieve a GPA of 3.7. A student with an SAT score of 600 would be predicted to achieve a GPA of 1.9, and a student with an SAT score of 0 would be predicted to earn a GPA of 0.61. (Actually, it is impossible to have an SAT score of 0, so, strictly speaking, this prediction is outside...
the range for which the equation is valid.) Colleges use this information to select students. They can use SAT scores to predict which applicants are likely to do better than others.

**Variance Accounted For**

An important property of the correlation coefficient is that by squaring it we obtain a measure of the proportion of the variability in y that is accounted for by x. Now, there is a certain amount of variability in the scores on variable y and also on variable x. When there is a correlation between x and y, we can predict the value of y when we know x. Another way of saying this is that some proportion of the variability of y can be explained by the effect of x. If the correlation is 1.00, we have accounted for all the variability in y when we know x because the square of 1.00 is 1.00. If the correlation is 0.00, we have accounted for none of the variability in y by knowing x because the square of 0.00 is 0.00.

When the correlation is other than 1.00 or 0.00, the proportion of variance accounted for is less than the value of the correlation. A correlation of
0.5 accounts for only 0.25 of the variance because the square of 0.5 is 0.25. To account for half the variance, you need a correlation of 0.71, because 0.71 squared is 0.5.

This concept is important to remember because researchers sometimes are impressed when they find a correlation of 0.5 in their data. They need to remember that such a correlation accounts for only one-fourth of the variability in their data. Three-fourths of the variability in \( y \) is not associated with \( x \).

The square of the correlation coefficient, \( r^2 \), is sometimes considered a measure of the goodness of fit of the data to the regression line. As \( r^2 \) approaches 1.00, the data fit the regression line better and better.

**Effect of Truncation of Range**

The size of the correlation coefficient is sensitive to the range of the variables measured. If you measure the correlation between two variables for a set of data that cover only part of the range over which the data vary, the correlation will be seriously underestimated. It is possible to think of the correlation as a measure of the pattern formed by the points around the line that relates the two variables. When the correlation is 1.00, there is no scatter and the data points form a straight line. When the correlation is 0.00, the data points make a completely unpredictable pattern. Now, when the correlation is somewhere in between these extremes, the points will form a sort of hot dog-shaped pattern. You can see from Figure A.1(c) that if you were to measure only the higher values of \( x \), the pattern of \( y \) scores would be rounder than if you measured the entire range of \( x \); that is, the shorter the pattern, the less it would look like a good fit to a straight line. Therefore, the correlation between \( x \) and \( y \) drops when you consider only part of the range of \( x \).
An example of this situation is given by the earlier case of the college that uses SAT scores as a basis for selecting students. Suppose there is a correlation of 0.74 between SAT score and first-year grade-point average (GPA). This correlation exists when you consider all students who might apply to and attend this college. The college, however, has more applicants than it can admit, and it naturally wants to admit the best students it can. Suppose the college accepts only students who score at least 1000 on the SAT. If we consider the correlation between SAT score and GPA only for students who score at least 1000 on the SAT, we find that the correlation drops to 0.42. This is not an artifact of the way we have selected the particular data because the correlation for those who fall below 1000 on the SAT is 0.41.

The drop in the correlation when the range is truncated is accompanied by a corresponding decrease in the slope of the regression line. The dashed lines in Figure A.2 show the regression lines for the two halves of the data considered separately. When the college accepts only students who score at least 1000 on the SAT, the slope of the regression line drops from 0.0022 to 0.0014. Using the regression line based only on those students who scored 1000 or better on the SAT, we would predict that a student scoring 1000 would achieve a GPA of 3.0 (instead of 2.8, based on the full range) and one who scored 1400 would achieve 3.5 (instead of 3.7). Thus, the college’s ability to predict GPAs from SAT scores decreases when the range is truncated.

The problem is that the college cannot admit students over the whole range just to determine the true correlation for all students who might be admitted. There isn’t room for them, and the students with lower SAT scores will tend to do less well. But, if the college uses SAT scores as a basis for admitting students and tries to assess the ability of the SAT score to predict GPA, it will be underestimating the degree of the relationship between SAT score and GPA. The college may conclude that SAT score is not a useful predictor of college success because its correlation of 0.42 accounts for only 18% of the variability in college grades. In fact, if the college had admitted all students who took the SAT, it would have found that there was a correlation of 0.74 between SAT scores and grades and that SAT scores accounted for 55% of the variability in college grades.

INFERENTIAL STATISTICS

Now we turn to techniques for allowing us to draw inferences about the population from a sample drawn from the population.

Sampling Distributions

Suppose you knew that the mean IQ of the population was 100 and its standard deviation was 15. If you were to select a person at random and had to guess his or her IQ, you would probably guess that it was 100 because that is the mean of the population. You would not be greatly surprised, however, if that randomly selected person had an IQ of 70 or 130. After all, one person out of a population might have any score in the population. If you had a group of 100 randomly selected people, however, you would be very surprised if the mean IQ of the group was 70 or 130. If it was 70, you would probably guess that you had somehow selected a class of slow learners instead of a random sample from the population;
if the mean was 130, you would likely think you had gotten the members of an honors class. This intuitive notion is related to the concept of a sampling distribution.

We know from descriptive statistics that we can describe a population by its mean and standard deviation. Now, suppose we take samples from the population and measure IQ. These samples of IQ will vary from one to another. See Figure A.3, in which we have taken three samples of size 5 from the population.

We can do various things with these samples. First, we can find the means of the samples. Our three samples have means of 92, 99, and 104.

![Image of IQ distribution and sample means](image)

**Figure A.3**

Relationship of population distribution, sample distributions, and sampling distributions.
We can also make a distribution of the means of the samples. This is a new
distribution, a distribution of sample means, and it is not to be confused
with the original population distribution. This distribution of the means of
samples from a population is called a sampling distribution.

The sampling distribution has three important properties. First, it has the
same mean as the original distribution. If the mean IQ in the population is
100, the mean of the sampling distribution will be 100. (The average sample
you select will have a mean of 100.)

Second, the sampling distribution has a smaller standard deviation than
the population distribution. The larger the size of the samples that are drawn
from the population, the smaller the standard deviation of the sampling dis-
tribution. As noted previously, the larger the sample you draw from a popula-
tion, the more you expect its mean to be close to the population mean. The
standard deviation of the sampling distribution is called the standard error of
the mean. The standard error of the mean is the standard deviation of the
population divided by the square root of the sample size:

$$\sigma_n = \frac{\sigma}{\sqrt{N}}$$

The standard error of the mean is thus inversely proportional to the
square root of the sample size: The larger the square root of the sample size,
the smaller the standard error of the mean. This is true because the square
root of $N$ is the denominator of a fraction. The larger the denominator of a
fraction is, the smaller the value of the fraction is. Therefore, increasing the
square root of $N$ by a factor of 2 cuts the standard error of the mean in half. Put-
ting it another way, large samples result in a smaller standard error of the mean
because the means of large samples are more similar to one another. The stan-
dard error of the mean of the sampling distribution in Figure A.3 is 6.7.

The third characteristic of the sampling distribution is that, as the sample
size becomes larger, the shape of the distribution approaches a normal dis-
tribution, regardless of the shape of the population from which the samples are
drawn. The population distribution in Figure A.3 is normal because the distri-
bution of IQ is normal. However, the sampling distribution will be normal no
matter what the original population looks like. Exercise A.1 on pages 422–423
illustrates the concept of sampling distributions and shows the effects of chang-
ing sample size. Turn there now to see the effect of sample size on the shape
and variability of sampling distributions.

**Testing Hypotheses**

We said in Chapter 1 that virtually all scientific research has the purpose of test-
ing a hypothesis. Usually this is a test of a theory-level hypothesis. But when we
test such a hypothesis, we must state it as a particular research hypothesis. The
research hypothesis is more specific than the theory-level hypothesis because it
must be stated in terms of the particular way the study was carried out.

In the Simcock and Hayne study discussed in Chapter 1, the theory-level
hypothesis was that the absence of verbal ability in young children prevents
them from coding their early memories in a form that can be remembered
later using language. The theory predicts that children will not be able to re-
member events for which they lacked the vocabulary required to describe
them. The research-level hypothesis was that particular children at particular ages would not remember particular experiences when tested in a particular way. But there may be many reasons why children may not remember events. For this reason the experiment has to be designed to take these other factors into account.

The difference between the theory-level hypothesis and the research hypothesis is that the research hypothesis must take into account other factors that the theory does not address, such things as whether the children played enough with the machine to remember it, or whether they were too tired or hungry to pay attention to the experience. Whether the research hypothesis does justice to the theory-level hypothesis was the subject of Chapter 7.

When we have collected the data in a study, we need to analyze them statistically, to see which unobserved population they are most likely to have come from. (The logic of hypothesis testing is the same no matter what statistic you use. The arithmetic procedure differs with the statistic—t test, analysis of variance, and so forth. We refer you to a statistics book for the details.) The statistical hypothesis is actually a restating of the research hypothesis into two different hypotheses. The first one is just the research hypothesis itself, but we are going to introduce a new name for it: the alternative hypothesis. It is generally written \( H_1 \) and is called "\( H \) sub one." The reason for this term will become clear in a moment. In our example, the alternative hypothesis is as follows: Children who had the vocabulary to describe events will recall them better than children who lacked the vocabulary.

The second hypothesis is the one that would be true if the alternative hypothesis were false. We call this one the null hypothesis. It is often written \( H_0 \) and is called "\( H \) sub oh." It is called the null hypothesis because it is an empty hypothesis, of no scientific interest to you. You set up the null hypothesis strictly for the purpose of rejecting it. It is a "straw man" hypothesis. In our example, the null hypothesis is as follows: Children who had the vocabulary to describe events will not recall them better than children who lacked the vocabulary.

Notice that the null hypothesis covers all possible exceptions to the research hypothesis. We could have stated the null hypothesis as follows: Children who had the vocabulary to describe events will recall them less well than children who lacked the vocabulary, or they will recall them equally well. In other words, either the null hypothesis is true, or the alternative hypothesis is true. Between the null hypothesis and the alternative hypothesis, you have covered all possible states of the world. A philosopher would say that the two alternatives are mutually exclusive and exhaustive.

The hypothesis that we are trying to prove is the alternative hypothesis. (Some people object to the phrase proving the alternative hypothesis because it implies that you have made the correct decision when, in fact, your decision may be incorrect. We use the expression here anyway because we believe that everyone should realize that science never proves anything once and for all.) We prove this hypothesis to be true by using a roundabout method of disproving the null hypothesis. If we have disproved the hypothesis that includes all possible outcomes that could happen if the research hypothesis were false, then the research hypothesis is left standing.

Now we know the reason for calling the research hypothesis the alternative hypothesis: The logic of the test is set up so that you try to reject the null
hypothesis. When you have done that, all there is left, if you have set up the hypotheses properly, is the alternative hypothesis. So the logic of hypothesis testing is to set up a straw man. When you have knocked it down, you have proven your research hypothesis. Here is a tricky but extremely important point. Your alternative hypothesis is the one that you want to be true. However, according to the logic of the statistical test, you cannot do this directly. You can only reject the null hypothesis, which leaves you with only one alternative: to accept the alternative hypothesis. If it turns out that your results are not statistically significant, then you fail to reject the null hypothesis. Thus, you must say that your alternative hypothesis was not accepted because you could not reject the null hypothesis.

The null hypothesis is often stated in the following way:

\[ H_0 : \mu_{HV} \leq \mu_{LV} \]

That is, the null hypothesis is that the mean of the population of those children who have the vocabulary (HV) is less than or equal to the mean of those who lack the vocabulary (LV). The alternative hypothesis is then stated as follows:

\[ H_1 : \mu_{HV} > \mu_{LV} \]

The alternative hypothesis is that the mean of the population of those who have the vocabulary is greater than the mean of those who lack the vocabulary.

In this example, the alternative hypothesis was that the experimental population had a higher mean than the controls. This is called a directional hypothesis because we predicted that the HV children would differ in one particular direction from the LV children. This gives rise to what is called a one-tailed hypothesis test. We are not interested in the case in which the HV children might be worse than the LV children in recall.

Sometimes we predict only that the two groups will differ from each other; we don’t know which group will be higher. This is a nondirectional hypothesis, and it gives rise to a two-tailed hypothesis test. The null and alternative hypotheses in this case would be stated as follows:

\[ H_0 : \mu_X = \mu_C \]
\[ H_1 : \mu_X \neq \mu_C \]

That is, the null hypothesis is that the mean of \( X \) equals the mean of \( C \), and the alternative hypothesis is that the mean of \( X \) does not equal the mean of \( C \). We would use a two-tailed hypothesis if we predicted that the rewarded children would differ from the controls but our theory did not predict in which direction. When you do a two-tailed hypothesis test, you reject the null hypothesis if the experimental group is sufficiently higher or sufficiently lower than the control group.

**Dealing with Uncertainty in Hypothesis Testing**

One consequence of the fact that data are inherently variable is that we must be prepared to deal with uncertainty in making decisions about those data. Suppose the psychology department at your college wants to evaluate the effectiveness of its undergraduate program. To do this, the department requires
a sample of majors to take the psychology section of the Graduate Record Exam (GRE). We know that the mean score on the GRE is 500 for all students who take the exam. Suppose that the mean for the psychology majors at your college is 530. Does this prove that your college's psychology graduates are reliably better than the national average?

The fact that the group scored higher than the national average does not prove that the college's students are better than the national average. Although it may be that the college's program was responsible for the difference, it is not certain. The students in this year's graduating class may have been better than usual, or they may have been lucky on the test, or any of a number of other factors may have been responsible. So, we are faced with a difference between our group and the population mean that we must interpret: Does an average score this high prove that your college's psychology majors are actually better than the national average? This average score will have its own sampling distribution, a sampling distribution of means. Our job is to decide whether the observed average score is likely to have come from the distribution of all possible samples of means from the population of scores on the GRE. So we need to make a decision based on this observed sample mean compared with the theoretical sampling distribution.

**Type I and Type II Errors**

Suppose that your college's psychology majors were actually the same as the average of the population of psychology majors. What would you expect to happen? This is the same as asking what the null hypothesis is. If we are studying the effectiveness of your college's psychology program (C) compared with that of the average psychology program (X), we would say that our null hypothesis is that your college's psychology majors are no better than average:

\[ H_0 : \mu_C \leq \mu_X \]

Our alternative hypothesis is that your college's psychology majors are better than average:

\[ H_1 : \mu_C > \mu_X \]

These two hypotheses set up two possible states of the world: Either the mean of your college's psychology majors is better than the average for all psychology majors in the country, or it is not. Note that the two states of the world concern not the outcome of your single study, but the outcome of all possible studies using exactly the same methods, type of students, and so on, that you did. In other words, the two states of the world are statements about whether your experimental hypothesis is true or false.

Corresponding to the two states of the world are two possible decisions you could make: Either your college's psychology majors are better, or they are not. These two sets of alternatives make it possible to consider four distinct situations. First, let us suppose that your college's psychology program is better. You might decide that it is better, or you might decide that it is not better. Second, suppose instead that the psychology program is not better. Here you could also decide that it is better or that it is not better. It is convenient to summarize these four possibilities in a 2 x 2 table (see Table A.4).
TABLE A.4
Four Possible Outcomes in Making a Decision Concerning Rejection of the Null Hypothesis

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null hypothesis is true</th>
<th>Alternative hypothesis is true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept null hypothesis</td>
<td>Correct decision</td>
<td>Wrong decision: Type II error</td>
</tr>
<tr>
<td>Reject null hypothesis</td>
<td>Wrong decision: Type I error</td>
<td>Correct decision (power of test)</td>
</tr>
</tbody>
</table>

FIGURE A.4
Four possible outcomes of decision making concerning the null hypothesis.

Looking at the table, you can see that if the program was better and you decided that it was better, you would be correct. You would also be correct if the program was not better and you decided that it was not better. Figure A.4 provides the same information in the form of a decision tree.

There are two ways of making errors. One kind of error occurs when the program is not better but you decide that it is. This is called a Type I error; you have rejected the null hypothesis when it is true. The other kind of error is to decide that the program is not better when it actually is. This is called a Type II error; you have accepted the null hypothesis when it is false.

**Alpha and Statistical Significance**

These four outcomes can be related to the sampling distributions of the means for the two situations. First, let us consider the sampling distribution of the means for the psychology majors when the null hypothesis is true.

The curve in Figure A.5(a) shows the relative probability of getting any particular group mean. If the null hypothesis is true, the distribution of means will have a mean of 500 and a certain standard deviation (standard error). We can see that sometimes the mean will be higher than 500 and sometimes it will be less than 500.
FIGURE A.5
Power of the test of the null hypothesis against the alternative hypothesis.

Figure A.5(b) shows the distribution of means when the alternative hypothesis is true (when the program is better). Notice that the distribution of differences has a mean that is 10 points higher. Most of the time the difference is greater than zero, but sometimes it is zero or less. What we need to do is set up a criterion on which to base our decision.

The way this is done is to decide how often we are willing to say that the program is better when in fact it is not. This is the same as setting a cutoff on the dimension of scores in Figure A.5(a) that will cause us to reject the null hypothesis a certain percentage of the time when it is true. This point is indicated by the vertical line that cuts through both curves in Figure A.5. The line divides the curve indicating the sampling distribution given the null hypothesis into two sections. The area under the curve represents the probability of various events given that the null hypothesis is true. The area to the right of the line cuts off a certain proportion of the curve. Suppose we decide to say that the null hypothesis is false and the alternative hypothesis is true whenever the mean score is greater than a certain amount, indicated by the vertical line. The section to the right of the line on the upper curve is the probability of deciding that the null hypothesis is false when in fact it is true. This is the probability of making a Type I error and is known as alpha. Alpha is the probability of deciding that the null hypothesis is false when it is actually true.

Usually, scientists prefer to make alpha a fairly small number, such as 0.05 or 0.01. The reason is that scientists believe that to decide that an experimental finding is true when it is not is a more serious error than it is to miss a true finding.

Alpha is also called the level of significance of an effect, or the statistical significance. It is common to say that a certain experimental result was significant at the 0.05 level. This means that the effect was large enough that the probability that it happened purely by chance was 0.05, or 1 in 20.
The Significance of Significance

Let's suppose we have found that left-handed people with hazel eyes and free earlobes are better at repeating the Pledge of Allegiance backward and that the effect is significant at the 0.05 level. What does that mean? First, there is something that it does not mean: It does not mean that the results are important. Although in common usage that is exactly what the term significance means, results can be statistically significant when they have no importance at all. Results may be statistically significant even when an experiment has to do with the most trivial question imaginable, as in our little example.

Statistical significance means simply that your results have a certain low probability of having been the result of chance. When a result is significant at the 0.05 level, there is a 0.05 probability, or 1 in 20, that the result occurred when the null hypothesis was true—that is, when there was actually no effect. The level of significance is the same as alpha, or the probability of a Type I error.

There is a second thing statistical significance does not mean: It does not mean that the effect of the independent variable was large. The reason is simple and comes from the fact that the level of significance of an effect depends partly on the number of observations on which the test was based. For a given effect size, the more observations there are, the greater the significance will be. Recall our discussion in Chapter 13 of research methods that focus on big effects. Suppose there were a true difference of 1 mm in the average height of students at two colleges. If you compared a sample of 10 students from each college you would not expect to see a statistically significant difference between the groups. The larger the samples you take, however, the more you would expect to find a significant difference. In fact, if you measured every single student in the two colleges, any difference at all would be significant by definition. The reason is that when you have measured the entire populations, there is zero probability that they do not differ (assuming no error in measuring their heights). For these reasons, the APA has recommended that hypothesis tests be supplemented by measures of effect size to help the reader interpret the results (Wilkinson, 1999).

Effect Size

As we have just seen, the sample size and the effect size both influence the statistical significance of a result. Let us take a closer look at effect size. Effect size is a straightforward concept; it is defined as the strength of the relationship between the independent and dependent variables in a study. One common measure of effect size that we have discussed already is the correlation coefficient, which can vary from \(-1.0\) through \(0\) to \(+1.0\). A zero correlation indicates no relationships between variables; positive or negative 1.0 indicates a perfect relationship. Correlations less than 0.2 are considered weak; 0.2 to 0.4, moderately weak; 0.4 to 0.6, moderate; 0.6 to 0.8, moderately strong; and 0.8 to 1.0, strong.

The distinction between effect size and significance is illustrated by the fact that any of these correlations can be significant or insignificant depending on the sample size. A correlation of 0.5 can be nonsignificant if it is based on too few observations. Statistics books provide tables showing the level of
significance of correlations of different sample sizes. Computerized statistical packages routinely print the significance of correlations.

For some studies, effect size can be measured by meaningful units at a practical level: Group 1 ate five grams more, or solved six more problems, or smoked seven fewer cigarettes per day than Group 2 did. Measures used in other studies may be harder to interpret. For these, effect size is commonly measured as the difference between the means of the groups, measured in terms of the variability within the groups. One widely used such measure is Cohen’s $d$, which is defined for the $t$ test as:

$$
Cohen's \ d = \frac{\bar{X}_1 - \bar{X}_2}{\sigma}
$$

where $\sigma$ is the (pooled estimate of) population standard deviation. When there is a lot of variability in the data, Cohen’s $d$ will be small; when variability is low, Cohen’s $d$ is large.

Other measures of effect size can be computed for many statistics, such as $F$. Further information can be found in most statistics books.

**Power**

Looking now at Figure A.5(b), we find that the same criterion has divided the curve that represents the situation given the alternative hypothesis into two portions. The portion to the left is the probability of deciding that the null hypothesis is true when it is actually false. This is the probability of a Type II error.

The rest of the area under the curve showing the alternative hypothesis shows us another important probability: the probability of rejecting the null hypothesis when it is actually false. This is one minus the probability of a Type II error and is known as the power of the test. This is a very important probability: the probability of deciding that you have an experimental effect when you actually do.

Notice something about the two curves. You could decide to make your alpha, or probability of a Type I error, smaller by moving the line dividing the two curves to the right. However, by decreasing alpha, you would also decrease the probability of accepting the alternative hypothesis when it is true. In other words, you would decrease the power of the test.

Three things influence the power of a test. The first is the value of alpha. The smaller your alpha level, the smaller your power. If you decide that you want to make it unlikely that you will make a Type I error, you must accept the fact that you will be more likely to make a Type II error.

The second thing that influences the power of a test is the size of your experimental effect. If you are able to make your psychology program much more effective—by giving the students more courses, better materials, higher motivation, or other advantages—you will increase their mean score. This will have the effect of moving the curve showing the alternative hypothesis to the right, making less overlap between the curves.

The third way to increase the power of a test is to increase the size of the two groups. This will have the effect of decreasing the variability of the sampling distribution, or the variability of the two curves in Figure A.5. This will also reduce the amount of overlap between them. Exercise A.1 is an illustration of this point.
Chi-Square

Both in everyday life and in the course of experimentation, situations often arise in which it is important to make decisions about whether one thing has happened more often (or more frequently) than another thing. The Chi-Square statistic \( (\chi^2) \) helps us to make these decisions about two or more categorical variables (see Chapter 5 for more information on variables). For example, imagine that a corporation is concerned about the color of a new plastic toy product, and is torn between making it in pink or in blue. So, the consumer psychologists at the corporation recruit a wide variety of children, show them the toy in both colors, and ask them to choose the one that they like best. Even though the toys are identical in every way except color, blue wins the contest by quite a bit. Is it possible that because of cultural stereotypes, the boys might have shown a reluctance to accept a pink toy? Because our question is about the frequency with which children choose the colored toys, the Chi-Square statistic is appropriate. Chi-Square wouldn’t be able to answer questions about how long the children played with the toys (duration), but it would be able to evaluate our frequency question well. In order to ask the question in our example more specifically, though, one might rephrase it for testing as follows: Is it possible that significantly more boys would vote for the blue toy than girls?

The Chi-Square enables us to decide whether a relationship exists between categorical variables, which in our example are votes for the color of the blue toy and sex of the child who voted. The Chi-Square statistic works on the expectation that two variables with equal probabilities should yield equal results if they are not related. So, in the case of our toy, if cultural stereotypes were not playing a role, we would expect that equal numbers of boys and girls should have voted in roughly equal numbers for the blue toy. Thus, according to the null hypothesis \( (H_0) \), our expected frequencies for these preferences should be roughly similar for boys and girls. However, the real question is whether the votes that we actually obtained for the toy preferences are different from our expectations. In other words, do our observed frequencies differ from the values we expected? Essentially, when the votes are actually counted, did more boys vote for the blue toy than girls?

To examine the voting habits of boys and girls in this example with Chi-Square, a \( 2 \times 2 \) contingency table must be created. First, we have to figure out our expected frequencies. If we know that the blue toy received 70% of the 200 votes cast, and also know that boys and girls voted in equal numbers, the expected frequency table will look like Table A.5.

We also have to examine the way that the children voted by also placing our observed frequencies into the contingency table (Table A.6).

Once we have both expected and observed frequencies, we need to examine the difference between them. We make this comparison initially by obtaining the difference between the observed and expected votes (Table A.7).

It is clear from these numbers that the blue toy received more votes from boys than girls, but it is not clear as to whether these additional votes are likely to have occurred by chance. So, how do we know whether this difference would be likely to occur in a random sample of voters even if there were no difference at all within the full population? Computing the
TABLE A.5
Expected Frequencies (from null hypothesis)

<table>
<thead>
<tr>
<th>Sex of Child</th>
<th>Votes for Blue Toy</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Girls</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>140</td>
<td>60</td>
</tr>
</tbody>
</table>

TABLE A.6
Observed Frequencies (sample results)

<table>
<thead>
<tr>
<th>Sex of Child</th>
<th>Votes for Blue Toy</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Girls</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>140</td>
<td>60</td>
</tr>
</tbody>
</table>

TABLE A.7
Observed minus Expected Frequencies

<table>
<thead>
<tr>
<th>Sex of Child</th>
<th>Votes for the Blue Toy</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boys</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80–70= 10</td>
<td>20–30= –10</td>
</tr>
<tr>
<td></td>
<td>Girls</td>
<td>60–70= –10</td>
<td>40–30= 10</td>
</tr>
</tbody>
</table>

Chi-Square statistic allows us to determine whether the sex of the voter is associated with a vote for the blue toy. The general formula is:

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

where \( O \) = observed frequency and \( E \) = expected frequency in each cell of the table. (Caution: \( \chi^2 \) isn’t accurate if any expected value is less than five.) So, each of the differences between the observed and the expected (as in Table A.7) is squared, then divided by the expected frequency, and finally added to the results from the other cells. To get an intuitive feel for this statistic, you might notice that the greater the difference between what is observed and what is expected, the larger the Chi-Square value. Statistical tables are available in which the significance level of any Chi-Square value can be found, though most statistics software will output both a value and its significance level. Since the present example (with one degree of freedom) yielded a Chi-Square value of 5.74, the tables show that it is significant at the 0.02 level. This low level shows that it is highly likely that a voter’s sex influenced whether or not they voted for the blue toy.
Analysis of Variance

Analysis of variance (ANOVA) is a powerful statistical method for analyzing experimental data. It is a flexible but complex method that can be adapted to a great variety of experimental designs. These characteristics make ANOVA difficult to summarize and to understand. Because ANOVA is so widely used, however, it is necessary to have an acquaintance with some of its basic concepts.

When one wants to analyze the data of an experiment with only two conditions, a *t* test is appropriate. Whenever there are more than two conditions, however, a *t* test is not appropriate because a *t* test can compare only two groups at a time. If there were three conditions, it would be necessary to compare the first with the second, the first with the third, and the second with the third. This would result in three separate statistical tests. Performing several tests would give you more chances to reject a true null hypothesis, so you would increase the likelihood of rejecting the null hypothesis when it is actually true.

ANOVA was developed to make it possible to test the null hypothesis that there is no difference among a number of conditions. Suppose that an experiment tested the effect of three different doses of marijuana on motor coordination. The null hypothesis would be that there is no difference among the three doses:

\[ H_0 : \mu_1 = \mu_2 = \mu_3 \]

The alternative hypothesis is that:

\[ H_1 : [\mu_1 \neq \mu_2 \neq \mu_3] \]

is not true. Note that the alternative hypothesis simply says that it is not true that all the means are equal. It does not say that all the means are unequal. For the alternative hypothesis to be true, it is necessary only that some combination of means not be equal to some other combination. For example, one situation that would make the null hypothesis false and the alternative hypothesis true could be that Conditions 1 and 2 differ significantly from Condition 3 in a three-condition experiment. It is entirely possible that neither Condition 1 nor Condition 2, considered alone, differs significantly from Condition 3, but that the subjects in conditions 1 and 2 considered together differ significantly from those in Condition 3.

Any experiment that has more than two conditions must be analyzed by ANOVA instead of a *t* test. Very simply, any ANOVA tests the significance of a difference among several conditions in an experiment by making two different estimates of the variability that you would expect to find in the data given that the null hypothesis is true. If there is no true difference among the groups, each estimate of variability should, on the average, be the same. Let us see where those two estimates of variability come from.

Two Estimates of the Variability in the Population

Suppose that the null hypothesis is true. In that case, because the experimental conditions have no effect, you would predict that all the individual data would have the same value. This value would be the mean of all the data, or the *grand mean*. Now, it should be obvious that all the data would not be exactly the same, because chance factors are nearly always operating to introduce variability to the data. Therefore, individual data points will vary about the mean of their respective groups. Just as the individuals will differ within a group, the means of the
data from the various groups will not have the same value, for the same reason. Thus, the group means will vary about the overall mean in a random way.

We can think of the variability of the subjects within each group as one estimate of the variability in the population, and the variability of the means of the groups as another estimate of the variability of the population. Thus, we have two different ways in which chance will affect the data in an experiment when there is no experimental effect (and the null hypothesis is therefore true): by producing differences between the means of the various conditions and by producing differences among the subjects within the various groups. These two ways in which chance affects the data that provide the basis for doing ANOVA.

Figure A.6 represents the data from a hypothetical experiment in which motor coordination was measured in three different groups after the subjects experienced three different doses of marijuana. Figure A.6(a) represents the data from all the subjects in the experiment. We see that there was some variability among the subjects, but we don’t know how the variability relates to the conditions. In Figure A.6(b), we see one possible situation. All the subjects in a particular group had exactly the same data, and each of these groups differed from the others. Here there is variability among the groups but no variability within the groups. Figure A.6(c) shows another way in which the variability in the experiment could be distributed. Here all the groups have exactly the same mean, and all the variability is within the groups. This example is admittedly highly artificial. Ordinarily, of course, there would be some variability within, and some between, the groups. In our example, there was either no variability within the groups or no variability among the groups. The point, however, is that we have two ways of estimating the variability in the population: one based on the variability within the groups and the other based on the variability between the groups.

The differences among the means of the groups and the differences among the subjects within the groups give us two separate estimates of the variability in the population. Sometimes one estimate will be larger, and sometimes the other will be larger. On the average, however, they should be equal if the independent variable in fact had no effect. Consider a ratio of the between-conditions variability to the within-conditions variability:

\[
\text{between-conditions variability/within-conditions variability}
\]

This ratio is called F. The value of F will be 1.0, on the average.

The term between groups can lead to a misunderstanding if one understands it to refer to two groups, rather than more than two. As discussed earlier, the null hypothesis in ANOVA is that all the groups are equal. Recall that no two groups need be significantly different from each other for this hypothesis to be false (and the alternative hypothesis to be true). It simply means that some group or combination of groups differs significantly from the rest.

If the null hypothesis is false and there is an experimental effect, the between-conditions variability will be larger than expected because the variability caused by the experimental effect will add to the random variability. Then the value of the F ratio will be greater than 1.0. The F ratio is central to ANOVA. Every ANOVA has at least one F ratio. The value of F obtained in the study is evaluated statistically against the value that would be expected to occur by chance alone. If the F ratio is larger than a certain value, the experimental effect is considered to be statistically significant.
Partitioning the Variance

We need to consider one more concept before discussing how to read an ANOVA summary table. The reason this technique is called *analysis of variance* is that ANOVA makes it possible to analyze all the sources of variability in an experiment. In other words, a set of data contains a certain amount of variability. Some of this variability comes from variability among the subjects: how people differ from one another regardless of the experimental conditions. Other variability in the data is caused by the experimental conditions: The independent variable caused the subjects to behave differently.

In a simple ANOVA, there are two sources of variance: between-groups variance and within-groups variance. In Figure A.6(a), we see a certain amount of
variability among the subjects, but we don’t yet know how much is between-
groups variance and how much is within-groups variance. Figure A.6(b) shows
a situation in which all the variance is between groups and there is none within
groups. Figure A.6(c) shows a different possibility—all the variance is within
groups and there is none between groups. Thus, in these highly artificial exam-
pl es, we could say that all the variance is either between- or within-groups vari-
ance. Ordinarily, of course, there would be some of each.

How to Read an ANOVA Summary Table
There are many different types of ANOVA, depending on the particular ex-
perimental design. You will need to consult a statistics book to be able to per-
form an ANOVA. Frequently, however, you will read a description of an
experiment that contains an ANOVA summary table. Fortunately, you can
read and interpret an ANOVA table without knowing how to perform the
ANOVA. The purpose of this section is to help you understand such a table.

It is actually possible to decipher a great deal about an experiment from an
ANOVA summary table. Consider Table A.8, which shows the analysis of a simple one-way ANOVA. The term one-way means that the experiment
contains only one independent variable (IV).

An ANOVA summary table always contains at least two rows of informa-
tion, one for each source of variance in the experiment. In addition, certain
totals are shown. We see in this case that there is a row labeled “Between
conditions.” This row shows the information about the variance resulting
from the different conditions, or levels, of the independent variable. The second
row is labeled “Within conditions.” This row indicates the information about
the variance of the data within the conditions, or groups.

The columns indicate sum of squares (SS), degrees of freedom (df), mean
square (MS), the F ratio (F), and probability (p). The sum of squares is a mea-
sure of the variability in the data. Some of the variance can be attributed to
the experimental variable. This is the between-conditions variance, which has
a value of 504 in this example. The rest of the variance can be attributed to
the variance within each condition. This is the within-conditions variance,
which is 2.51 in this example. Notice that the sum of these values gives us
the total variability in the experiment.

Degrees of freedom is a quantity that depends on the number of groups,
subjects, and the like. The total degrees of freedom, 39 in our example, is
one less than the number of observations; therefore, we can conclude that
there were 40 observations in the experiment. Because this was a between-
subjects experiment, and therefore each subject contributed one observation,
we know that there were 40 subjects. The number of degrees of freedom be-
 tween conditions, three in this example, is one less than the number of condi-
tions; thus, we know that there were four conditions. The mean square is the
sum of squares divided by the number of degrees of freedom in the same row.

All these figures are used to determine the value of F, which is the ratio of the
mean square between subjects to the mean square within subjects. Remember
that if there is an experimental effect, the variance attributable to the conditions
(the mean square between conditions) will be larger than the variance attributable
to the subjects (the mean square within conditions). If the F ratio is sufficiently
greater than 1.0, it is considered significant. This is determined by looking in a
table, or it may be printed out automatically by certain statistical programs.
ANOVA summary tables differ in their structure depending on the design of the experiment. Certain things can always be counted on, however. First, there is a row in the table for each source of variance. Each independent variable is a source of variance. In Table A.8, the first row indicates the independent variable. In a two-way ANOVA, both independent variables are sources of variance. The interaction between the two variables is also a source of variance.

Second, at least one of the sources of variance serves as an error term. The purpose of the error term is to form the denominator of the F ratio. In a simple ANOVA, such as the one summarized in Table A.8, the within-groups variance serves as the error term. The error term is generally in the last row.

Third, there is always at least one F ratio computed from two mean squares. The numerator of the ratio is the particular effect being tested—the between-groups effect (the effect of the independent variable), for example. The denominator of the ratio is the error term. Therefore, the F ratios in the

---

**TABLE A.8**

ANOVA Summary Table

<table>
<thead>
<tr>
<th>Sources</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between conditions</td>
<td>504</td>
<td>3</td>
<td>168</td>
<td>24</td>
<td>&lt;.025</td>
</tr>
<tr>
<td>Within conditions</td>
<td>251</td>
<td>36</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>755</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Because there were four groups and 40 subjects, we infer that there were 10 subjects per group.

---

**TABLE A.9**

ANOVA Summary Table: Within-Subjects Design

<table>
<thead>
<tr>
<th>Sources</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subjects</td>
<td>167.33</td>
<td>9</td>
<td>18.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between conditions</td>
<td>423.27</td>
<td>2</td>
<td>211.64</td>
<td>76.13</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>50.07</td>
<td>18</td>
<td>2.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>640.67</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

There were four groups ($df = K - 1$) and MS = SS/df. $F = \frac{MS \text{ between}}{MS \text{ residual}}$.
There were three conditions of A  
\((df = I - 1)\)

There were 30 observations of B  
\((df = J - 1)\)

**TABLE A.10**

ANOVA Summary Table: Factorial Design

<table>
<thead>
<tr>
<th>Sources</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st IV</td>
<td>264.48</td>
<td>2</td>
<td>132.24</td>
<td>76.13</td>
<td>&lt;01</td>
</tr>
<tr>
<td>2nd IV</td>
<td>1.93</td>
<td>2</td>
<td>0.97</td>
<td>0.13</td>
<td>n.s.</td>
</tr>
<tr>
<td>A x B</td>
<td>404.18</td>
<td>4</td>
<td>101.05</td>
<td>13.37</td>
<td>&lt;01</td>
</tr>
<tr>
<td>Within cells (error)</td>
<td>340.17</td>
<td>45</td>
<td>7.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1010.76</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note significant.
NOTE: Because there were nine conditions, we refer that there were 54/9 = 6 subjects per condition.

\[ df = (I-1)(J-1) = (3-1)(3-1) = 4 \]

\[ MS = SS/df \]

There were 54 subjects  
(totat \(df = N - 1\)).

\[ F = \frac{MS}{MS \text{ error}} \]

Table are usually computed as the ratio of the mean square in one of the upper rows to the mean square in the last row.

Fourth, if there is a row with the word *subjects* in the label, we know that we have a within-subjects (repeated-measures) design. Table A.9 shows such a design. Here, the numerator of the \(F\) ratio is found in the between-conditions row, as before. The denominator of the \(F\) ratio (error term) is once again in the last row, but here it is called the *residual*. Because there are only three rows of information, we know that we have a one-way, repeated-measures ANOVA. We can see that there were 30 observations because there are 29 total degrees of freedom. Also, we know that there were three conditions because there are two degrees of freedom between conditions.

Table A.10 gives the summary information from another experiment. Here we see one row indicating the information for an independent variable A, labeled “Between conditions of A,” and a second row labeled “Between conditions of B.” This tells us that there were two independent variables. This means we have a two-way ANOVA and the experiment had a factorial design. With two-way ANOVA, we see a third row for information about the interaction between the two independent variables, labeled “A X B.”

**WEB-BASED WORKSHOPS ON RESEARCH METHODS AND STATISTICS**

Wadsworth Publishing Company maintains Web-based workshops on research methods and statistics. These workshops give a different slant on the material in this book.

www.cengage.com/psychology/workshops

For this chapter, there are 18 workshops on various aspects of statistics.
EXERCISES

A.1 Hypothesis Testing and Power

This exercise uses the computer program given in the Instructor’s Manual. The program takes samples from one of two distributions. The first distribution has a mean of 100 and a standard deviation of 16. The mean of the second is 101, and it has the same standard deviation. When you run the exercise, you will be presented with samples from either one distribution or the other; you will not know which.

Suppose you know that the mean IQ of the general population is 100 and the standard deviation is 16. (Actually it is 15, but 16 is more convenient for the program.) You also know that people with first names that have exactly seven letters have a mean IQ of 101 and the same standard deviation. Now, suppose you are told only the mean IQ of a sample of people, all of whom either have seven letters in their first names or have other than seven letters in their first names. Your job is to guess whether your sample has seven-letter names or other-than-seven-letter names.

Your best strategy is to guess 100 whenever the mean of the sample is less than 100.50. When it is equal to or greater than 100.50, it is at least an even bet that the sample is from the 101 population.

This exercise is an illustration of hypothesis testing. Your null hypothesis is that the sample came from the general population—those people who have names other than seven letters long and have a mean IQ of 100. The alternative hypothesis is that the sample came from the population of those who have seven letters in their names and have an average IQ of 101.

After each trial, the program will tell you whether you have made a correct guess or whether you have made an error. You can make a correct guess in two ways. First, you can guess that your sample comes from the general population when it does. This would be an example of correctly accepting the null hypothesis. Second, you could guess that the sample comes from the seven-letter-name population when it actually does. This would be an example of correctly rejecting the null hypothesis.

You could also make an error in two ways. You could guess that the sample came from the seven-letter-name population when it came from the general population. This would be a Type I error because you rejected a true null hypothesis. Alternatively, you could guess that the sample came from the general population when it actually came from the seven-letter-name population. This would be a Type II error because you accepted the null hypothesis when it was false.

You will see as you go through the exercise that you will sometimes make errors even when you make the best possible decision. This illustrates that hypothesis testing is a matter of probability—in this case, a probability of making certain kinds of errors.

At first, you will be given the mean of samples of size four; in other words, four individuals will be selected from the population. Later, you will be given the means of samples of size 400. As the sample size gets larger, the means of the samples will tend to be closer to the means of the populations from which they are sampled, and you will be able to make more accurate judgments. This illustrates the effect of sample size on the probability of making a correct decision.

The reason for this effect can be seen by examining the equation for the standard error of the mean. The standard error of the mean is the standard deviation of the distribution of sample means:

$$\sigma_x = \frac{\sigma}{\sqrt{N}}$$

With a sample size of 4:

$$\sigma_x = \frac{16}{\sqrt{4}} = \frac{16}{2} = 8$$

When the sample size is 400, the standard error of the mean is one-tenth as large:

$$\sigma_x = \frac{16}{\sqrt{400}} = \frac{16}{20} = 0.8$$

The smaller standard error of the mean means that there is less variability in the means of your sample, so they tend to fall nearer the mean of the population. Thus, you are more often able to guess correctly which population you are sampling from.

This is an illustration of the concept of statistical power (see page 413). Power is the probability of rejecting the null hypothesis when it is false, or the
probability of deciding that you have an experimental effect when in fact you do. This probability is one minus the probability of a Type II error:

\[
\text{power} = 1 - P(\text{Type II error})
\]

In the IQ example, the probability of guessing 101 when it is in fact 101 is the power of your test. Power increases as sample size increases. The theoretical power of your decision is 0.52 when the sample size is four, but it increases to 0.73 when the sample size is 400. (These numbers assume that you guessed that the sample came from the population with a mean of 101 whenever the mean of the sample was 100.5 or greater.)

The computer will print out your obtained power. Your actual power will differ from the theoretical power because the data are empirical data subject to chance factors. If your empirical data differ markedly from the theoretical values, you may want to repeat the exercise.

REQUIRED:
a. What was your observed power when the sample size was four? When it was 400?
b. What else does power depend on?

A.2 Scatterplots

Professor Cora Late teaches psychological statistics to undergraduates at a large university. She believes that performance on statistics examinations is related to competence in algebra. She designs a study to test the hypothesis that students who are more proficient in algebra will do better in statistics. At the beginning of the term, she administers a 15-item algebra pretest to the population of undergraduates enrolled in her introductory statistics course. At the end of the term, she administers a 20-item comprehensive statistics examination to these same students.

REQUIRED:
a. Option A: Use the first 20 subjects from the population data set in Appendix C. Column F contains the subjects’ algebra scores, and Column G contains the subjects’ statistics examination scores. List the paired values of algebra and statistics exam scores for each of the 20 subjects. Option B: Using the procedures outlined in Chapter 10, draw a random sample of 20 subjects from the population data set in Appendix C. Column F contains the subjects’ algebra scores, and Column G contains the subjects’ statistics exam scores. List the paired values of algebra and statistics exam scores for each of the 20 subjects.
b. Construct a scatterplot of the paired values of algebra and statistics exam scores for each of the 20 subjects.
c. Is Professor Late justified in her belief? Explain.

OPTIONAL:
a. Calculate the Pearson correlation coefficient, \( r \), for the obtained sample of paired raw scores.
b. Interpret the sign and size of \( r \) in terms of the scatterplot.
c. Determine the equation of the regression line for predicting statistics performance from algebra scores.
d. Plot the regression line determined in (f) on the scatterplot constructed in (b).
e. What statistics score would you predict for a student with an algebra score of eight?

A.3 Identify Directional and Nondirectional Hypotheses

A. \( H_0 : \mu_1 = \mu_2 \)  D. \( H_1 : \mu_1 \neq \mu_2 \)
B. \( H_0 : \mu_1 \leq \mu_2 \)  E. \( H_1 : \mu_1 < \mu_2 \)
C. \( H_0 : \mu_1 \geq \mu_2 \)  F. \( H_1 : \mu_1 > \mu_2 \)

For each of the following statements, identify the correct null hypothesis (A, B, or C) and the correct alternative hypothesis (D, E, or F).
a. Introductory algebra students who are taught with hand-held calculators (Group 1) for a 15-week period will have different scores on tests of computational skills (taken without the use of a calculator) from introductory algebra students who are taught without calculators (Group 2).
b. New graduate male nurses whose orientation program uses a preceptor (an individual tutor) as a major component (Group 1) will exhibit higher performance levels than will new graduate male nurses in a traditional orientation program (Group 2).
c. Kindergarten students who have a volunteer parent pool assisting their teachers (Group 1) in the classroom will show different
achievement from kindergarten students whose teachers do not have extra assistance in the classroom (Group 2).
d. Students who receive key images, or pictures, with new vocabulary words (Group 1) will show greater acquisition and retention of the definition of words than will students who are left to their own strategies for learning new words (Group 2).

A.4 Interpret an ANOVA Summary Table

REQUIRED:
a. Complete the missing information in Table A.11.
b. Is this a one-way or a two-way ANOVA?
c. Was this a within-subjects or a between-subjects design?
d. How many levels of A were there?
e. How many subjects were in the experiment?
f. How many subjects were in each group?
g. Was the effect of A significant? If so, at what level?

**TABLE A.11**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between A</td>
<td></td>
<td>1</td>
<td>160</td>
<td></td>
<td>.05</td>
</tr>
<tr>
<td>Within groups</td>
<td>1230</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5166</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A.5 Interpret an ANOVA Summary Table

REQUIRED:
a. Complete the missing information in Table A.12.
b. Is this a one-way or a two-way ANOVA?
c. Was this a within-subjects or a between-subjects design?
d. How many levels of A were there?
e. How many levels of B were there?
f. How many subjects were in this experiment?
g. How many subjects were in each group?
h. Was the effect of B significant? If so, at what level?
i. Was the interaction significant? If so, at what level?

**TABLE A.12**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between A</td>
<td>160</td>
<td>1</td>
<td>160</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Between B</td>
<td>160</td>
<td>2</td>
<td></td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>AxB</td>
<td></td>
<td>2</td>
<td>12</td>
<td></td>
<td>.75</td>
</tr>
<tr>
<td>Within groups</td>
<td>384</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>728</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Not significant.

A.6 Interpret an ANOVA Summary Table

REQUIRED:
a. Complete the missing information in Table A.13.
b. Is this a one-way or a two-way ANOVA?
c. Was this a within-subjects or a between-subjects experiment?
d. How many levels of A were there?
e. How many subjects were in the experiment?
f. Was the F test significant?

**TABLE A.13**

<table>
<thead>
<tr>
<th>Sources</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subjects</td>
<td>162</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between A</td>
<td>456</td>
<td>2</td>
<td>228</td>
<td></td>
<td>.01</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>672</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A.7 Interpret an ANOVA Summary Table

REQUIRED:
a. Based on Table A.14, how many subjects were there?
b. How many independent variables were there?
c. How many groups were there?
d. Did each subject experience all conditions?
e. Is this a one-way or a two-way ANOVA?
f. Show where the value of F came from.
g. Were the results significant at the 0.05 level?

**TABLE A.14**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between conditions</td>
<td>504</td>
<td>3</td>
<td>168</td>
<td>24</td>
<td>.025</td>
</tr>
<tr>
<td>Within conditions</td>
<td>251</td>
<td>36</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>755</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>