In this part, we'll begin by examining the differences between the two major branches of statistics—*descriptive statistics*, which help us summarize and describe data that we've collected, and *inferential statistics*, which help us make inferences from samples to the populations from which they were drawn. Then we will consider basic statistical techniques used in analyzing the results of research.

Because this part is designed to prepare you to comprehend statistics reported in research reports, the emphasis is on understanding their meanings—not on computations. Thus, we will take a conceptual look at statistics and consider computations only when they are needed to help you understand concepts.

**Important Note**

The topics in this part are highly interrelated; in many topics, I have assumed that you have mastered material in the earlier topics. Thus, you are strongly advised to read the topics in the order presented.
**TOPIC 37  DESCRIPTIVE AND INFERENTIAL STATISTICS**

**Descriptive statistics** help us summarize data so they can be easily comprehended. For example, suppose we administered a test to all 362 freshmen enrolled in a university. An unordered list of the scores would be difficult to process mentally. However, if we prepare a frequency distribution such as that in Table 1, we can easily see how the scores are distributed. For example, the figure clearly indicates that a majority had scores of 14 through 16, with a scattering above and below these levels.

Table 1  *Frequency distribution with percentages*

<table>
<thead>
<tr>
<th>Score (X)</th>
<th>Frequency (f)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>5</td>
<td>1.4</td>
</tr>
<tr>
<td>19</td>
<td>9</td>
<td>2.5</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>6.6</td>
</tr>
<tr>
<td>17</td>
<td>35</td>
<td>9.7</td>
</tr>
<tr>
<td>16</td>
<td>61</td>
<td>16.9</td>
</tr>
<tr>
<td>15</td>
<td>99</td>
<td>27.3</td>
</tr>
<tr>
<td>14</td>
<td>68</td>
<td>18.8</td>
</tr>
<tr>
<td>13</td>
<td>29</td>
<td>8.0</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>5.8</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>362</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>

The frequencies in Table 1 are descriptive statistics; they describe how many students earned each score. The percentages are also descriptive; they describe how many students per one hundred had each score. These and other descriptive statistics such as averages are described in this part of the book.

Now let’s suppose that for the sake of efficiency, instead of testing all 362 freshmen, we sampled at random (by drawing names out of a hat) only 100 to be tested. Would we obtain exactly the same results as we would if we tested all freshmen? In all likelihood, no. As you probably recall from Topics 17 through 19, random sampling produces random errors called *sampling errors*. **Inferential statistics** help us draw inferences about the effects of sampling errors on our results. They are defined as statistical techniques that help us generalize from samples to the populations from which the samples were drawn.

One type of inferential statistic you may already be familiar with is a *margin of error*. When reporting the results of public opinion polls in the media, reporters frequently cite margins of error to help us interpret results in light of sampling error. For example, a recent poll indicated that approval of the president was at 52% with a margin of error of ±2 (i.e., plus/minus 2) percentage points. This means we can be highly confident that the level of approval in the population is between 50% and 54% (that is, within two points of the 52% observed in the sample).

An important family of inferential statistics consists of *significance tests*, which help us decide whether differences that we observe (such as differences in the reading achievement of samples of boys and girls) are reliable. The next topic will help you understand the general purpose of significance testing, and in later sections, we will consider three popular tests of significance.

Because inferential statistics help us evaluate results in light of sampling errors, it follows that if we do not sample, we do not need inferential statistics. For example, if we conduct a census (a study in which all members of a population are included), the descriptive values that we obtain such as percentages are values that are free of sampling errors.

We distinguish between values obtained from a sample and values obtained from a census by using the terms **parameters** for values from a census and **statistics** for values from studies in which samples were examined. Thus, percentages, averages, and frequencies are classified as parameters when they result from a census, but they are classified as statistics when they are based on a sample. Remember the first letters:

Samples yield Statistics, and Populations yield Parameters.
EXERCISE ON TOPIC 37

1. Which branch of statistics helps us summarize data so they can be easily comprehended?

2. According to Figure 1 in this topic, how many subjects had a score of 19?

3. What is the name of the statistic that describes how many subjects per 100 have a certain characteristic?

4. Which branch of statistics helps us draw inferences about the effects of sampling errors on our results?

5. If we test a random sample instead of all members of a population, is it likely that the sample results will be the same as the results we would have obtained by testing the population?

6. Is a margin of error a descriptive or an inferential statistic?

7. Do we perform significance tests with inferential or descriptive statistics?

8. By studying populations, do we obtain statistics or parameters?

9. By studying samples, do we obtain statistics or parameters?

Question for Discussion

10. Keep your eye out for a report of a poll in which a margin of error is reported. Copy the exact words and bring it to class for discussion.

For Students Who Are Planning Research

11. Will you be reporting descriptive statistics? (Note that statistics often are not reported in qualitative research. See Topics 9 and 10.)

12. Will you be reporting inferential statistics? (Note that they are needed only if you have sampled.)
Suppose we drew random samples of engineers and psychologists, administered a self-report measure of sociability, and computed the mean (the most commonly used average) for each group. Furthermore, suppose the mean for engineers is 65.00 and the mean for psychologists is 70.00. Where did the five-point difference come from? There are three possible explanations:

1. Perhaps the population of psychologists is truly more sociable than the population of engineers, and our samples correctly identified the difference. (In fact, our research hypothesis may have been that psychologists are more sociable than engineers—which now appears to be supported by the data.)

2. Perhaps there was a bias in procedures. By using random sampling, we have ruled out sampling bias, but other procedures such as measurement may be biased. For example, maybe the psychologists were contacted during December, when many social events take place and the engineers were contacted during a gloomy February. The only way to rule out bias as an explanation is to take physical steps to prevent it. In this case, we would want to make sure that the sociability of both groups was measured in the same way at the same time.

3. Perhaps the populations of psychologists and engineers are the same but the samples are unrepresentative of their populations because of random sampling errors. For instance, the random draw may have given us a sample of psychologists who are more sociable, on the average, than their population.

The third explanation has a name—it is the null hypothesis. The general form in which it is stated varies from researcher to researcher. Here are three versions, all of which are consistent with each other:

**Version A of the null hypothesis:**
The observed difference was created by sampling error. (Note that the term sampling error refers only to random errors—not errors created by a bias.)

**Version B of the null hypothesis:**
There is no true difference between the two groups. (The term true difference refers to the difference we would find in a census of the populations, that is, the difference we would find if there were no sampling errors.)

**Version C of the null hypothesis:**
The true difference between the two groups is zero.

Significance tests determine the probability that the null hypothesis is true. (We will be considering the use of specific significance tests in Topics 41—42 and 48-50.) Suppose for our example we use a significance test and find that the probability that the null hypothesis is true is less than 5 in 100; this would be stated as $p < .05$, where $p$ obviously stands for probability. Of course, if the chances that something is true is less than 5 in 100, it’s a good bet that it’s not true. If it’s probably not true, we reject the null hypothesis, leaving us with only the first two explanations that we started with as viable explanations for the difference.

There is no rule of nature that dictates at what probability level the null hypothesis should be rejected. However, conventional wisdom suggests that .05 or less (such as .01 or .001) is reasonable. Of course, researchers should state in their reports the probability level they used to determine whether to reject the null hypothesis.

Note that when we fail to reject the null hypothesis because the probability is greater than .05, we do just that: We "fail to reject" the null hypothesis and it stays on our list of possible explanations; we never "accept" the null hypothesis as the only explanation—remember, there are three possible explanations and failing to reject one of them does not mean that you are accepting it as the only explanation.

An alternative way to say that we have rejected the null hypothesis is to state that the difference is statistically significant. Thus, if we state that a difference is statistically significant at
the .05 level (meaning .05 or less), it is equivalent to stating that the null hypothesis has been rejected at that level.

When you read research reported in academic journals, you will find that the null hypothesis is seldom stated by researchers, who assume that you know that the sole purpose of a significance test is to test a null hypothesis. Instead, researchers tell you which differences were tested for significance, which significance test they used, and which differences were found to be statistically significant. It is more common to find null hypotheses stated in theses and dissertations since committee members may wish to make sure that the students they are supervising understand the reason they have conducted a significance test.

As we consider specific significance tests in the next three parts of this book, we'll examine the null hypothesis in more detail.

**EXERCISE ON TOPIC 38**

1. How many explanations were there for the difference in sociability between psychologists and engineers in the example in this topic?

2. What does the null hypothesis say about sampling error?

3. Does the term *sampling error* refer to *random errors* or to *bias*?

4. The null hypothesis says that the true difference equals what value?

5. What is used to determine the probabilities that null hypotheses are true?

6. For what does p < .05 stand?

7. Do we reject the null hypothesis when the probability of its truth is high or when it is low?

8. What do we do if the probability is greater than .05?

9. What is an alternative way of saying that we have rejected the null hypothesis?

10. Are you more likely to find a null hypothesis stated in a journal article or in a thesis?

**Question for Discussion**

11. We all use probabilities in everyday activities to make decisions. For example, before we cross a busy street, we estimate the odds that we will get across the street safely. Briefly describe one other specific use of probability in everyday decision making.

**For Students Who Are Planning Research**

12. Will you need to test the null hypothesis in your research? Explain.
There are four scales (or levels) at which we measure. The lowest level is the \textbf{nominal} scale. This may be thought of as the "naming" level. For example, when we ask subjects to name their marital status, they will respond with words—not numbers—that describe their status such as "married," "single," "divorced," etc. Notice that nominal data do not put subjects in any particular order. There is no logical basis for saying that one category such as "single" is higher or lower than any other.

The next level is \textbf{ordinal}. At this level, we put subjects in order from high to low. For instance, an employer might rank order applicants for a job on their professional appearance. Traditionally, we give a rank of 1 to the subject who is highest, 2 to the next highest, and so on. It is important to note that ranks do not tell us by how much subjects differ. If we are told that Janet has a rank of 1 and Frank has a rank of 2, we do not know if Janet's appearance is greatly superior to Frank's or only slightly superior. To measure the \textit{amount} of difference among subjects, we use the next levels of measurement.

Measurements at the \textbf{interval} and \textbf{ratio} levels have equal distances among the scores they yield. For example, when we say that Jill weighs 120 pounds and Sally weighs 130 pounds, we know by \textit{how much} the two subjects differ. Also, note that a 10-pound difference represents the same amount regardless of where we are on the scale. For instance, the difference between 120 and 130 pounds is the same as the difference between 220 and 230 pounds.

The ratio scale is at a higher level than the interval scale because the ratio has an absolute zero point that we know how to measure. Thus, \textit{weight} is an example of the ratio scale because it has an absolute zero that we can measure.

The interval scale, while having equal intervals like the ratio scale, does not have an absolute zero. The most common examples of interval scales are scores obtained using objective tests such as multiple-choice tests of achievement. It is widely assumed that each multiple-choice test item measures a single point's worth of the trait being measured and that all points are equal to all other points—making it an interval scale (just as all pounds are equal to all other pounds of weight). However, such tests do not measure at the ratio level because the zero on such tests is arbitrary—not absolute. To see this, consider someone who gets a zero on a multiple-choice final examination. Does the zero mean that the student has absolutely no knowledge of or skills in the subject area? Probably not. He or she probably has some knowledge of simple facts, definitions, and concepts, but the test was not designed to measure at the skill level at which the student is operating. Thus, a score of zero indicates only that the student knows nothing \textit{on that test}—not that the student has zero knowledge of the content domain.

Here's a summary of the levels:

<table>
<thead>
<tr>
<th>Lowest Level</th>
<th>Scale</th>
<th>Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textbf{Nominal}</td>
<td>naming</td>
<td></td>
</tr>
<tr>
<td>\textbf{Ordinal}</td>
<td>ordering</td>
<td></td>
</tr>
<tr>
<td>\textbf{Interval}</td>
<td>equal interval without absolute zero</td>
<td></td>
</tr>
<tr>
<td>\textbf{Ratio}</td>
<td>equal interval with absolute zero</td>
<td></td>
</tr>
</tbody>
</table>

For those of you who like to use mnemonics when memorizing material, try learning this environmentally friendly phrase:

\textbf{No Oil In Rivers}

The first letters—\textbf{NOIR}—are the first letters of the scales in order from lowest to highest.

At which level should we measure? First, some variables are inherently nominal in nature. For example, when we need to know subjects' gender or state of residence, nominal data is the natural choice. Second, many novice researchers overuse the ordinal scale. For instance, if we want to measure reading ability, it usually would be much better to use a carefully constructed
standardized test (which measures at the interval level) than having teachers rank order students in terms of their reading ability. Remember, measuring at the interval level gives you more information because it tells you by how much students differ. Also, as you will learn when we explore statistics, you can do more interesting and powerful types of analyses when you measure at the interval rather than the ordinal level. Thus, when planning instruments for a research project, if you are thinking in terms of having subjects ranked (for ordinal measurement), you would be well advised to consider whether there is an alternative at the interval level.

The choice between interval and ratio depends solely on whether it is possible to measure with an absolute zero. When it is possible, we usually do so. For the purposes of statistical analysis, interval and ratio data are treated in the same way.

The level at which we measure has important implications for data analysis, so you will find references to scales of measurement throughout our discussion of statistics.

EXERCISE ON TOPIC 39

1. If we ask subjects to name the country in which they were born, we are using what scale of measurement?
2. Which two scales of measurement have equal distances among the scores they yield?
3. If we have a teacher rank students according to their oral language skills, we are using which scale of measurement?
4. Which scale of measurement has an absolute zero that is measured?
5. Which scale of measurement is at the lowest level?
6. Objective, multiple-choice achievement tests are usually assumed to measure at what level?
7. If we measure in such a way that we find out which subject is most honest, which is the next most honest, and so on, we are measuring at what scale of measurement?
8. The number of minutes of overtime work that employees perform is an example of what scale of measurement?
9. Weight measured in pounds is an example of which scale of measurement?

Question for Discussion

10. Name a trait that inherently lends itself to nominal measurement. Explain your answer.

For Students Who Are Planning Research

11. List the measures you will be using, and name the scale of measurement for each one.
We obtain nominal data when we classify subjects according to names (words) instead of quantities. For example, suppose we asked a population of 540 teachers which candidate they each prefer for a school board vacancy and found that 258 preferred Smith and 282 preferred Jones. The 258 and 282 are frequencies, whose symbol is \( f \); we can also refer to them as numbers of cases, whose symbol is \( N \).

We can convert the numbers of cases into percentages by dividing the number who prefer each candidate by the number in the population and multiplying by 100. Thus, for Smith, the calculations are:

\[
258 \div 540 = .478 \times 100 = 47.8\% 
\]

When reporting percentages, it's a good idea to also report the underlying numbers of cases, which is done in Table 1.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>52.2%</th>
<th>47.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones</td>
<td>( (N = 282) )</td>
<td>( (N = 258) )</td>
</tr>
<tr>
<td>Smith</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 is an example of univariate analysis. We are analyzing how people vary (hence, we use the root variate) on only one variable (hence, we use the prefix uni-).

We can examine a relationship between two nominal variables by conducting a bivariate analysis. Perhaps we want to know whether there is a relationship between teachers' gender and their preferences for candidates. Table 2 shows the results of a bivariate analysis of these variables.

<table>
<thead>
<tr>
<th></th>
<th>Jones</th>
<th>Smith</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>66.4%</td>
<td>33.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td>( (N = 85) )</td>
<td>( (N = 43) )</td>
<td>( (N = 128) )</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>47.8%</td>
<td>52.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>( (N = 197) )</td>
<td>( (N = 215) )</td>
<td>( (N = 412) )</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Teachers' preferences by gender

The data in Table 2 clearly indicate that there is a relationship between gender and preference for either Jones or Smith. A larger percentage of males than females prefers Jones, but a larger percentage of females than males prefers Smith. Another way to put this is that teachers' gender is predictive of their preferences. For instance, by knowing that a teacher is male, we would predict that he is more likely to vote for Jones than Smith.

Notice that in this population of teachers, there are many more female teachers than male teachers. When this is the case, we can be misled by examining only the numbers of cases (for example, 85 males for Jones versus 197 females for Jones). Notice that, in fact, a majority of the smaller population of males is in favor of Jones but only a minority of the larger population of females is in favor of him. With percentages, legitimate comparisons of groups of unequal size are possible. This is because percentages convert numbers of cases to a common scale with a base of 100. (The percentage of 66.4\% of males for Jones indicates that every 100 males, 66.4\% of them are in favor of Jones while the percentage of 47.8\% of females for Jones indicates that for every 100 females, only 47.8\% of them are in favor of Jones.)

In academic writing, some researchers report the proportions instead of the percentages. For example, a percentage of 47.8\% in favor of Smith in Table 1 corresponds to a proportion of .478 or .48. (Proportions are calculated in the same way as percentages are except that we do not multiply by 100.) Since a proportion has a base of 1, a proportion of .48 means that for every one subject, 48 hundredths of each subject favors Smith. Clearly, proportions are harder to comprehend than percentages. When you encounter proportions in literature, it's a good idea to convert them mentally to percentages. That is, think of .48 as 48\% (the percentage you get by multiplying by 100).

In the next section, we will consider how to analyze nominal data in studies in which we have sampled at random and need to take account of random sampling errors.
EXERCISE ON TOPIC 40

1. If 400 people in a population of 1,000 are Democrats, what percentage are Democrats?

2. When reporting a percentage, is it a good idea to also report the underlying number of cases?

3. Do we use univariate or bivariate analyses to examine relationships among nominal variables?

4. Percentages convert numbers of cases to a common scale with what base?

5. What is the base for a proportion?

6. Are percentages or proportions easier for most people to comprehend?

Question for Discussion

7. Be on the lookout for a report in the popular press in which percentages are reported. Bring a copy to class. Be prepared to discuss whether the frequencies are also reported and whether it is a univariate or bivariate analysis.

For Students Who Are Planning Research

8. Will you be measuring anything at the nominal level? Explain.

TOPIC 41  INTRODUCTION TO THE CHI SQUARE TEST

Suppose we drew at random a sample of 200 members of a professional association of sociologists and asked them whether they were in favor of a proposed change to their bylaws. The results are shown in Table 1. But do these observed results reflect the true results that we would have obtained if we had questioned the entire population? Remember that the null hypothesis (see Topic 38) says that the observed difference was created by random sampling errors; that is, in the population, the true difference is zero. Put another way, the observed difference \( n = 120 \) vs. \( n = 80 \) is an illusion created by chance errors.

The usual test of the null hypothesis when we are considering frequencies (that is, number of cases or \( n \)) is chi square, whose symbol is:

\[
\chi^2
\]

It turns out that after doing some computations, which are beyond the scope of this book, for the data in Table 1, the results are:

\[
\chi^2 = 4.00, df = 1, p < .05
\]

What does this mean for a consumer of research who sees this in a report? The values of chi square and degrees of freedom \( (df) \) were calculated solely to obtain the probability that the null hypothesis is correct. That is, chi square and degrees of freedom are not descriptive statistics that you should attempt to interpret. Rather, think of them as substeps in the mathematical procedure for obtaining the value of \( p \). Thus, the consumer of research should concentrate on the fact that \( p \) is less than .05. As you probably recall from Topic 38, when the probability \( (p) \) that the null hypothesis is correct is .05 or less, we reject the null hypothesis. (Remember, when the probability that something is true is less than 5 in 100—a low probability—conventional wisdom suggests that we should reject it as being true.) Thus, the difference we observe in Table 1 was probably not created by random sampling errors; therefore, we can say that the difference is statistically significant at the .05 level.

So far, we have concluded that the difference we observed in the sample was probably not created by sampling errors. So where did the difference come from? Two possibilities remain:

1. Perhaps there was a bias in procedures such as the person asking the question in the survey leading the respondents by talking enthusiastically about the proposed change in the bylaws. If we are convinced that adequate measures were taken to prevent procedural bias, we are left with only the next possibility as a viable explanation.

2. Perhaps the population of sociologists is, in fact, in favor of the proposed change, and this fact is correctly identified by studying the random sample.

Now let’s consider some results from a survey in which the null hypothesis was not rejected. Table 2 shows the numbers and percentages of subjects in a random sample from a population of teachers who prefer each of three methods for teaching reading. In the table, there are three differences (30 for A versus 27 for B, 30 for A versus 22 for C, and 27 for B versus 22 for C). The null hypothesis says that this set of differences was created by random sampling errors; in other

\*We are using the term true results here to stand for the results of a census of the entire population. The results of a census are true in the sense that they are free of sampling errors. Of course, there may also be measurement errors, which we are not considering here.
words, it says that there is no true difference in the population; we have observed a difference only because of sampling errors. The results of the chi square test for the data in Table 2 are:

\[ \chi^2 = 1.214, df = 2, p > .05 \]

Using the decision rule that \( p \) must be equal to or less than .05 to reject the null hypothesis, we fail to reject the null hypothesis, which is called a statistically insignificant result. In other words, the null hypothesis must remain on our list as a viable explanation for the set of differences we observed by studying a sample.

In this topic, we have considered the use of chi square in a univariate analysis in which we classify each subject in only one way (such as which candidate each prefers). In the next topic, we’ll consider its use in bivariate analysis in which we classify each subject in two ways (such as which candidate each prefers and the gender of each) in order to examine a relationship between the two.

**EXERCISE ON TOPIC 41**

1. When we study a sample, are the results called the true results or the observed results?

2. According to the null hypothesis, what created the difference in Table 1 in this topic?

3. What is the name of the test of the null hypothesis used when we are analyzing frequencies?

4. As a consumer of research, should you try to interpret the value of \( df \)?

5. What is the symbol for probability?

6. If you read that a chi square test of a difference yielded \( up \) of less than 5 in 100, what should you conclude about the null hypothesis on the basis of conventional wisdom?

7. Does \(< .05 \text{ or } > .05\) usually lead a researcher to declare a difference to be statistically significant?

8. If we fail to reject a null hypothesis, is the difference in question statistically significant?

9. If we have a statistically insignificant result, does the null hypothesis remain on our list of viable hypotheses?

**Question for Discussion**

10. Briefly describe a hypothetical study in which it would be appropriate to conduct a chi square test for univariate data.

**For Students Who Are Planning Research**

11. Will you be conducting a chi square test? Explain.
In this topic, we will examine the use of the chi square test in a bivariate analysis — that is, when each subject is classified in terms of two variables in order to examine the relationship between them. Let's look at an example. Suppose we conducted an experiment in which three methods of job training were tried with welfare recipients. Random samples of recipients were drawn for each method, and the number who obtained jobs by the end of the training sessions was determined. The resulting data are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Method A</th>
<th>Method B</th>
<th>Method C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>n = 20</td>
<td>n = 15</td>
<td>n = 9</td>
</tr>
<tr>
<td></td>
<td>(66.7%)</td>
<td>(51.7%)</td>
<td>(31.0%)</td>
</tr>
<tr>
<td>No</td>
<td>n = 10</td>
<td>n = 14</td>
<td>n = 20</td>
</tr>
<tr>
<td></td>
<td>(33.3%)</td>
<td>(48.3%)</td>
<td>(69.0%)</td>
</tr>
</tbody>
</table>

Clearly, the data suggest that there is a relationship between which method of job training was used and the outcome (whether or not subjects got jobs). Based on the random samples, it appears that Method A is superior to Methods B and C and that Method B is superior to Method C. A stumbling block in our interpretation of these results is the null hypothesis, which says that there is no true difference (that is, if all members of the population had been studied, we would have found no differences among the three methods). For instance, quite by the luck of the random draw, recipients who were more employable to begin with (before treatment) were assigned to Method A, while the less employable, by chance, were assigned to the other two methods. We can test the null hypothesis by using chi square.

For the data shown in Table 1, this result would be shown in a report on the experiment:

As you know from the previous topics, we reject the null hypothesis when the odds that it is true are equal to or less than .05. Thus, for this data, we reject the null hypothesis and declare the result to be significant at the .05 level. We have concluded that the observed differences that suggest a relationship between method of training and job placement are too great to be attributed to random errors.

Now let's consider what we mean by the .05 level in more detail than we have up to this point. When we reject the null hypothesis at exactly the .05 level (that is, \( p = .05 \)), there are 5 chances in 100 that the null hypothesis is correct; thus, we are taking 5 chances in 100 of being wrong by rejecting null hypotheses at this level. In other words, we can never be certain that we have made the correct decision when rejecting the null hypothesis. It is always possible that the null hypothesis is true (in this case, there are 5 in 100 chances that it is true), and that we are making a mistake by rejecting it. This possibility is called a Type I Error. When we use the .05 level, the odds of making a Type I Error are 5 in 100; when we use the .01 level, the odds of making this type of error are 1 in 100; and when we use the .001 level, the odds of making it are 1 in 1,000.

When we fail to reject the null hypothesis, as we did in Topic 41 for the data in Table 2, we also are taking a chance that we are wrong. That is, perhaps the null hypothesis should have been rejected, but the significance test failed to lead us to the correct decision. This mistake is called a Type II Error. In review, these are the two types of errors:

- **Type I Error**: Rejecting the null hypothesis when it is, in fact, a correct hypothesis.
- **Type II Error**: Failing to reject the null hypothesis when it is, in fact, an incorrect hypothesis.

At first, this discussion of errors may make significance tests such as chi square seem weak — after all, we can be wrong no matter what decision we make. But let's look at the big picture. Once we decide to sample at random (the desirable way to sample because it's free from bias), we are open to the possibility that random errors have influenced our results. From this point on, we can never be certain. Instead, we must use probabilities to make decisions. We use them in such a way that we **minimize** the probability that we are
wrong. To do this, we usually emphasize minimizing the probability of a Type I error by using a low probability such as .05 or less. By using a low probability, we will infrequently be wrong in rejecting the null hypothesis.

**EXERCISE ON TOPIC 42**

1. What is the name of the type of analysis when each subject is classified in terms of two variables in order to examine the relationship between them?

2. What decision have we made about the null hypothesis if a chi square test leads us to the conclusion that the observed differences that suggest a relationship between two variables are too great to be attributed to random errors?

3. If \( p = .05 \) for a chi square test, chances are how many in 100 that the null hypothesis is true?

4. When we use the .01 level, what are the odds of making a Type I error?

5. What is the name for the error we make when we fail to reject the null hypothesis when it is, in fact, an incorrect hypothesis?

6. What is the name for the error we make when we reject the null hypothesis when it is, in fact, a correct hypothesis?

7. Why is random sampling desirable even though it creates errors?

**Questions for Discussion**

8. Are both of the variables in Table 1 in this topic nominal? Explain.

9. Briefly describe a hypothetical study in which it would be appropriate to conduct a chi square test on bivariate data.

**For Students Who Are Planning Research**

10. Will you be using a chi square test in a bivariate analysis? Explain.
TOPIC 43  SHAPES OF DISTRIBUTIONS

One way to describe quantitative data is to prepare a frequency distribution such as that shown in Topic 37 (see page 91). It is easier to see the shape of the distribution if we prepare a figure called a frequency polygon. This figure is a frequency polygon for the data in Topic 37:

The smooth, bell-shaped curve in Figure 2 has a special name; it is the normal curve. As the name "normal" suggests, it is the common shape that is regularly observed. Many things in nature are normally distributed—the weights of grains of sand on a beach, the heights of women (or men), the annual amounts of rainfall in most areas, and so on. The list is almost limitless. Many social and behavioral scientists also believe that mental traits of humans probably are also normally distributed.¹

Some distributions are skewed—that is, they have a tail to the left or right. Figure 3 shows a distribution that is skewed to the right (that is, the tail is to the right); it is said to have a positive skew. An example of a distribution with a positive skew is income. Most people earn relatively small amounts, so the curve is high on the left. Small numbers of rich and very rich people create a tail to the right.

Figure 4 is skewed to the left; it has a negative skew. We would get a negative skew, for example, if we administered a test of basic math skills to a large sample of college seniors. Most would do very well and get almost perfect scores, but a small scattering will get lower scores for a variety of reasons such as misunderstanding the directions for marking their answers, not feeling well the day the test was administered, and so on.

¹Because measures of mental traits are far from perfect, it is difficult to show conclusively that mental traits are normally distributed. However, many norm-referenced tests do yield normal distributions when large representative national samples are tested.

Figure 1  Frequency polygon for data on page 91

A frequency polygon is easy to read. For example, a score of 20 has a frequency (f) of 5, which is why the curve is low at a score of 20. A score of 15 has a frequency of 99, which is why the curve is high at 15.

Notice that the curve in Figure 1 is fairly symmetrical with a high point in the middle and dropping off on the right and left. When very large samples are used, the curve often takes on an even smoother shape, such as the one shown in Figure 2.

Figure 2  The normal curve

Figure 3  A distribution with a positive skew

Figure 4  A distribution with a negative skew
While there are other shapes, the three shown here are the ones you are most likely to encounter. Whether a distribution is basically normal or skewed affects how quantitative data at the interval and ratio levels are analyzed, which we will consider in the next topic.

Figure 4  A distribution with a negative skew

EXERCISE ON TOPIC 43

1. According to Figure 1, about how many subjects had a score of 14?

2. In Figure 1, are the frequencies on the vertical or horizontal axis?

3. Which of the curves discussed in this topic is symmetrical?

4. If a distribution has some extreme scores on the right (but not on the left) it is said to have what type of skew?

5. If a distribution is skewed to the left, does it have a positive or negative skew?

6. In most populations, income has what type of skew?

7. Does a distribution with a tail to the right have a positive or negative skew?

Question for Discussion

8. Name a population and a variable that might be measured. Speculate on whether the distribution would be normal or skewed.

For Students Who Are Planning Research

9. Do you anticipate that any of your distributions will be highly skewed? Explain.
TOPIC 44  THE MEAN, MEDIAN, AND MODE

The most frequently used average is the \textbf{mean}, which is the \textit{balance point} in a distribution. Its computation is simple—just sum (add up) the scores and divide by the number of scores. The most common symbol for the mean in academic journals is \( M \) (for the mean of a population) or \( m \) (for the mean of a sample). The symbol preferred by statisticians is \( \bar{X} \) which is pronounced "X-bar."

Because the mean is very frequently used as the average, let's consider its \textit{formal definition}, which is \textit{the value around which the deviations sum to zero}. You can see what this means by considering the scores in Table 1. When we subtract the mean of the scores (which is 4.0) from each of the other scores, we get the deviations (whose symbol is \( x \)). If we sum the deviations, we get zero, as shown in Table 1.

<table>
<thead>
<tr>
<th>( X ) minus</th>
<th>( M ) equals</th>
<th>( x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.0</td>
<td>-3.0</td>
</tr>
<tr>
<td>1</td>
<td>4.0</td>
<td>-3.0</td>
</tr>
<tr>
<td>1</td>
<td>4.0</td>
<td>-3.0</td>
</tr>
<tr>
<td>2</td>
<td>4.0</td>
<td>-2.0</td>
</tr>
<tr>
<td>2</td>
<td>4.0</td>
<td>-2.0</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>4.0</td>
<td>2.0</td>
</tr>
<tr>
<td>7</td>
<td>4.0</td>
<td>3.0</td>
</tr>
<tr>
<td>8</td>
<td>4.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

\textbf{Sum of the deviations (}x\textbf{) = 0.0

Note that if you take \textit{any set of scores}, compute their mean, and follow the steps in Table 1, the sum of the deviations will always equal zero.\(^1\)

Considering the formal definition, you can see why we also informally define the mean as the \textit{balance point} in a distribution. The positive and negative deviations balance each other out.

A major drawback of the mean is that it is drawn in the direction of extreme scores. Consider the following two sets of scores and their means.

\textbf{Scores for Group A:} 1, 1, 1, 2, 3, 6, 7, 8, 8  
\( M = 4.11 \)

\textbf{Scores for Group B:} 1, 2, 2, 3, 4, 7, 9, 25, 32  
\( M_f = 9.44 \)

Notice that in both sets there are nine scores and the two distributions are very similar except for the scores of 25 and 32 in Group B, which are much higher than the others and, thus, create a skewed distribution. (To review skewed distributions, see Topic 43.) Notice that the two very high scores have greatly pulled up the mean for Group B; in fact, the mean for Group B is more than twice as high as the mean for Group A because of the two high scores.

When a distribution is highly skewed, we use a different average, the \textbf{median}, which is defined as the \textit{middle score}. To get an \textit{approximate median}, put the scores in order from low to high as they are for Groups A and B above, and then count to the middle. Since there are nine scores in Group A, the median (middle score) is 3 (five scores up from the bottom). For Group B, the median (middle score) is 4 (five scores up from the bottom), which is more representative of the center of this skewed distribution than the mean, which we noted was 9.44. Thus, one use of the median is to describe the average of skewed distributions. Another use is to describe the average of ordinal data, which we'll explore in Topic 46.

A third average, the \textbf{mode}, is simply the \textit{most frequently occurring score}. For Group B, there are more scores of 2 than any other score; thus, 2 is the mode. The mode is sometimes used in informal reporting but is very seldom used in formal reports of research.

Because there is more than one type of average, it is vague to make a statement such as, "The average is 4.11." Rather, we should indicate the specific type of average being reported with statements such as, "The mean is 4.11."

\(^1\)It might be slightly off from zero if you use a rounded mean such as using 20.33 as the mean when its precise value is 20.3333333333.
Note that a synonym for the term *averages* is **measures of central tendency.** Although the latter is seldom used in reports of scientific research, you may encounter it in other research and statistics textbooks.

**EXERCISE ON TOPIC 44**

1. Which average is defined as the *most frequently occurring score*?
2. Which average is defined as the *balance point* in a distribution?
3. Which average is defined as the *middle score*?
4. What is the formal definition of the mean?
5. How is the mean calculated?
6. Should the mean be used for highly skewed distributions?
7. Should the median be used for highly skewed distributions?
8. What is a synonym for the term *averages*?

**Question for Discussion**

9. Suppose a fellow student gave a report in class and said, "The average was 25.88." For what additional information should you ask? Why?

**For Students Who Are Planning Research**

10. Do you anticipate calculating measure(s) of central tendency? If so, which one(s) are you likely to use? Explain your choice(s).
TOPIC 45 THE MEAN AND STANDARD DEVIATION

Often, a distribution of scores is described with only two statistics: the mean to describe its average, and the standard deviation (whose symbol is S or SD for a population, and s or sd for a sample) to describe its variability. What do we mean by variability? It refers to the amount by which subjects vary or differ from each other. Let’s see what this means by considering three groups, all of which have the same mean but different standard deviations.

**Group A:** 0, 5, 10, 15, 20, 25, 30  
* M = 15.00, S = 10.00

**Group B:** 14, 14, 14, 15, 16, 16, 16  
* M = 15.00, S = 0.93

**Group C:** 15, 15, 15, 15, 15, 15, 15  
* M = 15.00, S = 0.00

Although the three groups are the same on the average, as indicated by the mean, they are very different in terms of variability. Notice that the differences among the scores of Group A (a score of 0 vs. a score of 5 vs. a score of 10 vs. a score of 15, etc.) are much greater than the differences among the scores of Group B (a score of 14 vs. a score of 14 vs. a score of 14 vs. a score of 15, etc.). At the extreme, when all the scores are the same, as in Group C, the standard deviation equals zero. Thus, you can see, the smaller the standard deviation, the smaller the variability.¹

The standard deviation has a special relationship to the normal curve (see Topic 43). If a distribution is normal, 68% of the subjects in the distribution lies within one standard deviation unit of the mean.² For example, if you read in a report that *M*=70 and *S*=10 for a normal distribution, you would know that 68% of the subjects have scores between 60 and 80 (that is, 70 - 10 = 60 and 70 + 10 = 80). This is illustrated in Figure 1.

In Figure 2, the mean is also 70, but the standard deviation is only 5. The smaller standard deviation in Figure 2 is reflected by the fact that the curve is narrower than in Figure 1. Yet, in both distributions, 68% of the cases lies within one standard deviation unit of the mean because they are both normal.

¹For those of you who are mathematically inclined, the computation of the standard deviation is illustrated in Appendix F. Considering how it is computed may give you a better feeling for its meaning.

²Note that within means on both sides of the mean, that is, the standard deviation plus/minus the mean.
EXERCISE ON TOPIC 45

1. Which average is usually reported when the standard deviation is reported?

2. What is meant by the term variability?

3. Is it possible for two groups to have the same mean but different standard deviations?

4. If everyone in a group has the same score, what is the value of the standard deviation for the scores?

5. What percentage of the subjects lies within one standard deviation unit of the mean in a normal distribution?

6. The middle 68% of the subjects in a normal distribution has scores between what two values if the mean equals 100 and the standard deviation equals 15?

7. If the mean of a normal distribution equals 50 and the standard deviation equals 5, what percentage of the subjects has scores between 45 and 50?

8. Does the 68% rule strictly apply if a distribution is not normal?

9. If the standard deviation for Group X is 14.55 and the standard deviation for Group Y is 20.99, which group has less variability in their scores?

10. Which group in question 9 has a narrower curve?

Question for Discussion

11. Examine a journal article in which a mean and standard deviation are reported. Does the author indicate whether the distribution is normal in shape? Does the 68% rule strictly apply? Explain.

For Students Who Are Planning Research

12. Will you be reporting means and standard deviations? Explain.
As you know from Topic 44, the **median** is the middle score in a distribution. Being in the middle, it always has 50% of the scores above it and 50% of the scores below it. For the scores in Figure 1, the median is 6.5 (halfway between the middle two scores of 6 and 7).

![Figure 1 Scores for Group A and their median](image)

The median is used instead of the mean when a distribution is highly skewed (see Topic 43). It is also used to describe the average of a set of ordinal data (that is, data that put subjects in order from high to low but do not have equal intervals among them; see Topic 39).

When the **median** is reported as the average, it is customary to report the range or interquartile range as a measure of variability. You should recall from Topic 45 that **variability** refers to the amount by which subjects vary or differ from each other.

The **range** is simply the highest score minus the lowest score. For the scores shown above, it is 12-1 = 11 (or we can say that the scores range from 1 to 12). For reasons that are beyond the scope of this discussion, measurement theory tells us that the more extreme the score, the more unreliable it is. Since the range is based on the two most extreme scores, it is an unreliable statistic. To get around this problem, we often use a modified version of the range—called the **interquartile range**. Since inter- means between and -quartile refers to quarters, interquartile range refers to the range between quarters. To find it, first divide the distribution into quarters, as shown in Figure 2. As you can see, the middle 50% of the subjects is between the values of 2.5 and 9.5. Since the formal definition of the interquartile range (IQR) is the range of the middle 50% of the subjects, we can calculate it as follows: 9.5 - 2.5 = 7.0. Thus, IQR = 7.0.

| 1, 1, 1, 2, 3, 4, 5, 6, 7, 8, 8, 9, 10, 11, 11, 12 | ↑ | ↑ | ↑ |
| 2.5 | 6.5 | 9.5 |

**Figure 2 Scores for Group A divided into quarters**

To see how the IQR helps us understand the variability in sets of data, let’s consider Figure 3, where the median is 43.0 and the interquartile range is 71.0 (83.5 - 12.5 = 71.0).

| 0, 5, 7, 10, 15, 22, 30, 41, 45, 57, 67, 78, 89, 92, 95, 99 | ↑ | ↑ | ↑ |
| 12.5 | 43.0 | 83.5 |

**Figure 3 Scores for Group B divided into quarters**

Thus, for Groups A and B, we might find the data presented in a research report as illustrated in Table 1.

| Table 1 Medians and interquartile ranges for two groups |
|----------------|----------------|----------------|
| Group | Median | Interquartile range |
| A     | 6.5    | 7.0             |
| B     | 43.0   | 71.0            |

As you can see, Table 1 indicates two things. First, Group A has a lower average than Group B (as indicated by the median of 6.5 for Group A vs. 43.0 for Group B). Second, Group A has less variability (as indicated by an interquartile range of 7.0 for Group A vs. 71.0 for Group B). Reconsideration of the scores shown in Figures 2 and 3 indicates that these results make sense. The middle score for Group A is much lower than the middle
score for Group B, and the differences among the scores for Group A (1 vs. 1 vs. 1 vs. 1 vs. 2, etc.) are much smaller than the differences among the scores for Group B (0 vs. 5 vs. 7 vs. 10 vs. 15, etc.)—indicating less variability in Group A than Group B.

EXERCISE ON TOPIC 46

1. If the median for a group of subjects is 34.00, what percentage of the subjects has scores below a score of 34?

2. Should the mean or median be used with ordinal data?

3. How do you calculate the range of a set of scores?

4. Is the range or the interquartile range a more reliable statistic?

5. What is the definition of the interquartile range?

6. Suppose you read that for Group X, the median equals 55.1 and the IQR equals 30.0, while for Group Y, the median equals 62.9 and the IQR equals 25.0. Which group has the higher average score?

7. Based on the information in question 6, the scores for which group are more variable?

8. Which two statistics discussed in this topic are measures of variability?

9. Which two statistics mentioned in this topic are measures of central tendency (that is, are averages)?

Question for Discussion

10. Name two circumstances under which the median is preferable to the mean.

For Students Who Are Planning Research

11. Do you anticipate reporting medians and interquartile ranges? Explain.
TOPIC 47  THE PEARSON CORRELATION COEFFICIENT

When we want to examine the relationship between two quantitative sets of scores (at the interval or ratio levels), we compute a correlation coefficient. The most widely used coefficient is the Pearson product-moment correlation coefficient, whose symbol is $r$. It is usually called simply Pearson's $r$.

Consider again the scores in Table 1, which we considered in Topic 25. As you can see, the employment test scores put subjects in roughly the same order as the ratings by supervisors. In other words, those who had high employment test scores (such as Joe and Jane) tended to have high supervisors' ratings, and those who had low test scores (such as John and Jake) tended to have low supervisors' ratings. This illustrates what we mean by a direct relationship (also called a positive relationship).

<table>
<thead>
<tr>
<th>Employee</th>
<th>Employment Test Scores</th>
<th>Supervisors' Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe</td>
<td>35</td>
<td>9</td>
</tr>
<tr>
<td>Jane</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>Bob</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>June</td>
<td>27</td>
<td>8</td>
</tr>
<tr>
<td>Leslie</td>
<td>25</td>
<td>7</td>
</tr>
<tr>
<td>Homer</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>Milly</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>Jake</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>John</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1  Direct relationship, $r = .89$

Note that the relationship in Table 1 is not perfect. For example, although Joe has a higher employment test score than Jane, Jane has a higher supervisor's rating than Joe. If the relationship were perfect, the value of the Pearson $r$ would be 1.00. Being less than perfect, its actual value is .89. As you can see in Figure 1, this value indicates a strong, direct relationship.

In an inverse relationship (also called a negative relationship), those who are high on one variable are low on the other. Such a relationship exists between the scores in Table 2. Those who are high on self-concept (such as Joe and Jane) are low on depression while those who are low on self-concept (such as Jake and John) are high on depression. However, the relationship is not perfect. The value of the Pearson $r$ for the relationship in Table 2 is -.86.

<table>
<thead>
<tr>
<th>Employee</th>
<th>Self-Concept Scores</th>
<th>Depression Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Jane</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Bob</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>June</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Leslie</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Homer</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Milly</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Jake</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>John</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

The relationships in Tables 1 and 2 are strong but, in each case, there are exceptions, which make the Pearson rs less than 1.00 and -.100. As the number and size of the exceptions increase, the values of the Pearson $r$ become closer to 0.00. Therefore, a value of 0.00 indicates the complete absence of a relationship. (See Figure 1.)

It is important to note that a Pearson $r$ is not a proportion and cannot be multiplied by 100 to get a percentage. For instance, a Pearson $r$ of .50 does not correspond to 50% of anything. To think about correlation in terms of percentages, we must

Figure 1  Values of the Pearson $r$
convert Pearson rs to another statistic, the coefficient of determination, whose symbol is \( r^2 \), which indicates how to compute it—simply square \( r \). Thus, for an \( r \) of .50, \( r^2 \) equals .25. If we multiply .25 by 100, we get 25%. What does this mean? Simply this: A Pearson \( r \) of .50 is 25% better than a Pearson \( r \) of 0.00. Table 3 shows selected values of \( r \), \( r^2 \), and the percentages you should think about when interpreting an \( r \).

<table>
<thead>
<tr>
<th>( r )</th>
<th>( r^2 )</th>
<th>Percentage better than zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>.90</td>
<td>.81</td>
<td>81%</td>
</tr>
<tr>
<td>.50</td>
<td>.25</td>
<td>25%</td>
</tr>
<tr>
<td>.25</td>
<td>.06</td>
<td>6%</td>
</tr>
<tr>
<td>-.25</td>
<td>.06</td>
<td>6%</td>
</tr>
<tr>
<td>-.50</td>
<td>.25</td>
<td>25%</td>
</tr>
<tr>
<td>-.90</td>
<td>.81</td>
<td>81%</td>
</tr>
</tbody>
</table>

Also called percentage of variance accounted for or percentage of explained variance.

EXERCISE ON TOPIC 47

1. "Pearson r" stands for what words?

2. When the relationship between two variables is perfect and inverse, what is the value of \( r \)?

3. Is it possible for a negative relationship to be strong?

4. Is an \( r \) of -.90 stronger than an \( r \) of .50?

5. Is a relationship direct or inverse when those with high scores on one variable have high scores on the other and those with low scores on one variable have low scores on the other?

6. What does an \( r \) of 1.00 indicate?

7. For a Pearson \( r \) of .60, what is the value of the coefficient of determination?

8. What do we do to a coefficient of determination to get a percentage?

9. A Pearson \( r \) of .70 is what percentage better than a Pearson \( r \) of 0.00?

Question for Discussion

10. Name two variables between which you would expect to get a strong, positive value of \( r \).

For Students Who Are Planning Research

11. Will you be reporting Pearson rs? If so, name the two variables that will be correlated for each value of \( r \).

'Note that the procedure for computing a Pearson \( r \) is beyond the scope of this book.
Suppose we have a research hypothesis that says "homicide investigators who take a short course on the causes of HIV will be less fearful of the disease than investigators who have not taken the course," and test it by conducting an experiment in which a random sample of investigators is assigned to take the course and another random sample is designated as the control group. Let's suppose that at the end of the experiment the experimental group gets a mean of 16.61 on a fear of HIV scale and the control group gets a mean of 29.67 (where the higher the score, the greater the fear of HIV). These means support our research hypothesis. But can we be certain that our research hypothesis is correct? If you've been reading the topics on statistics in order from the beginning, you already know that the answer is "no" because of the null hypothesis, which says that there is no true difference between the means; that is, the difference was created merely by the chance errors created by random sampling. (These errors are known as sampling errors) Put another way, unrepresentative groups may have been assigned to the two conditions quite at random.

The $t$ test is often used to test the null hypothesis regarding the observed difference between two means. For the example we are considering, a series of computations (which are beyond the scope of this book) would be performed to obtain a value of $t$ (which, in this case, is 5.38) and a value of degrees of freedom (which, in this case, is $df = 179$). These values are not of any special interest to us except that they are used to get the probability ($p$) that the null hypothesis is true. In this particular case, $p$ is less than .05. Thus, in a research report, you may read a statement such as this:

The difference between the means is statistically significant ($t = 5.38, df = 179, p < .05$).

As you know from Topic 38, the term statistically significant indicates that the null hypothesis has been rejected. You should recall that when the probability that the null hypothesis is true is .05 or less (such as .01 or .001), we reject the null hypothesis. (When something is unlikely to be true because it has a low probability of being true, we reject it.)

Having rejected the null hypothesis, we are in a position to assert that our research hypothesis probably is true (assuming no procedural bias was allowed to affect the results, such as testing the control group immediately after a major news story on a famous person with AIDS, while testing the experimental group at an earlier time).

What leads a $t$ test to give us a low probability? Three things:

1. **Sample size.** The larger the sample, the less likely that an observed difference is due to sampling errors. (You should recall from the sections on sampling that larger samples provide more precise information.) Thus, when the sample is large, we are more likely to reject the null hypothesis than when the sample is small.

2. **The size of the difference between means.** The larger the difference, the less likely that the difference is due to sampling errors. Thus, when the difference between the means is large, we are more likely to reject the null hypothesis than when the difference is small.

3. **The amount of variation in the population.** You should recall from Topic 22 that when a population is very heterogeneous (has much variability) there is more potential for sampling error. Thus, when there is little variation (as indicated by the standard deviations of the sample), we are

You probably recall that we prefer random sampling because it precludes any bias in the assignment of subjects to the groups and because we can test for the effect of random errors with significance tests; we cannot test for the effects of bias.

To test the null hypothesis between two medians, the median test is used; it is a specialized form of the chi square test, whose results you already know how to interpret.

Sometimes researchers leave out the abbreviation $df$ and present the result as $f(179) = 5.38, p < .05$. 

113
more likely to reject the null hypothesis than when there is much variation.

A special type of $t$ test is also applied to correlation coefficients. Suppose we drew a random sample of 50 students and correlated their hand size with their GPAs and got an $r$ of .19. The null hypothesis says that the true correlation in the population is 0.00—that we got .19 merely as the result of sampling errors. For this example, the $t$ test indicates that $p > .05$. Since the probability that the null hypothesis is true is greater than 5 in 100, we do not reject the null hypothesis; we have a statistically insignificant correlation coefficient. (In other words, for $n = 50$, an $r$ of .19 is not significantly different from an $r$ of 0.00.) When reporting the results of the $t$ test for the significance of a correlation coefficient, it is conventional not to mention the value of $t$. Rather, researchers usually indicate only whether or not the correlation is significant at a given probability level.

**EXERCISE ON TOPIC 48**

1. What does the null hypothesis say about the difference between two sample means?

2. Is the value of $t$ usually of any special interest to consumers of research?

3. Suppose you read that for the difference between two means, $t = 2.000$, $df = 20$, $p > .05$. Using conventional standards, should you conclude that the null hypothesis should be rejected?

4. Suppose you read that for the difference between two means, $t = 2.859$, $df = 40$, $p < .01$. Using conventional standards, should you conclude that the null hypothesis should be rejected?

5. Based on the information in question 4, should you conclude that the difference between the means is statistically significant?

6. When we use a large sample, are we more or less likely to reject the null hypothesis than when we use a small sample?

7. When the size of the difference between means is large, are we more or less likely to reject the null hypothesis than when the size of the difference is small?

8. If we read that for a sample of 92 subjects, $r = .41$, $p < .001$, should we reject the null hypothesis?

9. Is the value of $r$ in question 8 statistically significant?

**Question for Discussion**

10. Of the three things that lead to a low probability, which one is most directly under the control of a researcher?

**For Students Who Are Planning Research**

11. Will you be conducting $t$ tests? Explain.