

8

The Basics of Statistical Analysis

While this chapter does not presuppose sophistication in statistical analysis, and is not designed to substitute for formal training, we do try to summarize *why* statistical analysis is necessary for testing hypotheses cross-culturally. And we do discuss the basic tools that you will need to understand research that includes simple statistical results, tables, and figures. In chapter 10 we discuss more complex statistical analyses involving multiple variables. Statistics cannot really be taught in two chapters of a book, but we try.

Descriptive statistics and inferential statistics have very different purposes. **Descriptive statistics** are used to summarize data to make them comprehensible. **Inferential statistics** allow researchers to generalize their sample results to a larger universe, assuming that the sampling is unbiased and the research design is appropriate. Let us first turn to some basic ways of summarizing data.

Descriptive Statistics

Consider the example of a classroom test taken by 131 students. One student asks how the class did in general. Suppose the professor were to hand out a piece of paper with 131 numbers on

Table 8.1. Summary of Classroom Grades
(frequency distribution and percentages)

<i>Scores</i>	<i>Frequency</i>	<i>Percentage</i>
90–100	20	15.3
80–89	42	32.1
70–79	53	40.4
60–69	11	8.4
<60	3	3.8
Total	131	100.0

it ordered in the sequence in which the individual papers were graded. Such a list would be hardly comprehensible. To be sure, the list would answer the question of how the class performed, but it doesn't answer the question in any clear way. Probably the real intent of the student's question was to find out how he or she did compared to everyone else. Most professors therefore usually put a frequency distribution on the board, showing how often a group of scores occurred. A frequency distribution provides a count of the number of people who got a particular score or one of a group of scores (A, B, C, etc., or a numerical range). When there are many possible scores, such as on a test with a hundred points, the professor could group the scores as shown in table 8.1. Looking down the frequency column tells you at a glance that more people got scores between 70 and 79 than any other range of scores. An alternative is to give percentages (see the "Percentage" column in table 8.1), which are just the frequency counts divided by the number of cases and multiplied by one hundred. A more graphic way of showing the information would be to show the frequency distribution as a bar chart as shown in figure 8.1. The vertical axis shows the frequency or number of people in each group of scores on the horizontal axis.

A measure of **central tendency** summarizes the data more succinctly. Such a measure conveys the center of the distribution with one number. We are very used to one measure of central tendency—what we call the average (the **mean** in the language of statistics). We are used to computing averages. You just add

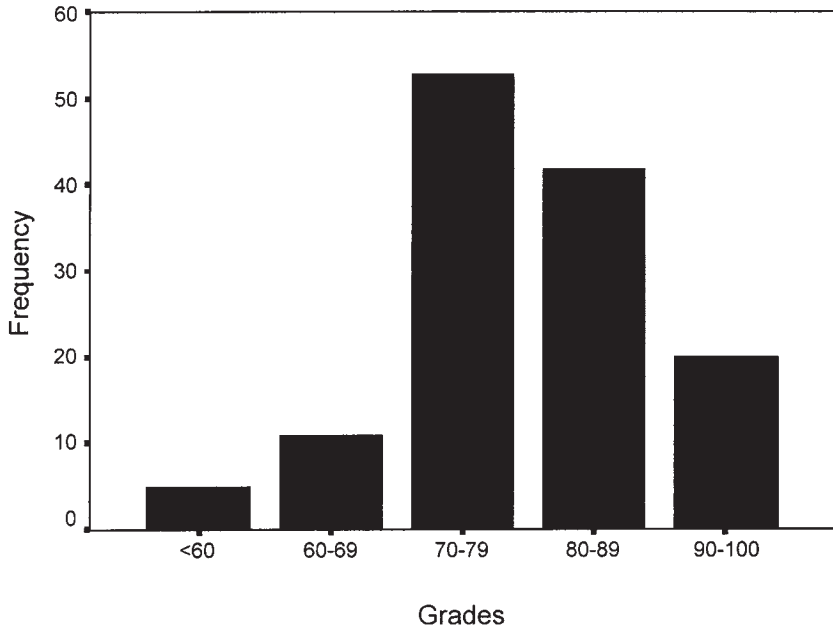


Figure 8.1. Bar graph of grade distribution

up all the scores and divide by the number of cases. For the data shown in table 8.1 and figure 8.1, the average score is 78.63. There are two other common measures of central tendency—the **median** and the **mode**. The *median* score is that number below and above which 50 percent of the scores fall. In other words, if we rank-order the scores, the median would be the score of the middle case. If we rank-order the actual grades behind the grouped scores in table 8.1, the median score is 79. The *mode* is the score with the highest frequency. If there are many different scores, the test score with the highest frequency doesn't make much sense to compute. But if we group the scores as shown in table 8.1 and figure 8.1, the modal group (with fifty-three scores) is the 70–79 range.

In the grade example just discussed, the mean, median, and grouped mode are almost identical. Why would we use one measure of central tendency over another to describe the center? The answer is that in most roughly symmetrical distributions,

Table 8.2. Yearly Income in Nine Households in Each of Three Communities

	Community 1		Community 2		Community 3	
	11,000		9,000		7,700	
	12,500		14,000		6,900	
	14,000		13,000		8,500	
	19,500		19,000		7,500	
	20,000		14,500		23,000	
	9,000		12,500		57,000	
	10,000		10,000		67,000	
	11,500		16,000		59,000	
	200,000		20,000		55,000	
Mean	34,167	Mean	14,222	Mean	32,400	
Median	12,500	Median	14,000	Median	23,000	

the measure of central tendency doesn't much matter. But it matters a lot if the distribution is skewed or not balanced, as the following example shows.

Let us suppose that we are interested in the typical income in three different communities. Table 8.2 shows the income of nine households in each community. (We normally would look at more households, but we are only looking at nine here to make the point easier to see.) Suppose we compute the mean (average) and the median scores for each community (see table 8.2). Let's look at the mean scores first. If we look only at the mean scores, we might mistakenly infer that the "typical" household in community 1 has a higher income (mean = \$34,167) than the "typical" household in community 2 (mean = \$14,222). So why do the median scores not suggest the same kind of difference? The median scores suggest that the "typical" household in community 1 makes \$12,500. The two measures of central tendency are quite different (about \$24,000 different)!

It is important to recognize then that the three measures of central tendency may work differently in different circumstances. If we look at the actual numbers again, we notice that one household in community 1 (the last row in the community 1 column) has an income well beyond any other household in the community. That household earns \$200,000, which is at least

\$180,000 more than the others; none of the others earns more than \$20,000. The mean is high for community 1 because when you compute a sum with one or a few very extreme numbers, the sum is heavily influenced by those numbers. The mean is analogous to the center of gravity (Senter 1969, 63–66). We all have had the experience of balancing on a seesaw with a friend about our size. What if we try to balance with a much heavier older child or a parent? We get pushed way up in the air. The only way to balance the seesaw is for the bigger person to move close to the balance point. (The lighter person could hardly move back!) If we think of the mean as being at the center of gravity or the point of balance, the mean will be influenced by a very skewed distribution that pulls it toward the extreme score(s) on one end. Notice that the median, which is the “middle” case, is not influenced by an extreme score. With regard to community 1, four households have more income than \$12,500 (\$14,000, 19,500, 20,000 and 200,000) and four households have less. It doesn’t matter whether the highest number is \$200,000 or \$21,000—the median remains the same. So, with a very skewed distribution, the median may give a better indication of where the center is. There is one circumstance where the median (as well as the mean) may be misleading. That instance is where there are few or no cases in the middle. Look at figure 8.2, which shows the distribution of income in community 3, grouped into \$5,000 ranges (from the data in table 8.2). The median for income in community 3 is \$23,000, but most households have either much lower income or much higher income. In this instance, we would be better off describing the shape of the distribution. This kind of distribution would be called bimodal; there are four households with very low incomes (\$5,000–\$9,999) and an equal number of households with incomes above \$45,000. A researcher with a bimodal distribution would describe it more accurately in terms of modes than in terms of the mean or median. (Even distributions that did not have exactly equal modes would still be discussed as bimodal if the pattern looked similar to that shown in figure 8.2.)

So far the scales we have summarized are interval and ratio scales. Recall that we said in chapter 4 that nominal scales only

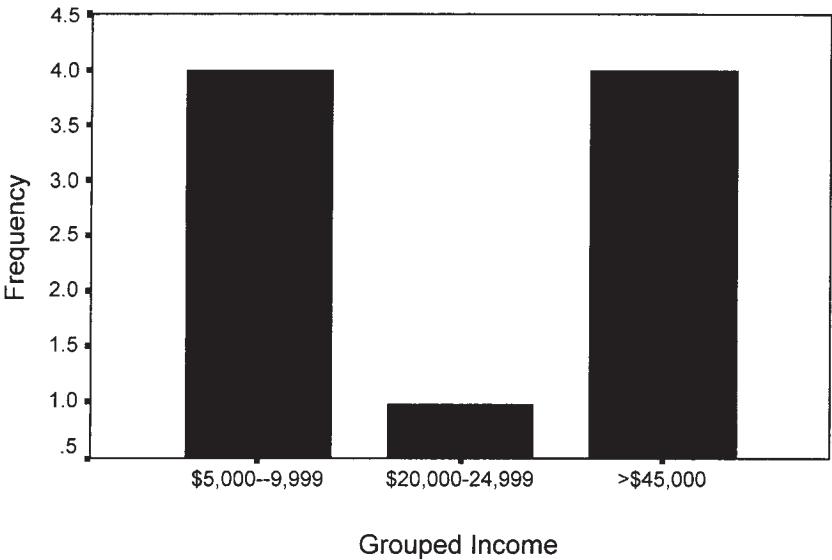


Figure 8.2. Distribution of income in community 3

convey difference and ordinal scales only convey order. Means and median scores are best reserved for interval and ratio scales. If you wish to summarize nominal scores or grouped ordinal scores for a set of cases, it is best to use frequencies or percentage summaries. A modal score can also be given for the scale score with the highest frequency.

One other common kind of descriptive statistic is a *measure of variability*. This kind of measure helps us understand how much the scores are spread out. Two classes can have average scores of 75 on a test. But in one class grades range from 20 to 100, while in the other class the grades range from 69 to 81. What we want is a measure conveying that the variability is wide in one class, but narrow in the other. The simplest measure of variability is a **range**. You can describe the grades in the first class as having a mean of 75 with a range of 80 points (100–20); the second class as having a mean of 75 with a range of 12 points (81–69). The main trouble with a range is that it is defined solely by the highest and lowest scores. If almost every score is clustered in the middle except for one or a few, the range would suggest a lot

more average variability than exists. A better measure of variability is to use the distance of every single score from the center and calculate some kind of average distance from the center. The most common such measures in statistics are the *variance* and the *standard deviation*. The variance is calculated as follows. The mean is subtracted from every score and the difference is squared. Then the squared distances are summed and the total is divided by the number of cases.¹ The result is the *variance*. If you take the square root of this number, you have the **standard deviation**. Why are the distances squared? Couldn't you just sum the distances? No, because if you subtract the mean from every score, some numbers will be positive and some will be negative; and if you sum them, they will cancel each other out. Squaring is one way to get rid of negative numbers before you sum them. To transform the variance to the original scale, you need to take the square root of the variance. Let's look again at the income variability of communities 1, 2, and 3, using the standard deviation. First look at table 8.2. Intuitively, community 2 looks like it has the tightest cluster of income. Community 3 is pretty variable because four households make less than \$9,000 and four make at least \$55,000. Community 1 is more variable because one household (with \$200,000) is far different from the others. Now let us compare the standard deviations for the three:

Standard Deviation

Community 1	62,310.21
Community 2	3,692.15
Community 3	38,692.14

The standard deviations convey the degree of variability we intuited. Community 2 has the least variability, community 3 the next, and community 1 has the most.

There is one other kind of descriptive statistic that is important—summarizing how strongly variables are related to one another. Because such summaries are usually calculated in the context of inferring a result to a larger universe of cases, we discuss them in the next section.

Inferential Statistics

When researchers look at a sample of cases, they usually are not just looking to describe what they observe. They usually want to be able to generalize to a wider set of cases. When political pollsters ask a sample of individuals whom they will vote for, they usually are not content to say that politician X wins in this sample. They are interested in predicting how all the voters will vote in the election; they want to be able to predict the winner in the election. Pollsters know that they cannot predict exactly, but they want to predict correctly within a small margin of error. When researchers are looking for a vaccine to prevent a deadly disease, they have to use clinical trials to see if the vaccinated are more likely to resist the disease than the unvaccinated, and then they can recommend use of the vaccine in the larger population. When we ask whether hunter-gatherers are more peaceful than agriculturalists, we are asking about the situation generally; we don't just mean among the cases we are looking at.

While we never can know for sure what is true in the larger set of cases, because it would be too time-consuming or expensive to find out, most inferential statistical techniques rest on a few very simple principles:

- It is assumed that the sample you have chosen is unbiased. The best way to achieve that is to select the sample cases randomly from the larger universe of cases you wish to generalize to (this is why sampling strategies are so important).
- A statistical “test of significance” will evaluate how likely it is that your result (or a better result) is due solely to chance. This is like playing devil's advocate. The reasoning goes like this. Let's pretend there is no difference or no relationship. What is the probability of getting the results you got (and better results) by accident? The lower the chance, the more we can believe that the result from the sample is correct.

These principles are simple. But learning statistics is complicated because different kinds of tests make different assumptions about the data, and different types of measures require different types of tests. While computer programs make computations easy, they do not readily tell you whether a particular type of test is appropriate. If you give numbers for your variables to your computer, it will compute anything, even if it is not appropriate.

Tests about Relationships between Two Nominal Variables

In chapter 1, we introduced the notion of contingency tables, which we discussed further in chapter 5. Such tables are useful for displaying how the sample cases are distributed in a cross-tabulation of the two nominal or categorical variables. Contingency tables may also be used to look at a relationship between two ordinal variables (that convey more or less of some quality or quantity), if there aren't more than five or so scale points on each variable. And contingency tables can be used to examine an association between one nominal and one ordinal variable. The simplest contingency table is referred to as a *two by two table* (two rows and two columns). If there were a perfect relationship between the two dichotomous variables, we would expect all the cases to fall into the two cells on one diagonal, as in the first table shown in box 8.1.

Box 8.1. Computing Exact Probabilities (Fisher's Exact Test)

To make it easy to compute the probability of a particular result, we are deliberately going to use a very small sample. Suppose we randomly draw two hunting-gathering societies and two agricultural societies. Our hypothesis is that the hunter-gatherers will tend to lack permanent settlements. *The alternative hypothesis, which the statistical test evaluates, is that there is no relationship between hunting-gathering and lack of permanent settlements.* Fisher's Exact Test computes the probability that our observed result is due to chance.

First, look at the result displayed immediately below. At first glance, the result looks perfect. Look at the hunter-gatherer (HG) row. Two of the two hunter-gatherer cases lack permanent settlements. And in the agricultural row, two of the two agricultural cases have permanent settlements. There are no exceptions.

Box 8.1. (continued)

	No Permanent Settlements	Permanent Settlements	Total
HG	2	0	2
Agric	0	2	2
Total	2	2	4

But is this result something we can trust, or is it due to chance? Here's where we can compute the chance of this result or a better one (in this case there is no better one) occurring by chance and chance alone. We see how many different ways the cases could fall by chance. There is only one constraint in this test. We must keep the totals the same. The totals, called the **marginals**, are bolded. In other words, we must have two cases fall in the hunter-gatherer row, two in the agricultural row, and two cases in each column.

Let us label the two hunter-gatherer cases "A" and "B" and the two agricultural cases "C" and "D." With the marginal constraints above, how many ways could the cases fall?

The table below shows each of the possible tables on the left (in bold). To the immediate right of each table are the possible combinations of A, B, C, and D that can make up this combination. Notice that for the perfect table we saw above, there is only one way the table can occur. But for the table below (1, 1, 1, 1), with one case in each cell, there are four different ways the cases could be distributed. The last table, which is directly opposite to the hypothesis expected, could only occur in one way.

2	0	AB	
0	2	CD	1

1	1	A	B	
1	1	C	D	
		A	B	
		D	C	
		B	A	
		C	D	4
		B	A	
		D	C	
0	2		AB	
2	0	CD		1

Total **6**

So, to return to the original question, what is the probability that the table we got at the very top of this box or a better one would occur by chance and chance alone? Since there are total of six ways that these four cases could distribute themselves (add up all the A, B, C, D tables), the probability of the top table occurring by chance is one out of six or 0.167. We would write under the following phrase under our observed table the phrase:

$$p = 0.167, \text{ one tail, by Fisher's Exact Test}$$

What the phrase means is: The probability of this strong a result or a stronger result occurring by chance in this direction (which is what the one-tail means) is one out of six or about seventeen times out of a hundred. The method of computation is Fisher's Exact Test. (For larger samples, you can use a formula or look up the exact probability in a published table.)

What does the "tail" mean exactly? Notice that when we figured out the various combinations, there were two perfect tables—the top and the bottom ones. If we started out with a hypothesis that did not predict a direction, that is, "Permanence of settlements is related to type of subsistence (hunting-gathering versus agriculture)," the appropriate probability would be to add up the chances of this or better outcomes. So, we would add the probabilities 0.167 and 0.167 for the two directions. Then we would say $p = 0.33$, two tails, by Fisher's Exact Test.

Displaying a contingency table is not enough. We can look at the first table in box 8.1 and say that the relationship looks perfect because all the cases are on the diagonal. But in this table there are only four cases and we might suspect that the result could be due to chance. Box 8.1 takes you through the steps of computing the probability that this "perfect" result is due to chance.

We now have our probability; it is $p = 0.167$. So is our hypothesis supported? To decide whether we accept or reject our hypothesis, we have to have a decision rule as to whether we will accept our hypothesis or not. It is best to set this rule *before* conducting your test. There is no right or wrong p -value. If it is a matter of life and death where you can't afford to be wrong, you might want to see a very low chance of an accidental association. However, when you make it harder to be wrong, you also make it harder to find associations that are probably true. Most social scientists accept the convention that if the probability of

the result occurring by chance is *less than or equal to* 0.05, we can accept the hypothesis. If the p -value is 0.05 or lower, we say that the association is **statistically significant**. If the p -value is greater than 0.05, most researchers would reject the hypothesis. Given this convention about what to conclude from p -values, we would say that the relationship displayed near the top of box 8.1 is not significant, because the p -value is *higher* than 0.05. Instead of writing the exact probability under our table we would probably write:

$$p > 0.05, \text{ one tail, by Fisher's Exact Test}$$

However, keep in mind that the example we are considering involves only four cases. Realistically, we would not test this hypothesis with four cases because it is not possible to get a statistically significant result with only four sample cases in a contingency table. The minimum number of cases in a contingency table that might provide a significant result is six, but the result has to be perfect. (If the table had been 3, 0, 0, 3, the p -value would have been 0.05 because there are twenty possible ways the six cases could be distributed and only one way the 3, 0, 0, 3 table could occur; hence the p -value would be one divided by twenty or 0.05.) If we really wanted to give this hypothesis a chance, we would probably choose a random sample of at least twenty cases, because there are almost always exceptions (cases that fall into the unpredicted cells because of measurement error, cultural lag, other causes, etc.). There are formulas and tables for looking up exact probabilities according to Fisher's Exact Test for two by two tables with up to fifty cases. But most researchers would use the **chi-square (χ^2) test** for the significance of a contingency table that contains more than twenty or thirty cases. This is assuming that the chance-expected values (the values that would occur with no relationship) are five or more in each cell. If the expected values are large enough, chi-square tests can be calculated also when contingency tables are larger than two by two.

Sometimes we want to know more than whether two variables are probably associated. We want to know how *strongly* they are associated. Usually we are looking for strong predictors

because only together would they account for most of the cases. So, for example, if we want to understand why people in some societies go to war more often than people in other societies, we would expect a presumably causal variable to predict many cases, not just a few. Let's look now at the top two rows of table 8.3. In each row a perfect table is displayed on the left and a "nothing" table is displayed on the right. The difference between the top row and the bottom row is only in the number of cases in each table, one hundred in the top row and twenty in the bottom row. Since the number of cases in the top row is one hundred, we use a chi-square test to evaluate the statistical significance of a result. (Consult a statistics book for the formula for computing chi-square and a table in which to look up the p -values for different values of chi-square. You will also need to know the degrees of freedom in a table. A two by two table has one degree of freedom because, given the row and column totals, once you know the number in *one* cell, you can calculate all the other cell values.) It is important to compute the test of significance first because the strength of an association is not relevant if the relationship is likely to be due to chance.

Notice that the first four left tables in the top row of table 8.3 can be described as statistically significant because the chi-square test gives associated p -values of less than or equal to 0.05. The first three tables are even more significant since they would be likely to occur by chance less than one time out of one thousand.

We can intuit that the table on the left shows a strong relationship because there are no exceptions—all the cases are on one of the diagonals. What we are looking for is a measure of association that will give a high score to the table on the left and a "zero relationship" score to the table on the extreme right. There are several measures of association for two by two tables. The different measures are usually based on different assumptions and they have different mathematical properties (so you cannot compare one measure of association with another). The most important thing to realize is that measures of association usually give scores (coefficients) that range between ± 1.00 and 0.00. A coefficient of -1.00 means that the association is

Table 8.3. Strength of Association (phi) and Tests of Significance

<table><tr><td>50</td><td>0</td></tr><tr><td>0</td><td>50</td></tr></table> phi = 1.00 $\chi^2 = 100, df = 1, p < 0.001$	50	0	0	50	<table><tr><td>45</td><td>5</td></tr><tr><td>5</td><td>45</td></tr></table> phi = 0.80 $\chi^2 = 64, df = 1, p < 0.001$	45	5	5	45	<table><tr><td>35</td><td>15</td></tr><tr><td>15</td><td>35</td></tr></table> phi = 0.40 $\chi^2 = 16, df = 1, p < 0.001$	35	15	15	35	<table><tr><td>30</td><td>20</td></tr><tr><td>20</td><td>30</td></tr></table> phi = 0.20 $\chi^2 = 4, df = 1, p < 0.05$	30	20	20	30	<table><tr><td>25</td><td>25</td></tr><tr><td>25</td><td>25</td></tr></table> phi = 0.00 $\chi^2 = 0.00; df = 1, n.s.$	25	25	25	25
50	0																							
0	50																							
45	5																							
5	45																							
35	15																							
15	35																							
30	20																							
20	30																							
25	25																							
25	25																							
<table><tr><td>10</td><td>0</td></tr><tr><td>0</td><td>10</td></tr></table> phi = 1.00 $p < 0.01, \text{ two tails}$ by Fisher's Exact Test	10	0	0	10	<table><tr><td>9</td><td>1</td></tr><tr><td>1</td><td>9</td></tr></table> phi = 0.80 $p < 0.01, \text{ two tails}$ by Fisher's Exact Test	9	1	1	9	<table><tr><td>7</td><td>3</td></tr><tr><td>3</td><td>7</td></tr></table> phi = 0.40 $n.s. (p > 0.05)$ by Fisher's Exact Test	7	3	3	7	<table><tr><td>6</td><td>4</td></tr><tr><td>4</td><td>6</td></tr></table> phi = 0.20 $n.s. (p > 0.05)$ by Fisher's Exact Test	6	4	4	6	<table><tr><td>5</td><td>5</td></tr><tr><td>5</td><td>5</td></tr></table> phi = 0.00 $n.s. (p > 0.05)$ by Fisher's Exact Test	5	5	5	5
10	0																							
0	10																							
9	1																							
1	9																							
7	3																							
3	7																							
6	4																							
4	6																							
5	5																							
5	5																							

perfectly negative, that one variable goes up as the other goes down; a +1.00 coefficient means that one variable goes up as the other goes up. The direction is meaningful only when the pair of categories on each variable can be ordered (e.g., present versus absent, high versus low).

The phi (ϕ) coefficient is a commonly used measure of association for two by two tables. As you can see across the top row of table 8.3, the phi coefficients show that the strength of the relationship is weaker as you go from left to right. Notice that the tables in the second row have exactly the same proportion of cases on the diagonal as the top row tables. Not surprisingly, the phi coefficients are the same in the second row. However, only the two left-hand tables are statistically significant! We used Fisher's Exact Test because the number of cases is relatively small, but we could have used chi-square since the expected value in each cell is at least five. This tells us that we cannot rely on the measure of association to draw an inference about statistical significance. We need to do the test of significance first.

While some might infer that it is better not to have a small sample because it is harder to get significant results, we would argue that it depends upon what you are looking to accomplish. If you are looking for strong predictors, a small sample may be advantageous. First, it is less work to code the cases for a small sample. Second, only strong associations will be significant.

One drawback of the phi coefficient is that it will only go to 1.00 (or -1.00) with a symmetrical relationship. For example, in table 8.4, the totals for the rows do not match the totals for the columns and the maximum phi coefficient possible is 0.67.

This may be fine for some purposes. After all, you may expect there to be exceptions, even with a strong relationship, and therefore you cannot expect the coefficient of association to be 1.00. But consider that each of the following theoretical models will produce different kinds of tables:

1. X is the only cause of Y (X is necessary and sufficient).
2. X is a sufficient cause of Y (that is, when X is present, Y is present; however, Y may have other causes and therefore Y may be present when X is absent).

Table 8.4. Contingency Table with Unequal Column and Row Totals

Variable X	Variable Y		Total
	Present	Absent	
Present	40	0	40
Absent	20	40	60
Total	60	40	100

Note: $\phi = 0.67$

3. X is a necessary cause of Y (that is, when X is absent, Y will be absent; however some additional factor is necessary for Y).

Notice that table 8.4 is more consistent with model 2 than with models 1 and 3. If Y has more than one cause, then we would expect exceptions to the table in one particular cell (when X is absent, Y may be present). Those exceptions do not mean that X and Y are not related in a causal way. If X were a necessary cause of Y, but not a sufficient cause, as model 3 suggests, we would expect a zero cell when X is absent and Y is present.

The point is that the statistical measure used should be consistent with the theoretical model you expect to apply. If you do not expect that X is the only cause of Y, you may wish to choose a coefficient of association that goes to 1.00 with just one zero cell. In a two by two table, the gamma coefficient of association (which can be used for an ordinal by ordinal association) may be a better choice than the phi coefficient if you want the coefficient to go to one when there is one zero cell.

Some other coefficients for contingency tables are:

- Lambda (based on how well you can predict Y from X, X from Y—the two coefficients may not be the same).
- Uncertainty coefficient (based on how much uncertainty about one variable is reduced by knowing the other).

Table 8.5 summarizes the appropriate statistics for different types of variables.

Table 8.5. Type of Statistical Analysis for the Relationship between Two Variables

	Nominal—Two Categories	Nominal—Three or More Categories	Ordinal	Interval
Nominal—two categories	Tests of significance: <ul style="list-style-type: none">• Fisher's Exact Test• Chi-square Measures of association: <ul style="list-style-type: none">• Phi• Yule's Q (gamma)• Lambda• Uncertainty Coefficient	Tests of significance: <ul style="list-style-type: none">• Chi-square Measures of association: <ul style="list-style-type: none">• Phi (but can get larger than 1.00)• Cramer's V• Lambda• Uncertainty Coefficient	Tests of significance: <ul style="list-style-type: none">• Mann-Whitney U• Kolmogorov-Smirnov Z	Tests of significance: <ul style="list-style-type: none">• t-test Measure of association: <ul style="list-style-type: none">• point-by-serial r
Nominal—three or more categories		Tests of significance: <ul style="list-style-type: none">• Chi-square Measures of association: <ul style="list-style-type: none">• Phi (but can get larger than 1.00)• Cramer's V• Lambda• Uncertainty Coefficient	Tests of significance: <ul style="list-style-type: none">• Kruskal-Wallis analysis of variance	Tests of significance: <ul style="list-style-type: none">• Analysis of variance

(continued)

Table 8.5. (continued)

Ordinal	Measures of association: <ul style="list-style-type: none">• Spearman's rho*• Kendall's tau*• Gamma	Treat as ordinal x ordinal
Interval or ratio		Measures of association: <ul style="list-style-type: none">• Pearson's r (for linear relationships)* Tests of fit for difference curves

Note: The statistical tests included do not exhaust all of those available.
*Can be tested for significance

Statistical Inference about Differences between Two Groups

Often a research question asks about differences between two samples or groups. Do males exhibit more aggression than females? Is one population taller than another? Do hunter-gatherers have lower population densities than agriculturalists? In all these questions, we have two nominal groups to compare. However, in contrast to using a contingency table that is most appropriate for the intersection between one nominal scale and another, the three questions above allow for the possibility of measuring the other variable on an interval or ratio scale. For example, you can compare the number of aggressive acts (in a particular period of time) in two groups of children, you can compare the average adult heights in two human populations, or you can compare the population densities of a sample of hunter-gatherers and a sample of agriculturalists.

Assuming that you find a difference in means between the two groups and that certain assumptions are met (see below), the most commonly used statistic to test the significance of the difference between means is the *t-test for independent samples* (each of the two groups consists of different individuals). As in most other statistical tests, the *t-test* evaluates statistical significance against the hypothesis that there is no difference between the groups. In other words, the *t-test* asks how likely is it that a

Table 8.6. A Hypothetical Comparison of Height (in inches) in a Sample of Adult Males Compared with Adult Females

<i>Females</i>	<i>Males</i>
66	69
72	65
60	75
65	74
62	66
71	68
63	71
64	67
65	70
Mean = 65.33	Mean = 69.44

difference of magnitude X (or a bigger difference) could occur by chance if the larger populations from which the samples are drawn actually have the same mean scores.

An example of data that would be appropriate for a t -test is the hypothetical comparison of height (in inches) in adult males compared with adult females shown in table 8.6.

If we perform a t -test for independent samples, we get the following results:

$$t = 2.361, df = 16, p = 0.03, \text{ two tails}$$

Once again, the p -value tells us the likelihood that this difference is due to chance and chance alone. Since the p -value is less than 0.05, the conventional level of significance, we can reject the hypothesis of no difference and accept the hypothesis that there is a significant difference in height between males and females. The p -value is given as two-tailed, which means that we have allowed for the possibility of a difference in height in either direction (males taller than females, females taller than males). The df (degree of freedom) is the total number of cases (eighteen) minus the number of groups (two).

Assumptions of the t -test:

1. The data are measured on interval or ratio scales.
2. The populations have “normal” distributions on the measured variables. This means that the frequency distributions for the variables in both populations are roughly bell-shaped. The modal height is at the center and the curve slopes down symmetrically with very few cases at the extremes.
3. The variances (or the standard deviations) for the two populations are roughly the same.
4. The cases were randomly selected from their respective populations.

The t -test is relatively robust; it can tolerate some violation of its assumptions. But you could use tests that do not require the same assumptions. Because the t -test assumes certain char-

acteristics or parameters (assumptions 2 and 3 above) about the populations the samples come from, the *t*-test is called a *parametric* test. **Parametric tests** make certain assumptions about the underlying distributions, such as that the distribution is normally distributed. **Nonparametric tests**, sometimes called “distribution-free” tests (Siegel 1956, 19), do not require parametric assumptions about the data.

Nonparametric tests to evaluate the difference between two groups are of course particularly useful if you have ordinal (rank-order) rather than interval measures. For example, in the height example above, if we did not have measurements by ruler, we could line the persons up by size and assign the shortest person the lowest rank (one) and the tallest person the highest rank (eighteen). One nonparametric test of difference between two groups, called the *Mann-Whitney U test*, evaluates the significance of the difference between one group’s average rank and the other group’s average rank. This test is analogous to the *t*-test, but makes no assumptions about the underlying distributions. If we perform a Mann-Whitney *U* test on the same data used in the *t*-test above, we find that the average rank of the males is 12.22 and the average rank of the females is 6.78. The Mann-Whitney *U* is 16 and $p < 0.02$, two tails. In this example, the Mann-Whitney *U* test gives a *p*-value that is similar to the *p*-value given by the *t*-test, and this is generally true if we compare the outcomes of the two tests on the same data (Siegel 1956, 126). That is, the Mann-Whitney *U* test can detect significant differences between two groups just about as well as the *t*-test, without having to make parametric assumptions. Another nonparametric test for the significance of the difference between independent groups is the *Kolmogorov-Smirnov test* (Siegel 1956, 127ff.).

Statistical Inferences about Differences among Three or More Groups

Suppose you want to compare more than two nominal groups on an interval variable. Perhaps you want to compare three or four populations on height, or compare population densities of

hunter-gatherers, pastoralists, horticulturalists, and intensive agriculturalists.

A parametric test can be used to test whether the means of more than two groups are significantly different. (If they are, we can say that the differences in means are unlikely to be due to chance.) This test is called a *one-way analysis of variance* (ANOVA).² But what if the parametric assumptions cannot be met? Fortunately, there is also a nonparametric equivalent called the *Kruskal-Wallis one-way analysis of variance*. Like the Mann-Whitney *U* test, the Kruskal-Wallis test uses rank-ordered scores. This test is also almost as powerful as the parametric analysis of variance (Siegel 1956).

With three or more groups the inferential statistics and the associated *p*-values only tell you whether or not the differences in the sets of scores are unlikely to be due to chance. However, you cannot infer without further testing where the difference lies. Suppose you have three groups, A, B, and C. Perhaps all three groups are significantly different from each other, but it is possible that only one group is different from the others. And, if one is different, it could be A, B, or C that is different from the others. Most computer programs have special routines for testing the difference between any pair of groups after the overall analysis of variance is performed.

Measures of Association and Tests of Significance for Interval and Ordinal Variables

All the associations (or differences) we have discussed so far involve at least one nominal variable. Now we turn to relationships with more continuous scales—interval or ratio scales, or ordinal scales with many ranks.

Interval Variables

To examine the relationship between two interval scales, it is important first to plot the cases on a graph with an X and a Y axis and look at the scatter plot of points. This is to evaluate whether or not a relationship looks linear. If it does, you mea-

sure the strength of the association in the usual way (i.e., by using *Pearson's r*); if the relationship looks curvilinear, you should use another, more appropriate measure (e.g., *eta*) for the strength of the relationship. (See any statistics text for information about these measures.) The convention is to put the dependent variable on the Y axis and the independent variable on the X axis. Bergmann's Rule (C. R. Ember, Ember, and Peregrine 2007, 199) suggests that human populations living in colder climates have more body mass (weight). The reasoning is that it is more adaptive in cold climates to have more body mass, because the body conserves heat better the more mass it has. To test this directional hypothesis, Roberts (1953) conducted a cross-cultural comparison and found support for Bergmann's Rule. We don't show his data here, but let us imagine that a plot of temperature and weight (one point for each population) would look something like the plot shown in figure 8.3. How would we describe this relationship? We could say that, in general, colder temperatures are associated with more weight, but we could also describe the

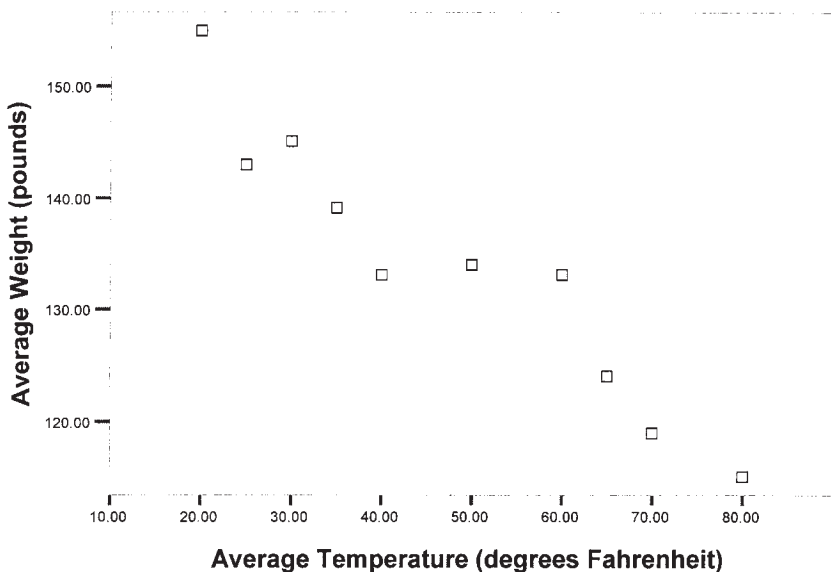


Figure 8.3. Scatter plot of average temperature and average weight (sample of eleven hypothetical populations)

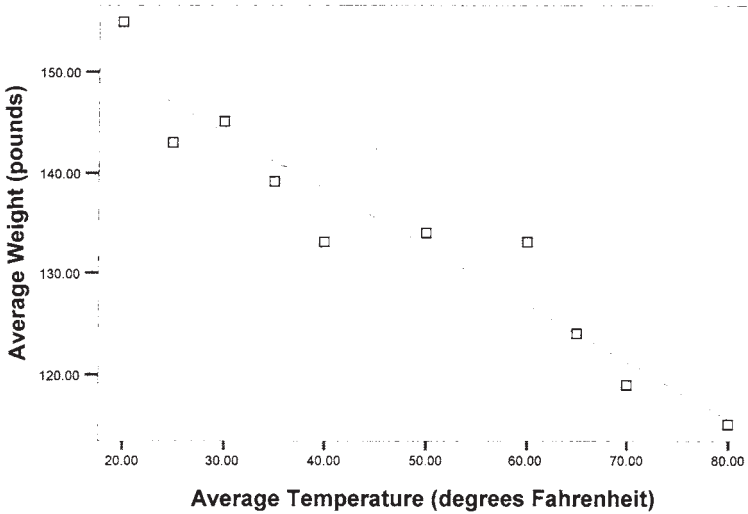


Figure 8.4. Line of best fit for average weight predicted by temperature (sample of eleven hypothetical populations)

relationship in statistical terms. If the relationship looks roughly linear, as this one does, linear regression is a method for getting a straight line that best fits a set of points. It is important to try linear regression only when the relationship looks roughly linear.³ The most common method used is called “least squares.” Basically we arrive at a straight line that minimizes the squared vertical deviations of all the points from the line. Most computers do such computations in a very short time. If we do the least-squares computation for X as a function of Y in our hypothetical data set (shown in figure 8.3), we get the line plotted in figure 8.4. If we want to predict Y (average weight) from X (average temperature), we can use the formula for the line or the line itself to say what the weight would be if the average temperature were 45° Fahrenheit.

But the line or the formula for the line doesn’t tell us how strong the linear relationship is. (A line of best fit can be calculated for any set of points, no matter how much the points deviate from the line.) Pearson’s r is a measure of the strength of a linear relationship. Just as with most other coefficients of association, Pearson’s r is 0.00 if there is no linear relationship and ± 1.00 if all the points

fall exactly on the line. (Remember that a minus sign would mean that as one variable goes up, the other goes down; a plus sign would mean that one variable goes up as the other goes up.) In our hypothetical data set, the r is -0.94 . (As average temperature increases, average weight decreases.) The coefficient is so strong (close to -1.00) because the points are not far off the line.

So far, everything we have described about linear regression is *descriptive*. That is, we have just described how you can get a formula for the best linear fit and a measure of the degree to which the points fall on a straight line. Neither of these things tells us whether the linear-looking relationship might be due to chance and chance alone. Remember that a best-fitting straight line can be drawn for any set of points, even one that does not look linear!

It is important then to test the significance of your measure of association. What is evaluated is the probability that the obtained coefficient (or a stronger one) could be due to chance if the true linear relationship were zero. In our example, the r could be positive or negative, but we would look for the one-tailed significance because the direction of the relationship was predicted. The p -value for our hypothetical data set is <0.0005 , one tail. This means that the likelihood of there being no negative linear relationship is less than five times in ten thousand.

What if the relationship is not linear? Figure 8.5 shows an example of a nonlinear relationship. If we didn't plot our data but just asked for a Pearson's r , we would have gotten an r of 0.00 because the line of best fit in figure 8.5 is flat. (A flat line means that the best predictor of Y for each value of X is the mean of Y .) If the variation on X doesn't help us predict the variation on Y any better than using the mean of Y , there appears to be no relationship between X and Y . But concluding that there is no relationship is obviously incorrect. There appears to be a strong relationship, but the nature of the relationship is not described by a straight line. Most statistical programs allow you to try to fit other curves to your data; the significance of these fits can also be determined. Figure 8.6 shows the fit of a quadratic equation, which appears to fit quite well. The associated p -value is 0.0005 , which means that the chance of getting this result (if there were no quadratic relationship) is only five in ten thousand.

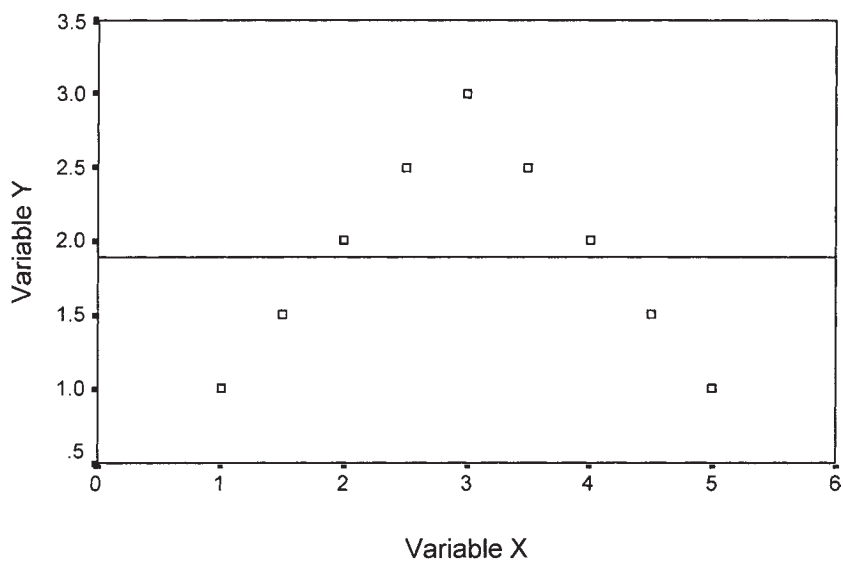


Figure 8.5. Linear fit to a nonlinear relationship

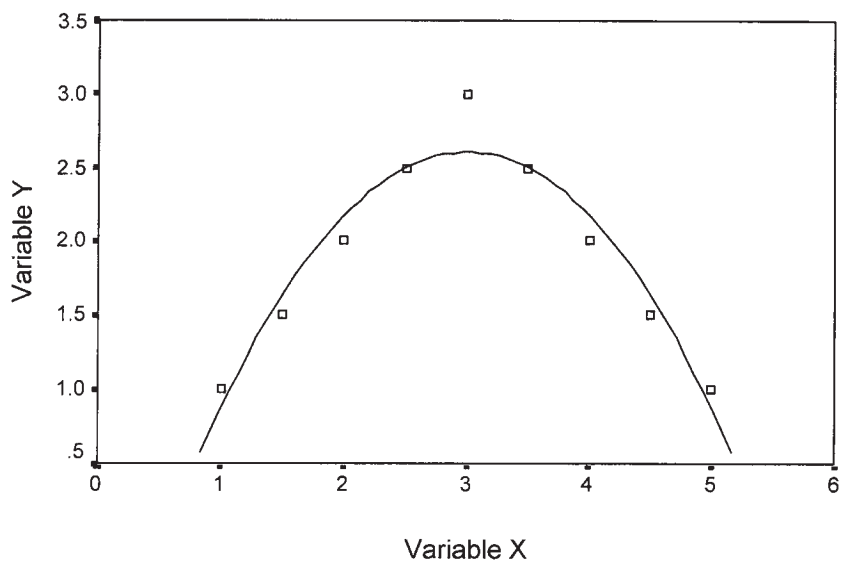


Figure 8.6. A quadratic fit to a nonlinear relationship

Ordinal Variables

Suppose you want to know if there is a relationship between two variables, each of which is measured on an ordinal scale. There are several nonparametric measures of association, each of which can be tested for statistical significance. Like Pearson's r , each of these measures provides a coefficient that varies between 0.00 and ± 1.00 . Tests of significance give probability values for the likelihood that these coefficients (or larger ones) would occur by chance, if there were no associations between the variables. The most commonly used measures are:

- *Spearman's rho*
- *Kendall's tau* (which is more appropriate when ties in rank are numerous)
- *Gamma* (which can also be used, as we have noted, to measure the association in a contingency table with ordered rows and columns)

Recall that we discussed gamma when we talked about measures of association for two by two tables. Phi coefficients can reach 1.00 only when all the cases are on one diagonal. Gamma will reach 1.00 with one zero cell. If your theoretical model suggests that X is *either* a necessary *or* a sufficient cause of Y (but not necessary *and* sufficient), gamma may be the appropriate choice. So, as we see in tables 8.7a and 8.7b, table 8.7a would have a rho, tau, and gamma of 1.00. But for table 8.7b, only gamma would be 1.00. Gamma goes to one when the cells on one side of the diagonal are empty, just as in table 8.7b where the cells below the diagonal are empty. Because of this characteristic, gamma usually gives you a higher coefficient than rho or tau for the same table. Your decision as to which measure to use for an association between ordinal variables should depend on what kind of association you theoretically expect, not on the fact that gamma would give you a larger coefficient.

Table 8.7a. All Cases on Diagonal

	Y Rank 1	Y Rank 2	Y Rank 3
X Rank 1	10		
X Rank 2		15	
X Rank 3			20

Table 8.7b. No Cases below Diagonal

	Y Rank 1	Y Rank 2	Y Rank 3
X Rank 1	8	2	2
X Rank 2		10	5
X Rank 3			20

Multivariate Analyses

All of the analyses we have discussed so far consider the relationship between two variables. Technically, two variable associations are called **bivariate** associations. Usually one is considered a possible cause and the other the possible effect. But often a researcher has reason to believe that more than one variable is possibly a cause. One of the most common situations is when previous research has supported a particular factor as a possible cause and a new researcher thinks that an additional factor is involved. The first step might be to see if the new factor also predicts by itself. If it does, the second step is to compare how well the new factor predicts, when its effect is compared with that of the previously suspected factor. In short, what we need this time around is an analysis that considers the possible independent effects of two or more variables on the dependent variable. Or, there may be other models that suggest how two or more variables are related. After the next chapter, we will provide a simple introduction to *multivariate* analyses.

Notes

1. The formulas for variance and standard deviation are divided by $n - 1$ if you are calculating them to estimate the population from a sample; they are divided by n if you are dealing with the population.
2. A two-way analysis of variance allows for evaluating the effects of two nominal variables on a dependent variable that is an interval scale, but we will not discuss this type of analysis here.
3. The computations can be done without computers, but now these computations are rarely done without computer assistance.

