Analysis of Quantitative Data

Statistics may also be regarded as a method of dealing with data. This definition stresses the view that statistics is a tool concerned with the collection, organization, and analysis of numerical facts or observations. . . . The major concern of descriptive statistics is to present information in a convenient, usable, and understandable form.

—Richard Runyon and Audry Haber, *Fundamentals of Behavioral Statistics,* p. 6.

If you read a research report or article based on quantitative data, you will probably see many charts, graphs, and tables full of numbers. Do not be intimidated by them. The author provides the charts, graphs, and tables to give you, the reader, a condensed picture of the data. The charts and tables allow you to see the evidence collected by the researcher and examine it for yourself. When you collect your own quantitative data, you will use similar techniques to reveal what is inside the data. You will need to organize and manipulate the quantitative data to get them to disclose things of interest about the social world. In this chapter, you will be introduced to the fundamentals of organizing and analyzing quantitative data. Its analysis is a complex field of knowledge. This chapter cannot substitute for a course in social statistics. It covers only the basic statistical concepts and data-handling techniques necessary to understand social research.

Data collected using the techniques in the past chapters are in the form of numbers. The numbers represent values of variables, which measure characteristics of participants, respondents, or other cases. The numbers are in a raw form on questionnaires, note pads, recording sheets, or computer files. We do several things to the raw data in order to see what they can say about the hypotheses: reorganize them into a form suitable for computer entry, present them in charts or graphs to summarize their features, and interpret or give theoretical meaning to the results.

DEALING WITH DATA

Coding Data

Before we examine quantitative data to test hypotheses, we must put them in a specific form. Data coding means systematically reorganizing raw data into a format that is easy to analyze using statistics software on computers. As with coding in content analysis, researchers create and consistently apply rules for transferring information from one form to another.[1]

Coding can be a simple clerical task when you have recorded the data as numbers on well-organized recording sheets, but it is very difficult if

From Chapter 12 of *Social Research Methods: Qualitative and Quantitative Approaches,* 7/e. W. Lawrence Neuman.

you want to code answers to open-ended survey questions into numbers in a process similar to latent content analysis. To code open-ended survey data or other data that are not already in the form of numbers requires a coding procedure and a codebook. The **coding procedure** is a set of rules stating that you will assign certain numbers to variable attributes. For example, you code males as 1 and females as 2, or for a Likert scale, you code strongly agree as 4, agree as 3, and so forth. You need a code for each category of all variables and missing information. The coding procedure explains in detail how you converted non-numerical information into numbers.

A **codebook** is a document (i.e., one or more pages) describing the coding procedure and the computer file location of data for variables in a specific format. When you code data, it is essential to create a well-organized, detailed codebook and make multiple copies of it. If you do not write down the details of the coding procedure or if you misplace the codebook, you have lost the key to the data and may have to recode them again.

You should begin to think about a coding procedure and codebook before you collect any data. For example, many survey researchers precode a questionnaire before interviewing or collecting data. Precoding involves placing the code categories (e.g., 1 for male, 2 for female) on the questionnaire and building the features of a codebook into it.[2] If you do not precode, your first step after collecting data is to create a codebook. You also must assign an identification number to each case to keep track of the cases. Next you transfer the information from each questionnaire into a computer-readable format.

Coding procedure A set of rules created by a quantitative researcher for assigning numbers to specific variable attributes, usually in preparation for statistical analysis and carefully recorded in a codebook.

Codebook A document that describes the procedure for coding variables and their location in a format that computers can use.

Data records The units or reports in computer-based data that contain information on the variables for a case.

Data field One or more columns in data organized for a computer representing the location of information on a specific variable.

Entering Data

Most computer programs designed for numerical data analysis require that the data be in a grid format. In the grid, each row represents a respondent, participant, or case. In computer terminology, these are called **data records**. Each data record is for a single case. A column or a set of columns represents specific variables. It is possible to go from a column and row location (e.g., row 7, column 5) back to the original source of data (e.g., a questionnaire item on marital status for respondent 8). A column or a set of columns assigned to a variable is called a **data field**, or simply *field*.

For example, you code survey data for three respondents in a format for computers like the start of a data file presented in Figure 1. People cannot easily read data in this format and without the codebook, it is worthless. The data file condenses answers to 50 survey questions for three respondents into three lines or rows. The raw data for many research projects look like this, except that there may be more than 1,000 rows, and the lines may be more than 100 columns long. For example, a 15-minute telephone survey of 250 students produces a grid of data that is 250 rows by 240 columns.

The codebook in Figure 1 states that the first two numbers are identification numbers. Thus, the example data are for the first (01), second (02), and third (03) respondents. Notice that we use zeros as placeholders to reduce confusion between 1 and 01. The 1s are always in column 2; the 10s are in column 1. The codebook states that column 5 contains the variable “gender”: Cases 1 and 2 are male and Case 3 is female. Column 4 tells us that Carlos interviewed Cases 1 and 2 and Sophia Case 3.

There are four ways to enter raw quantitative data into a computer:

1. *Code sheet.* Gather the information, then transfer it from the original source onto a grid

FIGURE 1 Coded Data for Three Cases and Codebook

EXCERPT FROM SURVEY QUESTIONNAIRE

Respondent ID ______________ Interviewer Name _________________

Note the Respondent's Gender: ____ Male ____ Female

1. The first question is about the President of the United States. Do you Strongly Agree, Agree, Disagree, Strongly Disagree, or Have No Opinion About the following statement:

   The President of the United States is doing a great job.

   ____ Strongly Agree ____ Agree ____ Disagree ____ Strongly Disagree ____ No Opinion

2. How old are you? ________

EXCERPT OF CODED DATA

```
                    Column
000000000111111111122222222223333333333444 ... etc. (tens)
123456789012345678901234567890123456789012 ... etc. (ones)
01 212736302 182738274 10239 18.82 3947461 ... etc.
02 213334821 124988154 21242 18.21 3984123 ... etc.
03 420123982 113727263 12345 17.36 1487645 ... etc.
etc.
```

Raw data for first three cases, columns 1 through 42.

EXCERPT FROM CODEBOOK

Column	*Variable Name*	*Description*
1–2	ID	Respondent identification number
3	BLANK	
4	Interviewer	Interviewer who collected the data: 1 = Susan 2 = Carlos 3 = Juan 4 = Sophia 5 = Clarence
5	Gender	Interviewer report of respondent's sex 1 = Male, 2 = Female
6	PresJob	The President of the United States is doing a great job. 1 = Strongly Agree 2 = Agree 3 = No Opinion 4 = Disagree 5 = Strongly Disagree Blank = Missing Information

format (code sheet). Next, enter what is on the code sheet into a computer line by line.

2. *Direct-entry method (including CATI).* As information is being collected, sit at a computer keyboard (or similar recording device) while listening to or observing the information and enter or have a respondent/participant enter the information him- or herself. To use the

direct-entry method, the computer must be preprogrammed to accept the information.

3. *Optical scan.* Gather the information and then enter it onto optical scan sheets (or have a respondent/participant enter the information) by filling in the correct "dots." Next use an optical scanner or reader to transfer the information into a computer.
4. *Bar code.* Gather the information and convert it into different widths of bars that are associated with specific numerical values; then use a bar-code reader to transfer the information into a computer.

Cleaning Data

Accuracy is extremely important when coding data (see Example Box 1, Example of Dealing with Data). Errors you make when coding or entering data into a computer threaten the validity of the measures and cause misleading results. If you have a perfect sample, perfect measures, and no errors in gathering data but make errors in the coding process or in entering data into a computer, you can ruin an entire research project.

After very careful coding, you must check the accuracy of coding, or "clean" the data. Often you want to code random sample of 10 to 15 percent of the data a second time. If you discover no coding errors in the recoded sample, you can proceed. If you find errors, you need to recheck all of the coding.

You can verify coding after the data are in a computer in two ways. **Possible code cleaning** (or *wild code checking*) involves checking the categories of all variables for impossible codes. For example, respondent gender is coded 1 = Male, 2 = Female. A 4 for a case found in the field for the gender variable indicates a coding error. A second method, **contingency cleaning** (or *consistency checking*), involves cross-classifying two variables and looking for logically impossible combinations. For example, you cross-classify school level by occupation. If you find a respondent coded never having passed the eighth grade and recorded as being a medical doctor, you must check for a coding error.

You can modify data in some ways after they are in a computer, but you cannot use more refined categories than those used collecting the original data. For example, you may group ratio-level income data into five ordinal categories, and you can collapse variable categories and combine information from several indicators to create a new index variable.

Direct-entry method Process of entering data directly into a computer by typing them without bar codes or optical scan sheets.

Possible code cleaning Clarifying data using a computer by searching for responses or answer categories that cannot have cases.

Contingency cleaning Flushing data using a computer in which the researcher reviews the combination of categories for two variables for logically impossible cases.

Descriptive statistics A general type of simple statistics used by researchers to describe basic patterns in the data.

Frequency distribution A table that shows the dispersion of cases into the categories of one variable, that is, the number or percent of cases in each category.

RESULTS WITH ONE VARIABLE

Frequency Distributions

The word *statistics* can refer to a set of collected numbers (e.g., numbers telling how many people live in a city) as well as a branch of applied mathematics we use to manipulate and summarize the features of numbers. Social researchers use both types of statistics. Here we focus on the second type: ways to manipulate and summarize numbers that represent data from a research project.

Descriptive statistics describe numerical data. We can categorize them by the number of variables involved: univariate, bivariate, or multivariate (for one, two, and three or more variables). Univariate statistics describe one variable (*uni*- refers to one; -*variate* refers to variable). The easiest way to describe the numerical data of one variable is with a **frequency distribution**. You can use the frequency

EXAMPLE BOX 1

Example of Dealing with Data

There is no good substitute for getting your hands dirty with the data. Here is an example of data preparation from a study I conducted with my students. My university surveyed about one-third of the students to learn their thinking about and experience with sexual harassment on campus. A research team drew a random sample and then developed and distributed a self-administered questionnaire. Respondents put answers on optical scan sheets that were similar to the answer sheets used for multiple-choice exams. The story begins with the delivery of more than 3,000 optical scan sheets.

After the sheets arrived, we visually scanned each one for obvious errors. Despite instructions to use pencil and fill in each circle neatly and darkly, we found that about 200 respondents used a pen, and another 200 were very sloppy or used very light pencil marks. We cleaned up the sheets and redid them in pencil. We also found about 25 unusable sheets that were defaced, damaged, or too incomplete (e.g., only the first 2 of 70 questions answered).

Next we read the usable optical scan sheets into a computer. We had the computer produce the number of occurrences, or frequency, of the attributes for each variable. Looking at them, we discovered several kinds of errors. Some respondents had filled in two responses for a question to which only one answer was requested or possible. Some had filled in impossible response codes (e.g., the numeral 4 for gender, when the only legitimate codes were 1 for male and 2 for female), and some had filled in every answer in the same way, suggesting that they did not take the survey seriously. For each case with an error, we returned to the optical scan sheet to see whether we could recover any information. If we could not recover information, we reclassified the case as a nonresponse or recoded a response as missing information.

The questionnaire had two contingency questions. For each, a respondent who answered "no" to one question was to skip the next five questions. We created a table for each question. We looked to see whether all respondents who answered "no" to the first question skipped or left blank the next five. We found about 35 cases in which the respondent answered "no" but then went on to answer the next five questions. We returned to each sheet and tried to figure out which the respondent really intended. In most cases, it appeared that the respondent meant the "no" but failed to read the instructions to skip questions.

Finally, we examined the frequency of attributes for each variable to see whether they made sense. We were very surprised to learn that about 600 respondents had marked "Native American" for the racial heritage question. In addition, more than half of those who had done so were freshmen. A check of official records revealed that the university enrolled a total of about 20 Native Americans or American Indians, and that over 90 percent of the students were White, non-Hispanic Caucasians. The percentage of respondents marking Black, African-American, or Hispanic-Chicano matched the official records. We concluded that some White Caucasian respondents had been unfamiliar with the term "Native American" for "American Indian." Apparently, they had mistakenly marked it instead of "White, Caucasian." Because we expected about 7 Native Americans in the sample, we recoded the "Native American" responses as "White, Caucasian." This meant that we reclassified Native Americans in the sample as Caucasian. At this point, we were ready to analyze the data.

distribution with nominal-, ordinal-, interval-, or ratio-level data. For example, I have data for 400 respondents and want to summarize the information on the gender at a glance. The easiest way is with a raw count or a percentage frequency distribution (see Figure 2). I can present the same information in graphic form.

Some common types of graphic representations are the histogram, bar chart, and pie chart. Bar charts or graphs are used for discrete variables. They can have a vertical or horizontal orientation with a small space between the bars. The terminology is not exact, but the **histogram** is

Histogram A graphic display of univariate frequencies or percentages, usually with vertical lines indicating the amount or proportion.

FIGURE 2 Examples of Univariate Statistics

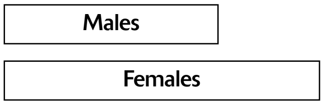
RAW COUNT FREQUENCY DISTRIBUTION

Gender	*Frequency*
Male	100
Female	300
Total	400

PERCENTAGE FREQUENCY DISTRIBUTION

Gender	*Percentage*
Male	25%
Female	75%
Total	100%

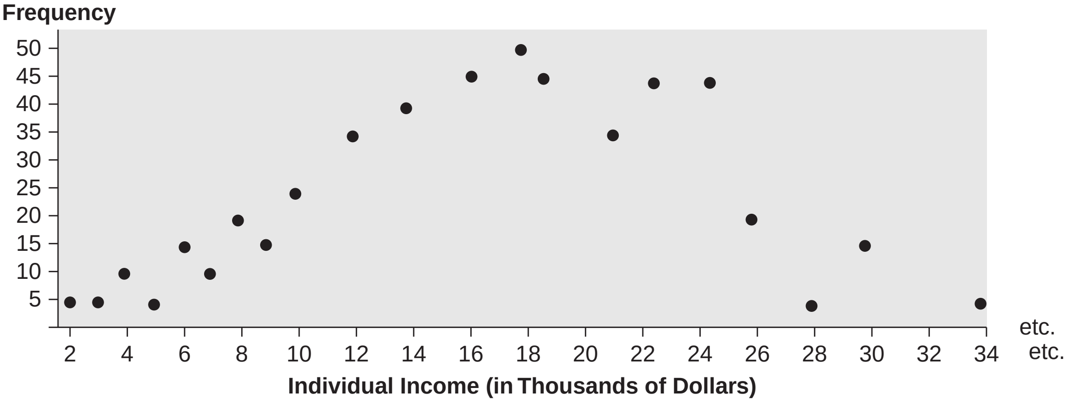
BAR CHART OF SAME INFORMATION



EXAMPLE OF GROUPED DATA FREQUENCY DISTRIBUTION

First Job Annual Income	*N*
Under $5,000	25
$5,000 to $9,999	50
$10,000 to $15,999	100
$16,000 to $19,999	150
$20,000 to $29,999	50
$30,000 and over	25
Total	400

EXAMPLE OF FREQUENCY POLYGON



usually a set of upright bar graphs for interval or ratio data.[3]

For interval- or ratio-level data, we often group the information into several categories. The grouped categories must be mutually exclusive. We also can plot interval- or ratio-level data in a **frequency polygon** with the number of cases or frequency along the vertical axis and the values of the variable or scores along the horizontal axis. A polygon appears when we connect the dots.

Measures of Central Tendency

Often, we want to summarize the information about one variable into a single number. To do this, we use three **measures of central tendency** (i.e. measures of the center of the frequency distribution: mean, median, and mode). Many people call them *averages*, a less precise or clear way of saying the same thing.

The **mode** is the easiest to use and we can use it with nominal, ordinal, interval, and ratio data. It is simply the most common or frequently occurring number. For example, the mode of the following list is 5: 6, 5, 7, 10, 9, 5, 3, 5. A distribution can have more than one mode. For example, the mode of this list is both 5 and 7: 5, 6, 1, 2, 5, 7, 4, 7. If the list gets long, it is easy to spot the mode in a frequency distribution; just look for the most frequent score. There is always at least one case with a score equal to the mode.

The **median** is the middle point. It is also the 50th percentile, or the point at which half the cases are above it and half below it. We can use it with ordinal-, interval-, or ratio-level data (but not nominal level). We can "eyeball" the mode, but computing a median requires a little more work. The easiest way is first to organize the scores from highest to lowest and then count to the middle. If there is an odd number of scores, it is simple. Seven people are waiting for a bus; their ages are 12, 17, 20, 27, 30, 55, 80. The median age is 27. Note that the median does not change easily. If the 55-year-old and the 80-year-old both got on one bus and the remaining people were joined by two 31-year-olds, the median remains unchanged. If there is an even number of scores, things are a bit more complicated. For example, six people at a bus stop have the following ages: 17, 20, 26, 30, 50, 70. The median is halfway between 26 and 30. Compute the median by adding the two middle scores together and dividing by 2 (26 + 30 = 56/2 = 28). The median age is 28, even though no person is 28 years old. Note that there is no mode in the list of six ages because each person has a different age.

The **mean** (also called the *arithmetic average*) is the most widely used measure of central tendency. We can use it only with interval- or ratio-level data.[4] To compute the mean, we add up all scores and then divide by the number of scores. For example, the mean age in the previous example is 17 + 20 + 26 + 30 + 50 + 70 = 213; 213/6 = 35.5. No one in the list is 35.5 years old, and the mean does not equal the median.

Changes in extreme values (very large or very small) can greatly influence the mean. For example, the 50-year-old and 70-year-old left and were replaced with two 31-year-olds. The distribution now looks like this: 17, 20, 26, 30, 31, 31. The median is unchanged: 28. The mean is 17 + 20 + 26 + 30 + 31 + 31 = 155; 155/6 = 25.8. Thus, the mean dropped a great deal when a few extreme values were removed.

If the frequency distribution forms a **normal distribution** or bell-shaped curve, the three measures of central tendency equal each other. If the distribution is a **skewed distribution** (i.e., more cases are in the upper or lower scores), then the three will not be equal. If most cases have lower scores with a few extreme high scores, the mean will be the highest, the median in the middle, and the mode the

Frequency polygon A graph of connected points showing how many cases fall into each category of a variable.

Measures of central tendency A class of statistical measures that summarizes information about the distribution of data for one variable into a single number.

Mode A measure of central tendency for one variable that indicates the most frequent or common score.

Median A measure of central tendency for one variable that indicates the point or score at which half of the cases are higher and half are lower.

Mean A measure of central tendency for one variable that indicates the arithmetic average, that is, the sum of all scores divided by the total number of them.

Normal distribution A bell-shaped frequency polygon for a dispersion of cases with a peak in the center and identical curving slopes on either side of the center; distribution of many naturally occurring phenomena and a basis of much statistical theory.

Skewed distribution A dispersion of cases among the categories of a variable that is not normal, that is, not a bell shape; instead of an equal number of cases on both ends, more are at one of the extremes.

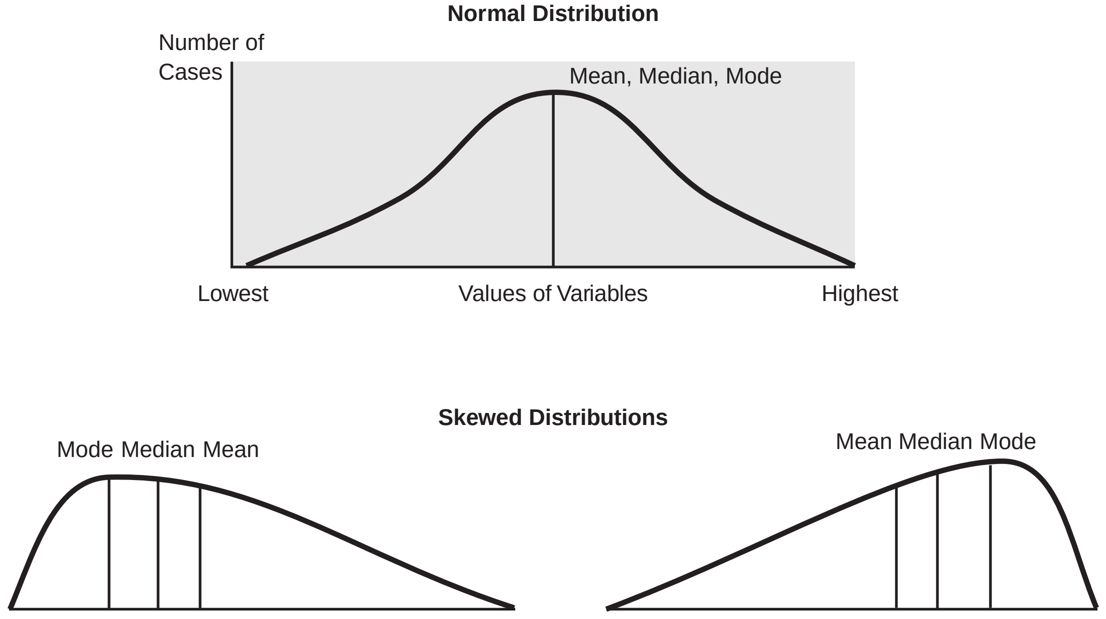


FIGURE 3 **Measures of Central Tendency**

lowest. If most cases have higher scores with a few extreme low scores, the mean will be the lowest, the median in the middle, and the mode the highest. In general, the median is best to use for a skewed distribution, although the mean is used in most other statistics (see Figure 3).

Measures of Variation

The measure of central tendency is a single-number summary of a distribution; however, the measures give only its center. Another characteristic of a distribution is its *spread, dispersion,* or *variability* around the center. Two distributions can have identical measures of central tendency but differ in their spread about the center. For example, seven people are at a bus stop in front of a bar. Their ages are 25, 26, 27, 30, 33, 34, 35. Both the median and the mean are 30. At a bus stop in front of an ice cream store, seven people have the identical median and mean, but their ages are 5, 10, 20, 30, 40, 50, 55. The ages of the group in front of the ice cream store are spread more from the center, or the distribution has more variability.

Variability has important social implications. For example, in city X, the median and mean family income is $37,600 per year, and it has zero variation. *Zero variation* means that every family has an income of exactly $37,600. City Y has the same median and mean family income, but 96 percent of its families have incomes of $14,000 per year and 4 percent have incomes of $350,000 per year. City X has perfect income equality whereas there is great inequality in city Y. If we do not know the variability of income in the two cities, we miss very important information.

We measure variation in three ways: range, percentile, and standard deviation. **Range** is the simplest. It consists of the largest and smallest scores. For example, the range for the bus stop in front of the bar is from 25 to 35, or 35 − 25 = 10 years. If

Range A measure of dispersion for one variable indicating the highest and lowest scores.

the 35-year-old got onto a bus and was replaced by a 60-year-old, the range would change to 60 − 25 = 45 years. Range has limitations because it only tells us the extreme high and low. For example, here are two groups of six with a range of 35 years: 30, 30, 30, 30, 30, 65 and 20, 45, 46, 48, 50, 55.

Percentiles tell us the score at a specific place within the distribution. One percentile you already studied is the median, the 50th percentile. Sometimes the 25th and 75th percentiles or the 10th and 90th percentiles are used to describe a distribution. For example, the 25th percentile is the score at which 25 percent of cases in the distribution have either that score or a lower one. The computation of a percentile follows the same logic as the median. If you have 100 people and want to find the 25th percentile, you rank the scores (i.e. measures in numbers of variables) and count up from the bottom until you reach number 25. If the total is not 100, you simply adjust the distribution to a percentage basis.

Standard deviation is the most difficult to compute measure of dispersion; it is also the most comprehensive and widely used. The range and percentile are for ordinal-, interval-, and ratio-level data, but the standard deviation requires an interval or ratio level of measurement. It is based on the mean and gives an "average distance" between all scores and the mean. People rarely compute the standard deviation by hand for more than a handful of cases because computers do it in seconds.

Look at the calculation of the standard deviation in Figure 4. If you add the absolute difference between each score and the mean (i.e., subtract each score from the mean), you get zero because the mean is equally distant from all scores. Also notice that the scores that differ the most from the mean have the largest effect on the sum of squares and on the standard deviation.

The standard deviation is of limited usefulness by itself. It is used for comparison purposes. For example, the standard deviation for the schooling of parents of children in class A is 3.317 years; for class B, it is 0.812; and for class C, it is 6.239. The standard deviation tells a researcher that the parents of children in class B are very similar, whereas those for class C are very different. In fact, in class B, the schooling of an "average" parent is less than a year above or below the mean for all parents, so the parents are very homogeneous. In class C, however, the "average" parent is more than six years above or below the mean, so the parents are very heterogeneous.

We use the standard deviation and the mean to create **z-scores**, which let you compare two or more distributions or groups. The z-score, also called a *standardized score,* expresses points or scores on a frequency distribution in terms of a number of standard deviations from the mean. Scores are in terms of their relative position within a distribution, not as absolute values (see Expansion Box 1, Calculating *Z*-Scores). Z-scores can tell us a lot. For example, Katy, a sales manager in firm A, earns $70,000 per year, whereas Mike in firm B earns $60,000 per year. Despite the $10,000 absolute income differences between them, the managers are paid equally relative to others in the same firm. Both Katy and Mike are paid more than two-thirds of other employees in each of their respective firms.

Here is another example of how to use z-scores. Hans and Heidi are twin brother and sister, but Hans is shorter than Heidi. Compared to other girls her age, Heidi is at the mean height; she has a z-score of zero. Likewise, Hans is at the mean height among boys his age. Thus, within each comparison group, the twins are at the same z-score, so they have the same relative height.

Z-scores are easy to calculate from the mean and standard deviation. For example, an employer interviews students from Kings College and Queens College. She learns that the colleges are similar and that both grade on a 4.0 scale, yet the mean grade-point average at Kings College is 2.62 with

Percentile A measure of dispersion for one variable that indicates the percentage of cases at or below a score or point.

Standard deviation A measure of dispersion for one variable that indicates an average distance between the scores and the mean.

Z-score A standardized location of a score in a distribution of scores based on the number of standard deviations it is above or below the mean.

FIGURE 4 The Standard Deviation

STEPS IN COMPUTING THE STANDARD DEVIATION

1. Compute the mean.
2. Subtract the mean from each score.
3. Square the resulting difference for each score.
4. Total up the squared differences to get the sum of squares.
5. Divide the sum of squares by the number of cases to get the variance.
6. Take the square root of the variance, which is the standard deviation.

EXAMPLE OF COMPUTING THE STANDARD DEVIATION

[8 respondents, variable = years of schooling]

Score	*Score – Mean*	*Squared (Score – Mean)*
15	15 – 12.5 = 2.5	6.25
12	12 – 12.5 = −0.5	.25
12	12 – 12.5 = −0.5	.25
10	10 – 12.5 = −2.5	6.25
16	16 – 12.5 = 3.5	12.25
18	18 – 12.5 = 5.5	30.25
8	8 – 12.5 = 4.5	20.25
9	9 – 12.5 = –3.5	12.25

Mean = 15 + 12 + 12 + 10 + 16 + 18 + 8 + 9 = 100, 100/8 = 12.5
Sum of squares = 6.25 + .25 + .25 + 6.25 + 12.25 + 30.25 + 20.25 + 12.25 = 88
Variance = Sum of squares/Number of cases = 88/8 = 11
Standard deviation = Square root of variance = $\sqrt{11}$ = 3.317 years.
Here is the standard deviation in the form of a formula with symbols.

Symbols:

X = SCORE of case
$\bar{X}$ = MEAN
Σ = Sigma (Greek letter) for sum, add together
N = Number of cases

Formula:[a]

$$\text{Standard deviation} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}}$$

[a] There is a slight difference in the formula depending on whether one is using data for the population or a sample to estimate the population parameter.

a standard deviation of .50, whereas the mean grade-point average at Queens College is 3.24 with a standard deviation of .40. The employer suspects that grades at Queens College are inflated. Suzette from Kings College has a grade-point average of 3.62; Jorge from Queens College has a grade-point average of 3.64. Both students took the same courses. The employer wants to adjust the grades for the grading practices of the two colleges (i.e., create standardized scores). She calculates z-scores by subtracting each student's score from the mean and then divides by the standard deviation. For example, Suzette's z-score is 3.62 − 2.62 = 1.00/.50 = 2, whereas Jorge's z-score is 3.64 − 3.24. = .40/.40 = 1. Thus, the employer learns that Suzette is two standard deviations above the mean in her college, whereas Jorge is only one standard deviation above the mean for his college. Although Suzette's absolute grade-point average is lower than Jorge's, relative to the students in each of their colleges, Suzette's grades are much higher than Jorge's.

EXPANSION BOX 1

Calculating *Z*- Scores

Personally, I do not like the formula for *z*-scores, which is:

Z-score = (Score – Mean)/Standard Deviation, or in symbols:

$$z = \frac{X - \bar{X}}{\delta}$$

where: X = score, $\bar{X}$ = mean, δ = standard deviation

I usually rely on a simple conceptual diagram that does the same thing and that shows what *z*-scores really do. Consider data on the ages of schoolchildren with a mean of 7 years and a standard deviation of 2 years. How do I compute the *z*-score of 5-year-old Miguel, or what if I know that Yashohda's *z*-score is a +2 and I need to know her age in years? First, I draw a little chart from –3 to +3 with zero in the middle. I will put the mean value at zero, because a *z*-score of zero is the mean and *z*-scores measure distance above or below it. I stop at 3 because virtually all cases fall within 3 standard deviations of the mean in most situations. The chart looks like this:

```
 |_____|_____|_____|_____|_____|_____|
-3    -2    -1     0    +1    +2    +3
```

Now, I label the values of the mean and add or subtract standard deviations from it. One standard deviation above the mean (+1) when the mean is 7 and standard deviation is 2 years is just 7 + 2, or 9 years. For a –2 *z*-score, I put 3 years. This is because it is 2 standard deviations, of 2 years each (or 4 years), lower than the mean of 7. My diagram now looks like this:

```
 1     3   5   7     9     11    13  age in years
 |_____|___|___|_____|_____|_____|
-3    -2  -1   0    +1    +2    +3
```

It is easy to see that Miguel, who is 5 years old, has a *z*-score of –1, whereas Yashohda's *z*-score of +2 corresponds to 11 years old. I can read from *z*-score to age, or age to *z*-score. For fractions, such as a *z*-score of –1.5, I just apply the same fraction to age to get 4 years. Likewise, an age of 12 is a *z*-score of +2.5.

RESULTS WITH TWO VARIABLES

A Bivariate Relationship

Univariate statistics describe a single variable in isolation. **Bivariate statistics** are much more valuable. They let us consider two variables together and describe the relationship between variables. Even simple hypotheses require two variables. Bivariate statistical analysis shows a **statistical relationship** between variables—that is, things that tend to appear together. For example, a relationship exists between water pollution in a stream and the fact that people who drink the water get sick. It is a statistical relationship between two variables: pollution in the water and the health of the people who drink it.

Statistical relationships are based on two ideas: covariation and statistical independence. **Covariation** means that things go together or are associated. To *covary* means to vary together; cases with certain values on one variable are likely to have certain values on the other one. For example, people with higher values on the income variable are likely to have higher values on the life expectancy variable. Likewise, those with lower incomes have lower life expectancy. This is usually

Univariate statistics Statistical measures that deal with one variable only.

Bivariate statistics Statistical measures that involve two variables only.

Statistical relationship Expression of whether two or more variables affect one another based on the use of elementary applied mathematics, that is, whether there is an association between them or independence.

Covariation The concept that two variables vary together, such that knowing the values on one variable provides information about values found on another.

stated in a shorthand way by saying that income and life expectancy are related to each other, or covary. We could also say that knowing one's income tells us one's probable life expectancy, or that life expectancy depends on income.

Statistical independence is the opposite of covariation. It means there is no association or no relationship between variables. If two variables are independent, cases with certain values on one variable do not have a special value on the other variable. For example, Rita wants to know whether number of siblings is related to life expectancy. If the variables are independent, then people with many brothers and sisters have the same life expectancy as those who are only children. In other words, knowing how many brothers or sisters someone has tells Rita nothing about the person's life expectancy.

We usually state hypotheses in terms of a causal relationship or expected covariation; if we use the null hypothesis, it is that there is independence. It is used in formal hypothesis testing and is frequently found in inferential statistics (to be discussed).

We use several techniques to decide whether a relationship exists between two variables. Three elementary ones are a scattergram, or a graph or plot of the relationship; a percentaged table; and measures of association, or statistical measures that express the amount of covariation by a single number (e.g., correlation coefficient). Also see Chart 1 on graphing data.

Statistical independence The absence of a statistical relationship between two variables, that is, when knowing the values on one variable provides no information about the values found on another variable; no association between the variable.

Scattergram A diagram to display the statistical relationship between two variables based on plotting each case's values for both of the variables.

Linear relationship An association between two variables that is positive or negative across the levels of variables; when plotted in a scattergram, the pattern of the association forms a straight line, without a curve.

The Scattergram

Definition of Scattergram. A **scattergram** (or *scatterplot*) is a graph on which you plot each case or observation. Each axis represents the value of one variable. It is used for variables measured at the interval or ratio level, rarely for ordinal variables, and never if either variable is nominal. There is no fixed rule for determining which variable (independent or dependent) to place on the horizontal or vertical axis, but usually the independent variable (symbolized by the letter X) goes on the horizontal axis and the dependent variable (symbolized by Y) on the vertical axis. The lowest value for each should be the lower left corner and the highest value should be at the top or to the right.

Constructing a Scattergram. Begin with the range of the two variables. Draw an axis with the values of each variable marked and write numbers on each axis (graph paper is helpful). Next label each axis with the variable name and put a title at the top. You are now ready to enter the data. For each case, find the value of each variable and mark the graph at a place corresponding to the two values. For example, you want to make a scattergram of years of schooling by number of children. You look at the first case to see years of schooling (e.g., 12) and number of children (e.g., 3). Then you go to the place on the graph where 12 for the "schooling" variable and 3 for the "number of children" variable intersect and put a dot for the case. You repeat this for each case until all are plotted on the scattergram.

The scattergram in Figure 5 is a plot of data for 33 women. It shows a negative relationship between the years of education the woman completed and the number of children she gave birth to.

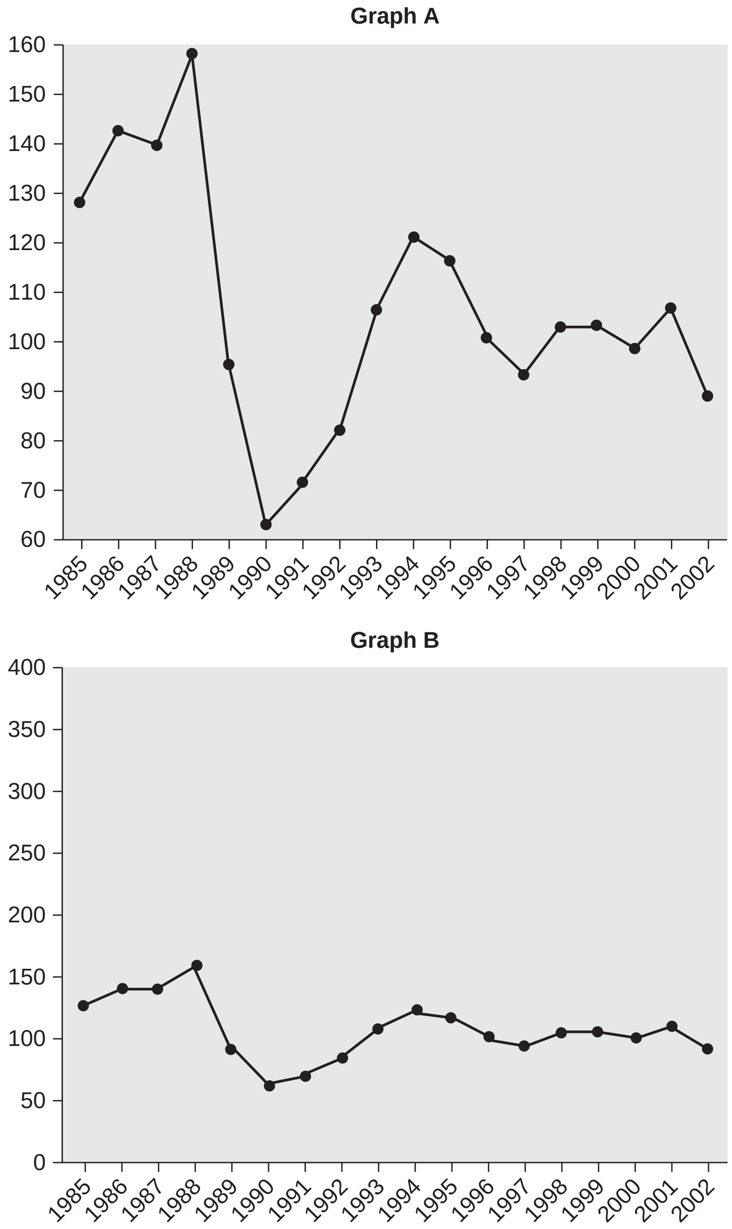
A scattergram shows us three aspects of a bivariate relationship: form, direction, and precision.

1. *Form.* Relationships can take three forms: independence, linear, and curvilinear. Independence or no relationship is the easiest to see. It looks like a random scatter with no pattern, or a straight line that is exactly parallel to the horizontal or vertical axis. A **linear relationship** means that a straight line can be visualized in the middle of a maze of cases running from one corner to another. A

CHART 1 Graphing Accurately

The pattern in graph A shows drastic change. A steep drop in 1990 is followed by rapid recovery and instability. The pattern in graph B is much more constant. The decline from 1989 to 1990 is smooth, and the other years are almost level. Both graphs are for identical data, the U.S. business failure rate from 1985 to 2002. The *X* axis (bottom) for years is the same.

The scale of the *Y* axis is 60 to 160 in graph A and 0 to 400 in graph B. The pattern in graph A only looks more dramatic because of the *Y* axis scale. When reading graphs, be careful to check the scale. Some people purposely choose a scale to minimize or dramatize a pattern in the data.



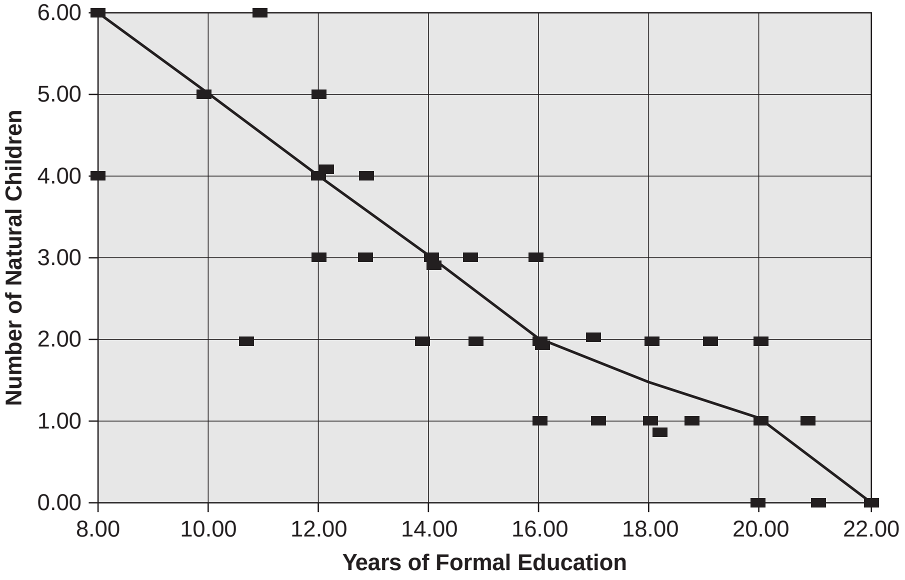


FIGURE 5 Example of a Scattergram: Years of Education by Number of Natural Children for 33 Women

curvilinear relationship means that the center of a maze of cases would form a U curve, right side up or upside down, or an S curve.

2. *Direction.* Linear relationships can have a positive or negative direction. The plot of a positive relationship looks like a diagonal line from the lower left to the upper right. Higher values on *X* tend to go with higher values on *Y,* and vice versa. The income and life expectancy example described a positive linear relationship. A negative relationship looks like a line from the upper left to the lower right. It means that higher values on one variable go with lower values on the other. For example, people with more education are less likely to have been arrested. If we look at a scattergram of data on a group of males that plots years of schooling (*X* axis) by number of arrests (*Y* axis), we see that most cases (or men) with many arrests are in the lower right because most of them completed fewer years of school. Most cases with few arrests are in the upper left because most have had more schooling. The imaginary line for the relationship can have a shallow or a steep slope. More advanced statistics provide precise numerical measures of the line's slope.

3. *Precision.* Bivariate relationships differ in their degree of precision. *Precision* is the amount of spread in the points on the graph. A high level of precision occurs when the points hug the line that summarizes the relationship. A low level occurs when the points are widely spread around the line. We can "eyeball" a highly precise relationship or use advanced statistics to measure the precision of a relationship in a way that is analogous to the standard deviation for univariate statistics.

Curvilinear relationship An association between two variables so that as the values of one variable increase, the values of the second show a changing pattern, for example, first decrease, then increase, and finally decrease; not a linear relationship.

Bivariate Tables

We use the bivariate **contingency table** in many situations. It presents the same information as a scattergram in a more condensed form. One advantage of it over the scattergram is that the data can be measured at any level of measurement, although interval and ratio data must be grouped.

The bivarate contingency table is based on **cross-tabulation** (i.e., tabulating two or more variables simultaneously). It is "contingent" because the cases in each category of a variable are distributed into each category of a second (or additional) variable. The table distributes cases into the categories of multiple variables at the same time and shows us how the cases, by category of one variable, are "contingent upon" the categories of other variables.

Constructing Percentaged Tables. Contingency tables made up of the counts of a case are of limited use because seeing patterns or variable relationships with the counts of cases is difficult. By "standardizing" data, or turning them into percentages, we can see patterns and relationships among variables more easily even if the counts of cases vary greatly. It is not difficult to construct a percentaged table, and there are ways to make it look professional. We first review the steps for constructing a table by hand. The same principles apply if a computer makes the table for you. We begin with the raw data (see data from an imaginary survey in Example Box 2, Raw Data and Frequency Distributions).

If you create a table by hand, you may find an intermediate step between raw data and the table useful (i.e., create a compound frequency distribution [CFD]). It is similar to the frequency distribution except that it is for each combination of the values of two variables. For example, you want to see the relationship between age and attitude about the legal age to drink alcohol. Age is a ratio measure, so you group it to treat the ratio-level variable as if it were ordinal. In percentage tables, we group ratio- or interval-level data to convert them into the ordinal level. Otherwise, we might have 50 categories for a variable and a table that is impossible to read.

The CFD has every combination of category. Age has four categories and Attitude three, so there are 3 × 4 = 12 rows. The steps to create a CFD are as follows:

1. Determine all possible combinations of variable categories.
2. Make a mark next to the combination category into which each case falls.
3. Add the marks for the number of cases in a combination category.

If there is no missing information problem, add the numbers of categories (e.g., all the "Agree"s, or all the "61 and Older"s). In the example, missing data are an issue. The four "Agree" categories in the CFD add to 37 (20 + 10 + 4 + 3), not 38, as in the univariate frequency distribution, because one of the 38 cases has missing information for age.

The next step is to set up the parts of a table (see Figure 6) by labeling the rows and columns. The independent variable usually is placed in the columns, but this convention is not always followed. Next, each number from the CFD is placed in a cell in the table that corresponds to the combination of variable categories. For example, the CFD shows that 20 of the under-30-year-olds agree (top number) as does Figure 6 (upper left cell).

Figure 6 is a raw count or frequency table. Its cells contain a count of the cases. It is easy to make but very difficult to interpret because the rows or columns can have different totals. What is of real interest is the relative size of cells compared to others.

Raw count tables can be converted into percentaged tables in three ways: percent by row, by column, and by total. The first two are often used to show relationships. The percent by total is almost never used and does not reveal relationships easily.

Is it best to percentage by row or column? Either can be appropriate. Here are the mechanics of making a percentage table. When calculating column percentages, compute each cell's percentage

Contingency table A summary format of the cross-tabulation of two or more variables showing bivariate quantitative data for variables in the form of percentages across rows or down columns for the categories of one variable.

Cross-tabulation The process of placing data for two variables in a contingency table to show the percentage or number of cases at the intersection of variable categories.

EXAMPLE BOX 2

Raw Data and Frequency Distributions

EXAMPLE OF RAW DATA

Case	*Age*	*Gender*	*Schooling*	*Attitude*	*Political Party, etc. . . .*
01	21	F	14	1	Democrat
02	36	M	8	1	Republican
03	77	F	12	2	Republican
04	41	F	20	2	Independent
05	29	M	22	3	Democratic Socialist
06	45	F	12	3	Democrat
07	19	M	13	2	Missing Information
08	64	M	12	3	Democrat
09	53	F	10	3	Democrat
10	44	M	21	1	Conservative

(Attitude scoring, 1 = Agree, 2 = No Opinion, 3 = Disagree)

TWO FREQUENCY DISTRIBUTIONS:
AGE AND ATTITUDE TOWARD CHANGING THE DRINKING AGE

Age Group	*Number of Cases*	*Attitude*	*Number of Cases*
Under 30	26		
30–45	30	Agree	38
46–60	35	No Opinion	26
61 and older	15	Disagree	40
Missing	3	Missing	5
Total	109	Total	109

COMPOUND FREQUENCY DISTRIBUTION:
AGE GROUP AND ATTITUDE TOWARD CHANGING THE DRINKING AGE

Age	*Attitude*	*Number of Cases*
Under 30	Agree	20
Under 30	No Opinion	3
Under 30	Disagree	3
30–45	Agree	10
30–45	No Opinion	10
30–45	Disagree	5
46–60	Agree	4
46–60	No Opinion	10
46–60	Disagree	21
61 and older	Agree	3
61 and older	No Opinion	2
61 and older	Disagree	10
	Subtotal	101
Missing on either variable		8
Total		109

FIGURE 6 Age Group by Attitude about Changing the Drinking Age, Raw Count Table

RAW COUNT TABLE (a)

	AGE GROUP (b)				
ATTITUDE (b)	*Under 30*	*30–45*	*46–60*	*61 and Older*	**TOTAL (c)**
Agree	20	10	4	3	37
No opinion	3 (d)	10	10	2	25
Disagree	3	5	21	10	39
Total (c)	26	25	35	15	101

Missing cases (f) = 8. (e)

THE PARTS OF A TABLE

(a) Give each table a *title,* which names variables and provides background information.
(b) Label the row and column variable and give a name to each of the variable categories.
(c) Include the totals of the columns and rows. These are called the **marginals.** They equal the univariate frequency distribution for the variable.
(d) Each number or place that corresponds to the intersection of a category for each variable is a **cell of a table.**
(e) The numbers with the labeled variable categories and the totals are called the **body of a table.**
(f) If there is missing information (cases in which a respondent refused to answer, ended interview, said, "don't know," etc.), report the number of missing cases near the table to account for all original cases.

of the column total. This includes the total column or **marginal**, which is the name for totals of a row or of a column variable. For example, look at the column marginals in Table 1. The first column total is 26 (there are 26 people under age 30), and the first cell of that column is 20 (there are 20 people under age 30 who agree). The percentage is 20/26 = 0.769, or 76.9 percent. Or, for the first number in the row marginal, which is 37, 37/101 = 0.366 = 36.6 percent. This tells you that 36.6 percent of cases agree. Except for rounding, the total should equal 100 percent.

Computing row percentages is similar. Compute the percentage of each cell as a percentage of the row total. For example, using the same cell with 20 in it, you now want to know what percentage 20 is of the row total of 37, or 20/37 = 0.541 = 54.1 percent. Percentaging by row or column gives different percentages for a cell unless the marginals are the same.

Row and column percentages let you address different questions. The row-percentaged table answers the question: Among those who want to lower the drinking age, what percentage comes from each age group? It says of respondents who agree, 54.1 percent are in the under-30 age group. The column-percentaged table addresses the question: Among those in each age group, what percentage holds different attitudes? It says that among

Marginal In a contingency table, the row of totals or the column of totals.

TABLE 1 Age Group by Attitude about Changing the Drinking Age, Percentaged Tables

COLUMN-PERCENTAGED TABLE

	AGE GROUP				
ATTITUDE	*Under 30*	*30–45*	*46–60*	*61 and Older*	**TOTAL**
Agree	76.9%	40.0%	11.4%	20.0%	36.6%
No opinion	11.5	40.0	28.6	13.3	24.8
Disagree	11.5	20.0	60.0	66.7	38.6
Total	99.9%	100%	100%	100%	100%
(N)	(26)*	(25)*	(35)*	(15)*	(101)*
Missing cases = 8					

ROW-PERCENTAGED TABLE

	AGE GROUP					
ATTITUDE	*Under 30*	*30–45*	*46–60*	*61 and Older*	**TOTAL**	***(N)***
Agree	54.1%	27%	10.8%	8.1%	100.0%	(37)*
No opinion	12.0	40.0	40.0	8.0	100.0	(25)*
Disagree	7.7	12.8	53.8	25.6	99.9	(39)*
Total	25.7%	24.8%	34.7%	14.9%	100.1%	(101)*
Missing cases = 8						

*For percentaged tables, provide the number of cases or *N* on which percentages are computed in parentheses near the total of 100%. This makes it possible to go back and forth from a percentaged table to a raw count table and vice versa.

those who are under 30, 76.9 percent agree. From the row percentages, you learn that a little over half of those who agree are under 30 years old. From column percentages, you learn that among the under-30 people, more than 75 percent agree. The first way of percentaging tells you about people with specific attitudes; the second tells you about people in specific age groups and lets you compare them.

Your hypothesis often tells you to look at either the row or column percentages. When beginning, you may want to calculate percentages each way and practice interpreting what each says. For example, your hypothesis is that a person's age affects his or her legal alcohol age attitude, and you are interested in the age of people most/least supportive. This suggests that you look at column percentages because you want to compare attitudes across the different age groups. However, if your interest is in describing the age makeup of groups of people with different attitudes, then row percentages are appropriate. Perhaps you want to buy TV advertising about the issue and you want to know what age group will be viewing the commercials. As Zeisel (1985:34) noted, whenever one factor in a cross-tabulation can be considered the cause of the other, the most illuminating percentage will be obtained by computing percentages in the direction of the causal factor. So, if age is your causal variable, create the percentage table by rows.

Unfortunately, there is no "industry standard" for putting the independent and dependent variable in a percentage table as row or column, or for percentage by row and column. A majority of researchers place the independent variable on the column and percentage by column, but a large minority put the independent variable as row and percentage by row.

Reading a Percentaged Table. Once you understand how to make a table, you will find it easier to read and figure out what the table says. To read a percentage table, first look at the title, the variable labels, and any background information. Next, look at the direction in which percentages have been computed: in rows or columns. Notice that the headings in Table 1 are the same. This is so because the same variables are used. It would be easier if headings included how the data are percentaged, but this is not done. Sometimes you will see abbreviated tables that omit the 100 percent total or the marginals, which adds to the confusion. When you create a table, it is best to include all the parts of a table and use clear labels.

When you read percentaged tables, you will make comparisons in the opposite direction from that in which percentages are computed. This sounds confusing but is simple in practice. A rule of thumb is to compare across rows if the table is percentaged down (i.e., by column) and to compare up and down in columns if the table is percentaged across (i.e., by row).

For example, in row-percentaged Table 1, compare columns or age groups. Most of those who agree are in the youngest group. The proportion saying they agree declines as age increases. Most no-opinion people are in the middle-age groups whereas those who disagree are older, especially in the 46-to-60 group. When reading column-percentaged Table 1, compare across rows. You can see that a majority of the youngest group agree, and they are the only group in which most people agree. Only 11.5 percent disagree, compared to a majority in the two oldest groups.

Seeing a relationship in a percentaged table takes practice. If there is no relationship in a table, the cell percentages look approximately equal across rows or columns. A linear relationship appears like larger percentages in the diagonal cells. If there is a curvilinear relationship, the largest percentages form a pattern across cells. For example, the largest cells might be the upper right, the bottom middle, and the upper left. It is easiest to see a relationship in a moderate-size table (9 to 16 cells) in which most cells have some cases (at least five are recommended) and the relationship is strong and precise.

Principles of reading a scattergram can help you see a relationship in a percentage table. Imagine a scattergram divided into 12 equal-size sections. The cases in each section correspond to the number of cases in the cells of a table that is superimposed onto the scattergram. You can think of the table as a condensed form of the scattergram. The bivariate relationship line in a scattergram corresponds to the diagonal cells in a percentaged table. Thus, a simple way to see strong relationships is to circle the largest percentage in each row (for row-percentaged tables) or column (for column-percentaged tables) and see whether a line appears.

The circle-the-largest-cell rule works—with one important caveat. The categories in the percentages table must be ordinal or interval and in the same order as in a scattergram. In scattergrams the lowest variable categories begin at the bottom left. If the categories in a table are not ordered the same way, the rule does not work.

For example, Table 2a looks like a positive relationship and Table 2b like a negative relationship. Both use the same data and are percentaged by row. The actual relationship is negative. Look closely: Table 2b has age categories ordered as in a scattergram. When in doubt, return to the basic difference between positive and negative relationships. A positive relationship means that as one variable increases, so does the other. A negative relationship means that as one variable increases, the other decreases.

Bivariate Tables without Percentages. Another kind of bivariate table condenses information—a measure of central tendency (usually the mean). You can use it when one variable is nominal or ordinal and another is measured at the interval or ratio level. The mean (or a similar measure) of the interval or ratio variable is presented for each category of the nominal or ordinal variable. Do not construct the measure of central tendency from the CFD. Instead, divide the cases into the ordinal or

TABLE 2A Age by Schooling

	YEARS OF SCHOOLING				
AGE	*0–11*	*12*	*13–14*	*16+*	TOTAL
Under 30	5%	25	30	40	100
30–45	15	25	40	20	100
46–60	35	45	12	8	100
61+	45	35	15	5	100

TABLE 2B Age by Schooling

	YEARS OF SCHOOLING				
AGE	*0–11*	*12*	*13–14*	*16+*	TOTAL
61+	45%	35	15	5	100
46–60	35	45	12	8	100
30–45	15	25	40	20	100
Under 30	5	25	30	40	100

TABLE 3 Attitude about Changing the Drinking Age by Mean Age of Respondent

DRINKING AGE ATTITUDE	MEAN AGE	(*N*)
Agree	26.2	(37)
No opinion	44.5	(25)
Disagree	61.9	(39)

Missing cases = 8

nominal variable categories; then calculate the mean for the cases in each variable category from the raw data. Table 3 shows the mean age of people in each of the attitude categories. The results suggest that the mean age of those who disagree is much higher than for those who agree or have no opinion.

Measures of Association

A *measure of association* is a single number that expresses the strength, and often the direction, of a relationship. It condenses information about a bivariate relationship into a single number. There are many measures of association. The correct one to use depends on the level of measurement of the data and specific research purposes. Many measures are identified by letters of the Greek alphabet. Lambda, gamma, tau, chi (squared), and rho are commonly used measures. The emphasis here is on interpreting the measures, not on their calculation. To understand each measure, you will need to complete at least one statistics course. Some measures of association, such as gamma, are for data measured at the ordinal level (see Expansion Box 2, Gamma). Other measures, such as the correlation coefficient, assume data measured at the ratio-level (see Expansion Box 3, Correlation).

Proportionate reduction in error A logic in many statistics that measures the strength of association between two variables. A strong association reduces most errors in predicting the dependent variable using information from the independent variable.

Most of the elementary measures discussed here follow a **proportionate reduction in error** logic. The logic asks how much does knowledge of one variable reduce the errors that are made when guessing the values of the other variable. *Independence* means that knowledge of one variable does not reduce the chance of errors on the other variable. Measures of association equal zero if the variables are independent.

If there is a strong association or relationship between the independent and dependent variable, we make few errors in predicting a dependent variable based on knowledge of the independent variable, or the proportion of errors reduced is large. A large number of correct guesses suggests that the measure of association is a nonzero number if an association exists between the variables. Table 4 describes five commonly used bivariate measures of association. Notice that most range from –1 to +1, with negative numbers indicating a negative relationship and positive numbers a positive relationship. A measure of 1.0 means a 100 percent reduction in errors, or perfect prediction.

EXPANSION BOX 2

Gamma

Gamma is a comparatively simple statistic that measures the strength of an association between two ordinal-level variables. This bivariate measure requires you to specify which variable is independent and which is dependent in a hypothesis. It illustrates the basic logic of other measures of association.

Gamma allows you to predict the rank of one variable based on knowledge of the rank of another variable. Essentially, it answers this question: If you know how I rank on variable 1, how good is your prediction of my rank on variable 2? For example, if you know my letter grade in mathematics, how accurately can you predict my grade in literature? Perfect prediction or the highest possible gamma is +1 or −1, depending on whether the ranks are the same (positive) or the opposite on another (negative). Perfect statistical independence of the two variables is a gamma of zero. The formula for calculating gamma uses data in the cells in the body of a cross-tabulation.

Let us look at a simple example using real data from a national sample of adults in United States in 2008 (the GSS). A total of 672 people were asked questions about their happiness and health. Many health care professionals and social scientists noted that emotional happiness is associated with being healthier, so we can test the hypothesis that happy people are healthier.

By looking at the raw count or frequency table, we see from the marginals that most people are pretty happy and more say they are in good health.

Would You Say Your Own Health in General Is:	*Taking All Things Together, How Would You Say Things Are these Days?*			
	Very Happy	Pretty Happy	Not Too Happy	Total
Excellent	63 A	100 D	19 G	182
Good	93 B	190 E	53 H	336
Fair or poor	27 C	77 F	50 I	154
Total	183	327	122	672

Gamma is based on the idea of "paired observations" (i.e., observations compared in terms of their relative rankings on the independent and dependent variables). Concordant (same-order) paired observations show a positive association, that is, when the member of the pair ranked higher on the independent variable is also ranked higher on the dependent variable. Discordant (inverse-order) paired observations show a negative association. The member of the pair ranked higher on the independent variable is ranked lower on the dependent variable

The formula for gamma is

$$\text{Gamma} = [(P-Q)/(P+Q)]$$

Where P = concordant and Q = discordant pairs.

Gamma ranges from −1.0 to zero to +1.0 and is a proportionate reduction in error statistic. If Gamma = 0 means the extra information provided by the independent variable does not help prediction. The higher the gamma, the more strength there is in predicting the dependent variable. Gamma can be positive or negative, giving a direction of the association between the variables. When there are more concordant pairs, gamma will be positive; when there are more discordant pairs, gamma will be negative.

Gamma compares cells that are concordant (i.e., same ranked) on the independent and dependent variables to those that are discordant (i.e., opposite ranked) and ignores tied cells (i.e., cells where the independent and dependent variable are ranked the same). The table shown on the left has nine cells. First, let us identify all "concordant" pairs of cells (each cell has a letter).

Cell A in the upper left and Cell F are concordant. Because they are along a diagonal from upper left to lower right, this is predicted in the hypothesis (i.e., very happy people have excellent health, pretty happy have good health, etc.). Other concordant pairs are E:I, B:F, and D:H for the same reason. In the opposite direction are discordant pairs center, G:E, E:C, D:B, and H:F. We multiply the number of cases in each pair. In the formula these are (A x (E + F + H + I)) + (D x (H + I)) + (B x (F + I)) + (E x I) for concordant pairs and (G x (B + E + C + F) + (D x (B + C)) + (H x (C + F)) + (E x C) for discordant pairs. Substituting the number of cases for each cell, this becomes (63 x (190 + 77 + 53 + 50)) + (100 x (53 + 50)) + (93 x

(*continued*)

EXPANSION BOX 2

(continued)

(77 + 50)) + (190 x 50) = 23310 + 10300 +11811 + 9500 = 54921 concordant pairs. Also (19 x (93 + 190 + 27 + 77) + (100 x (93 + 27)) + (53 x (77 + 27)) + (190 x 27) = 7353 + 5512 + 12000 + 5130 = 29995 discordant pairs. Putting this into the formula, (54921 – 29995)/(54921 + 29995) = 0.2935. Computers usually do the calculations for us. A gamma of .2935 suggests a weak positive relationship or that health and happiness tend to go together somewhat.

Interpreting gamma (+ means positive relation, – means negative relation):

GAMMA	MEANING
0.00 to 0.24	No relationship
0.25 to 0.49	Weak relationship (positive or negative)
0.50 to 0.74	Moderate relationship (positive or negative)
0.75 to 1.00	Strong relationship (positive or negative)

EXPANSION BOX 3

Correlation

The formula for a correlation coefficient (rho) looks awesome to most people. Calculating it by hand, especially if the data have multiple digits, can be a very long and arduous task. Nowadays, computers do the calculation. However, the problem with relying on computers to do the work is that a researcher may not understand what the coefficient means. Here is a short, simplified example to show how it is done.

The purpose of a correlation coefficient is to show how much two variables "go together" or covary. Ideally, the variables have a ratio level of measurement (some use variables at the interval level). To calculate the coefficient, we first convert each score on a variable into its *z*-score. This "standardizes" the variable based on its mean and standard deviation. Next we multiply the *z*-scores for each case together. This tells us how much the variables for a case vary together—cases with high *z*-scores on both variables are much larger, while those low on both are much smaller. Finally, we divide the sum of the multiplied *z*-scores by the number of cases. It yields a type of "average" covariation that has been standardized. In short, a correlation coefficient is the product of *z*-scores added together and then divided by the number of cases. It is always between +1.0 and –1.0 and summarizes scattergram information about a relationship into a single number.

Let us look at the correlation between the age and price for five small bottles of red wine. First, anyone who is brave or lacks math-symbol phobia can look at one of the frequently used formulas for a correlation coefficient:

$$(\Sigma\ [z\text{-score}_1][z\text{-score}_2])/N$$

where: Σ = sum, *z*-score$_1$ = *z*-score for 1st variable (see Expansion Box 12.1), *z*- score$_2$ = *z*-score for 2nd variable, *N* = number of cases

Here is how to calculate a correlation coefficient without directly using the formula:

			(DIFFERENCE)		SQUARED DIFF.		*Z*-SCORES		*Z*-SCORE
WINE	AGE	PRICE	*Age*	*Price*	*Age*	*Price*	*Age*	*Price*	*Product*
A	2	$10	–2	–5	4	25	–1.43	–0.70	1.00
B	3	5	–1	–10	1	100	–1.41	1.00	
C	5	20	+1	+5	1	25	0.71	+0.70	0.50
D	6	25	+2	+10	4	100	+1.43	+1.41	2.00
E	4	15	0	0	0	0	0.00	0.00	0.00
Total	20	$75			10	250			4.50

(*continued*)

EXPANSION BOX 3 (continued)

Mean: Age = 4; Price = $15
Variance: Age = 10/5 = 2; Price = 250/5 = 50.
Stnd. Dev.: Age = square root of 2 = 1.4; Price = square root of 50 = 7.1
Correlation: 4.50/5 = .90

Step 1: Calculate the mean and standard deviation for each variable. (For the standard deviation, first subtract each score from its mean, next square the difference, sum squared differences, and then divide the sum by the number of cases for the variance. Then take the square root of the variance.)
Step 2: Convert each score for the variables into their *z*-scores. (Just subtract each score from its mean and divide by its standard deviation.)
Step 3: Multiply the *z*-scores together for each case.
Step 4: Sum the products of *z*-scores and then divide by the number of cases.

TABLE 4 Five Measures of Association

Lambda is used for nominal-level data. It is based on a reduction in errors based on the mode and ranges between zero (independence) and 1.0 (perfect prediction or the strongest possible relationship).

Gamma is used for ordinal-level data. It is based on comparing pairs of variable categories and seeing whether a case has the same rank on each. Gamma ranges from −1.0 to +1.0 with zero meaning no association.

Tau is also used for ordinal-level data. It is based on a different approach than gamma and takes care of a few problems that can occur with gamma. Actually, there are several statistics named tau (it is a popular Greek letter), and the one here is Kendall's tau. Kendall's tau ranges from −1.0 to +1.0, with zero meaning no association.

Rho is also called Pearson's product moment correlation coefficient (named after the famous statistician Karl Pearson and based on a product moment statistical procedure). It is the most commonly used measure of correlation, the correlation statistic people mean if they use the term *correlation* without identifying it further. It can be used only for data measured at the interval or ratio level. Rho is used for the mean and standard deviation of the variables and tells how far cases are from a relationship (or regression) line in a scatterplot. Rho ranges from −1.0 to +1.0 with zero meaning no association. If the value of rho is squared, sometimes called *R*-squared (R^2), it has a unique proportion reduction in error meaning. *R*-squared tells how the percentage in one variable (e.g., the dependent) is accounted for, or explained by, the other variable (e.g., the independent). Rho measures linear relationships only. It cannot measure nonlinear or curvilinear relationships. For example, a rho of zero can indicate either no relationship or a curvilinear relationship (see Expansion Box 3).

Chi-square has two different uses. It can be used as a measure of association in descriptive statistics like the others listed here or in inferential statistics. As a measure of association, chi-square can be used for nominal and ordinal data. It has an upper limit of infinity and a lower limit of zero, meaning no association (see Expansion Box 3).

(*continued*)

TABLE 4 continued

SUMMARY OF MEASURES OF ASSOCIATION

Measure	*Greek Symbol*	*Type of Data*	*High Association*	*Independence*
Lambda	λ	Nominal	1.0	0
Gamma	γ	Ordinal	+1.0, –1.0	0
Tau (Kendall's)	τ	Ordinal	+1.0, –1.0	0
Rho	ρ	Interval, ratio	+1.0, –1.0	0
Chi-square	χ^2	Nominal, ordinal	Infinity	0

MORE THAN TWO VARIABLES

Statistical Control

Demonstrating an association between two variables is an important first step for understanding the data. However, it is not sufficient for you to say that an independent variable causes a dependent variable. In addition to temporal order and association, we must eliminate alternative explanations that can make the hypothesized relationship spurious. Experimental researchers do this by choosing a research design that physically controls potential alternative explanations for results (i.e., that threaten internal validity).

In nonexperimental research, we can statistically control for alternative explanations with control variables (discussed shortly). We examine the control variables with multivariate tables and statistics that help us decide whether a bivariate relationship might be spurious. We can also show the relative size of the effect of multiple independent variables on a dependent variable.

A **control variable** is a third (or fourth or fifth) variable that represents an alternative explanation for a two-variable relationship. It is a "control" in that is adjusts for, or takes into account, the effects of variables other than the primary independent and dependent variable of a hypothesis. For example, your bivariate table shows that taller teenagers like baseball more than shorter ones do. But the bivariate relationship between height and attitude toward baseball might be spurious. Why is this; because you suspect that teenage males are taller than females and you suspect that males like baseball more than females do? To test whether the relationship is actually due to height, you must control for gender. By controlling for gender, you are statistically removing their effect. Once you do this, you can see whether the bivariate relationship between height and attitude toward baseball remains or whether the association between height and baseball attitude was really due to gender.

Control variable A "third" factor that shows whether a bivariate relationship holds up to alternative explanations; can occur before or between other variables.

You can "control for" a third variable by seeing whether the bivariate relationship persists within categories of the control variable. For example, you control for gender, and the relationship between height and baseball attitude persists. This means that tall males and tall females both like baseball more than short males and short females do. In other words, the control variable has no effect. When this is so, the bivariate relationship is not spurious, and the control variable (suspected alternative explanation) has no effect.

What if the bivariate relationship weakens or disappears after you control for gender? It means that tall males are no more likely than short males to like baseball, and tall females are no more likely to like baseball than short females. It indicates that the initial bivariate relationship is spurious and suggests that the third variable (in this case gender), not height, is the true cause of differences in attitudes toward baseball.

Statistical control is a central idea used in many advanced statistical techniques. A measure of association such as the correlation coefficient only suggests a relationship. Until you consider control variables, the bivariate relationship might be spurious. This is why researchers are cautious in interpreting bivariate relationships until they have considered control variables.

After you introduce control variables, you see the **net effect** of an independent variable, that is, the effect of the independent variable "net of," or in spite of, the control variable. We briefly look at two ways to introduce control variables: trivariate percentaged tables and multiple regression analysis.

The Elaboration Model of Percentaged Tables

Constructing Trivariate Tables. To meet the conditions needed for causality, we want to "control for" or see whether an alternative explanation eliminates a causal relationship. If an alternative explanation accounts for a relationship, then the bivariate relationship may be spurious. We operationalize alternative explanations as third or control variables.

You can consider such third variables by statistically introducing control variables in trivariate or three-variable tables. Trivariate tables differ only slightly from bivariate tables. In a sense, they consist of multiple bivariate tables. A trivariate table consists of a separate bivariate table of the independent and dependent variables created for each category of the control variable. The multiple tables of your independent and dependent variable, one for each control variable category, are its **partials**. The tables partial out the effects based on the control variable. The number of partials depends on the number of categories in the control variable. Partial tables look just like bivariate tables, but they use a subset of the cases. Only cases with a specific value on the control variable are in the partial. Thus, you can combine the partials to restore the initial bivariate table without a control variable.

Trivariate tables have three limitations. First, they are difficult to interpret if a control variable has more than four categories. Second, control variables can be at any level of measurement, but you must group interval-level or ratio-level control variables (i.e., convert them to the ordinal level). Finally, the total number of cases is a limiting factor because the cases are divided among cells in partials. The number of cells in the partials equals the number of cells in the bivariate relationship multiplied by the number of categories in the control variable. For example, a control variable has three categories, and a bivariate table has 12 cells, so the partials have $3 \times 12 = 36$ cells. An average of five cases per cell is recommended, so $5 \times 36 = 180$ cases at minimum are required.

Like bivariate table construction, a trivariate table begins with a CFD but a three-way instead of a two-way CFD. An example of a trivariate table with "gender" as a control variable for the bivariate relation in Table 1 is shown in Table 5.

As with the bivariate tables, each combination in the CFD represents a cell in the final (here the partial) table. Each partial table has the variables in an initial bivariate table. For three variables, three bivariate tables are logically possible. In the example of Table 5, the combinations are (1) gender by attitude, (2) age group by attitude, and (3) gender by age group. The partials are set up on the basis of the initial bivariate relationship. The independent variable in each is age group, the dependent variable is attitude, and gender is the control variable. Thus, the trivariate table consists of a pair of partials, each showing the age/attitude relationship for a given gender.

Your theory and understanding of the social world suggest both the hypothesis in the initial bivariate relationship and which variables might be alternative explanations (i.e., the control variables).

As with bivariate tables, the CFD provides the raw count for cells (partials here). You convert them

Net effect The result of one variable (usually independent) on another (usually dependent) after the impact of control variables that affects both has been statistically removed.

Partials In contingency tables for three variables, tables between the independent and dependent variables for each category of a control variable.

TABLE 5 CFD and Tables for a Trivariate Analysis

COMPOUND FREQUENCY DISTRIBUTION FOR TRIVARIATE TABLE

MALES			**FEMALES**		
Age	*Attitude*	*Number of Cases*	*Age*	*Attitude*	*Number of Cases*
Under 30	Agree	10	Under 30	Agree	10
Under 30	No opinion	1	Under 30	No opinion	2
Under 30	Disagree	2	Under 30	Disagree	1
30–45	Agree	5	30–45	Agree	5
30–45	No opinion	5	30–45	No opinion	5
30–45	Disagree	2	30–45	Disagree	3
46–60	Agree	2	46–60	Agree	2
46–60	No opinion	5	46–60	No opinion	5
46–60	Disagree	11	46–60	Disagree	10
61 and older	Agree	3	61 and older	Agree	0
61 and older	No opinion	0	61 and older	No opinion	2
61 and older	Disagree	5	61 and older	Disagree	5
	Subtotal	51	Subtotal		50
Missing on either variable		4	Missing on either variable		4
Number of males		55	Number of females		54

PARTIAL TABLE FOR MALES

	AGE GROUP				
ATTITUDE	*Under 30*	*30–45*	*46–60*	*61 and Older*	**TOTAL**
Agree	10	5	2	3	20
No Opinion	1	5	5	0	11
Disagree	2	2	11	5	20
Total	13	12	18	8	51
Missing cases = 4					

PARTIAL TABLE FOR FEMALES

	AGE GROUP				
ATTITUDE	*Under 30*	*30–45*	*46–60*	*61 and Older*	**TOTAL**
Agree	10	5	2	0	17
No Opinion	2	5	5	2	14
Disagree	1	3	10	5	19
Total	13	13	17	7	50
Missing cases = 4					

into percentages in the same way as for a bivariate table (i.e., divide cells by the row or column total). For example, in the partial table for females, the upper left cell has a 10. The row percentage for that cell is 10/17 = 58 percent.

The **elaboration paradigm** is a system for reading percentaged trivariate tables.[5] It describes five possible patterns that might emerge after you add a control variable. The patterns describe how the partial tables compare to the initial bivariate table, or how the original bivariate relationship changes after you add the control variable (see Example Box 3, Summary of Elaboration Paradigm). The examples of patterns presented here show strong cases. You will need to use advanced statistics when the differences are not as obvious.

Of the five patterns, the **replication pattern** is the easiest to understand. It occurs when the partials replicate or reproduce the same relationship that existed in the bivariate table before considering the control variable, and means that the control variable has no effect. The **specification pattern** is the next easiest pattern. It occurs when one partial replicates the initial bivariate relationship but other partials do not. For example, you find a strong (negative) bivariate relationship between automobile accidents and college grades. You control for gender and discover that the relationship holds only for males (i.e., the strong negative relationship was in the partial for males, not for females). This is the *specification* because you specify the category of the control variable in which the initial relationship persists.

The control variable has a large effect in both the interpretation and explanation patterns. In both, the bivariate table shows a relationship that disappears or greatly weakens in the partials. In other words, you saw a relationship between the independent and dependent variables in a bivariate table, but the relationship disappears and the variables appear to be independent in the partial tables. You cannot distinguish between the two patterns by looking at the tables alone. The difference between the patterns depends on the location of the control variable in the causal order of variables. Theoretically, a control variable can be in one of two places, either between the original independent and dependent variables (i.e., the control variable is intervening), or before the original independent variable.

The **interpretation pattern** describes the situation in which the control variable intervenes between the original independent and dependent variables. For example, you examine a relationship between religious upbringing and abortion attitude. Political ideology is a control variable. You reason that religious upbringing affects current political ideology and abortion attitude. You theorize that political ideology is logically prior to an attitude about a specific issue, such as abortion. Thus, religious upbringing causes political ideology, which in turn has an impact on abortion attitude. The control variable is an intervening variable, which helps you interpret the meaning of the complete relationship.

The **explanation pattern** looks the same as the interpretation pattern. The difference is the temporal order of the control variable. In the explanation pattern, a control variable comes before the independent variable in the initial bivariate relationship. For example, the original relationship is between religious upbringing and abortion attitude, but now gender is the control variable. Gender comes before

Elaboration paradigm A system for describing patterns evident among tables when the bivariate contingency table is compared with partials after the control variable has been added.

Replication pattern An arrangement in the elaboration paradigm in which the partials show the same relationship as in a bivariate contingency table of the independent and dependent variable alone.

Specification pattern An arrangement in the elaboration paradigm in which the bivariate contingency table shows a relationship; one of the partial tables but others do not.

Interpretation pattern An arrangement in the elaboration paradigm in which the bivariate contingency table shows a relationship, but the partials show no relationship and the control variable is intervening in the causal explanation.

Explanation pattern A pattern in the elaboration paradigm in which the bivariate contingency table shows a relationship, but the partials show no relationship, and the control variable occurs prior to the independent variable.

EXAMPLE BOX 3

Summary of the Elaboration Paradigm

Pattern Name	*Pattern Seen When Comparing Partials to the Original Bivariate Table*
Replication	Relationship in both partials is same as in bivariate table.
Specification	Bivariate relationship is seen only in one of the partial tables.
Interpretation	Bivariate relationship weakens greatly or disappears in the partial tables (control variable is intervening).
Explanation	Bivariate relationship weakens greatly or disappears in the partial tables (control variable is before independent variable).
Suppressor variable	No bivariate relationship exists; relationship appears in partial tables only.

EXAMPLES OF ELABORATION PATTERNS

Replication (percentages)

	BIVARIATE TABLE			PARTIALS			
				Control = Low		*Control = High*	
	Low	*High*		*Low*	*High*	*Low*	*High*
Low	85%	15%	Low	84%	16%	86%	14%
High	15%	85%	High	16%	84%	14%	86%

Interpretation or Explanation (percentages)

	BIVARIATE TABLE			PARTIALS			
				Control = Low		*Control = High*	
	Low	*High*		*Low*	*High*	*Low*	*High*
Low	85%	15%	Low	45%	55%	55%	45%
High	15%	85%	High	55%	45%	45%	55%

Specification (percentages)

	BIVARIATE TABLE			PARTIALS			
				Control = Low		*Control = High*	
	Low	*High*		*Low*	*High*	*Low*	*High*
Low	85%	85%	Low	95%	5%	50%	50%
High	15%	15%	High	5%	95%	50%	50%

Suppressor Variable (percentages)

	BIVARIATE TABLE			PARTIALS			
				Control = Low		*Control = High*	
	Low	*High*		*Low*	*High*	*Low*	*High*
Low	54%	46%	Low	84%	16%	14%	86%
High	46%	54%	High	16%	84%	86%	14%

religious upbringing because one's gender is fixed at birth. The explanation pattern changes how a researcher explains the results. It implies that the initial bivariate relationship is spurious.

The **suppressor variable pattern** occurs when the bivariate tables suggest independence but a relationship appears in one or both of the partials. For example, religious upbringing and abortion attitude are independent in a bivariate table. Once you introduce the control variable region of the country, you see that religious upbringing is associated with abortion attitude in the partial tables. The control variable suppressed the true relationship, and the true relationship appears in the partials.

Multiple Regression Analysis

Multiple regression is a popular statistical technique whose calculation is beyond the level of this book. Although by using appropriate statistics software you can compute multiple regression quickly, a background in statistics is needed to prevent you from making errors in its calculation and interpretation. Multiple regression requires interval- or ratio-level data.

Multiple regression's great advantage is its ability to adjust for several control variables (i.e., alternative explanations) simultaneously. With percentaged tables, you can rarely use more than one control variable at a time. In addition, multiple regression is widely used, and you are likely to encounter it when reading research reports or articles. Multiple regression results tell the reader two things. First, it tells the overall predictive power of the set of independent and control variable on the dependent variable. A statistic, R-squared (R^2), tells us how well a set of variables "explains" a dependent variable. *Explain* here means making fewer errors when predicting the dependent variable scores on the basis of information about the independent variables. A good model with several variables might account for, or explain, a large percentage of variation in a dependent variable. For example, an R^2 of 0.50 means that knowing the independent and control variables improves the accuracy of predicting the dependent variable by 50 percent and that you would make one-half as many errors in predicting the dependent variable with the variable as you would not knowing about the independent and control variables.

Second, multiple regression results give the direction and size of the effect of each variable on a dependent variable. The effect is measured precisely with a numerical value. The higher the value, the larger the effect of a variable on predicting the dependent variable. The sign (positive or negative) of the effect tells you the direction of the impact on the dependent variable. For example, you can see how five independent or control variables simultaneously affect a dependent variable with all variables controlling for the effects of one another. This is especially valuable for testing theories that state that multiple independent variables cause one dependent variable.

We measure effect of an independent or control variable on the dependent variable by using a standardized regression coefficient or the Greek letter beta (ß). It is similar to a correlation coefficient, and ranges from zero to +0.99 or –0.99 with zero meaning no effect. We can perform statistical tests to determine the statistical significance (discussed later in this chapter) of a coefficient. The beta coefficient for two variables equals the correlation coefficient.

We use the beta regression coefficient to determine whether control variables have an effect. For example, the bivariate correlation between *X* and *Y* is 0.75. Next, we statistically add four control variables. If the beta remains at 0.75, the four control variables have no effect. However, if the beta for *X* and *Y* becomes smaller (e.g., drops to 0.20), the control variables have an effect on the dependent variable.

Consider an example of regression analysis with age, income, education, and region as independent variables. The dependent variable is a score on a political ideology index. The multiple regression results show that income and religious attendance have large effects, education and region minor effects, and age no effect. All independent variables together have a 38 percent accuracy in predicting a person's political

Supressor variable pattern Occurs when the bivariate tables suggest independence but a relationship appears in one or both partials.

EXAMPLE BOX 4

Example of Multiple Regression Results

DEPENDENT VARIABLE IS POLITICAL IDEOLOGY INDEX (HIGH SCORE MEANS VERY LIBERAL)

Independent Variable	*Standardized Regression Coefficients*
Region = South	−.19
Age	.01
Income	−.44
Years of education	.23
Religious attendance	−.39
$R^2 = .38$	

ideology (see Example Box 4, Example of Multiple Regression Results).[6] The example suggests that high income, frequent religious attendance, and a southern residence are positively associated with conservative opinions, whereas having more education is associated with liberal opinions. The impact of income is more than twice the size of the impact of living in a southern region.

Chart 2 summarizes the types and techniques of descriptive statistics. Next we turn our attention to inferential statistics.

INFERENTIAL STATISTICS

The Purpose of Inferential Statistics

The statistics discussed so far in this chapter are descriptive statistics. But we often want to do more than just describe; we want to test hypotheses, to find out whether sample results hold true in a population, and decide whether results (e.g., between the mean scores of two groups) are big enough to indicate that a relationship truly exists and is not due to chance alone. **Inferential statistics** build on probability theory to test hypotheses formally, permit inferences from a sample to a population, and test whether descriptive results are likely to be due to random factors or to a real relationship. This section explains the basic ideas of inferential statistics but does not deal with inferential statistics in any detail. This area is more complex than descriptive statistics and requires a background in statistics.

Inferential statistics rely on principles from probability sampling by which we use a random process (e.g., a random-number table, random computer process) to select cases from the entire population. Inferential statistics are a precise way to talk about how confident we can be when inferring from the results in a sample to the population.

You have already encountered inferential statistics if you have read or heard about "statistical significance" or results "significant at the 0.05 level." We use them to conduct various statistical tests (e.g., a *t*-test or an *F*-test). We use statistical significance in formal hypothesis testing, which is a precise way to decide whether to accept or to reject a null hypothesis.[7]

Inferential statistics A branch of applied mathematics based on random sampling that allows researchers to make precise statements about the level of confidence they can have that measures in a sample are the same as a population parameter.

Statistical significance The likelihood that a finding or statistical relationship in a sample's results is due to random factors rather than to the existence of an actual relationship in the entire population.

Statistical Significance

The term *statistically significant results* means that the results are not likely to be due to chance factors. **Statistical significance** indicates the probability of finding a relationship in the sample when there is none in the population. Because probability samples involve a random process, it is always possible that sample results will differ from a population parameter. We want to estimate the odds that sample results are due to a true population parameter or to chance factors of random sampling. With some probability theory from mathematics and specific statistical tests, we can tell whether the results (e.g., an association, a difference between two means, a regression coefficient) are likely to be produced by random error in random sampling

CHART 2 Summary of Major Types of Descriptive Statistics

TYPE OF TECHNIQUE	STATISTICAL TECHNIQUE	PURPOSE
Univariate	Frequency distribution, measures of central tendency, standard deviation, *z*-score	Describe one variable.
Bivariate	Correlation, percentage table, chi-square	Describe a relationship or the association between two variables
Multivariate	Elaboration paradigm, multiple regression	Describe relationships among several variables, or see how several independent variables have an effect on a dependent variable.

or are likely to show effects actually occurring in the social world.

Statistical significance tells us only what is likely. It cannot prove anything with absolute certainty. It states that particular outcomes are more or less probable. Statistical significance is not the same as practical, substantive, or theoretical significance. Results can be statistically significant but theoretically meaningless or trivial. For example, two variables can have a statistically significant association due to coincidence with no logical connection between them (e.g., length of fingernails and ability to speak French).

Levels of Significance

We usually express statistical significance in terms of levels (e.g., a test is statistically significant at a specific level) rather than giving the specific probability. The **level of statistical significance** (usually .05, .01, or .001) is an easy way of talking about the likelihood that results are due to chance factors, that is, that a relationship appears in the sample when there is none in the population. When we say that results are significant at the .05 level, we mean the following:

- Results like these are due to chance factors only 5 in 100 times.
- There is a 95 percent chance that the sample results are not due to chance factors alone but reflect the population accurately.
- The odds of such results based on chance alone are .05, or 5 percent.
- One can be 95 percent confident that the results are due to a real relationship in the population, not chance factors.

These all say the same thing in different ways. This may sound a bit like the discussion of sampling distributions and the central limit theorem in the chapter on sampling. It is no accident! Both are based on probability theory, which we use to link sample data to a population. Probability theory lets us predict what happens in the long run over many events when a random process is used. In other words, it allows us to make precise predictions over many situations in the long run but not for a specific situation. Because we have just one sample and we want to infer to the population, probability theory helps us estimate the odds that our particular sample represents the population. We cannot know for certain unless we have the whole population, but probability theory lets us state our confidence: how likely it is that the sample shows one thing while something else is true in the population.

Level of statistical significance A set of numbers that researchers use as a simple way to measure the degree to which a statistical relationship results from random factors rather than the existence of a true relationship among variables.

For example, a sample shows that college men and women differ in how many hours they study. Is the result due to having an unusual sample, and in reality there is no difference in the population, or does it reflect a true difference between the men and women? (See Example Box 5, Chi-Square.)

Type I and Type II Errors

The logic of statistical significance rests on whether chance factors might have produced the results. You may ask, why use the .05 level? We use it to mean a 5 percent chance that randomness could cause the results. Why not use a more certain standard—for example, a 1 in 1,000 probability of random chance? This gives a smaller chance that randomness versus a true relationship caused the results.

Type I Error The mistake made in saying that a relationship exists when in fact none exists; a false rejection of a null hypothesis.

Type II Error The mistake made in saying that a relationship does not exist when in fact it does; false acceptance of a null hypothesis.

There are two answers to this way of thinking. The simple answer is that the scientific community has informally agreed to use .05 as a rule of thumb for most purposes. Being 95 percent confident of results is the accepted standard for explaining the social world. A second, more complex answer involves a trade-off between making Type I and Type II errors. We can make two kinds of logical mistakes. A **Type I error** occurs when we say that a relationship exists when in fact none exists. It means falsely rejecting a null hypothesis. A **Type II error** occurs when we say that a relationship does not exist, when in fact it does. It means falsely accepting a null hypothesis (see Table 6). Of course, we want to avoid both errors and say a relationship is in the data only when it does indeed exist and there is no relationship only when there really is none. However, we face a dilemma: As the odds of making one type of error decline, the odds of making the opposite error increase.

You may find the ideas of Type I and Type II errors difficult at first, but the same logical dilemma appears outside research settings. For example, a jury can err by deciding that an accused person is guilty when in fact he or she is innocent, or the jury

EXAMPLE BOX 5

Chi-Square

The chi-square (χ^2) is used in two ways. This creates confusion. As a *descriptive statistic,* it tells us the strength of the association between two variables; as an *inferential statistic,* it tells us the probability that any association we find is likely to be due to chance factors. The chi-square is a widely used and powerful way to look at variables measured at the nominal or ordinal level. It is a more precise way to tell whether there is an association in a bivariate percentaged table than by just "eyeballing" it.

Logically, we first determine "expected values" in a table. We do this based on information from the marginals alone. Recall that marginals are frequency distributions of each variable alone. An expected value can be thought of as our "best guess" without examining the body of the table. Next we consider the data to see how much differs from the "expected value." If they differ a lot, then there may be an association between the variables. If the data in a table are identical or very close to the expected values, then the variables are not associated; they are independent. In other words, *independence* means "what is going on" in a table is what we would expect based on the marginals alone. Chi-square is zero if there is independence increases as the association gets stronger. If the data in the table greatly differ from the expected values, then we know something is "going on" beyond what we would expect from the marginals alone (i.e., an association between the variables). See the example of an association between height and grade.

EXAMPLE BOX 5
(continued)

Raw or Observed Data Table

STUDENT HEIGHT		GRADE IN RESEARCH METHODS			TOTAL
		C	B	A	
Tall		30	10	10	50
Medium		10	30	10	50
Short		30	20	50	100
	Total	70	60	70	200

Expected Values Table

Expected value = (Column total × Row total)/Grand total). EXAMPLE (70 × 50)/200 = 17.5

STUDENT HEIGHT		GRADE IN RESEARCH METHODS			TOTAL
		C	B	A	
Tall		17.5	15.0	17.5	50.0
Medium		17.5	15.0	17.5	50.0
Short		35.0	30.0	35.0	100.0
	Total	70.0	60.0	70.0	200.0

Difference Table

Difference = (Observed – Expected). EXAMPLE (30 – 17.5) = 12.5

STUDENT HEIGHT		GRADE IN RESEARCH METHODS			TOTAL
		C	B	A	
Tall		12.5	–5.0	–7.5	0.0
Medium		–7.5	15.0	–7.5	0.0
Short		–5.0	–10.0	15.0	0.0
	Total	0.0	0.0	0.0	0.0

Chi-square = Sum of each difference squared, then divided by the expected value of the cell. Example: 12.5 squared = 156.25, divided by 17.5 = 8.93.

Chi-square = 1st row (8.93 + 1.67 + 3.21) +
2nd row (3.21 + 15 + 3.21) +
3rd row (.71 + 3.33 + 6.43) = 45.7

Because chi-square is not zero, the data are not independent; there is an association. The chi-square coefficient cannot tell us the direction (e.g., negative) of the association. For inferential statistics, we need to use a chi-square table or computer program to evaluate the association (i.e., to see how likely such a large chi-square is to occur by chance alone). Without going into all the details about the chi-square table, this association is rare; it occurs by chance less than 1 in 1,000 times. For a table with nine cells, a chi-square of 45.7 is significant at the .001 level.

TABLE 6 Type I and Type II Errors

WHAT THE RESEARCHER SAYS	TRUE SITUATION IN THE WORLD	
	No Relationship	*Causal Relationship*
No relationship	No error	Type II error
Causal relationship	Type I error	No error

can err by deciding that a person is innocent when in fact she or he is guilty. The jury does not want to make either error. It does not want to jail the innocent or to free the guilty, but it must make a judgment using limited information. Likewise, a pharmaceutical company has to decide whether to sell a new drug. The company can err by stating that the drug has no side effects when, in fact, it has the side effect of causing blindness, or it can err by holding back a drug because of fear of serious side effects when in fact there are none. The company does not want to make either error. If it makes the first error, the company will face lawsuits and injure people. The second error will prevent the company from selling a drug that may cure illness and produce profits.

Combining the ideas of statistical significance and the two types of error together: If you are overly cautious and set a very high level of significance, you are likely to make one type of error. For example, you use the .0001 level. You attribute the results to chance only if they are so rare that they would occur by chance only 1 in 10,000 times. Such a high standard means that you are most likely to err by saying results are due to chance when in fact they are not. You may falsely accept the null hypothesis when there is a causal relationship (a Type II error). By contrast, if you are a risk-taking researcher and set a low level of significance, such as .10, your results indicate that a relationship would occur by chance 1 in 10 times. You are likely to err by saying that a causal relationship exists, when in fact random factors (e.g., random sampling error) actually cause the results. You are likely to falsely reject the null hypothesis (Type I error). In sum, the .05 level is a compromise between Type I and Type II errors.

This section has outlined the basics of inferential statistics. The statistical techniques are precise and rely on the relationship between sampling error, sample size, and central limit theorem. The power of inferential statistics is their ability to let us state, with specific degrees of certainty, that specific sample results are likely to be true in a population. For example, you conduct statistical tests and learn that a relationship is statistically significant at the .05 level. You can state that the sample results are probably not due to chance factors. Indeed, there is a 95 percent chance that a true relationship exists in the social world. Tests for inferential statistics are useful but limited. The data must come from a random sample, and tests consider only sampling errors. Nonsampling errors (e.g., a poor sampling frame or a poorly designed measure) are not considered. Do not be fooled into thinking that such tests offer easy, final answers. See the discussion presented in Expansion Box 4, Statistical Programs on Computers.

CONCLUSION

This chapter discussed organizing quantitative data to prepare them for analysis and then analyzing them (organizing data into charts or tables, or summarizing them with statistical measures). We use statistical analysis to test hypotheses and answer research questions. You saw how data must first be coded and then analyzed using univariate or bivariate statistics. Bivariate relationships might be spurious, so control variables and multivariate analyses are often necessary. You also saw some basics about inferential statistics.

Beginning researchers sometimes believe they have done something wrong if their results do not

EXPANSION BOX 4

Statistical Programs on Computers

Almost every social researcher who needs to calculate many statistics does so with a computer program. One can calculate some statistics using a basic spreadsheet program, such as Excel. Unfortunately, spreadsheets are designed for accounting and bookkeeping functions; they include statistical functions but are clumsy and limited for that purpose. There are many computer programs designed for calculating general statistics. The marketplace can be confusing to a beginner for products rapidly evolve with changing computer technology. One or two decades ago, one had to know a computer language or do simple programming to have a computer calculate statistics.

In recent years, the software has become less demanding for a user. The most popular programs in the social sciences are Minitab, Microcase, and Stastical Package for the Social Sciences (SPSS). Others include Statistical Analysis System (SAS), BMPD (bought by SPSS, Inc.), STATISTICA by StratSoft, and Strata. Many began as simple, low-cost programs for research purposes. Today private corporations own many of these and are interested in selling a sophisticated set of software products to many diverse corporate and government users.

The most widely used program for statistics in the social sciences is SPSS. Its advantages are that social researchers have used it extensively for more than three decades, it includes many ways to manipulate quantitative data, and it contains most statistical measures. Its disadvantage is that it can take a long time to learn because of its many options and complex statistics. Also, it is expensive to purchase except for an inexpensive, “stripped down” student version included with a textbook or workbook.

As computer technology makes using statistics programs easier, the danger increases that some people will use the programs but not understand statistics or what the programs are doing. These people can easily violate basic assumptions required by a statistical procedure, use the statistics improperly, and produce results that are pure nonsense yet look very technically sophisticated.

support a hypothesis. There is nothing wrong with rejecting a hypothesis. The goal of scientific research is to produce knowledge that truly reflects the social world, not to defend pet ideas or hypotheses. Hypotheses are theoretical guesses based on limited knowledge; they need to be tested. Excellent-quality research can find that a hypothesis is wrong, and poor-quality research can support a hypothesis. Good research depends on high-quality methodology, not on supporting a specific hypothesis.

Good research means guarding against possible errors or obstacles to true inferences from data to the social world. Errors can enter into the research process and affect results at many places: research design, measurement, data collection, coding, calculating statistics and constructing tables, or interpreting results. Even if you can design, measure, collect, code, and calculate without error, you must also complete another step in the research process: interpret the tables, charts, and statistics, and answer the question: What does it all mean? The only way to assign meaning to facts, charts, tables, or statistics is to use theory, insight, and understanding.

Data, tables, or computer output alone cannot answer research questions. The facts do not speak for themselves. As a researcher, you must return to your theory (i.e., concepts, relationships among concepts, assumptions, theoretical definitions) and give the results meaning. Do not lock yourself into the ideas with which you began. There is room for creativity, and new ideas are generated by trying to figure out what results really say. It is important to be careful in designing and conducting research so that you can look at the results as a reflection of something in the social world and not worry about whether they are due to an error or an artifact of the research process itself.

Before we leave quantitative research, we must present one last issue. Journalists, politicians, and others increasingly use statistical results to make a point or bolster an argument. This has not produced increased accuracy or clarity in public debate. More often, it has increased confusion; this makes knowing what statistics can and cannot do essential. The cliché that you can prove anything with statistics is false; however, some people can and do misuse statistics to pretend to prove anything. Through ignorance or conscious deceit, some people use statistics to fool others. The best way to protect yourself from being misled by statistics is not to ignore them or hide from the numbers but to understand the research process and statistics, think about what you hear, and ask questions.

We turn next to qualitative research. The logic and purpose of qualitative research differ from those of the quantitative, positivist approach of the past chapters. It is less concerned with numbers, hypotheses, and causality and more concerned with words, norms and values, and meaning.

KEY TERMS

bivariate statistics
codebook
coding procedure
contingency cleaning
contingency table
control variable
covariation
cross-tabulation
curvilinear relationship
data field
data records
descriptive statistics
direct-entry method
elaboration paradigm
explanation pattern
frequency distribution
frequency polygon
histogram
inferential statistics
interpretation pattern
level of statistical significance
linear relationship
marginal
mean
measures of central tendency
median
mode
net effect
normal distribution
partials
percentile
possible code cleaning
proportionate reduction in error
range
replication pattern
scattergram
skewed distribution
specification pattern
standard deviation
statistical independence
statistical relationship
statistical significance
suppressor variable pattern
Type I error
Type II error
univariate statistics
z-score

REVIEW QUESTIONS

1. What is a codebook, and how is it used in research?
2. How do researchers clean data and check their coding?
3. Describe how researchers use optical scan sheets.
4. In what ways can a researcher display frequency distribution information?
5. Describe the differences between mean, median, and mode.
6. What three features of a relationship can be seen from a scattergram?
7. What is a covariation, and how is it used?
8. When can a researcher generalize from a scattergram to a percentaged table to find a relationship among variables?
9. Discuss the concept of control as it is used in trivariate analysis.
10. What does it mean to say “statistically significant at the .001 level,” and what type of error is more likely, Type I or Type II?

NOTES

1. Practical advice on coding and handling quantitative data comes from survey research. See discussions in Babbie (1998:366–372), Backstrom and Hursh-Cesar (1981:309–400), Fowler (1984:127–133), Sonquist and Dunkelberg (1977:210–215), and Warwick and Lininger (1975:234–291).

2. Note that coding gender as 1 = Male, 2 = Female, or as 0 = Male, 1 = Female, or reversing the gender for numbers is arbitrary. The only reason one uses numbers instead of letters (e.g., M and F) is that many computer programs work best with all numbers. Sometimes coding data as a zero can create confusion, so the number 1 is usually the lowest value.

3. For discussions of many different ways to display quantitative data, see Fox (1992), Henry (1995), Tufte (1983, 1991), and Zeisel (1985:14–33).

4. Other statistics measure special types of means for ordinal data and for other special situations, which are beyond the level of discussion in this book.

5. On the elaboration paradigm and its history, see Babbie (1998:400–409) and Rosenberg (1968).

6. Beginning students and people outside the social sciences are sometimes surprised at the low (10 to 50 percent) predictive accuracy in multiple regression results. There are three responses to this. First, a 10 to 50 percent reduction in errors is really not bad compared to purely random guessing. Second, positivist social science is still developing. Although the levels of accuracy may not be as high as those of the physical sciences, they are much higher than for any explanation of the social world possible 10 or 20 years ago. Finally, the theoretically important issue in most multiple regression models is less the accuracy of overall prediction than the effects of specific variables. Most hypotheses involve the effects of specific independent variables on dependent variables.

7. In formal hypothesis testing, we test the null hypothesis and usually want to reject the null because rejection of the null indirectly supports the alternative hypothesis to the null, the one we deduce from theory as a tentative explanation. The null hypothesis was discussed in Chapter 6.