APPENDIX F

QUANTITATIVE DATA ANALYSIS

"Oh no, not data analysis and statistics!" We now hit the chapter that you may have been fearing all along, the chapter on data analysis and the use of statistics. This appendix describes what you need to do after your data have been collected. You now need to analyze what you have found, interpret it, and decide how to present your data so that you can most clearly make the points you wish to make.

What you probably dread about this appendix is something that you either sense or know from a previous course: Studying data analysis and statistics will lead you into that feared world of mathematics. We would like to state at the beginning, however, that you have relatively little to fear. The kind of mathematics required to perform the data analysis tasks in this chapter is minimal. If you can add, subtract, multiply, and divide and are willing to put some effort into carefully reading the chapter, you will do well in the statistical analysis of your data. In fact, it is our position that the analysis of your data will require more in the way of careful and logical thought than in mathematical skill. One helpful way to think of statistics is that it consists of a set of tools that you will use to examine your data to help you answer the questions that motivated your research in the first place. Right now, the toolbox that holds your statistical tools is fairly empty (or completely empty). In the course of this chapter, we will add some fundamental tools to that toolbox. We would also like to note at the beginning that the kinds of statistics you will use on criminological data are very much the same as those used by economists, psychologists, political scientists, sociologists, and other social scientists. In other words, statistical tools are statistical tools, and all that changes is the nature of the problem to which those tools are applied.

This appendix will introduce several common statistics in social research and highlight the factors that must be considered in using and interpreting statistics. Think of it as a review of fundamental social statistics, if you have already studied them, or as an introductory overview, if you have not.

Two preliminary sections lay the foundation for studying statistics. In the first, we will discuss the role of statistics in the research process, returning to themes and techniques you already know. In the second preliminary section, we will outline the process of acquiring data for statistical analysis. In the rest of the chapter, we will explain how to describe the distribution of single variables and the relationships among variables. Along the way, we will address ethical issues related to data analysis. This appendix will be successful if it encourages you to see statistics responsibly and evaluate them critically

Learning Objectives

1. Identify the types of graphs and statistics that are appropriate for analysis of variables at each level of measurement.
2. List the guidelines for constructing frequency distributions.
3. Discuss the advantages and disadvantages of using each of the three measures of central tendency.
4. Understand the difference between the variance and the standard deviation.
5. Define the concept of skewness and explain how it can influence measures of central tendency.
6. Explain how to calculate percentages in a cross-tabulation table and how to interpret the results.
7. Discuss the three reasons for conducting an elaboration analysis.
8. Write a statement based on inferential statistics that reports the confidence that can be placed in a statistical statement of a population parameter.
9. Define the statistics obtained in a multiple regression analysis and explain their purpose.

and gives you the confidence necessary to seek opportunities for extending your statistical knowledge.

It should be noted that, in this appendix, we focus primarily on the use of statistics for descriptive purposes. Those of you looking for a more advanced discussion of statistical methods used in criminal justice and criminology should seek other textbooks (e.g., Bachman and Paternoster, 2017). Although many colleges and universities offer social statistics in a separate course, we don't want you to think of this appendix as something that deals with a different topic than the rest of the book. Data analysis is an integral component of research methods, and it's important that any proposal for quantitative research include a plan for the data analysis that will follow data collection.

INTRODUCING STATISTICS

Statistics play a key role in achieving valid research results in terms of measurement, causal validity, and generalizability. Some statistics are useful primarily to describe the results of measuring single variables and to construct and evaluate multi-item scales. These statistics include **frequency distributions**, graphs, measures of central tendency and variation, and reliability tests. Other statistics are useful primarily in achieving causal validity, by helping us describe the association among variables and control for, or otherwise take into account, other variables.

Frequency distributions: Numerical display showing the number of cases, and usually the percentage of cases (the relative frequencies), corresponding to each value or group of values of a variable.

Cross-tabulation (cross-tab): A bivariate (two-variable) distribution showing the distribution of one variable for each category of another variable.

Descriptive statistics: Statistics used to describe the distribution of and relationship among variables.

Inferential statistics: Mathematical tools for estimating how likely it is that a statistical result based on data from a random sample is representative of the population from which the sample is assumed to have been selected.

Cross-tabulation is one technique for measuring association and controlling other variables and is introduced in this appendix. All these statistics are called **descriptive statistics** because they are used to describe the distribution of and relationship among variables.

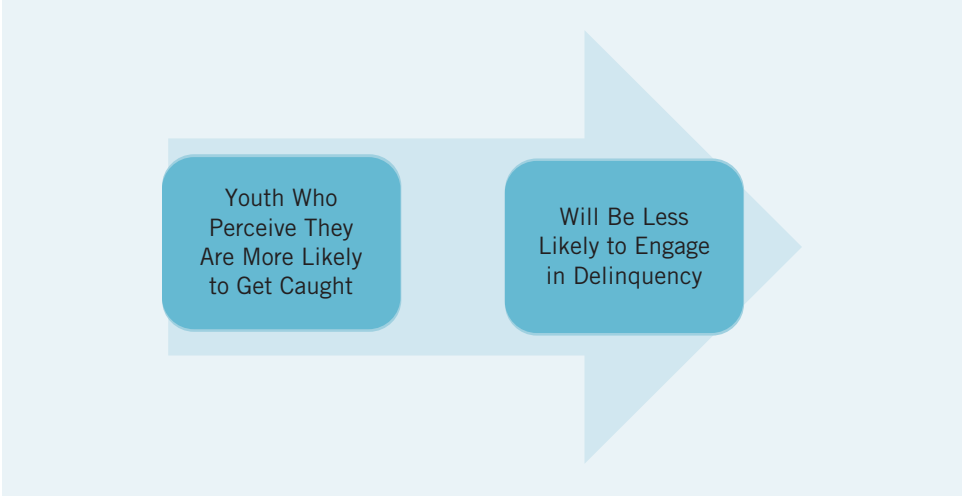
You learned in Chapter 5 that it is possible to estimate the degree of confidence that can be placed in generalizations for a sample and for the population from which the sample was selected. The statistics used in making these estimates are called **inferential statistics**, and they include confidence intervals, to which you were exposed in Chapter 5. In this appendix we will refer only briefly to inferential statistics, but we will emphasize later in the appendix their importance for testing hypotheses involving sample data.

Criminological theory and the results of prior research should guide our statistical plan or analytical strategy, as they guide the choice of other research methods. In other words, we want to use the statistical strategy that will best answer our research question. There are so many particular statistics and so many ways for them to be used in data analysis that even the best statistician can become lost in a sea of numbers if she is not using prior research and theorizing to develop a coherent analysis plan. It is also important for an analyst to choose statistics that are appropriate to the level of measurement of the variables to be analyzed. As you learned in Chapter 4, numbers used to represent the values of variables may not actually signify different quantities, meaning that many statistical techniques will be inapplicable. Some statistics, for example, will be appropriate only when the variable you are examining is measured at the nominal level. Other kinds of statistics will require interval-level measurement. To use the right statistic, then, you must be very familiar with the measurement properties of your variables (and you thought that stuff would go away!).

Case Study: The Causes of Delinquency

In this appendix, we will use research on the causes of delinquency for our examples. More specifically, our data will be a subset of a much larger study of a sample of approximately 1,200 high school students selected from the metropolitan and suburban high schools

Exhibit F.1 Hypothesis for Perceived Fear of Being Caught and Delinquency



of a city in South Carolina. These students, all of whom were in the 10th grade, completed a questionnaire that asked about such things as how they spent their spare time; how they got along with their parents, teachers, and friends; their attitudes about delinquency; whether their friends committed delinquent acts; and their own involvement in delinquency. The original research study was designed to test specific hypotheses about the factors that influence delinquency. It was predicted that delinquent behavior would be affected by such things as the level of supervision provided by parents, the students' own moral beliefs about delinquency, their involvement in conventional activities such as studying and watching TV, their fear of getting caught, their friends' involvement in crime, and whether these friends provided verbal support for delinquent acts. All these hypotheses were derived from extant criminological theory, theories we have referred to throughout this book. One specific hypothesis, derived from deterrence theory, predicts that youths who believe they are likely to get caught by the police for committing delinquent acts are less likely to commit delinquency than others. This hypothesis is shown in Exhibit F.1. The variables from this study that we will use in our appendix examples are displayed in Exhibit F.2.

PREPARING DATA FOR ANALYSIS

If you have conducted your own survey or experiment, your quantitative data must be prepared in a format suitable for computer entry. You learned in Chapter 8 that questionnaires and interview schedules can be precoded to facilitate data entry by representing each response with a unique number. This method allows direct entry of the precoded responses into a computer file, after responses are checked to ensure that only one valid answer code has been circled (extra written answers can be assigned their own numerical codes). Most survey research organizations now use a database management program to control data entry. The program prompts the data entry clerk for each response, checks the response to ensure that it is a valid response for that variable, and then saves the response in the data file. Not all studies have used precoded data entry, however, and individual researchers must enter the data

Exhibit F.2 List of Variables for Class Examples of Causes of Delinquency

Variable	SPSS Variable Name	Description
Gender	V1	Sex of respondent.
Age	V2	Age of respondent.
TV	V21	Number of hours per week the respondent watches TV.
Study	V22	Number of hours per week the respondent spends studying.
Supervision	V63	Do parents know where respondent is when he or she is away from home?
Friends think theft wrong	V77	How wrong do respondent's best friends think it is to commit petty theft?
Friends think drinking wrong	V79	How wrong do respondent's best friends think it is to drink liquor under age?
Punishment for drinking	V109	If respondent was caught drinking liquor under age and taken to court, how much of a problem would it be?
Cost of vandalism	V119	How much would respondent's chances of having good friends be hurt if he or she was arrested for petty theft?
Parental supervision	PARSUPER	Added scale from items that ask respondent if parents know where he or she is and whom he or she is with when away from home. A high score indicates high parental supervision.
Friend's opinion	FROPINON	Added scale that asks respondent if his or her best friends thought that committing various delinquent acts was all right. A high score means more support by friends for committing delinquent acts.
Friend's behavior	FRBEHAVE	Added scale that asks respondent how many of his or her best friends commit delinquent acts.
Certainty of punishment	CERTAIN	Added scale that measures how likely respondent thinks it is that he or she will be caught by police if he or she were to commit delinquent acts. A high score indicates youth perceive a greater probability of being caught.
Morality	MORAL	Added scale that measures how morally wrong respondent thinks it is to commit diverse delinquent acts. A high score means respondent has strong moral inhibitions.
Delinquency	DELINQ1	An additive scale that counts the number of times respondent admits to committing a number of different delinquent acts in the past year. The higher the score, the more delinquent acts she or he committed.

themselves. This is an arduous and time-consuming task, but not for us if we use secondary data. After all, we get the data only after they have been coded and computerized.

Data cleaning: The process of checking data for errors after the data have been entered in a computer file.

Of course, numbers stored in a computer file are not yet numbers that can be analyzed with statistics. After the data are entered, they must be checked carefully for errors, a process called **data cleaning**. If a data entry program has been used and programmed to flag invalid values, the cleaning process is much easier. If data are read in from a text file, a computer program must be written that defines which variables are coded in which columns, attaches meaningful labels to the codes, and distinguishes values representing missing data. The

procedures for doing so vary with each specific statistical package. We used the Windows version of the Statistical Package for the Social Sciences (SPSS) for the analysis in this chapter; you will find examples of SPSS commands required to define and analyze data on the Student Study Site for this text, **edge.sagepub.com/bachmanfrccj5e**.

DISPLAYING UNIVARIATE DISTRIBUTIONS

The first step in data analysis is usually to display the variation in each variable of interest in what are called *univariate frequency distributions*. For many descriptive purposes, the analysis may go no further. Frequency distributions and graphs of frequency distributions are the two most popular approaches for displaying variation; both allow the analyst to display the distribution of cases across the value categories of a variable. Graphs have the advantage over numerically displayed frequency distributions because they provide a picture that is easier to comprehend. Frequency distributions are preferable when exact numbers of cases with particular values must be reported, and when many distributions must be displayed in a compact form.

No matter which type of display is used, the primary concern of the data analyst is to accurately display the distribution's shape—that is, to show how cases are distributed across the values of the variable. Three features of the shape of a distribution are important: **central tendency**, **variability**, and **skewness** (lack of symmetry). All three of these features can be represented in a graph or in a frequency distribution.

Central tendency: A feature of a variable's distribution, referring to the value or values around which cases tend to center.

Variability: A feature of a variable's distribution; refers to the extent to which cases are spread out through the distribution or clustered in just one location.

Skewness: A feature of a variable's distribution, referring to the extent to which cases are clustered more at one or the other end of the distribution rather than around the middle.

These features of a distribution's shape can be interpreted in several different ways, and they are not all appropriate for describing every variable. In fact, all three features of a distribution can be distorted if graphs, frequency distributions, or summary statistics are used inappropriately.

A variable's level of measurement is the most important determinant of the appropriateness of particular statistics. For example, we cannot talk about the skewness (lack of symmetry) of a qualitative variable (measured at the nominal level). If the values of a variable cannot be ordered from lowest to highest, if the ordering of the values is arbitrary, we cannot say whether the distribution is symmetric, because we could just reorder the values to make the distribution more (or less) symmetric. Some measures of central tendency and variability are also inappropriate for qualitative variables.

The distinction between variables measured at the ordinal level and those measured at the interval or ratio level should also be considered when selecting statistics to use, but social researchers differ on just how much importance they attach to this distinction. Many social researchers think of ordinal variables as imperfectly measured interval-level variables and believe that in most circumstances statistics developed for interval-level variables also provide useful summaries for ordinal variables. Other social researchers believe that variation in ordinal variables will often be distorted by statistics that assume an interval level of measurement. We will touch on some of the details of these issues in the following sections on particular statistical techniques.

We will now examine graphs and frequency distributions that illustrate these three features of shape. Summary statistics used to measure specific aspects of central tendency and variability will be presented in a separate section. There is a summary statistic for the measurement of skewness, but it is used only rarely in published research reports and will not be presented here.

Graphs

It is true that a picture often is worth a thousand words. Graphs can be easy to read, and they very nicely highlight a distribution's shape. They are particularly useful for exploring

data, because they show the full range of variation and identify data anomalies that might be in need of further study. And good, professional-looking graphs can now be produced relatively easily with software available for personal computers. There are many types of graphs, but the most common and most useful are bar charts and histograms. Each has two axes, the vertical axis (*y*-axis) and the horizontal axis (*x*-axis), and labels to identify the variables and the values with tick marks showing where each indicated value falls along the axis. The vertical *y*-axis of a graph is usually in frequency or percentage units, whereas the horizontal *x*-axis displays the values of the variable being graphed. There are different kinds of graphs you can use to descriptively display your data, depending upon the level of measurement of the variable.

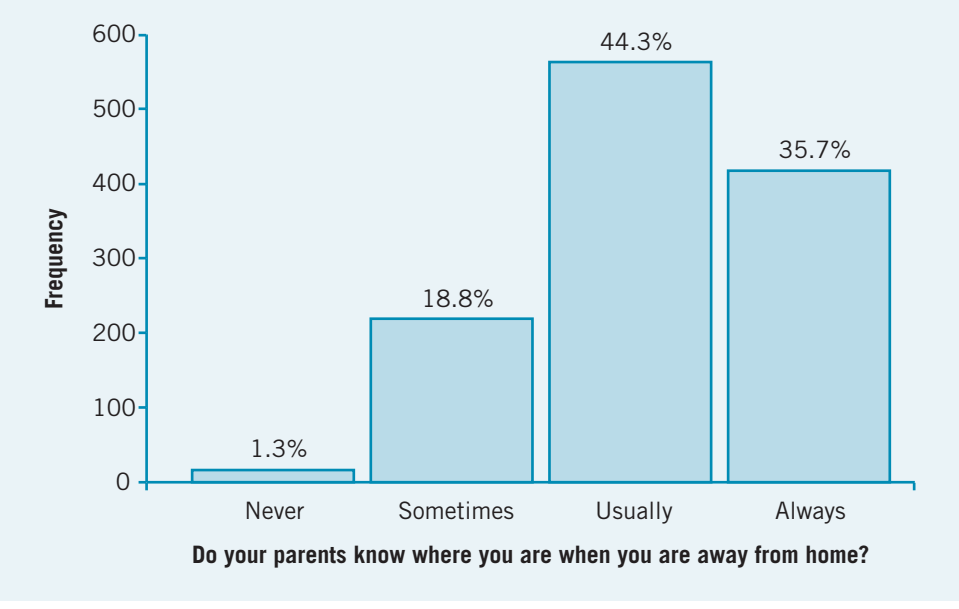
Bar chart: A graphic for qualitative variables in which the variable's distribution is displayed with solid bars separated by spaces.

Percentage: Relative frequencies, computed by dividing the frequency of cases in a particular category by the total number of cases, and multiplying by 100.

Mode: The most frequent value in a distribution, also termed the probability average.

A **bar chart** contains solid bars separated by spaces. It is a good tool for displaying the distribution of variables measured at the nominal level and other discrete categorical variables, because there is, in effect, a gap between each of the categories. In our study of delinquency, one of the questions asked of respondents was whether their parents knew where the respondents were when the respondents were away from home. We graphed the responses to this question in a bar chart, which is shown in Exhibit F.3. In this bar chart we report both the frequency count for each value and the **percentage** of the total that each value represents. The chart indicates that very few of the respondents (only 16, or 1.3%) reported that their parents "never" knew where the respondents were when the respondents were not at home. Almost one half (562, or 44.3%) of the youths reported that their parents "usually" knew where the respondents were. What you can also see, by noticing the height of the bars above "usually" and "always," is that most youths report that their parents provide very adequate supervision. You can also see that the most frequent response was "usually" and the least frequent was "never." Because the response "usually" is the most frequent value, it is called the **mode** or modal response. With ordinal data like these, the mode is the most appropriate measure of central tendency (more about this later).

Exhibit F.3 Bar Chart Showing Youths' Reponses on Parents Knowing Where They Are



Notice that the cases tend to cluster in the two values of "usually" and "always"; in fact, about 80% of all cases are found in those two categories. There is not much variability in this distribution, then.

A **histogram** is like a bar chart, but it has bars that are adjacent, or right next to each other, with no gaps. This is done to indicate that data displayed in a histogram, unlike the data in a bar chart, are quantitative variables that vary along a continuum (see the discussion of levels of measurement for variables in Chapter 4). Exhibit F.4 shows a histogram from the delinquency dataset we are using. The variable being graphed is the number of hours per week the respondent reported to be studying. Notice that the cases cluster at the low end of the values. In other words, there are a lot of youths who spend between 0 and 15 hours per week studying. After that, there are only a few cases at each different value, with "spikes" occurring at 25, 30, 38, and 40 hours studied. This distribution is clearly not symmetric. In a symmetric distribution there is a lump of cases or a spike with an equal number of cases to the left and right of that spike. In the distribution shown in Exhibit F.4, most of the cases are at the left end of the distribution (i.e., at low values), and the distribution trails off on the right side. The ends of a histogram like this are often called the tail of a distribution. In a symmetric distribution, the left and right tails are approximately the same length. As you can clearly see in Exhibit F.4, however, the right tail is much longer than the left tail. When the tails of the distribution are uneven, the distribution is said to be asymmetrical or skewed. A skew is either positive or negative. When the cases cluster to the left and the right tail of the distribution is longer than the left, as in Exhibit F.4, our variable distribution is **positively skewed**. When the cases cluster to the right side and the left tail of the distribution is long, our variable distribution is **negatively skewed**.

If graphs are misused, they can distort, rather than display, the shape of a distribution. Compare, for example, the two graphs in Exhibit F.5. The first graph shows that high school seniors reported relatively stable rates of lifetime use of cocaine between 1980 and 1985. The

Histogram: A graphic for quantitative variables in which the variable's distribution is displayed with adjacent bars.

Positively skewed: Describes a distribution in which the cases cluster to the left and the right tail of the distribution is longer than the left.

Negatively skewed: A distribution in which cases cluster to the right side, and the left tail of the distribution is longer than the right.

Exhibit F.4 Histogram

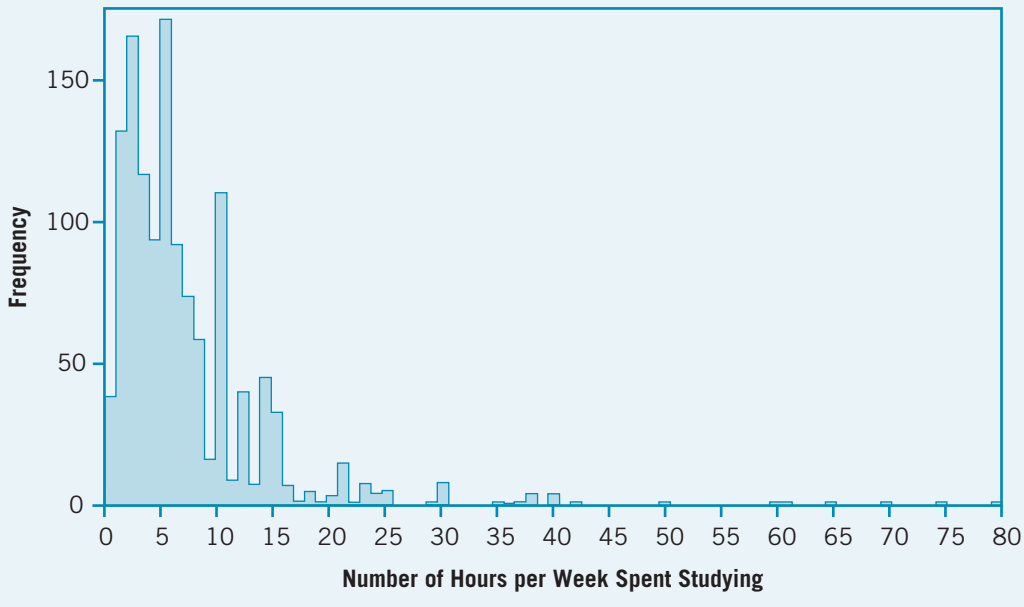
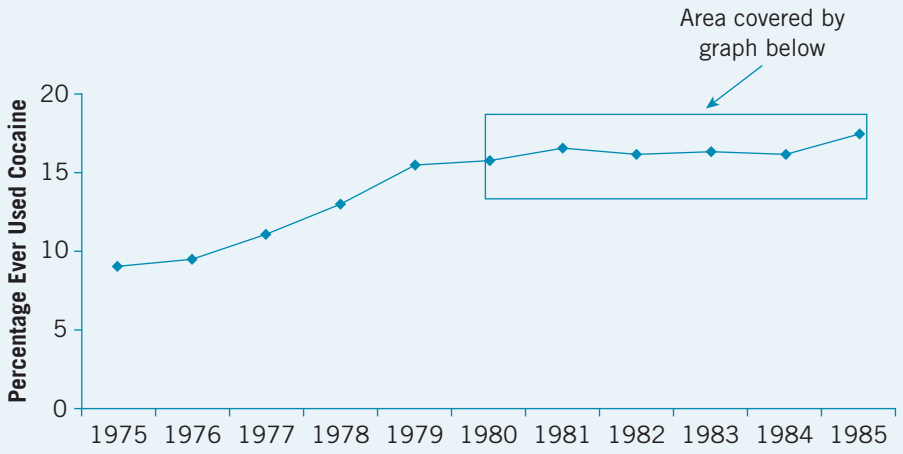
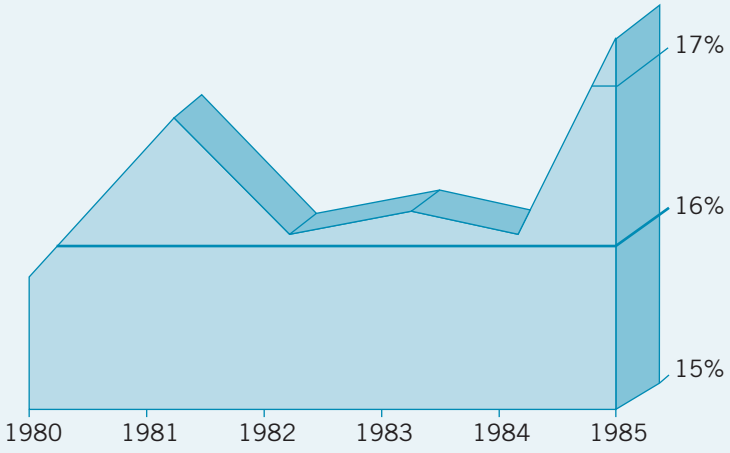


Exhibit F.5 Two Graphs of Cocaine Usage



A. University of Michigan Institute for Social Research, Time Series for Lifetime Prevalence of Cocaine Use



B. Final Stages of Construction

Source: James D. Orcutt and J. Blake Turner. "Shocking Numbers and Graphic Accounts." *Social Problems*, 40(2): 190–206. Copyright © 1993, The Society for the Study of Social Problems. Reprinted with permission from Oxford University Press.

second graph, using exactly the same numbers, appeared in a 1986 *Newsweek* article on the coke plague (Orcutt and Turner, 1993). To look at this graph, you would think that the rate of cocaine usage among high school seniors increased dramatically during this period. But, in fact, the difference between the two graphs is due simply to changes in how the graphs are drawn. In the "plague" graph (B), the percentage scale on the vertical axis begins at 15 rather than 0, making what was about a one-percentage-point increase look very big indeed. In addition, omission from the plague graph of the more rapid increase in reported usage between 1975 and 1980 makes it look as if the tiny increase in 1985 were a new, and thus more newsworthy, crisis.

Adherence to several guidelines (Tufte, 1983) will help you spot these problems and avoid them in your own work:

- The difference between bars will be exaggerated if you cut off the bottom of the vertical axis and display less than the full height of the bars. Instead, begin the graph of a quantitative variable at 0 on both axes. It may at times be reasonable to violate this guideline, as when an age distribution is presented for a sample of adults, but in this case be sure to mark the break clearly on the axis.
- Bars of unequal width, including pictures instead of bars, can make particular values look as if they carry more weight than their frequency warrants. Always use bars of equal width.
- Either shortening or lengthening the vertical axis will obscure or accentuate the differences in the number of cases between values. The two axes usually should be of approximately equal length.
- Avoid chart junk that can confuse the reader and obscure the distribution's shape (a lot of verbiage, numerous marks, lines, lots of cross-hatching, etc.).

Frequency Distributions

A frequency distribution displays the number, the percentage (the relative frequencies), or both for cases corresponding to each of a variable's values or a group of values. The components of the frequency distribution should be clearly labeled, with a title, a stub (labels for the values of the variable), a caption (identifying whether the distribution includes frequencies, percentages, or both), and perhaps the number of missing cases. If percentages are presented rather than frequencies (sometimes both are included), the total number of cases in the distribution (the **Base *N***) should be indicated (see Exhibit F.6). Remember that a percentage is simply a relative frequency. A percentage shows the frequency of a given value relative to the total number of cases times 100.

Base *N*: The total number of cases in a distribution.

Ungrouped Data

Constructing and reading frequency distributions for variables with few values is not difficult. In Exhibit F.6, we created the frequency distribution from the variable "Punishment for Drinking" found in the delinquency dataset (see Exhibit F.2). For this variable, the study

Exhibit F.6 Frequency Distribution

How much of a problem would it be if you went to court for drinking liquor under age?

Value	Frequency (*f*)	Percentage (%)
No problem at all	14	1.1
Hardly any problem	53	4.2
A little problem	196	15.4
A big problem	421	33.1
A very big problem	588	46.2
Total	1,272	100.0

asked the youths to respond to the following question: "How much of a problem would it be if you went to court for drinking liquor under age?" The frequency distribution in Exhibit F.6 shows the frequency for each value and its corresponding percentage.

As another example of calculating the frequencies and percentages, suppose we had a sample of 25 youths and asked them their gender. From this group of 25 youths, 13 were male and 12 were female. The frequency of males (symbolized here by *f*) would be 13 and the frequency of females would be 12. The percentage of males would be 52%, calculated by *f*/ the total number of cases × 100 (13/25 × 100 = 52%). The percentage of females would be 12/25 × 100 = 48%.

In the frequency distribution shown in Exhibit F.6, you can see that only a very small number (14 out of 1,272) of youths thought that they would experience "no problem" if they were caught and taken to court for drinking liquor under age. You can see that most—in fact, 1,009—of these youths, or 79.3% of them, thought that they would have either "a big problem" or "a very big problem" with this. If you compare Exhibit F.6 to Exhibit F.3, you can see that a frequency distribution (see Exhibit F.6) can provide much of the same information as a graph about the number and percentage of cases in a variable's categories. Often, however, it is easier to see the shape of a distribution when it is graphed. When the goal of a presentation is to convey a general sense of a variable's distribution, particularly when the presentation is to an audience not trained in statistics, the advantages of a graph outweigh those of a frequency distribution.

Exhibit F.6 is a frequency distribution of an ordinal-level variable; it has a very small number of discrete categories. In Exhibit F.7, we provide an illustration of a frequency distribution with a continuous quantitative variable. This variable is one we have already looked at and graphed from the delinquency data, the number of hours per week the respondent spent studying. Notice that this variable, like many continuous variables in criminological research, has a large number of values. Although this is a reasonable frequency distribution to construct—you can, for example, still see that the cases tend to cluster in the low end of the distribution and are strung way out at the upper end—it is a little difficult to get a good sense of the distribution of the cases. The problem is that there are too many values to easily comprehend. It would be nice if we could simplify distributions like these that have a large number of different values. Well, we can. We can construct what is called a **grouped frequency distribution**.

Grouped frequency distribution: A frequency distribution in which the data are organized into categories, either because there are more values than can be easily displayed or because the distribution of the variable will be clearer or more meaningful.

Grouped Data

Many frequency distributions, such as those in Exhibit F.7, and many graphs require grouping of some values after the data are collected. There are two reasons for grouping:

1. There are more than 15–20 values to begin with, a number too large to be displayed in an easily readable table.
2. The distribution of the variable will be clearer or more meaningful if some of the values are combined.

Inspection of Exhibit F.7 should clarify these reasons. In this distribution it is very difficult to discern any shape, much less the central tendency. What we would like to now do to make the features of the data more visible is change the values into intervals of values, or a range of values. For example, rather than having five separate values of 0, 1, 2, 3, and 4 hours studied per week, we can have a range of values or an interval for the first value, such as 0–4 hours studied. Then we can get a count or frequency of the number of cases (and percentage of the total) that fall within that interval.

Exhibit F.7 Frequency Distribution With Continuous Quantitative Data: Hours Studied per Week

Value	Frequency (*f*)	Percentage (%)
0	38	3.0
1	132	10.4
2	165	13.0
3	116	9.1
4	94	7.4
5	171	13.4
6	92	7.2
7	73	5.7
8	58	4.6
9	16	1.3
10	110	8.6
11	9	0.7
12	40	3.1
13	7	0.6
14	45	3.5
15	32	2.5
16	7	0.6
17	5	0.4
18	4	0.3
19	1	0.1
20	15	1.2
21	8	0.6
22	1	0.1
23	1	0.1
24	4	0.3
25	5	0.4
29	1	0.1
30	8	0.6
35	1	0.1
37	1	0.1

(Continued)

(Continued)

Value	Frequency (*f*)	Percentage (%)
40	4	0.3
42	1	0.1
50	1	0.1
60	1	0.1
61	1	0.1
65	1	0.1
70	1	0.1
75	1	0.1
80	1	0.1
Total	1,272	100.0

Once we decide to group values, or categories, we have to be sure that in doing so we do not distort the distribution. Adhering to the following guidelines for combining values in a frequency distribution will prevent many problems:

- Categories should be logically defensible and preserve the distribution's shape.
- Categories should be mutually exclusive and exhaustive, so every case is classifiable in one and only one category.
- The first interval must contain the lowest value, and the last interval must contain the highest value in the distribution.
- Each interval width, the number of values that fall within each interval, should be the same size.
- There should be between 7 and 13 intervals. This is a tough rule to follow. The key is not to have so few intervals that your data are clumped or clustered into only a few intervals (you will lose too much information about your distribution) and not to have so many intervals that the data are not much clearer than an ungrouped frequency distribution.

Let us use the data in Exhibit F.7 on the number of hours studied by these youths to create a grouped frequency distribution. We will follow a number of explicit steps:

Step 1. Determine the number of intervals you think you want. This decision is arbitrary, but try to keep the number of intervals you have in the 7–13 range. For our example, let us say we initially decided we wanted to have 10 intervals. (Note, if you do your frequency distribution and it looks too clustered or there are too many intervals, redo your distribution with a different number of intervals.) Don't worry; there are no hard and fast rules for the correct number of intervals, and constructing a grouped frequency distribution is as much art as science. Just remember that the frequency distribution you make is supposed to convey information about the shape and central tendency of your data.

Step 2. Decide on the width of the interval (symbolized by w_i). The interval width is the number of different values that fall into your interval. For example, an interval width of 5 has five different values that fall into it, say, the values 0, 1, 2, 3, and 4 hours studied. There is a simple formula to approximate what your interval width should be given the number of intervals you decided on in the first step: Determine the range of the data, where the range is simply the highest score in the distribution minus the lowest score. In our data, with the number of hours studied, the range is 80 because the high score is 80 and the low score is 0, so range = 80 – 0 = 80. Then determine the width of the interval by dividing the range by the number of intervals you want from Step 1. We wanted 10 intervals, so our interval width would be $w_i = 80/10 = 8$. We should therefore have an interval width of 8. If you use this simple formula for determining your interval width and you end up with a decimal, say 8.2 or 8.6, then simply round up or down to an integer.

Step 3. Make your first interval so that the lowest value falls into it. Our lowest value is 0 (for studied 0 hours per week), so our first interval begins with the value 0. Now, if the beginning of our first interval is 0 and we want an interval width of 8, is the last value of our interval 7 (with a first interval of 0–7 hours), or is the last value of our interval 8 (with a first interval of 0–8 hours)? One easy way to make a grouped frequency distribution is to do the following: Take the beginning value of your first interval (in our case, it is 0), and add the interval width to that value (8). This new value is the first value of your next interval. What we know, then, is that the first value of our first interval is 0, and the first value of our second interval is 8 (0–?, 8–?). This must mean that the last value to be included in our first interval is one less than 8, or 7. Our first interval, therefore, includes the range of values 0–7. If you count the number of different values in this interval, you will find that it includes eight different values (0, 1, 2, 3, 4, 5, 6, 7). This is our interval width of 8.

Step 4. After your first interval is determined, the next intervals are easy. They must be the same width and not overlap (mutually exclusive). You must make enough intervals to include the last value in your variable distribution. The highest value in our data is 80 hours per week, so we construct the grouped frequency distribution as follows:

0–7
8–15
16–23
24–31
32–39
40–47
48–55
56–63
64–71
72–79
80–87

Notice that in order to include the highest value in our data (80 hours) we had to make 11 intervals instead of the 10 we originally decided upon in Step 1. No problem. Remember, the number of intervals is arbitrary and this is as much art as science.

Step 5. Count the number or frequency of cases that appear in each interval and their percentage of the total. The completed grouped frequency distribution is shown in Exhibit F.8. Notice that this grouped frequency distribution conveys the important features of the distribution of these data. Most of the data cluster at the low end of the number of hours studied. In fact, more than two thirds of these youths studied less than 8 hours per week. Notice also that the frequency of cases thins out at each successive interval. In other words, there is a long right tail to this distribution, indicating a positive skew because fewer youths studied a high number of hours. Notice also that the distribution was created in such a way that the interval widths are all the same, and each case falls into one and only one interval (i.e., the intervals are exhaustive and mutually exclusive). We would have run into trouble if we had two intervals like 0–7 and 7–14, because we would not know where to place those youths who spent 7 hours a week studying. Should we put them in the first or second interval? If the intervals are mutually exclusive, as they are here, you will not run into these problems.

Exhibit F.8 Example of a Grouped Frequency Distribution From Hours Studied

Value	Frequency (f)	Percentage (%)
0–7	881	69.26
8–15	317	24.92
16–23	42	3.30
24–31	18	1.42
32–39	2	0.16
40–47	5	0.39
48–55	1	0.08
56–63	2	0.16
64–71	2	0.16
72–79	1	0.08
80–87	1	0.08
Total	1,272	100.00

Note: Total may not equal 100.0% due to rounding error.

SUMMARIZING UNIVARIATE DISTRIBUTIONS

Summary statistics, sometimes called descriptive statistics, focus attention on particular aspects of a distribution and facilitate comparison among distributions. For example, suppose you wanted to report the rate of violent crimes for each city in the United States with over 100,000 in population. You could report each city's violent crime rate, but it is unlikely that two cities would have the same rate, and you would have to report approximately 200 rates, one for each city. This would be a frequency distribution that many, if not most, people would find difficult to comprehend. One way to interpret your data for your audience would be to provide a summary measure that indicates what the average violent crime rate is in large U.S.

cities. That is the purpose of the set of summary statistics called measures of *central tendency*. You would also want to provide another summary measure that shows the variability or heterogeneity in your data—in other words, a measure that shows how different the scores are from each other or from the central tendency. That is the purpose of the set of summary statistics called measures of *variation* or *dispersion*. We will discuss each type of measurement in turn.

Measures of Central Tendency

Central tendency is usually summarized with one of three statistics: the mode, the median, or the mean. For any particular application, one of these statistics may be preferable, but each has a role to play in data analysis. To choose an appropriate measure of central tendency, the analyst must consider a variable's level of measurement, the skewness of a quantitative variable's distribution, and the purpose for which the statistic is used. In addition, the analyst's personal experiences and preferences inevitably will play a role.

Mode

The mode is the most frequent value in a distribution. For example, refer to the data in Exhibit F.8, which shows the grouped frequency distribution for the number of hours studied. The value with the greatest frequency in those data is the interval 0–7 hours; this is the mode of that distribution. Notice that the mode is the most frequently occurring value; it is not the frequency of that value. In other words, the mode in Exhibit F.8 is 0–7 hours; the mode is not 881, which is the frequency of the modal category. To show how the mode can also be thought of as the value with the highest probability, refer to Exhibit F.9. Suppose you had this grouped frequency distribution but knew nothing else about each of the 1,272 youths in the study. If you were to pick a case at random from the distribution of 1,272 youths and were asked how many hours the youth studied per week, what would your best guess be? Well, since 881 of the 1,272 youths fall into the first interval of 0–7 hours studied, the probability that a randomly selected youth studied from 0 to 7 hours would be .696 (881/1,272). This is higher than the probability of any other interval. It is the interval with the highest probability because it is the interval with the greatest frequency or mode of the distribution. When a variable distribution has one case or interval that occurs more often than the others, it is called a **unimodal distribution**. The ordinal variable of "parents knowing kids' whereabouts" in Exhibit F.3 is also unimodal. The category with the highest percentage is "usually."

Unimodal distribution: A distribution of a variable in which there is only one value that is the most frequent.

Bimodal distribution: A distribution that has two nonadjacent categories with about the same number of cases, and these categories have more cases than any other categories.

Sometimes a distribution has more than one mode because there are two values that have the highest frequency. This distribution would be called **bimodal**. Some distributions are trimodal in that there are three distinctively high frequency values. When there is no frequency much higher than another, it is even possible to have a distribution without a mode. In saying that there is no mode, though, you are communicating something very important about the data: that no case is more common than the others. Another potential problem with the mode is that it might happen to fall far from the main clustering of cases in a distribution. It would be misleading in this case, then, to say simply that the variable's central tendency was the same as the modal value.

Nevertheless, there are occasions when the mode is very appropriate. Most important, the mode is the only measure of central tendency that can be used to characterize the central tendency of variables measured at the nominal level. In Exhibit F.9 we have the frequency distribution of the conviction offense for 1,000 offenders convicted in a criminal court. The central tendency of the distribution is property offense, because more of the 1,000 offenders were convicted of a property crime than any other crime. For the variable "type of offense convicted of," the most common value is property crime. The mode also is often referred to in descriptions of the shape of a distribution. The terms *unimodal* and *bimodal* appear frequently,

Exhibit F.9 Frequency Distribution of Offense for 1,000 Convicted Offenders

Type of Offense	Frequency (f)
Violent	125
Drug	210
Property	480
Public order	100
Other	85
Total	1,000

as do descriptive statements such as "The typical (most probable) respondent was in her 30s." Of course, when the issue is determining the most probable value, the mode is the appropriate statistic.

Median

The **median** is the score in the middle of a rank-ordered distribution. It is, then, the score or point that divides the distribution in half (the 50th percentile). The median is inappropriate for variables measured at the nominal level because their values cannot be put in ranked order (remember, there is no "order" to nominal-level data), and so there is no meaningful middle position. To determine the median, we simply need to do the following. First, rank-order the values from lowest to highest. Because the median is a positional measure, we then have to find the position of the median in the rank order of scores by using the following simple formula:

Median: The position average, or the point that divides a distribution in half (the 50th percentile).

$$\frac{N+1}{2}$$

where N is equal to the total number of cases.

Exhibit F.10 Hypothetical Rate of Violent Crime for Selected U.S. Cities

	City	Number of Violent Crimes per 100,000
1.	Atlanta	3,571
2.	Boston	1,916
3.	Cleveland	1,530
4.	Dallas	1,589
5.	Los Angeles	2,059
6.	New Orleans	1,887
7.	New York	1,861
8.	Philadelphia	1,322
9.	San Francisco	1,461
1.	Atlanta	3,571
2.	Boston	1,916
3.	Cleveland	1,530
4.	Dallas	1,589
5.	Los Angeles	2,059
6.	New Orleans	1,887
7.	New York	1,861
8.	Philadelphia	1,322

In Exhibit F.10, we first list a sample of 17 U.S. cities and their rate of violent crime. We are going to calculate the median from two samples taken from this list, one sample of nine cities and another sample of eight cities.

The first sample of nine cities is shown in Exhibit F.10a.

In this sample of nine cities, we first must find the median position, which is determined by (9 + 1)/2 = 10/2 = 5. The median violent crime rate, then, is in the fifth position in this rank order. Starting either at the top of the scores and counting down to the fifth position or at the bottom and counting up, we find that in the fifth position is the score 1,861 violent crimes per 100,000, which is the median violent crime rate for these nine U.S. cities. Now, let us find the median in the second list, which has only eight cities that are rank ordered in Exhibit F.10b.

Now our median position is: (8 + 1)/2 = 9/2 = 4.5. Because we now have to find the value of the median between the fourth and fifth positions, we have to find the average of the values that fall in these two positions. The score at the fourth position is 1,861, and the score at the fifth is 1,887. The value of the median can now be found by adding these two scores and dividing by 2. The median rate of violent crime for this sample of eight cities, then, is equal to (1,861 + 1,887)/2 = 1,874 violent crimes per 100,000 population.

This tells us that 50% of the cities have violent crime rates lower than 1,874 and 50% of the cities have violent crime rates higher than 1,874.

Because the median is the score at the 50th percentile, we can also identify it in a frequency distribution by finding the value corresponding to a cumulative percentage of 50. We

Exhibit F.10a Sample of Nine Cities From Exhibit F.10

Rank	Crime Rate
1	1,322
2	1,461
3	1,530
4	1,589
5	1,861
6	1,887
7	1,916
8	2,059
9	3,571

Exhibit F.10b Sample of Eight Cities From Exhibit F.10

Rank	Crime Rate
1	1,322
2	1,530
3	1,589
4	1,861
5	1,887
6	1,916
7	2,059
8	3,571

show you how to do this in Exhibit F.11. These data are a repeat of the data in Exhibit F.7, and show the number of hours studied for the youths in the delinquency dataset.

To find the 50th percentile, we simply added a new column to these data, labeled "cumulative percentage." Cumulative percentages are found by taking the percentage of the interval percentage plus all others below it. So the first value (3.0%) would be entered as the first cumulative percentage, because there are no other intervals below the first. This cumulative percentage simply means that 3% of the youths studied for 0 hours per week. Then we add the percentage in the next value (10.4%) to this to arrive at a cumulative percentage of 13.4%. This means that 13.4% of the youths studied for 1 hour per week or less. This becomes the second entry in the cumulative percentage column. We continue adding each adjacent percentage value until we reach 50%. There is a cumulative percentage of 56.3% at the value of 5 hours per week. The median number of hours studied per week, then, is 5 hours. Of the respondents, 50% studied less than 5 hours per week, and 50% studied more than 5 hours per week.

Mean

The **mean** is simply the arithmetic average of all scores in a distribution. It is computed by adding up the value of all the cases and dividing by the total number of cases, thereby taking into account the value of each case in the distribution:

Mean: The arithmetic, or weighted, average, computed by adding up the value of all the cases and dividing by the total number of cases.

$$\text{Mean} = \text{Sum of value of all cases / number of cases}$$

The symbol for the mean is $\overline{X}$ (pronounced "X-bar"). In algebraic notation, the equation is

$$\overline{X} = \frac{\sum_1^N x_i}{N}$$

where x_i is a symbol for each ith score and i's go from 1 to N; N is the total number of cases. What the algebraic equation says to do is to sum all scores, starting at the first score and continuing until the last, or Nth, score; then divide this sum by the total number of cases (N).

Exhibit F.11 Frequency Distribution With Continuous Quantitative Data: Hours Studied per Week

Value	Frequency (*f*)	Percentage (%)	Cumulative Percentage
0	38	3.0	3.0
1	132	10.4	13.4
2	165	13.0	26.4
3	116	9.1	35.5
4	94	7.4	42.9
5	141	13.4	56.3 (includes 50th percentile)
6	92	7.2	
7	73	5.7	
8	58	4.6	
9	16	1.3	
10	110	8.6	
11	9	0.7	
12	40	3.1	
13	7	0.6	
14	45	3.5	
15	32	2.5	
16	7	0.6	
17	5	0.4	
18	4	0.3	
19	1	0.1	
20	15	1.2	
21	8	0.6	
22	1	0.1	
23	1	0.1	
24	4	0.3	
25	5	0.4	
29	1	0.1	
30	8	0.6	
35	1	0.1	
37	1	0.1	

Value	Frequency (*f*)	Percentage (%)	Cumulative Percentage
40	4	0.3	
42	1	0.1	
50	1	0.1	
60	1	0.1	
61	1	0.1	
65	1	0.1	
70	1	0.1	
75	1	0.1	
80	1	0.1	
Total	1,272	100.0	

We will calculate the mean rate of violent crime for the nine U.S. cities listed in Exhibit F.10a:

$$\overline{X} = \frac{(1{,}322 + 1{,}461 + 1{,}530 + 1{,}589 + 1{,}861 + 1{,}887 + 1{,}916 + 2{,}059 + 3{,}571)}{9} = 1{,}910.7$$

The mean rate of violent crime for these nine U.S. cities, then, is 1,910.7 violent crimes per 100,000 population. When calculating the mean, we do not have to first rank-order the scores. The mean takes every score into account, so it does not matter if we add 3,571 first, in the middle, or last.

Computing the mean requires adding up the values of the cases, so it makes sense to compute a mean only if the values of the cases can be treated as actual quantities—that is, if they reflect an interval or ratio level of measurement, or if they are ordinal and we assume that ordinal measures can be treated as intervals. It would make no sense, however, to calculate the mean for the variable racial or ethnic status. Imagine a group of four people in which there were two Caucasians, one African American, and one Hispanic. To calculate the mean you would need to solve the equation (Caucasian + Caucasian + African American + Hispanic)/4 = ? Even if you decide that Caucasian = 1, African American = 2, and Hispanic = 3 for data entry purposes, it still does not make sense to add these numbers, because they do not represent real numerical quantities. In other words, just because you code Caucasian as "1" and African American as "2," that does not mean that African Americans possess twice the race or ethnicity that Caucasians possess. To see how numerically silly this is, note that we could just as easily have coded African Americans as "1" and Caucasians as "2." Now, with one arbitrary flip of our coding scheme, Caucasians have twice as much race or ethnicity as African Americans. Thus, both the median and the mean are *not* appropriate measures of central tendency for variables measured at the nominal level.

Median or Mean?

Both the median and the mean are used to summarize the central tendency of quantitative variables, but their suitability for a particular application must be carefully assessed.

The key issues to be considered in this assessment are the variable's level of measurement, the shape of its distribution, and the purpose of the statistical summary. Consideration of these issues will sometimes result in a decision to use both the median and the mean and

will sometimes result in neither measure being seen as preferable. But in many other situations, the choice between the mean and median will be clear-cut as soon as the researcher takes the time to consider these three issues.

Level of measurement is a key concern, because to calculate the mean, we must add up the values of all the cases, a procedure that assumes the variable is measured at the interval or ratio level. So even though we know that coding Agree as 2 and Disagree as 3 does not really mean that Disagree is one unit more of disagreement than Agree, the mean assumes this evaluation to be true. Calculation of the median requires only that we order the values of cases, so we do not have to make this assumption. Technically speaking, then, the mean is simply an inappropriate statistic for variables measured at the ordinal level (and you already know that it is completely meaningless for nominal variables). In practice, however, many social researchers use the mean to describe the central tendency of variables measured at the ordinal level, for the reasons outlined earlier.

The shape of a variable's distribution should also be taken into account when deciding whether to use the median or the mean. When a distribution is perfectly symmetric (i.e., when the distribution is bell shaped), the distribution of values below the median is a mirror image of the distribution of values above the median, and the mean and median will be the same. But the values of the mean and median are affected differently by skewness, or the presence of cases with extreme values on one side of the distribution but not the other side. The median takes into account only the number of cases above and below the median point, not the value of these cases, so it is not affected in any way by extreme values. The mean is based on adding the value of all the cases, so it will be pulled in the direction of exceptionally high (or low) values. When the value of the mean is larger than the median, we know that the distribution is skewed in a positive direction, with proportionately more cases with lower than higher values. When the mean is smaller than the median, the distribution is skewed in a negative direction.

The differential impact of skewness and/or outliers on the median and the mean can be illustrated with a simple thought exercise. Let's assume your class has 20 people and we ask you each to tell us your family of origin's family income for the past year. We determine that the mean income for the families for your class members is $72,000. We also find that the median income is $54,000, which tells us that 50% of the families make less than $54,000 and 50% of families make more. Now imagine one of Bill Gate's kids enrolls in the class. Bill Gates is estimated to make over $3.5 billion annually. Wow. That makes the mean income for the class $166,735,238. Clearly, this figure does not represent the "typical" family income any longer. Notice that despite Bill Gates's child entering the class, the median family income would still remain $54,000. As you can see, the median now becomes a much better measure to use when describing the "typical" family income!

Measures of Variation

You have learned that central tendency is only one aspect of the shape of a distribution. Although the measure of center is the most important aspect for many purposes, it is still just a piece of the total picture. A summary of distributions based only on their central tendency can be very incomplete, even misleading. For example, three towns might have the same mean and median crime rate but still be very different in their social character due to the shape of the crime distributions. We show three distributions of community crime rates for three different towns in Exhibit F.12. If you calculate the mean and median crime rate for each town, you will find that the mean and median crime rate is the same for all three. In terms of its crime rate, then, each community has the same central tendency.

As you can see, however, there is something very different about these towns. Town A is a very heterogeneous town; crime rates in its neighborhoods are neither very homogeneous nor clustered at either the low or high end. Rather, the crime rates in its neighborhoods are

RESEARCH IN THE NEWS

Median Lifetime Earnings

If you are feeling a bit overwhelmed and wondering whether going to college was worth it, a story from the *Washington Post* will lift your spirits. It highlights a study that utilized census data to investigate the lifetime earnings of people by their level of education. The study also examined the difference in lifetime earnings across many different college majors. If you are taking this class, you are probably not getting your major to make millions of dollars, but to help people and improve society in some way, right? The article presents a bar graph of the "median" lifetime earnings by college major. While engineering and computer science majors are at that top of the pack in terms of earnings, criminal justice and criminology majors are above many majors.

Questions About the Article?

1. Why do you think the research presented median earnings rather than mean earnings over the lifetime?
2. What other statistics would you like to know from this article?

Source: Guo, Jeff. 2014. "Want Proof College is Worth It?" *Washington Post*, September 29. http://www.washingtonpost.com/news/storyline/wp/2014/09/29/ want-proof-college-is-worth-it-look-at-this-list-of-the-highest-paying-majors/

Exhibit F.12 Neighborhood Crime Rates in Three Different Towns

Town A	Town B	Town C
19.5	58.1	8.9
28.2	59.7	15.4
35.7	60.1	18.3
41.9	62.7	21.9
63.2	63.2	63.2
75.8	63.9	103.5
92.0	64.2	104.2
95.7	64.5	110.7
109.4	65.0	105.3

spread out from one another. Crime rates in these neighborhoods are, then, very diverse. Town B is characterized by neighborhoods with very homogeneous crime rates; there are no real high or low crime areas, because the rate in each neighborhood is not far from the overall mean of 62.4 crimes per 1,000. Town C is characterized by neighborhoods with either very low crime rates or very high crime rates. Crime rates in the first four neighborhoods are much lower than the mean (62.4 crimes per 1,000), whereas those in the last four neighborhoods are much higher than the mean. Although they share identical measures of central tendency, these three towns have neighborhood crime rates that are very different.

The way to capture these differences is with statistical measures of variation. Four popular measures of variation are the range, the interquartile range, the variance, and the standard

deviation (which is the most popular measure of variability). To calculate each of these measures, the variable must be at the interval or ratio level. Statistical measures of variation are used infrequently with qualitative variables, so statistical measures will not be presented here.

Range

Range: The true upper limit in a distribution minus the true lower limit (or the highest rounded value minus the lowest rounded value, plus one).

Outlier: An exceptionally high or low value in a distribution.

The **range** is a simple measure of variation, calculated as the highest value in a distribution minus the lowest value:

$$\text{Range} = \text{Highest value} - \text{Lowest value}$$

It often is important to report the range of a distribution, to identify the whole range of possible values that might be encountered. However, because the range can be drastically altered by just one exceptionally high or low value (called an **outlier**), it does not do an adequate job of summarizing the extent of variability in a distribution. For our three towns in Exhibit F.12, the range in crime rates for Town A is 89.9 (109.4 – 19.5), for Town B it is 6.9 (65.0 – 58.7), and for Town C it is 106.4 (115.3 – 8.9).

Interquartile Range

Interquartile range: The range in a distribution between the end of the first quartile and the beginning of the third quartile.

Quartiles: The points in a distribution corresponding to the first 25% of the cases, the first 50% of the cases, and the top 25% of the cases.

A version of the range statistics, the **interquartile range**, avoids the problem created by unusually high or low scores in a distribution. It is the difference between the scores at the first and third quartiles. **Quartiles** are the points in a distribution corresponding to the first 25% of the cases (the first quartile), the first 50% of the cases (the second quartile), and the first 75% of the cases (the third quartile). You already know how to determine the second quartile, corresponding to the point in the distribution covering half of the cases; it is another name for the median. The first and third quartiles are determined in the same way, but by finding the points corresponding to 25% and 75% of the cases, respectively.

Variance

If the mean is a good measure of central tendency, then it would seem that a good measure of variability would be the distance each score is away from the mean. Unfortunately, we cannot simply take the average distance of each score from the mean. One property of the mean is that it exactly balances negative and positive distances from it, so if we were to sum the difference between each score in a distribution and the mean of that distribution, it would always sum to zero. What we can do, though, is to square the difference of each score from the mean so the distance retains its value. This is the notion behind the variance as a measure of variability.

Variance: A statistic that measures the variability of a distribution as the average squared deviation of each score from the mean of all scores.

The **variance** is the average square deviation of each case from the mean, so it takes into account the amount by which each case differs from the mean. The equation to calculate the variance is:

$$s^2 = \frac{\Sigma(x - \overline{X})^2}{N - 1}$$

In words, this formula says to take each score and subtract the mean, then square this difference, then sum all these differences, and then divide this sum by N or the total number of scores. Calculations for the variance for the crime rate data from Town A in Exhibit F.12 are shown in the table that follows.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
19.5	(19.5 ~ 62.4) = –42.9	1,840.41
28.2	(28.2 ~ 62.4) = –34.2	1,169.64
35.7	(35.7 ~ 62.4) = –26.7	712.89
41.9	(41.9 ~ 62.4) = –20.5	420.25
63.2	(63.2 ~ 62.4) = –0.8	0.64
75.8	(75.8 ~ 62.4) = 13.4	179.56
92.0	(92.0 ~ 62.4) = 29.6	876.16
95.7	(95.7 ~ 62.4) = 33.3	1,108.89
109.4	(109.4 ~ 62.4) = 47.0	2,209.00
	$\Sigma(x - \bar{X}) = 0$	$\Sigma(x - \bar{X})^2 = 8{,}517.44$

We can now determine that the variance is

$$s^2 = \frac{8{,}517.44}{8} = 1{,}064.68$$

The variance of these data, then, is 1,064.68. In "squared deviation units," the variance tells us the amount of variation the distribution has around its mean. We had to square the original deviation units before summing them, because $\Sigma(x - \bar{X}) = 0$. For most people, however, it is difficult to grasp "squared deviation units." For this reason, we typically take the square root of this value, called the standard deviation, to bring the variable back to its original units of measurement.

Standard Deviation

The **standard deviation** is simply the square root of the variance. It is the square root of the average squared deviation of each case from the mean:

Standard deviation: The square root of the average squared deviation of each case from the mean.

$$s = \sqrt{\frac{\Sigma(x - \bar{X})^2}{N - 1}}$$

To find the standard deviation, then, simply calculate the variance and take the square root. For our example, the standard deviation is

$$s = \sqrt{1{,}064.68} = 32.62$$

This value tells us that, on average, the neighborhood crime rates in Town A vary 32.62 around their mean of 62.4.

The standard deviation has mathematical properties that make it the preferred measure of variability in many cases. In particular, the calculation of confidence intervals around sample statistics, which you learned about in Chapter 5, relies on an interesting property of normal curves. Areas under the normal curve correspond to particular distances from the

mean, expressed in standard deviation units. If a variable is normally distributed, 68% of the cases will lie between plus and minus 1 standard deviation from the distribution's mean, and 95% of the cases will lie between 1.96 standard deviations above and below the mean. Cases that fall beyond plus or minus 1.96 standard deviations from the mean are termed outliers. Because of this property, the standard deviation tells us quite a bit about a distribution, if the distribution is normal. This same property of the standard deviation enables us to infer how confident we can be that the mean (or some other statistic) of a population sampled randomly is within a certain range of the sample mean (see Chapter 5).

CROSS-TABULATING VARIABLES

Most data analyses focus on relationships among variables to test hypotheses or just to describe or explore relationships. For each of these purposes, we must examine the association among two or more variables. Cross-tabulation (cross-tab) is one of the simplest methods for doing so. A cross-tabulation displays the distribution of one variable for each category of another variable; it can also be called a *bivariate distribution*. Cross-tabs also provide a simple tool for statistically controlling one or more variables while examining the associations among others. In this section, you will learn how cross-tabs used in this way can help test for spurious relationships and evaluate causal models. Cross-tabulations are usually used when both variables are measured at either the nominal or the ordinal level—that is, when the values of both variables are categories.

We are going to provide a series of examples of cross-tabulations from our delinquency data. In our first example, the independent variable we are interested in is the youth's gender (V1, see Exhibit F.2), and the dependent variable is the youth's self-reported involvement in delinquent behavior (DELINQ1). To use the delinquency variable in a cross-tabulation, however, we first need to recode it into a categorical variable. We will make three approximately equal categories of self-reported delinquency: low, medium, and high. Using the SPSS recode command, we will create another variable called DELINQ2 using the following recode commands:

$$(0 - 2 = 1)$$
$$(3 - 13 = 2)$$
$$(14 - 118 = 3)$$

Anyone who reported from none to two delinquent acts is now coded as 1, or low delinquency; anyone reporting from three to 13 delinquent acts is now coded as 2, or medium delinquency; and anyone reporting 14 or more delinquent acts is now coded as 3, or high delinquency. If you were to do a frequency distribution of this new variable, DELINQ2, you would see that there are three approximately equal groups.

We are interested in the relationship between gender and delinquency because a great deal of delinquency theory would predict that males are more likely to be delinquent than females. The gender of the youth is the independent variable, and the level of self-reported delinquency is the dependent variable.

Exhibit F.13 shows the cross-tabulation of gender with DELINQ2. Some explanation of this table is in order. Notice that there are two values of gender (male and female) that comprise the values in the two rows of the table, and three values of delinquency (low, medium, and high) that comprise the values in the three columns of the table. Cross-tabulations are usually referred to by the number of rows and columns the table has. Our cross-tabulation in Exhibit F.13 is a 2 × 3 (pronounced "two-by-three") table because there are two rows

Exhibit F.13 Cross-Tabulation of Respondents' Gender by Delinquency

		Self-Reported Delinquency			
		Low	Medium	High	Total
Gender	Female	275 40.4%	182 26.8%	223 32.8%	680 100%
	Male	175 29.6%	166 28.0%	251 42.4%	592 100%
	Total	450	348	474	1,272

and three columns. Notice also that there are values at the end of each row and at the end of each column. These totals are referred to as the *marginals* of the table. These **marginal distributions** provide the sum of the frequencies for each column and each row of the table. For example, there are 680 females in the data and 592 males. These row marginals should sum to the total number of youths in the dataset: 1,272. There are 450 youths who are low in delinquency, 348 youths who are medium in delinquency, and 474 youths who are high on the delinquency variable. These column marginals should also sum to the total number of youths in the dataset: 1,272.

Marginal distributions: The summary distributions in the margins of a cross-tabulation that correspond to the frequency distribution of the row variable and of the column variable.

Now notice that there are 2 × 3 or 6 data entries in the table (let us ignore the percentages for now). These data entries are called the *cells* of the cross-tabulation and represent the joint distribution of the two variables: gender and delinquency. The table in Exhibit F.13 has six cells for the joint distribution of two levels of gender with three levels of delinquency. In other words, notice where the value for female converges with the value of low for delinquency. You see a frequency number of 275 in this cell. This frequency is how many times there is the joint occurrence of a female and low delinquency; it shows that 275 females were also low in delinquency. Moving to the cell to the right of this, we see that there are 182 females who were medium in delinquency, and moving to the right again we see that there are 223 females who were high in delinquency. The sum of these three numbers is equal to the total number of females, 680. The row for the males shows the joint distribution of males with each level of delinquency.

What we would like to know is whether there is a relationship or an association between gender and delinquency. In other words, are males more likely to be delinquent than females? Because raw frequencies can provide a deceptive picture, we determine whether there is any relationship between our independent and dependent variables by looking at the percentages. Keep in mind that the idea in looking at relationships is that we want to know if variation on the independent variable has any effect on the dependent variable. To determine this, what we always do in cross-tabulation tables is to calculate our percentages on each value of the independent variable. For example, notice that in Exhibit F.13, gender is our independent variable. We calculated our percentages so that for each value of gender the percentages sum to 100% at the end of each row. The percentages for both females and males, therefore, sum to 100% at the end of the row. Now we take a given category of the dependent variable and ask what percentage of each independent variable value falls into that category of the dependent variable. Another way to say this is that we calculate our percentages on the independent variable and compare them to percentages on the dependent variable. We compare the percentages for different levels of the independent variable on the same category or level of the dependent variable.

In Exhibit F.13, for example, notice that 40.4% of the female youths were low in delinquency, but only 29.6% of the males were low. This tells us that females are more likely to be

low in delinquency than males. Now let us look at the *high* category. We can see that 32.8% of the females were high in delinquency and 42.4% of the males were high. Together, this tells us that females are more likely to be low in delinquency and males are more likely to be high in delinquency. There is, then, a relationship between gender and delinquency. Also notice that the independent variable was the row variable and the dependent variable was the column variable. It does not always have to be this way; the independent variable could just as easily have been the column variable. The important general rule to remember is to always calculate your percentages on the levels of the independent variable (e.g., use marginal totals for the independent variable as denominators), and compare percentages on a level of the dependent variable.

In Exhibit F.14, we report the same data as in Exhibit F.13, this time switching the rows and the columns. Now, the independent variable (gender) is the column variable, so we calculate our percentage going down each of the two columns. We then compare percentages across rows. For example, we still see that 40.4% of the females were low in delinquency, whereas only 29.6% of the males were. And 42.4% of the males were high in delinquency, but only 32.8% of the females were high in delinquency.

Describing Association

A cross-tabulation table reveals four aspects of the association between two variables:

- *Existence*. Do the percentage distributions vary at all among categories of the independent variable?
- *Strength*. How much do the percentage distributions vary among categories of the independent variable?
- *Direction*. For quantitative variables, do values on the dependent variable tend to increase or decrease with an increase in value of the independent variable?
- *Pattern*. For quantitative variables, are changes in the percentage distribution of the dependent variable fairly regular (simply increasing or decreasing), or do they vary (perhaps increasing, then decreasing, or perhaps gradually increasing, then rapidly increasing)?

Exhibit F.14 Cross-Tabulation of Respondents' Delinquency by Gender

		Gender		
		Female	Male	Total
Self-Reported Delinquency	Low	275 40.4%	175 29.6%	450
	Medium	182 26.8%	166 28.0%	348
	High	223 32.8%	251 42.4%	474
	Total	680 100%	592 100%	1,272

Exhibit F.15 Cross-Tabulation of Respondents' Morals by Delinquency

		Self-Reported Delinquency			
		Low	Medium	High	Total
Morals	Weak	20 5.6%	79 22.3%	256 72.1%	355 100%
	Medium	170 33.8%	185 36.8%	148 29.4%	503 100%
	Strong	260 62.8%	84 20.3%	70 16.9%	414 100%
	Total	450	348	474	1,272

Exhibit F.14 shows that an association exists between delinquency and gender, although we can say only that it is a modest association. The percentage difference at the low and high ends of the delinquency variables is approximately 10 percentage points.

We provide another example of a cross-tabulation in Exhibit F.15. This is a 3 × 3 table that shows the relationship between how morally wrong a youth thinks delinquency is (the independent variable) and his or her self-reported involvement in delinquency (the dependent variable). This table reveals a very strong relationship between moral beliefs and delinquency. We can see that 5.6% of youths with weak moral beliefs are low on delinquency; this increases to 33.8% for those with medium beliefs and to 62.8% for those with strong moral beliefs. At the high end, over two thirds (72.1%) of those youths with weak moral beliefs are high in delinquency, 29.4% of those with medium moral beliefs are high in delinquency, and only 16.9% of those youths with strong moral beliefs are high in delinquency. Clearly, then, having strong moral beliefs serves to effectively inhibit involvement in delinquent behavior. This is exactly what control theory would have us believe.

Exhibit F.15 shows an example of a negative relationship between an independent and a dependent variable. As the independent variable increases (i.e., as one goes from weak to strong moral beliefs), the likelihood of delinquency decreases (one becomes less likely to commit delinquency). The independent and dependent variables move in opposite directions, so this is a negative relationship. The pattern in this table is close to what is called monotonic. In a **monotonic relationship**, the value of cases consistently increases (or decreases) on one variable as the value of cases increases (or decreases) on the other variable. *Monotonic* is often defined a bit less strictly, with the idea that as the value of cases on one variable increases (or decreases), the value of cases on the other variable tends to increase (or decrease), and at least does not change direction. This describes the relationship between moral beliefs and delinquency. Delinquency is most likely when moral beliefs are low, less likely when moral beliefs are medium, and least likely when moral beliefs are strong.

Monotonic relationship: A pattern of association in which the value of cases on one variable increases or decreases fairly regularly across the categories of another variable.

We present another cross-tabulation table for you in Exhibit F.16. This table shows the relationship between the variable "number of hours studied" and the variable "certainty of punishment" (see Exhibit F.2). Both variables were originally continuous variables that we recoded into three approximately equal groups for this example. We hypothesize that those youths who study more will have a greater perceived risk of punishment than those who study less, so hours studied is our independent variable and certainty is the dependent variable. Comparing levels of hours studied for those with high certainty, we see that there is not much variation. Of those who did not study very much (0–3 hours), 39.2% were high in perceived certainty. Of those who studied from 4 to 6 hours, 35.6% were high in perceived certainty, and 40.3% of those who studied more than 7 hours per week were high in perceived certainty.

Exhibit F.16 Cross-Tabulation of Respondents' Hours Studied and Perceived Certainty of Punishment

		Certainty of Punishment			
		Low	Medium	High	Total
Number of Hours Studied	0–3 Hours	126 27.9%	148 32.8%	177 39.2%	451 100%
	4–6 Hours	117 32.8%	113 31.7%	127 35.6%	357 100%
	7 + Hours	129 27.8%	148 31.9%	187 40.39%	464 100%
	Total	372	409	491	1,272

Much the same levels prevail at low levels of perceived certainty. Those who do not study very much are no more or less likely to perceive a low certainty of punishment than those who study a lot. Variation in the independent variable, then, is not related to variation in the dependent variable. It looks like there is no association between the number of hours a youth studies and the extent to which he or she thinks punishment for delinquent acts is certain.

You will find when you read research reports and journal articles that social scientists usually make decisions about the existence and strength of association on the basis of more statistics than just percentage differences in a cross-tabulation table. A **measure of association** is a type of descriptive statistic used to summarize the strength of an association. There are many measures of association, some of which are appropriate for variables measured at particular levels. One popular measure of association in cross-tabular analyses with variables measured at the ordinal level is **gamma**. As with many measures of association, the possible values of gamma vary from –1, meaning the variables are perfectly associated in a negative direction; to 0, meaning there is no association of the type that gamma measures; to +1, meaning there is a perfect positive association of the type that gamma measures.

Measure of association: A type of descriptive statistic that summarizes the strength of an association.

Gamma: A measure of association sometimes used in cross-tabular analyses.

Inferential statistics are used in deciding whether it is likely that an association exists in the larger population from which the sample was drawn. Even when the association between two variables is consistent with the researcher's hypothesis, it is possible that the association was just due to chance or to the vagaries of sampling on a random basis. (Of course, the problem is even worse if the sample is not random.) It is conventional in statistics to avoid concluding that an association exists in the population from which the sample was drawn unless the probability that the association was due to chance is less than 5%. In other words, a statistician normally will not conclude that an association exists between two variables unless he or she can be at least 95% confident that the association was not due to chance. This is the same type of logic that you learned about in Chapter 5, which introduced the concept of 95% confidence limits for the mean. Estimation of the probability that an association is not due to chance will be based on one of several inferential statistics, **chi-square** being the one used in most cross-tabular analyses. The probability is customarily reported in a summary form such as "$p < .05$," which can be translated as "the probability that the association was due to chance is less than 5 out of 100 [5%]."

Chi-square: An inferential statistic used to test hypotheses about relationships between two or more variables in a cross-tabulation.

Statistical significance: A relationship that is not likely to be due to chance, judged by a criterion set by the analyst (often that the probability is less than 5 out of 100, or $p < .05$).

When an association passes muster in this way, when the analyst feels reasonably confident (at least 95% confident) that it was not due to chance, it is said that the association is statistically significant. **Statistical significance** means that an association is not likely to be due to chance, according to some criterion set by the analyst. Convention (and the desire to

avoid concluding that an association exists in the population when it does not) dictates that the criterion be a probability less than 5%.

But statistical significance is not everything. You may remember from Chapter 5 that sampling error decreases as sample size increases. For this same reason, an association is less likely to appear on the basis of chance in a larger sample than in a smaller sample. In a table with more than 1,000 cases, such as those involving the delinquency data-set, the odds of a chance association are often very low indeed. For example, with our table based on 1,272 cases, the probability that the association between gender and delinquency (see Exhibit F.14) was due to chance was less than 1 in 1,000 ($p < .001$)! The association in that table was fairly weak, as indicated by a gamma of .20. Even weak associations can be statistically significant with such a large sample, which means that the analyst must be careful not to assume that just because a statistically significant association exists, it is therefore important. In a large sample, an association may be statistically significant but still be too weak to be substantively significant or important. All this boils down to another reason for evaluating carefully both the existence and the strength of an association.

Controlling for a Third Variable

Cross-tabulation can also be used to study the relationship between two variables while controlling for other variables. We will focus our attention on controlling for a third variable in this section, but we will say a bit about controlling for more variables at the section's end. We will examine three different uses for three-variable cross-tabulation: identifying an intervening variable, testing a relationship for spuriousness, and specifying the conditions for a relationship. Each of these uses for three-variable cross-tabs helps determine the validity of our findings, either by evaluating criteria for causality (nonspuriousness and identification of a causal mechanism) or by increasing our understanding of the conditions required for a relationship to hold, an indication of the cross-population generalizability of the findings. All three uses are aspects of **elaboration analysis**—the process of introducing control variables into a bivariate relationship in order to better understand—to elaborate the relationship (Rosenberg, 1968). We will examine the gamma and chi-square statistics for each table in this analysis.

Elaboration analysis: The process of introducing a third variable into an analysis in order to better understand—to elaborate—the bivariate (two-variable) relationship under consideration; additional control variables also can be introduced.

Intervening Variables

We have already discovered that females are less likely to be delinquent than males (see Exhibit F.14). Finding this relationship between gender and delinquency is just the beginning of our work, however. What we would now like to know and investigate is why this relationship exists. What is it about females that makes them less likely to commit delinquent acts than males? Let us first rule out strictly biological factors and explore some possible social reasons for this gender difference in delinquency. One possibility is that because they are more closely supervised than males, females have fewer opportunities to be delinquent. In other words, females are under more strict parental supervision, and it is because they are under more strict supervision that they are less likely than males to commit delinquency. This possible relationship is shown in Exhibit F.17. Notice that in this relationship the variable "parental supervision" intervenes between gender and delinquency. It explains why females are at lower risk for delinquency compared to males. To determine whether parental supervision intervenes in the relationship between gender and delinquency and whether it explains this relationship, we must examine the relationship between gender and delinquency while controlling for difference in parental supervision. If parental supervision intervenes in the gender-delinquency relationship, the effect of controlling for this third variable would be to eliminate, or at least substantially reduce, the original relationship between gender and delinquency.

Exhibit F.17 Cross-Tabulation of Respondents' Gender by Delinquency Within Levels of Parental Supervision

Weak Parental Supervision					
		Self-Reported Delinquency			
Gender		Low	Medium	High	Total
	Female	26.1%	27.9%	46.0%	337
	Male	23.2%	27.2%	49.6%	427
	Total				764
$\chi^2 = 1{,}220$ ($p > .05$), Gamma = .067					

Strong Parental Supervision					
		Self-Reported Delinquency			
Gender		Low	Medium	High	Total
	Female	54.4%	25.7%	19.8%	343
	Male	46.1%	30.3%	23.6%	165
	Total				508
$\chi^2 = 3{,}193$ ($p > .05$), Gamma = .136					

Subtables: Tables describing the relationship between two variables within the discrete categories of one or more other control variables.

To examine this possibility, we first recode the parental supervision variable (PARSUPER; see Exhibit F.2) into two approximately equal levels: weak supervision and strong supervision. We then look at two **subtables** of the gender-delinquency relationship: once under the condition of weak parental supervision and once under the condition of strong parental supervision (see Exhibit F.17). For ease of presentation, we will report only the cell percentages and not the frequencies. What we see is that once parental supervision is controlled, there is no real relationship between gender and delinquency. That is, if males and females have the same amount of supervision from their parents, they do not differ that much in their risk of being delinquent. For example, among females with weak parental supervision, 46.0% are high in delinquency, and among males with weak parental supervision, 49.6% are high in delinquency. There is less than four percentage points' difference between males and females in their risk of being high delinquents under these conditions. Among those with strong parental supervision, 19.8% of the females were high in delinquency and 23.6% of the males were high, less than four percentage points' difference.

This percentage analysis is borne out by the chi-square tests and measures of association. Under both the weak and strong levels of parental supervision, the relationship between gender and delinquency is not significant, and gamma is only .067 when supervision is weak and .136 when supervision is strong. In neither case is the obtained gamma very different from zero (indicating no relationship). Collectively, these results would lead us to the conclusion that parental supervision intervenes in the relationship between gender and delinquency. A very important reason females are less delinquent than males, therefore, is that females are under stricter supervision from their parents than are males, and strong parental supervision leads to a reduced risk of delinquency.

Extraneous Variables

Another reason for introducing a third variable into a bivariate relationship is to see whether the original relationship is spurious due to the influence of an **extraneous variable**, which is a variable that causes both the independent and dependent variables. The only reason the independent and dependent variables are related, therefore, is that they both are the effects of a common cause (another independent variable).

Extraneous variable: A variable that influences both the independent and dependent variables so as to create a spurious association between them that disappears when the extraneous variable is controlled.

Exhibit F.18 shows what a spurious relationship would look like. In this case, the relationship between *x* and *y* exists only because both are the effects of the common cause *z*. Controlling for *z*, therefore, will eliminate the *x*-*y* relationship. Ruling out possible extraneous variables will help considerably strengthen the conclusion that the relationship between the independent and dependent variables is causal, particularly if all the variables that seem to have the potential for creating a spurious relationship can be controlled.

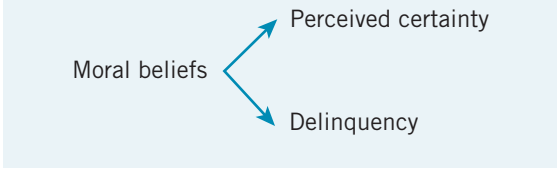
Notice that if a variable is acting as an extraneous variable, then controlling for it will cause the original relationship between the independent and dependent variables to disappear or substantially diminish. This was also the empirical test for an intervening variable. Therefore, the difference between intervening and extraneous variables is a logical one and not an empirical one. In both instances, controlling for the third variable will cause the original relationship to diminish or disappear. There should, therefore, be sound theoretical grounds for suspecting that a variable is acting as an intervening variable, explaining the relationship between the independent and dependent variables.

Exhibit F.18 Example of a Spurious Relationship

As an example of a possible extraneous relationship, we will look at the association between a youth's perception of the certainty of punishment and self-reported involvement in delinquency. Deterrence theory should lead us to predict a negative relationship between perceived certainty and delinquency. Indeed, this is exactly what we observe in our delinquency data. We will not show you the cross-tabulation table, but when we looked at the relationship between perceived certainty and delinquency, we found that 53.2% of youth who were low in certainty were high in delinquency; 39.1% of those who perceived medium certainty were high in delinquency; and only 23.6% of those who perceived a high certainty of punishment were high in delinquency. Youth who believed they would get caught if they engaged in delinquency, then, were less likely to be delinquent. The gamma value for this table was –.382, indicating a moderate negative relationship between perceived certainty and delinquency, exactly what deterrence theory would lead us to expect.

Someone may reasonably argue, however, that this discovered negative relationship may not be causal but instead may be spurious. It could be suggested that what is actually behind this relationship is the extraneous variable, moral beliefs. The argument is that those with strong moral inhibitions against committing delinquent acts think that punishment for morally wrongful actions is certain *and* refrain from delinquent acts. Thus, the observed negative relationship between perceived certainty and delinquency is really due to the positive effect of moral beliefs on perceived certainty and the negative effect of moral beliefs on delinquency (see Exhibit F.19). If moral beliefs are actually the causal factor at work, then controlling for them will eliminate or substantially reduce the original relationship between perceived certainty and delinquency.

Exhibit F.19 A Spurious Relationship Between x and y



To look at this possibility, we examined the relationship between perceived certainty and delinquency under three levels of moral beliefs (weak, medium, and strong). The crosstabulations are shown in Exhibit F.20. What we can see is that in each of the subtables there is a negative and significant association between the perceived certainty of punishment and delinquency. In two of the three subtables, however, the relationship is weaker than what was in the original table (there the gamma was –.382); we obtained gammas of –.271 and –.197. Under the condition of strong moral beliefs, however, the original relationship is unchanged. What we would conclude from this elaboration analysis is that the variable "moral beliefs" is not acting as a very strong extraneous variable. Although some of the relationship between perceived risk and delinquency is due to their joint relationship with moral beliefs, we cannot dismiss the possibility that the perceived certainty of punishment has a causal influence on delinquent behavior.

Exhibit F.20 Cross-Tabulation of Perceived Risk by Delinquency Within Levels of Moral Beliefs

Weak Moral Beliefs					
		Self-Reported Delinquency			
		Low	Medium	High	Total
Perceived Certainty	Low	3.8%	14.7%	81.4%	156
	Medium	8.6%	27.3%	64.1%	128
	High	4.2%	29.6%	66.8%	71
	Total				355
$\chi^2 = 13.646$ ($p<.001$), Gamma = –.271					

Medium Moral Beliefs					
		Self-Reported Delinquency			
		Low	Medium	High	Total
Perceived Certainty	Low	22.0%	42.5%	35.4%	127
	Medium	33.9%	35.6%	30.5%	174
	High	41.1%	34.2%	24.8%	202
	Total				503
$\chi^2 = 13.646$ ($p<.001$), Gamma = –.197					

Strong Moral Beliefs					
		Self-Reported Delinquency			
		Low	Medium	High	Total
Perceived Certainty	Low	42.7%	28.1%	29.2%	89
	Medium	58.9%	17.7%	23.4%	107
	High	72.9%	18.3%	8.7%	218
	Total				414
$\chi^2 = 13.646$ ($p<.001$), Gamma = –.393					

Specification

By adding a third variable to an evaluation of a bivariate relationship, the data analyst can also specify the conditions under which the bivariate relationship occurs. A **specification** occurs when the association between the independent and dependent variables varies across the categories of one or more other control variables—that is, when the original relationship is stronger under some condition or conditions of a third variable and weaker under others.

Specification: A type of relationship involving three or more variables in which the association between the independent and dependent variables varies across the categories of one or more other control variables.

In criminology, social learning theory would predict that youths who are exposed to peers who provide verbal support for delinquency are at greater risk for their own delinquent conduct. We found support for this hypothesis in our delinquency dataset. We examined this relationship by recoding into two approximately equal groups the variable FROPINON (see Exhibit F.2). The first group had weak verbal support from peers, whereas the second group had strong verbal support. Among those youths who reported that their peers provided only weak verbal support for delinquency, 15% were highly delinquent. Among those with strong verbal support from peers, nearly 58% were highly delinquent. The gamma value for this relationship was .711, a very strong positive relationship. Clearly, then, having friends give you verbal support for delinquent acts (e.g., "it's okay to steal") puts you at risk for delinquency.

It is entirely possible, however, that this relationship exists only when friends' verbal support is backed up by their own behavior. That is, verbal support from our peers might not affect our delinquency when they do not themselves commit delinquent acts or when they commit only a very few. In this case, their actions (inaction in this case) speak louder than their words, and their verbal support does not influence us. When they also commit delinquent acts, however, the verbal support of peers carries great weight.

We looked at this possibility to examine the relationship between friends' verbal support for delinquency and a youth's own delinquency within two levels of friends' behavior (FRBEHAVE; see Exhibit F.2). We recoded FRBEHAVE into two approximately equal groups. In the first group,

Exhibit F.21 Cross-Tabulation of Friends' Verbal Support by Delinquency Within Levels of Friends' Delinquent Behavior

Few Delinquent Friends					
		Self-Reported Delinquency			
		Low	Medium	High	Total
Friends' Verbal Support	Weak	67.3%	23.8%	8.9%	437
	Strong	44.3%	34.3%	21.4%	140
	Total				577
$\chi^2 = 27.374$ (p>.001), Gamma = .416					

Many Delinquent Friends					
		Self-Reported Delinquency			
		Low	Medium	High	Total
Perceived Certainty	Weak	30.5%	38.5%	31.0%	174
	Strong	7.5%	24.8%	67.4%	521
	Total				695
$\chi^2 = 87.508$ (p>.001), Gamma = .608					

fewer of one's friends are delinquent (few delinquent friends) than the other (many delinquent friends). This attempt to specify the relationship between friends' opinions and a youth's own delinquency is shown in Exhibit F.21. What we see is a little complex. When only a few of a youth's friends are committing delinquent acts, their verbal support still has a significant and positive effect on self-reported delinquency. The gamma value in this subtable is .416, which is moderately strong but less than the original gamma of .771. When many of a youth's friends are delinquent, however, the positive relationship between peers' verbal support and self-reported delinquency is much stronger, with a gamma of .608. The behavior of our peers, then, only weakly specifies the relationship between peer opinion and delinquency. Clearly, then, what our peers say about delinquency matters, even if they are not committing delinquent acts all the time themselves.

Regression analysis: A statistical technique for characterizing the pattern of a relationship between two quantitative variables in terms of a linear equation and for summarizing the strength of this relationship.

Correlation analysis: A standardized statistical technique that summarizes the strength of a relationship between two quantitative variables in terms of its adherence to a linear pattern.

REGRESSION AND CORRELATION

Our goal in introducing you to cross-tabulation has been to help you think about the associations among variables and to give you a relatively easy tool for describing association. To read most statistical reports and to conduct more sophisticated analyses of social data, you will have to extend your statistical knowledge. Many statistical reports and articles published in the social sciences use statistical techniques called **regression analysis** and **correlation analysis** to describe the associations among two or more quantitative variables. The terms actually refer to different aspects of the same technique. Statistics based on regression and correlation are used frequently in social science and have many advantages over cross-tabulation—as well as some disadvantages.

We provide only a brief overview of this approach here. Take a look at Exhibit F.22. It's a plot, termed a *scatterplot*, of the bivariate relationship between two interval/ratio-level variables. The variables were obtained from a U.S. state-level dataset. The dependent variable, presented on the *y*-axis (vertical) is the murder rate per 100,000 population, and the independent variable, presented on the *x*-axis (horizontal), is the poverty rate (percentage of each state's population living under the poverty level).

You can see that the data points in the scatterplot tend to run from the lower left to the upper right of the chart, indicating a positive relationship. States with higher levels of poverty also tend to have higher rates of murder. This regression line is the "best fitting" straight line for this relationship—it is the line that lies closest to all the points in the chart, according to certain criteria. But you can easily see that quite a few points are pretty far from the regression line.

How well does the regression line fit the points? In other words, how close does the regression line come to the points? (Actually, it's the square of the vertical distance, on the *y*-axis, between the points and the regression line that is used as the criterion.) The **correlation coefficient**, also called *Pearson's r*, or just *r*, gives one answer to that question. The value of *r* for this relationship is .60, which indicates a moderately strong positive linear relationship (if it were a negative relationship, *r* would have a negative sign). The value of *r* is 0 when there is absolutely no linear relationship between the two variables, and it is 1 when all the points representing all the cases lie exactly on the regression line (which would mean that the regression line describes the relationship perfectly).

Correlation coefficient (r): A summary statistic that varies from 0 to 1 or −1, with 0 indicating the absence of a linear relationship between two quantitative variables and 1 or −1 indicating that the relationship is completely described by the line representing the regression of the dependent variable on the independent variable.

So the correlation coefficient does for two interval/ratio-level variables what gamma does for a cross-tabulation table: It is a summary statistic that tells us about the strength of the association between the two variables. Values of *r* close to 0 indicate that the relationship is weak; values of *r* close to ±1 indicate the relationship is strong—in between there is a lot of room for judgment. You will learn in a statistics course that r^2 is often used instead of *r*. Exhibit F.23 provides an overview of how to interpret the values of *r*. Although not all possible values of *r* are displayed in Exhibit F.23, it highlights how the use of adjectives can describe various values between 0 and 1.

Exhibit F.22 Example of a Positive Relationship. Scatterplot of Murder Rate (dependent variable) and Poverty Rate (independent variable) in U.S. States, 2010.

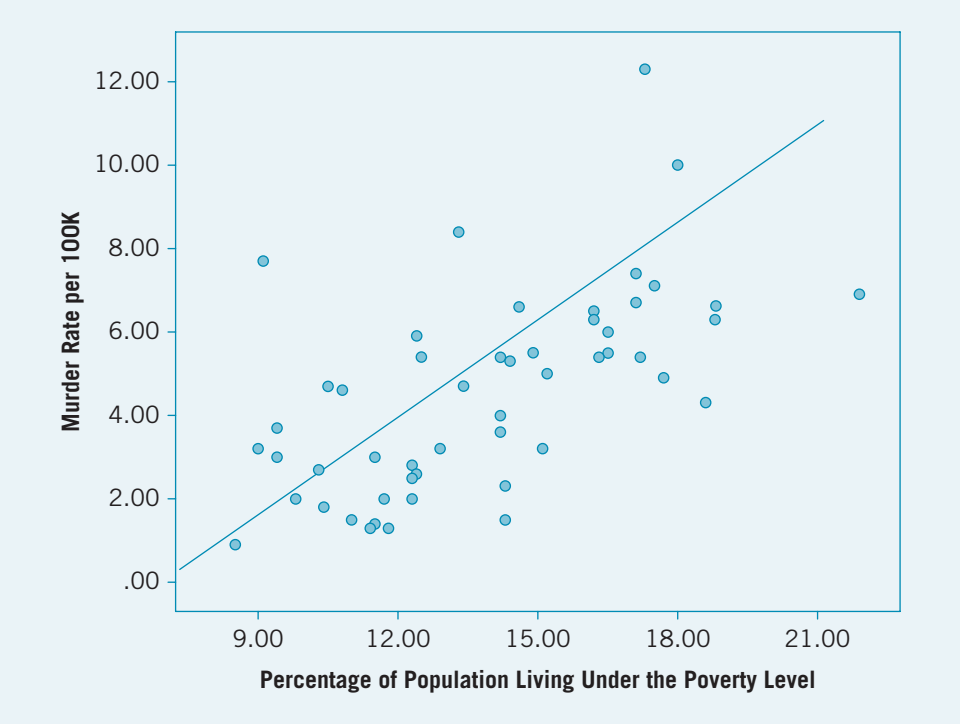
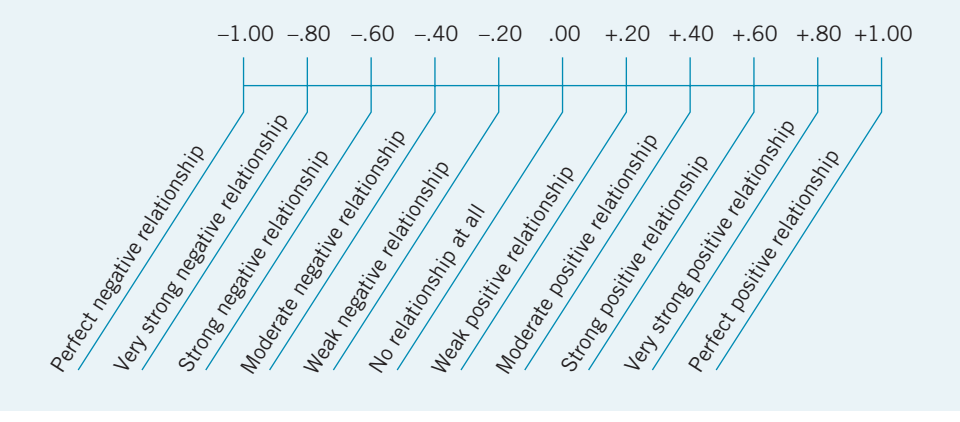


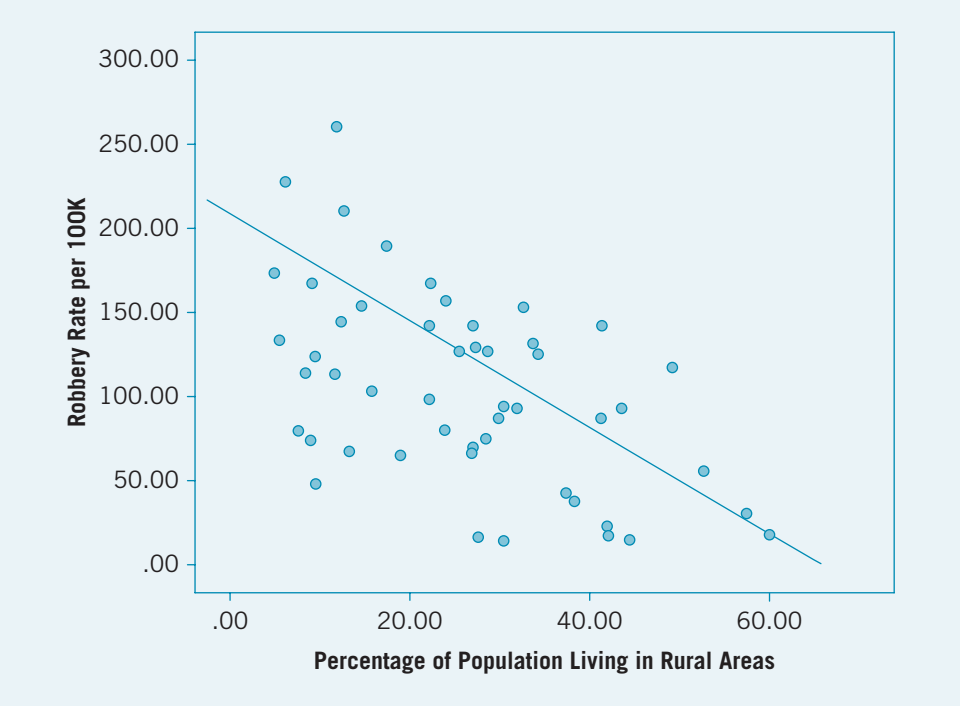
Exhibit F.23 A Guide to Interpreting Strong to Weak Relationships



Source: Frankfort-Nachmais and Leon-Guerrero (2006:230). Reprinted with permission from SAGE Publications, Inc.

An example of a negative relationship is shown in Exhibit F.24, where we provide a scatterplot of the robbery rate in states (dependent variable) on the *y*-axis and the percentage of each state's population that resides in rural areas as the independent variable (*x*-axis). You can

Exhibit F.24 Example of a Negative Relationship. Scatterplot of Robbery Rate (dependent variable) and Percentage Rural (independent variable) in U.S. States, 2010.



see here a clear negative relationship; a state that has a higher percentage of its population residing in rural areas will tend to have lower robbery rates. The correlation coefficient for this relationship is $r = -.53$, indicating a moderate negative relationship.

You can also use correlation coefficients and regression analysis to study simultaneously the association among three or more variables. Let's use the murder rate as the dependent variable to illustrate. In a *multiple regression analysis*, you could test to see whether several other variables in addition to poverty are associated simultaneously with the murder rate—that is, whether the variables have independent effects on murder after statistically controlling for each other.

Controlling for the geography in a state is also important for predicting murder rates, so we will be including percentage rural in our equation. We also know that robberies sometimes have lethal outcomes, so controlling for the robbery rate is also important. Let's examine what a multiple regression equation would look like predicting the murder rate using the poverty rate, the percentage rural, and the robbery rate as the three independent variables. Interpreting regression output is way beyond the scope of this text; we are simply going to examine the standardized regression coefficients, called *betas*, and their significance level for this illustration. Results are displayed in Exhibit F.25.

First, look at the numbers under the Beta Coefficient heading. Beta coefficients are standardized statistics that indicate how strong the linear relationship is between the dependent variable (murder rate, in this case) and each independent variable, while the other independent variables are controlled. Like the correlation coefficient (r), values of beta range

Exhibit F.25 Multiple Regression Predicting the Murder Rate in States Using Poverty, the Divorce Rate, and the Robbery Rate as Independent Variables

Variable	Beta Coefficient	Significance Level
Percentage poor	.462	$P = 001$
Percentage rural	–.001	$p = 382$
Robbery rate	.630	$p = 001$
R^2	.68	
N	50	

from 0, when there is no linear association, to ±1.0, when the association falls exactly on a straight line. You can see in the beta column that rural population is not significantly related to the murder rate when the other variables are controlled. Both the percentage poor and the robbery rate, however, are still significant predictors of murder. R^2 (r-squared) is a model fit statistic and tells us, when multiplied by 100, the percentage of the dependent variable's variation that is explained by all the independent variables in the model. In this model, we learn from R^2 that the three independent variables together explain, or account for, 68% of the total variation in murder rates. Our goal is to explain as much variation as possible of the 100%, so explaining over two-thirds of the variation is not bad!

You will need to learn more about when correlation coefficients and regression analysis are appropriate (e.g., both variables have to be quantitative, and the relationship has to be linear [not curvilinear]), but that's for another time and place. To learn more about correlation coefficients and regression analysis, you should take an entire statistics course. For now, this short introduction will enable you to make sense of more of the statistical analyses you find in research articles. You can also learn more about these techniques with the tutorials on the text's study site.

ANALYZING DATA ETHICALLY: HOW NOT TO LIE ABOUT RELATIONSHIPS

When the data analyst begins to examine relationships among variables in some real data, social science research becomes most exciting. The moment of truth, it would seem, has arrived. Either the hypotheses are supported or not. But, in fact, this is also a time to proceed with caution and to evaluate the analyses of others with even more caution. Once large datasets are entered into a computer, it becomes very easy to check out a great many relationships; when relationships are examined among three or more variables at a time, the possibilities become almost endless.

This range of possibilities presents a great hazard for data analysis. It becomes very tempting to search around in the data until something interesting emerges. Rejected hypotheses are forgotten in favor of highlighting what's going on in the data. It is not wrong to examine data for unanticipated relationships; the problem is that inevitably some relationships between variables will appear just on the basis of chance association alone. If

CAREERS AND RESEARCH

Dana Hunt, PhD, Principal Scientist

Source: Dana Hunt

In the study site video for this chapter, Dana Hunt discusses two of the many lessons she has learned about measurement in a decades-long career in social research. Hunt received her BA in sociology from Hood College in Pennsylvania and then earned her PhD in sociology at the University of Pennsylvania. After teaching at Hood for several years, she took an applied research position at National Development and Research Institutes (NDRI) in New York City. NDRI's description on its website gives you an idea of what drew the attention of a talented young social scientist.

Founded in 1967, NDRI is a nonprofit research and educational organization dedicated to advancing scientific knowledge in the areas of drug and alcohol abuse, treatment, and recovery; HIV, AIDS, and HCV (hepatitis C virus); therapeutic communities; youth at risk; and related areas of public health, mental health, criminal justice, urban problems, prevention, and epidemiology.

Hunt moved from New York to the Boston area in 1990, where she is now a principal scientist at Abt Associates, Inc., in Cambridge. Abt's website description conveys the scope of the research projects the company directs.

Abt Associates applies scientific research, consulting, and technical assistance expertise on a wide range of issues in social, economic, and health policy; international development; clinical trials; and registries. One of the largest for-profit government and business research and consulting firms in the world, Abt Associates delivers practical, measurable, high-value-added results.

Two of Hunt's major research projects in recent years are the nationwide Arrestee Drug Abuse Monitoring Program for the Office of National Drug Control Policy and a study of prostitution and sex trafficking demand reduction for the National Institute of Justice.

you search hard and long enough, it will be possible to come up with something that really means nothing.

A reasonable balance must be struck between deductive data analysis to test hypotheses and inductive analysis to explore patterns in a dataset. Hypotheses formulated in advance of data collection must be tested as they were originally stated; any further analyses of these hypotheses that involve a more exploratory strategy must be labeled as such in research reports. Serendipitous findings do not need to be ignored, but it must be reported that they were serendipitous. Subsequent researchers can try to deductively test the ideas generated by our explorations.

We also have to be honest about the limitations of using survey data to test causal hypotheses. The usual practice for those who seek to test a causal hypothesis with nonexperimental survey data is to test for the relationship between the independent and dependent variables, controlling for other variables that might possibly create spurious relationships. This is what we did by examining the relationship between the perceived certainty of punishment and delinquency while controlling for moral beliefs. But finding that a hypothesized relationship is not altered by controlling for just one variable does not establish that the relationship is causal, nor does controlling for two, three, or many more variables. There always is a possibility that some other variable that we did not think to control, or that was not even measured in the survey, has produced a spurious relationship between the independent and dependent variables in our hypothesis (Lieberson, 1985). We must always think about the possibilities and be cautious in our causal conclusions.

CONCLUSION

This appendix has demonstrated how a researcher can describe phenomena in criminal justice and criminology, identify relationships among them, explore the reasons for these relationships, and test hypotheses about them. Statistics provide a remarkably useful tool for developing our understanding of the social world, a tool that we can use to test our ideas and generate new ones.

Unfortunately, to the uninitiated, the use of statistics can seem to end debate right there; you cannot argue with the numbers. But you now know better than that. The numbers will be worthless if the methods used to generate the data are not valid, and the numbers will be misleading if they are not used appropriately, taking into account the type of data to which they are applied. And even assuming valid methods and proper use of statistics, there is one more critical step, for the numbers do not speak for themselves. Ultimately, it is how we interpret and report the numbers that determines their usefulness.

GLOSSARY

Bar chart: A graphic for qualitative variables in which the variable's distribution is displayed with solid bars separated by spaces.

Base *N*: The total number of cases in a distribution.

Bimodal distribution: A distribution that has two nonadjacent categories with about the same number of cases, and these categories have more cases than any other categories.

Central tendency: A feature of a variable's distribution; refers to the value or values around which cases tend to center.

Chi-square: An inferential statistic used to test hypotheses about relationships between two or more variables in a cross-tabulation.

Correlation analysis: A standardized statistical technique that summarizes the strength of a relationship between two quantitative variables in terms of its adherence to a linear pattern.

Correlation coefficient *(r)*: A summary statistic that varies from 0 to 1 or –1, with 0 indicating the absence of a linear relationship between two quantitative variables and 1 or –1 indicating that the relationship is completely described by the line representing the regression of the dependent variable on the independent variable.

Cross-tabulation (cross-tab): A bivariate (two-variable) distribution showing the distribution of one variable for each category of another variable.

Data cleaning: The process of checking data for errors after the data have been entered in a computer file.

Descriptive statistics: Statistics used to describe the distribution of and relationship among variables.

Elaboration analysis: The process of introducing a third variable into an analysis in order to better understand—to elaborate—the bivariate (two-variable) relationship under consideration; additional control variables also can be introduced.

Extraneous variable: A variable that influences both the independent and dependent variables so as to create a spurious association between them that disappears when the extraneous variable is controlled.

Frequency distributions: Numerical display showing the number of cases, and usually the percentage of cases (the relative frequencies), corresponding to each value or group of values of a variable.

Gamma: A measure of association sometimes used in cross-tabular analyses.

Grouped frequency distribution: A frequency distribution in which the data are organized into categories, either because there are more values than can be easily displayed or because the distribution of the variable will be clearer or more meaningful.

Histogram: A graphic for quantitative variables in which the variable's distribution is displayed with adjacent bars.

Inferential statistics: Mathematical tools for estimating how likely it is that a statistical result based on data from a random sample is representative of the population from which the sample is assumed to have been selected.

Interquartile range: The range in a distribution between the end of the first quartile and the beginning of the third quartile.

Marginal distributions: The summary distributions in the margins of a cross-tabulation that correspond to the frequency distribution of the row variable and of the column variable.

Mean: The arithmetic or *weighted* average, computed by adding up the value of all the cases and dividing by the total number of cases.

Measure of association: A type of descriptive statistic that summarizes the strength of an association.

Median: The position average, or the point that divides a distribution in half (the 50th percentile).

Mode: The most frequent value in a distribution, also termed the *probability average*.

Monotonic relationship: A pattern of association in which the value of cases on one variable increases or decreases fairly regularly across the categories of another variable.

Negatively skewed: A distribution in which cases cluster to the right side and the left tail of the distribution is longer than the right.

Outlier: An exceptionally high or low value in a distribution.

Percentage: Relative frequencies, computed by dividing the frequency of cases in a particular category by the total number of cases and multiplying by 100.

Positively skewed: Describes a distribution in which the cases cluster to the left and the right tail of the distribution is longer than the left.

Quartiles: The points in a distribution corresponding to the first 25% of the cases, the first 50% of the cases, and the top 25% of the cases.

Range: The true upper limit in a distribution minus the true lower limit (or the highest rounded value minus the lowest rounded value, plus one).

Regression analysis: A statistical technique for characterizing the pattern of a relationship between two quantitative variables in terms of a linear equation and for summarizing the strength of this relationship.

Skewness: A feature of a variable's distribution; refers to the extent to which cases are clustered more at one or the other end of the distribution rather than around the middle.

Specification: A type of relationship involving three or more variables in which the association between the independent and dependent variables varies across the categories of one or more other control variables.

Standard deviation: The square root of the average squared deviation of each case from the mean.

Statistical significance: An association that is not likely to be due to chance, judged by a criterion set by the analyst (often that the probability is less than 5 out of 100 or $p < .05$).

Subtables: Tables describing the relationship between two variables within the discrete categories of one or more other control variables.

Unimodal distribution: A distribution of a variable in which there is only one value that is the most frequent.

Variability: A feature of a variable's distribution, referring to the extent to which cases are spread out through the distribution or clustered in only one location.

Variance: A statistic that measures the variability of a distribution as the average squared deviation of each score from the mean of all scores.

HIGHLIGHTS

- Data collection instruments should be precoded for direct entry, after verification, into a computer. All data should be cleaned during the data entry process.
- Use of secondary data can save considerable time and resources but may limit data analysis possibilities.
- Bar charts, histograms, and frequency polygons are useful for describing the shape of distributions. Care must be taken with graphic displays to avoid distorting a distribution's apparent shape.
- Frequency distributions display variation in a form that can be easily inspected and described. Values should be grouped in frequency distributions in a way that does not alter the shape of the distribution. Following several guidelines can reduce the risk of problems.

- Summary statistics are often used to describe the central tendency and variability of distributions. The appropriateness of using the mode, mean, and median for a description varies with a variable's level of measurement, the distribution's shape, and the purpose of the summary.
- The variance and standard deviation summarize variability around the mean. The interquartile range is usually preferable to the range to indicate the interval spanned by cases, due to the effect of outliers on the range. The degree of skewness of a distribution is usually described in words rather than with a summary statistic.
- Cross-tabulations should normally be divided into percentages within the categories of the independent variable. A cross-tabulation can be used to determine the existence, strength, direction, and pattern of an association.
- Elaboration analysis can be used in cross-tabular analysis to test for spurious and intervening relationships and to identify the conditions under which relationships occur.
- Inferential statistics are used with sample-based data to estimate the confidence that can be placed in a statistical estimate of a population parameter. Estimates of the probability that an association between variables may have occurred on the basis of chance are also based on inferential statistics.
- Regression analysis is a statistical method for characterizing the relationship between two or more quantitative variables with a linear equation and for summarizing the extent to which the linear equation represents that relationship. Correlation coefficients summarize the fit of the relationship to the regression line.

EXERCISES

1. Create frequency distributions from lists in the Federal Bureau of Investigation (FBI) Uniform Crime Reports on characteristics of arrestees in at least 100 cases (cites). You will have to decide on grouping schemes for the distribution of data for variables such as race, age, and crime committed, and how to deal with outliers in the frequency distribution.
   a. Decide what summary statistics to use for each variable of interest. How well were the features of each distribution represented by the summary statistics? Describe the shape of each distribution.
   b. Propose a hypothesis involving two of these variables, and develop a cross-tabulation to evaluate the support for this hypothesis.
   c. Describe each relationship in terms of the four aspects of an association, after making percentages within each table within the categories of the independent variable. Which hypotheses appear to have been supported?
2. Become a media critic. For the next week, scan a newspaper or some magazines for statistics related to crime or criminal victimization. How many can you find using frequency distributions, graphs, and the summary statistics introduced in this chapter? Are these statistics used appropriately and interpreted correctly? Would any other statistics have been preferable or useful in addition to those presented?
3. The table that follows shows a frequency distribution of "trust in people" as produced by SPSS with the General Social Survey data. As you can see, the table includes abbreviated labels for the variable and its response choices, as well as the raw frequencies and three percentage columns. The first percentage column (Percentage) shows the percentage in each category; the next percentage column (Valid Percentage) is based on the total number of respondents who gave valid answers (3,929 in this instance). It is the Valid Percentage column that normally should be used to construct a frequency distribution for presentation. The last percentage column is Cumulative Percentage, adding up the valid percentages from top to bottom.

Redo the table for presentation, using the format of the frequency distributions presented in the text.

		Frequency	Percentage	Valid Percentage	Cumulative Percentage
Valid	CAN TRUST	1279	28.4		
	CANNOT TRUST	2458	54.5		
	DEPENDS	192	4.3		
	Total	3929	87.1		
		Frequency	Percentage		
Missing	NAP	575	12.7		
	NA	6	.1		
	Total	581	12.9		
Total		4510	100.0		

DEVELOPING A RESEARCH PROPOSAL

Use the General Social Survey data to add a pilot study to your proposal. A pilot study is a preliminary effort to test out the procedures and concepts that you have proposed to research.

1. Review the GSSCRJ2K variable list, and identify some variables that have at least some connection to your research problem. If possible, identify one variable that might be treated as independent in your proposed research and one that might be treated as dependent.
2. Request frequencies for these variables.
3. Request a cross-tabulation of the dependent variable by the independent variable (if you were able to identify any). If necessary, recode the independent variable to three or fewer categories.
4. Write a brief description of your findings and comment on their implications for your proposed research. Did you learn any lessons from this exercise for your proposal?

WEB EXERCISES

1. Search the web for a crime-related example of statistics. The Bureau of Justice Statistics is a good place to start: www.ojp.usdoj.gov/bjs/. Using the key terms from this chapter, describe the set of statistics you have identified. What phenomena does this set of statistics describe? What relationships, if any, do the statistics identify?
2. Do a web search for information on a criminological subject that interests you. How much of the information that you find relies on statistics as a tool for understanding the subject? How do statistics allow researchers to test their ideas about the subject and generate new ideas? Write your findings in a brief report, referring to the websites that you found.

ETHICS EXERCISES

1. Review the frequency distributions and graphs in this chapter. Change one of these data displays so that you are "lying with statistics."
2. Consider the relationship between gender and delinquency that is presented in Exhibit F.13. What third variable do you think should be controlled in

the analysis to better understand the basis for this relationship? How might criminal justice policies be affected by finding out that this relationship was due to differences in teacher expectations rather than to genetic differences in violence propensity?

SPSS OR EXCEL EXERCISES

	Data for Exercise
Dataset	**Description**
Youth.sav	This dataset is from a random sample of students from schools in a southern state. While not representative of the United States, it covers a variety of important delinquent behaviors and peer influences.
	Variables for Exercise
Variable Name	**Description**
delinquency	An interval/level variable that measures self-reported delinquency.
D1	A binary variable based on the number of delinquent acts a respondent reported. A 0 indicates that the respondent reported 1 or fewer acts, while 1 indicates 2 or more.
	Variables for Exercise
Variable Name	**Description**
lowcertain_bin	Binary indicator for whether the respondent felt there was certainty that he or she would be punished for delinquent behaviors, where 1 = low certainty and 0 = high certainty.
certain	A scale indicating how likely the individual feels it is that he or she will be punished for delinquent behavior. High values indicate high certainty.

1. For this exercise let's take a look at whether a person's expectation of punishment after Delinquency is associated with the number of deviant behaviors a student engages in, as measured by the variable Delinquency.
   a. Run a frequency of the dependent variable, delinquency, and answer the following questions:
      i. What level of measurement is this item?
      ii. What forms of descriptive analysis are appropriate?
      iii. How would you best represent this data in a graph?
   b. Based on your responses to Part 1a, conduct all appropriate descriptive analyses. Be sure to describe what you can about the data's distribution and what measures of central tendency are most appropriate. If one or another measure may produce misleading results, be sure to caution the reader why.
2. D1 measures delinquency differently than the interval/ratio level variable called Delinquency. Is the variable D1 appropriate for use in an ordinary least squares (OLS) regression analysis? Why or why not? If you have been taught them, consider how it will or will not meet different assumptions of OLS.
3. Repeat Part 1 for the independent variable, which is called lowcertain_bin. Again, describe the variable and how you would go about presenting it. Remember that you are required to conduct only the analyses that are appropriate.
4. On to the actual analysis! First, let's compare mean delinquency scores for lowcertain_bin and the

delinquency variable. This can be done under analyze->means->compare means.

a. What is the difference between the two group means?

b. What do these results suggest substantively?

5. Second, let's estimate a linear regression model. This can be accessed by selecting analyze->regression->linear.

   a. What are your results? How do they compare with your results in Part 4?

   b. Do you notice any similarities between your regression coefficient and the results from Part 4? Think carefully about why this is the case—would this similarity apply to all independent variables in a regression model or just binary ones?

   c. Test your answer to Part 5b by running the regression model again, but this time use the continuous version of lowcertain_bin, named "certain." High values on this measure indicate high levels of certainty, which is the inverse of the original measure.

      i. How have your results changed?

      ii. Do these results lead to substantively similar conclusions?

6. Return to your answer for Part 2. How sound do you think these specific analyses are, given that they are all based on the analysis of means? If you think they may be biased, explain how they are biased and any ideas you might have for overcoming them.