CHAPTER 8

Elementary Quantitative Data Analysis

Research|Social Impact Link 8.1
Read more about quantitative analysis and society.

"Show me the data," says your boss. Presented with a research conclusion, most people—not just bosses—want evidence to support it; presented with piles of data, you the researcher need to uncover what it all means. To handle the data gathered by your research, you need to use straightforward methods of data analysis.

In this chapter, we introduce several common statistics used in social research and explain how they can be used to make sense of the "raw" data gathered in your research. Such **quantitative data analysis**, using numbers to discover and describe patterns in your data, is the most elementary use of social statistics.

2 Why Do Statistics?

A **statistic**, in ordinary language usage, is a numerical description of a population, usually based on a sample of that population. (In the technical language of mathematics, a *parameter* describes a population, and a *statistic* specifically describes a sample.) Some statistics are useful for describing the results of measuring single variables or for constructing and evaluating multi-item scales. These statistics include frequency distributions, graphs, measures of central tendency and variation, and reliability tests. Other statistics are used primarily to describe the association among variables and to control for other variables, and thus, to enhance the causal validity of our conclusions. Cross-tabulation, for example, is one simple technique for measuring association and controlling other variables; it is introduced in this chapter. All of these statistics are termed **descriptive statistics**, because they describe the distribution of and relationship among variables. Statisticians also use **inferential statistics** to estimate the degree of confidence that can be placed in generalizations from a sample to the population from which the sample was selected.

Quantitative data analysis: Statistical techniques used to describe and analyze variation in quantitative measures.

Video Link 8.1
Watch a clip about research and social problems.

Statistic: A numerical description of some feature of a variable or variables in a sample from a larger population.

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

Inferential statistics: Statistics used to estimate 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.

Case Study: The Likelihood of Voting

In this chapter, we use for examples some data from the 2010 General Social Survey (GSS) on voting and other forms of political participation. What influences the likelihood of voting? Prior research on voting in both national and local settings provides a great deal of support for one hypothesis: The likelihood of voting increases with social status (Milbrath & Goel 1977:92–95; Salisbury 1975:326; Verba & Nie 1972:126). We will find out whether this hypothesis was supported in the 2010 GSS and examine some related issues.

The variables we use from the 2010 GSS are listed in Exhibit 8.1. We use these variables to illustrate particular statistics throughout this chapter.

2 How to Prepare Data for Analysis

Our analysis of voting in this chapter is an example of what is called **secondary data analysis**. It is secondary because we received the data secondhand. A great many high-quality datasets are available for reanalysis from the Inter-University Consortium for Political and Social Research at the University of Michigan (1996), and many others can be obtained from the government, individual researchers, and other research organizations (see Appendix C).

If you have conducted your own survey or experiment, your quantitative data must be prepared in a format suitable for computer entry. Questionnaires or other

Secondary data analysis: Analysis of data collected by someone other than the researcher or the researcher's assistants.

Exhibit 8.1 List of GSS 2010 Variables for Analysis of Voting

Variable[a]	SPSS Variable Name	Description
Social Status		
Family income	INCOME4R	Family income (in categories)
Education	EDUCR6	Years of education completed (6 categories)
	EDUC4	Years of education completed (4 categories)
	EDUC3	Years of education, trichotomized
Age	AGE4	Years old (categories)
Gender	SEX	Sex
Marital status	MARITAL	Married, never married, widowed, divorced
Race	RACED	White, minority
Politics	PARTYID3	Political party affiliation
Voting	VOTE08D	Voted in 2004 presidential election (yes/no)
Political views	POLVIEWS3	Liberal, moderate, conservative
Interpersonal trust	TRUSTD	Believe other people can be trusted

a. Some variables recoded.

data entry forms can be designed to facilitate this process (Exhibit 8.2). Data from such a form can be entered online, directly into a database, or first on a paper form and then typed or even scanned into a computer database. Whatever data entry method is used, the data must be checked carefully for errors—a process called **data cleaning**. Most survey research organizations now use a database management program to monitor data entry so that invalid codes can be corrected immediately. After data are entered, a computer program must be written to "define the data." A data definition program identifies the variables that are coded in each column or range of columns, attaches meaningful labels to the codes, and distinguishes values representing missing data. The procedures vary depending on the specific statistical package used.

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

2 What Are the Options for Displaying Distributions?

The first step in data analysis is usually to discover the variation in each variable of interest. How many people in the sample are married? What is their typical income? Did most of them complete high school? Graphs and frequency distributions are the two most popular display formats. Whatever format is used, the primary

Exhibit 8.2 Data Entry Procedures

OMB Control No: 6691-0001
Expiration Date: 04/30/07

Bureau of Economic Analysis
Customer Satisfaction Survey

1. Which data products do you use?	*Frequently (every week)*	*Often (every month)*	Infrequently	Rarely	Never	Don't know or not applicable
GENERAL DATA PRODUCTS	(On a scale of 1-5, please circle the appropriate answer.)					
Survey of Current Business	5	4	3	2	1	N/A
CD-ROMs	5	4	3	2	1	N/A
BEA website (www.bea.gov)	5	4	3	2	1	N/A
STAT-USA website (www.stat-usa.gov)	5	4	3	2	1	N/A
Telephone access to staff	5	4	3	2	1	N/A
E-Mail access to staff	5	4	3	2	1	N/A
INDUSTRY DATA PRODUCTS						
Gross Product by Industry	5	4	3	2	1	N/A
Input-Output Tables	5	4	3	2	1	N/A
Satellite Accounts	5	4	3	2	1	N/A
INTERNATIONAL DATA PRODUCTS						
U.S. International Transactions (Balance of Payments)	5	4	3	2	1	N/A
U.S. Exports and Imports of Private Services	5	4	3	2	1	N/A
U.S. Direct Investment Abroad	5	4	3	2	1	N/A
Foreign Direct Investment in the United States	5	4	3	2	1	N/A
U.S. International Investment Position	5	4	3	2	1	N/A
NATIONAL DATA PRODUCTS						
National Income and Product Accounts (GDP)	5	4	3	2	1	N/A
NIPA Underlying Detail Data	5	4	3	2	1	N/A
Capital Stock (Wealth) and Investment by Industry	5	4	3	2	1	N/A
REGIONAL DATA PRODUCTS						
State Personal Income	5	4	3	2	1	N/A
Local Area Personal Income	5	4	3	2	1	N/A
Gross State Product by Industry	5	4	3	2	1	N/A
RIMS II Regional Multipliers	5	4	3	2	1	N/A

Central tendency: The most common value (for variables measured at the nominal level) or the value around which cases tend to center (for a quantitative variable).

Variability: The extent to which cases are spread out through the distribution or clustered around just one value.

Skewness: The extent to which cases are clustered more at one or the other end of the distribution of a quantitative variable rather than in a symmetric pattern around its center. Skew can be positive (a right skew), with the number of cases tapering off in the positive direction, or negative (a left skew), with the number of cases tapering off in the negative direction.

concern of the analyst is to display accurately the distribution's shape; that is, to show how cases are distributed across the values of the variable.

Three features are important in describing the shape of the distribution: (1) central tendency, (2) variability, and (3) skewness (lack of symmetry). All three features can be represented in a graph or in a frequency distribution.

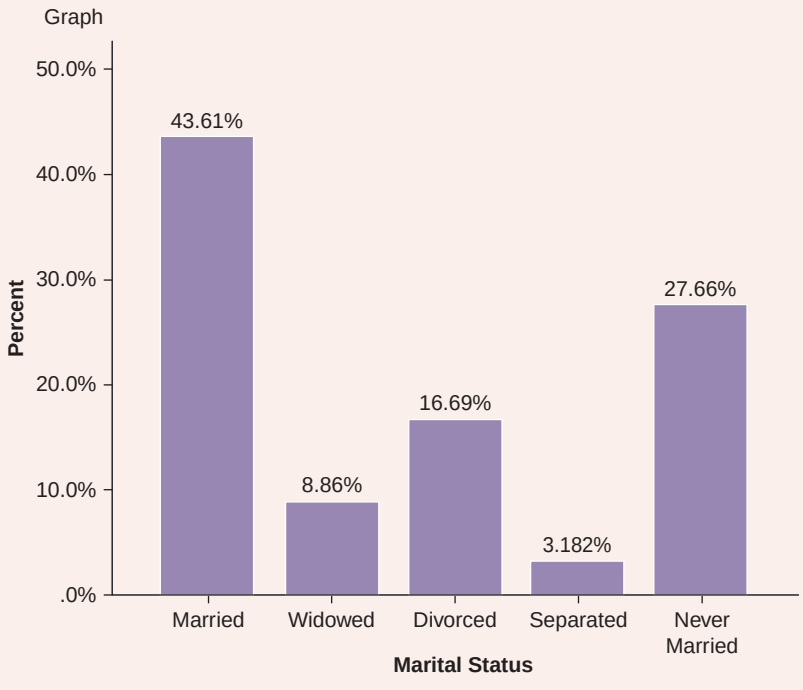
We now examine graphs and frequency distributions that illustrate the three features of shape. Several summary statistics used to measure specific aspects of central tendency and variability are presented in a separate section.

Graphs

There are many types of graphs, but the most common and most useful for the statistician are bar charts, histograms, and frequency polygons. Each has two axes, the vertical axis (the *y*-axis) and the horizontal axis (the *x*-axis), and labels to identify the variables and the values, with tick marks showing where each indicated value falls along each axis.

A **bar chart** contains solid bars separated by spaces. It is a good tool for displaying the distribution of variables measured in discrete categories (e.g., nominal variables such as religion or marital status), because such categories don't blend into each other. The bar chart of marital status in Exhibit 8.3 indicates that about half of adult Americans were married at the time of the survey. Smaller percentages

Exhibit 8.3 Bar Chart of Marital Status



were divorced, separated, widowed, or never married. The most common value in the distribution is *married.* There is a moderate amount of variability in the distribution, because the half who are not married are spread across the categories of widowed, divorced, separated, and never married. Because marital status is not a quantitative variable, the order in which the categories are presented is arbitrary, and there is no need to discuss skewness.

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

Histograms, in which the bars are adjacent, are used to display the distribution of quantitative variables that vary along a continuum that has no necessary gaps. Exhibit 8.4 shows a histogram of years of education from the 2010 GSS data. The distribution has a clump of cases centered at 12 years. The distribution is skewed because there are more cases just above the central point than below it.

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

In a **frequency polygon**, a continuous line connects the points representing the number or percentage of cases with each value. It is easy to see in the frequency polygon of years of education in Exhibit 8.5 that the most common value is 12 years (high school completion) and that this value also seems to be the center of the distribution. There is moderate variability in the distribution, with many cases having more than 12 years of education and almost one-third having completed at least 4 years of college (16 years). The distribution is highly skewed in the negative direction, with few respondents reporting less than 10 years of education.

Frequency polygon: A graphic for quantitative variables in which a continuous line connects data points representing the variable's distribution.

Exhibit 8.4 Histogram of Years of Education

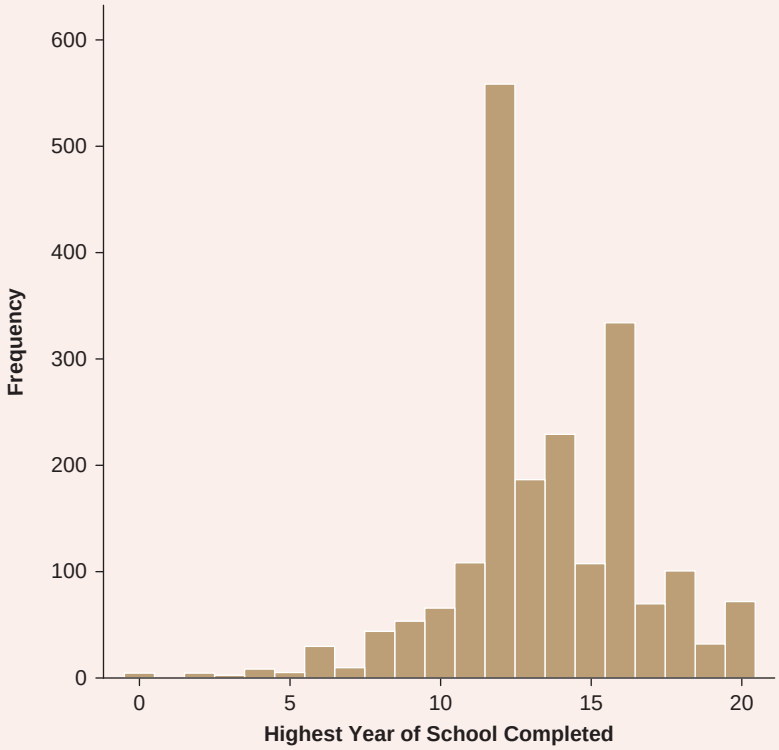
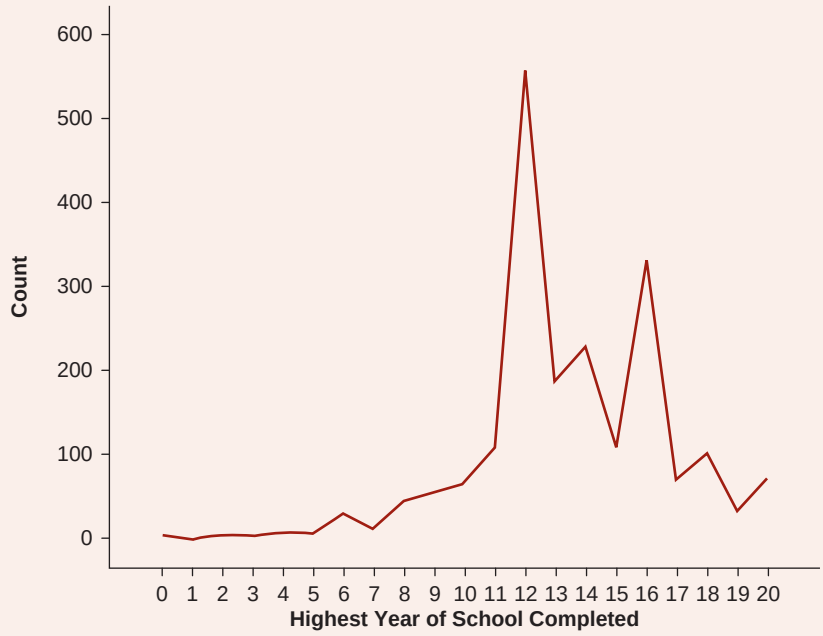


Exhibit 8.5 Frequency Polygon of Years of Education



Video Link 8.2
Watch for information on data visualization.

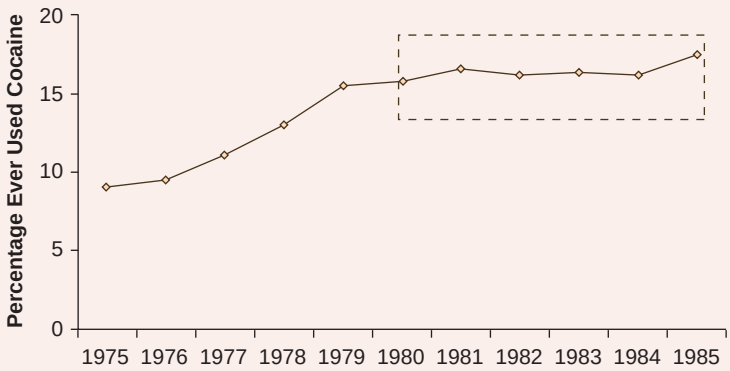
If graphs are misused, they can distort rather than display the shape of a distribution. Compare, for example, the two graphs in Exhibit 8.6. The first graph shows that high school seniors reported relatively stable rates of lifetime use of cocaine between 1980 and 1985.

The second graph, using exactly the same numbers, appeared in a 1986 *Newsweek* article on "the coke plague" (Orcutt & Turner 1993). To look at this graph, you would think that the rate of cocaine usage among high school seniors had increased dramatically during this period. But, in fact, the difference between the two graphs is due simply to changes in how the graphs were drawn. In the *Newsweek* graph, the percentage scale on the vertical axis begins at 15 rather than at 0, making what was about a 1 percentage point increase look very big indeed. In addition, omission from this 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. Finally, these numbers report "lifetime use," not current or recent use; such numbers can drop only when anyone who has used cocaine dies. The graph is, in total, grossly misleading.

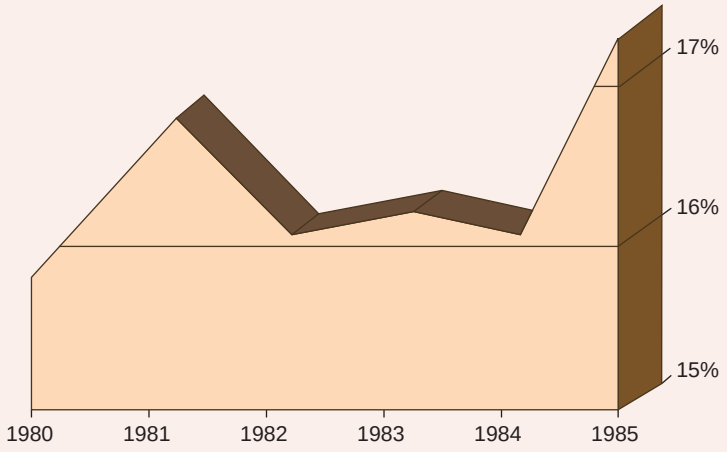
Adherence to several guidelines (Tufte 1983; Wallgren, Wallgren, Persson, Jorner, & Haaland 1996) will help you to spot such problems and to avoid them in your own work:

- Begin the graph of a quantitative variable at 0 on both axes. The difference between bars can be misleadingly exaggerated by cutting off the bottom of the vertical axis and displaying less than the full height of the bars. 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.
- Always use bars of equal width. Bars of unequal width, including pictures instead of bars, can make particular values look as if they carry more weight than their frequency warrants.
- Ensure that the two axes, usually, are of approximately equal length. Either shortening or lengthening the vertical axis will obscure or accentuate the differences in the number of cases between values.
- Avoid "chart junk"—a lot of verbiage or excessive marks, lines, lots of cross-hatching, and the like. It can confuse the reader and obscure the shape of the distribution.

Exhibit 8.6 Two Graphs of Cocaine Usage



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



B. *Newsweek*, "A Coke Plague"

Frequency Distributions

Another good way to present a univariate (one-variable) distribution is with a **frequency distribution**. A frequency distribution displays the number, **percentage** (the relative frequencies), or both corresponding to each of a variable's values. A frequency distribution will usually be labeled with a title, a stub (labels for the values), a caption, 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 number *N***) should be indicated (Exhibit 8.7).

Constructing and reading frequency distributions for variables with few values is not difficult. The frequency distribution of voting in Exhibit 8.7, for example, shows that 68.7% of the respondents eligible to vote said they voted and that 25.9% reported they did not vote. The total number of respondents to this question was 2,023, although 2,044 actually were interviewed. The rest were ineligible to vote,

Frequency distribution: 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.

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

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

Exhibit 8.7 Frequency Distribution of Voting in the 2008 Presidential Election

Value	Frequency	Valid Percentage
Voted	1390	72.4
Did not vote	530	27.6
Ineligible	103	
Don't know	12	
No answer	9	
Total	2044	100.0
N		(1920)

just refused to answer the question, said they did not know whether they had voted, or gave no answer.

When the distributions of variables with many values (for instance, age) are to be presented, the values must first be grouped. Exhibit 8.8 shows both an ungrouped and a grouped frequency distribution of age. You can see why it is so important to group the values, but we have to be sure that in doing so, we do not distort the distribution. Follow these two rules, and you'll avoid problems:

1. Categories should be logically defensible and should preserve the shape of the distribution.
2. Categories should be mutually exclusive and exhaustive so that every case is classifiable in one and only one category.

2 What Are the Options for Summarizing Distributions?

Summary statistics describe particular features of a distribution and facilitate comparison among distributions. We can, for instance, show that average income is higher in Connecticut than in Mississippi and higher in New York than in Louisiana. But if we just use one number to represent a distribution, we lose information about other aspects of the distribution's shape. For example, a measure of central tendency (such as the mean or average) would miss the point entirely for an analysis about differences in income inequality among states. A high average income could as easily be found in a state with little income inequality as in one with much income inequality; the average says nothing about the distribution of incomes. For this reason, analysts who report summary measures of central tendency usually also report a summary measure of variability or present the distributions themselves to indicate skewness.

In the News

Research in the News

GENERAL SOCIAL SURVEY SHOWS INFIDELITY ON THE RISE

Since 1972, about 12% of married men and 7% of married women have said each year that they have had sex outside their marriage. However, the lifetime rate of infidelity for men over age 60 increased from 20% in 1991 to 28% in 2006, while for women in this age group it increased from 5% to 15%. Infidelity has also increased among those under age 35: from 15% to 20% among young married men and from 12% to 15% among young married women. On the other hand, couples appear to be spending slightly more time with each other.

Source: Parker-Pope, Tara. 2008. Love, sex, and the changing landscape of infidelity. *The New York Times,* October 28:D1.

Exhibit 8.8 Grouped Versus Ungrouped Frequency Distributions

Ungrouped		Grouped	
Age	Percentage	Age	Percentage
18	0.5%	18–19	1.7%
19	1.2	20–29	16.7
20	1.2	30–39	17.8
21	1.7	40–49	18.0
22	.9	50–59	18.5
23	1.9	60–69	14.8
24	1.5	70–79	7.1
25	2.4	80–89	5.4
26	1.4		100.0%
27	2.1		(2041)
28	1.5		
29	2.2		
30	2.0		
31	2.1		
32	1.6		
33	1.8		
34	1.6		
35	2.2		
36	1.5		
37	1.9		
38	1.5		
39	1.7		
40	1.8		
41	1.9		
42	1.7		
43	2.1		
44	1.9		
45	1.8		
46	1.9		
. . .	. . .		

Encyclopedia Link 8.1
Read about when to use measures of central tendency.

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.

Mode

The **mode** is the most frequent value in a distribution. In a distribution of Americans' religious affiliations, Protestant Christian is the most frequently occurring value—the largest single group. In an age distribution of college students, 18- to 22-year-olds are by far the largest group and, therefore, the mode. One silly, but easy, way to remember the definition of the *mode* is to think of apple pie *á la mode,* which means pie with a big blob of vanilla ice cream on top. Just remember, the mode is where the big blob is—the largest collection of cases.

Mode (probability average): The most frequent value in a distribution; also termed the probability average.

The mode is also sometimes termed the **probability average**, because being the most frequent value, it is the most probable. For example, if you were to pick a case at random from the distribution of age (Exhibit 8.8), the probability of the case being in his or her 50s would be 18.5%—the most probable value in the distribution.

Bimodal: A distribution in which two nonadjacent categories have about the same number of cases and these categories have more cases than any others.

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

The mode is used much less often than the other two measures of central tendency, because it can so easily give a misleading impression of a distribution's central tendency. One problem with the mode occurs when a distribution is **bimodal**, in contrast to being **unimodal**. A bimodal distribution has two categories with a roughly equal number of cases and clearly more cases than the other categories. In this situation, there is no single mode.

Nevertheless, there are occasions when the mode is very appropriate. 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 addition, because it is the most probable value, it can be used to answer questions such as which ethnic group is most common in a given school.

Median

The **median** is the position average, or the point that divides the distribution in half (the 50th percentile). Think of the median of a highway—it divides the road exactly in two parts. To determine the median, we simply array a distribution's values in numerical order and find the value of the case that has an equal number of cases above and below it. If the median point falls between two cases (which happens if the distribution has an even number of cases), the median is defined as the average of the two middle values and is computed by adding the values of the two middle cases and dividing by 2. The median is not appropriate for variables that are measured at the nominal level; their values cannot be put in order, so there is no meaningful middle position.

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

The median in a frequency distribution is determined by identifying the value corresponding to a cumulative percentage of 50. Starting at the top of the years of education distribution in Exhibit 8.9, for example, and adding up the percentages, we find that we reach 44.2% in the 12-years category and then 69.8% in the 13- to 15-years category. The median is therefore 13 to 15.

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

Mean

The **mean** is just the arithmetic average. (Many people, you'll notice, use the word *average* a bit more generally to designate everything we've called central tendency.)

In calculating a mean, any higher numbers pull it up, and any lower numbers pull it down. Therefore, it takes into account the values of each case in a distribution—it is a weighted average. (The median, by contrast, only depends on whether the numbers are higher or lower compared to the middle, not *how* high or low.)

The mean is computed by adding up the values of all the cases and dividing the result by the total number of cases, thereby taking into account the value of each case in the distribution:

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

In algebraic notation, the equation is $X=\sum x_i/N$. For example, to calculate the mean value of 8 cases, we add the values of all the cases ($\sum x_i$ and divide by the number of cases (N):

$$(28+117+42+10+77+51+64+55)/8 = 55.5$$

Exhibit 8.9 Years of Education Completed

Years of Education	Percentage
Less than 8	5.7%
8–11	11.2
12	27.3
13–15	25.6
16	16.3
17 or more	13.8
	100.0
	(2044)

Computing the mean obviously 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 we assume that an ordinal measure can be treated as an interval (which is a fairly common practice). It makes no sense to calculate the mean of a qualitative (nominal) variable such as religion, for example. Imagine a group of four people in which there were 2 Protestants, 1 Catholic, and 1 Jew. To calculate the mean, you would need to solve the equation (Protestant + Protestant + Catholic + Jew) / 4 = *?*. Even if you decide that Protestant = 1, Catholic = 2, and Jew = 3 for data entry purposes, it still doesn't make sense to add these numbers because they don't represent quantities of religion. In general, certain statistics (such as the mean) can apply only if there is a high enough level of measurement.

Median or Mean?

Because the mean is based on adding the value of all the cases, it will be pulled in the direction of exceptionally high (or low) values. In a positively skewed distribution, the value of the mean is larger than the median—more so the more extreme the skew. For instance, in Seattle, the presence of Microsoft owner Bill Gates—possibly the world's richest person—probably pulls the mean wealth number up quite a bit. One extreme case can have a disproportionate effect on the mean.

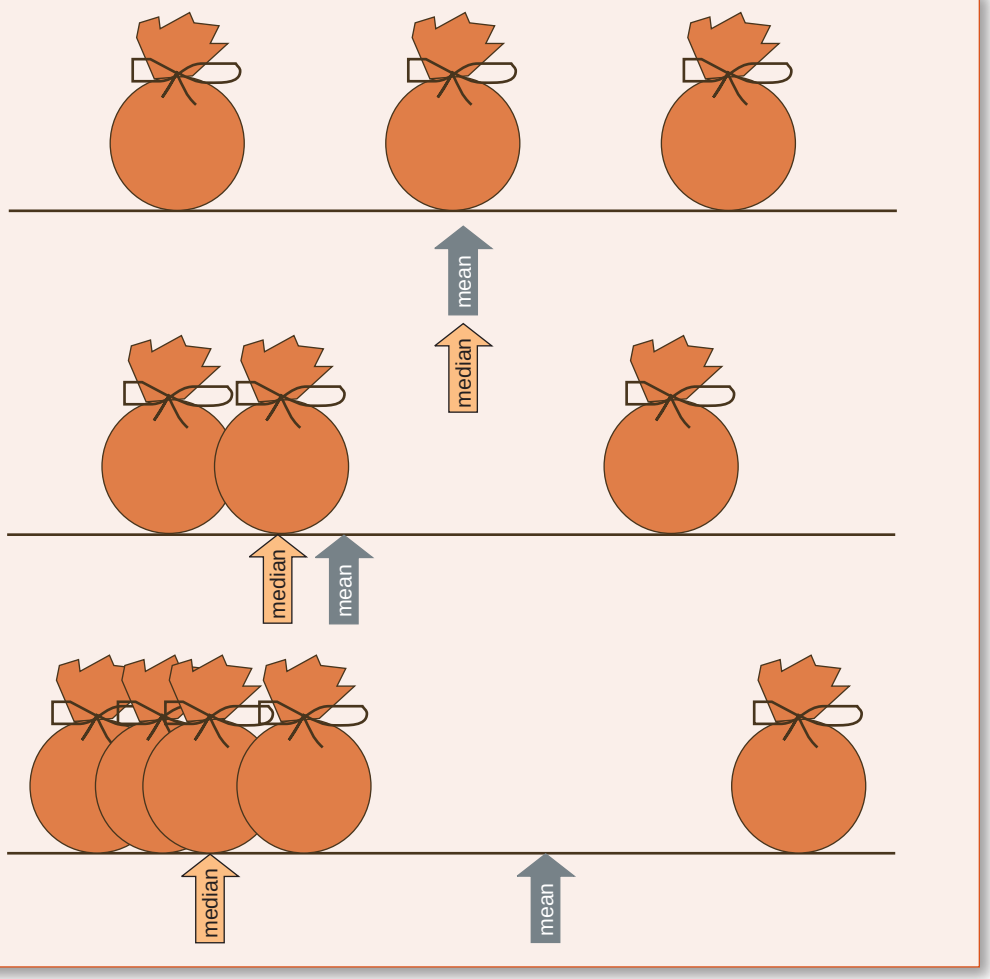
This differential impact of skewness on the median and mean is illustrated in Exhibit 8.10. On the first balance beam, the cases (bags) are spread out equally, and the median and mean are in the same location. On the second balance beam, the median corresponds to the value of the middle case, but the mean is pulled slightly upward toward the value of the one case with an unusually high value. On the third beam, the mean is clearly pulled up toward an unusual value. In some distributions the two measures will have markedly different values, and in such instances usually the median is preferred. (Income is a very common variable that is best measured by the median, for instance.)

Measures of Variation

Central tendency is only one aspect of the shape of a distribution—the most important aspect for many purposes but still just a piece of the total picture. The distribution, we have seen, also matters. It is important to

Research|Social Impact Link 8.2
Read about measures of variation.

Exhibit 8.10 The Mean as a Balance Point



know that the median household income in the United States is a bit over $50,000 a year, but if the variation in income isn't known—the fact that incomes range from zero up to hundreds of millions of dollars—we haven't really learned much. Measures of variation capture how widely and densely spread income (for instance) is. Four popular measures of variation for quantitative variables are the range, the interquartile range, the variance, and the standard deviation (which is the single most popular measure of variability). Each conveys a certain kind of information, with strengths and weaknesses. Statistical measures of variation are used infrequently with qualitative variables and are not presented here.

Range: The true upper limit in a distribution minus the true lower limit (or the highest rounded value minus the lowest rounded value, plus 1).

Range

The **range** is the simplest measure of variation, calculated as the highest value in a distribution minus the lowest value, plus 1:

Range = Highest value − Lowest value + 1

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 altered drastically by just one exceptionally high or low value—termed an **outlier**—it's not a good summary measure for most purposes.

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

Interquartile Range

The **interquartile range** avoids the problem outliers create by showing the range where most cases lie. **Quartiles** are the points in a distribution that correspond to the first 25% of the cases, the first 50% of the cases, and the first 75% of the cases. You already know how to determine the 2nd quartile, corresponding to the point in the distribution covering half of the cases—it is another name for the median. The interquartile range is the difference between the 1st quartile and the 3rd quartile (plus 1).

Interquartile range: The range in a distribution between the end of the 1st quartile and the beginning of the 3rd quartile.

Quartiles: The points in a distribution corresponding to the first 25% of the cases, the first 50% of the cases, and the first 75% of the cases.

Variance

Variance, in its statistical definition, is the average squared deviation of each case from the mean; you take each case's distance from the mean, square that number, and take the average of all such numbers. Thus, variance takes into account the amount by which each case differs from the mean. The variance is mainly useful for computing the standard deviation, which comes next in our list here. An example of how to calculate the variance, using the following formula, appears in Exhibit 8.11:

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

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

Symbol key: $\bar{X}$ = mean; N = number of cases; Σ = sum over all cases; X_i = value of case i on variable X.

The variance is used in many other statistics, although it is more conventional to measure variability with the closely related standard deviation than with the variance.

Standard Deviation

Very roughly, the **standard deviation** is the distance from the mean that covers a clear majority of cases (about two-thirds). More precisely, 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:

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

Exhibit 8.11 Calculation of the Variance

Case #	Score (X_i)	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	21	–3.27	10.69
2	30	5.73	32.83
3	15	–9.27	85.93
4	18	–6.27	39.31
5	25	0.73	0.53
6	32	7.73	59.75
7	19	–5.27	27.77
8	21	–3.27	10.69
9	23	–1.27	1.61
10	37	12.73	162.05
11	26	1.73	2.99
			434.15

Mean: $\bar{X}$ = 267/11 = 24.27

Sum of squared deviations = 434.15

Variance: σ^2 = 434.15/11 = 39.47

Variance: A statistic that measures the variability of a distribution as 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.

Normal distribution: A symmetric distribution shaped like a bell and centered around the population mean, with the number of cases tapering off in a predictable pattern on both sides of the mean.

Symbol key: $\bar{X}$ = mean; N = number of cases; Σ = sum over all cases; X_i = value of case on i variable X; $\sqrt{}$ = square root.

The standard deviation has mathematical properties that make it the preferred measure of variability in many cases, particularly when a variable is normally distributed. A graph of a **normal distribution** looks like a bell, with one "hump" in the middle, centered around the population mean, and the number of cases tapering off on both sides of the mean (Exhibit 8.12). A normal distribution is symmetric: If you were to fold the distribution in half at its center (at the population mean), the two halves would match perfectly. If a variable is normally distributed, 68% of the cases (almost exactly two-thirds) will lie between ±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.

So the standard deviation, in a single number, tells you quickly about how wide the variation is of any set of cases, or the range in which most cases will fall. It's very useful.

2 How Can We Tell Whether Two Variables Are Related?

Audio Link 8.1
Listen to an example of normal distribution.

Univariate distributions are nice, but they don't say how variables relate to each other—for instance, if religion affects education or if marital status is related to income. To establish cause, of course, one's first task is to show an association between independent and dependent variables (cause and effect). **Cross-tabulation** is a simple, easily understandable first step in such quantitative data analysis. Cross-tabulation displays the distribution of one variable within each category of another variable; it can also be termed a *bivariate distribution,* since it shows two variables at the same time. Exhibit 8.13 displays the cross-tabulation of voting by income so that we can see if the likelihood of voting increases as income goes up.

Exhibit 8.12 The Normal Distribution

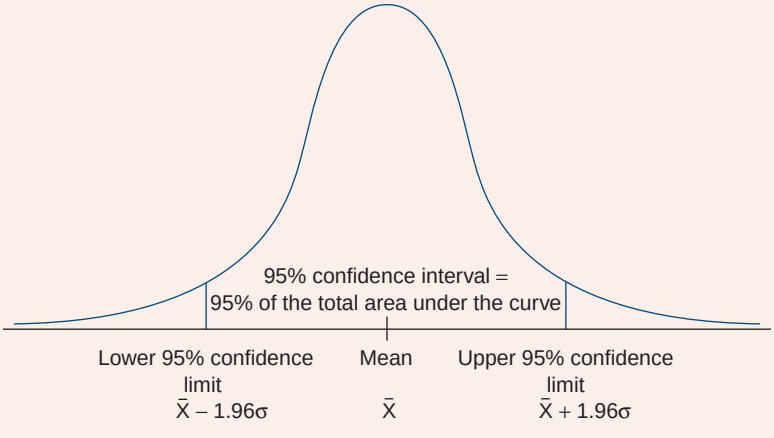


Exhibit 8.13 Cross-Tabulation of Voting in 2008 by Family Income: Cell Counts and Percentages

Voting	<$20,000	$20,000–$39,999	$40,000–$74,999	$75,000+
Cell Counts	**Family Income**			
Voted	228	266	334	404
Did not vote	187	124	109	56
Total (n)	(415)	(390)	(443)	(460)
Percentages				
Voted	55	68	75	88
Did not vote	45	32	25	12
Total	100	100	100	100

The "crosstab" table is presented first (the upper part) with frequencies and then again (the lower part) with percentages. The *cells* of the table are where row and column values intersect; for instance, the first cell is where < $20,000 meets Voted; 228 is the value. Each cell represents cases with a unique combination of values of the two variables. The independent variable is usually the column variable, listed across the top; the dependent variable, then, is usually the row variable. This format isn't necessary, but social scientists typically use it.

Cross-tabulation (crosstab): In the simplest case, a bivariate (two-variable) distribution showing the distribution of one variable for each category of another variable; can also be elaborated using three or more variables.

Reading the Table

The first (upper) table in Exhibit 8.13 shows the raw number of cases with each combination of values of voting and family income. It is hard to look at the table in this form and determine whether there is a relationship between the two variables. What we really want to know is the likelihood, for any level of income, that someone voted. So we need to convert the cell frequencies into percentages. Percentages show the likelihood per 100 (*per cent* in Latin) that something occurs. The second table, then, presents the data as percentages within the categories of the independent variable (the column variable, in this case). In other words, the cell frequencies have been converted into percentages of the column totals (the *N* in each column). For example, in Exhibit 8.13, the number of people earning less than $20,000 who voted is 228 out of 415, or 55%. Because the cell frequencies have been converted to percentages of the column totals, the numbers add up to 100 in each column but not across the rows.

Note carefully: You must *always* calculate percentages within levels of the independent variable—adding numbers down the columns in our standard format. In this example, we want to know the chance that a person with an income of less than $20,000 voted, so we calculate what percentage of those people voted. Then we *compare* that to the chance that people of other income levels voted. Calculating percentages across the table, by contrast, will not show the effect of the independent variable on voting. To repeat, *always* calculate percentages within levels of the independent variable (think: with**in** the **in**dependent variable).

To read the percentage table, compare the percentage distribution of voting/not voting across the columns. Start with the lowest income category (in the left column). Move slowly from left to right, looking at

each distribution down the columns. As income increases, you will see that the percentage who voted also increases, from 55% of those with annual incomes under $20,000 (in the first cell in the first column) up to 88% of those with incomes of $75,000 or more (the last cell in the body of the table in the first row). This result is consistent with the hypothesis: It seems that higher income is moderately associated with a greater likelihood of voting.

Now look at Exhibit 8.14, which relates gender (as the independent variable) to voting (the dependent variable). The independent variable is listed across the top, and the percentages have been calculated, correctly, down the columns with values of the independent variable. Does gender affect voting? As you look down the first column, you see that 68.6% of men voted; then, in the second column, 75.3% of women voted. Gender did, in this table, have some effect on voting. Women were more likely to vote.

Some standard practices should be followed in formatting percentage tables (crosstabs): When a table is converted to percentages, usually just the percentages in each cell should be presented, and not the number of cases in each cell. Include 100% at the bottom of each column (if the independent variable is the column variable) to indicate that the percentages add up to 100, as well as the base number (*N*) for each column (in parentheses). If the percentages add up to 99 or 101 due to rounding error, just indicate so in a footnote. As noted already, there is no requirement that the independent variable always be the column variable, although consistency within a report or paper is a must. If the independent variable is the row variable, we calculate percentages in the cells of the table on the row totals (the *N* in each row), and the percentages add up to 100 across the rows.

Exhibit 8.15 shows two different tables. The top half shows voting by education—that is, the likelihood that a person with a given level of education voted in 2008. Look first at the voting distribution for high school graduation: The percent voting has jumped to over 68%—a significant change from the percentage for grade school completers. As you move across to the numbers for some college, then college graduates, it becomes obvious that education has a major effect on a person's likelihood of voting.

Now try looking at the lower table, which is a bit more complex, since it shows several levels of the dependent variable, family income. Try to see the effect that education has on income. Among the 283 grade school graduates surveyed (the first column on the left), you can see that 52.7%—more than half—have incomes under $20,000 a year. Shifting to the high school graduates, the number in that lowest-income category has clearly fallen: The distribution has shifted some toward the higher income results. With some college, that trend continues; and for college graduates, you can see that 50.3% of them—more than half!—are making over $75,000 a year. That's more than double (50.3 to 24.0) the percent of people who only did some college.

Encyclopedia Link 8.2
Read an overview of correlation.

So, education has a powerful effect on a person's chances for making a high income—which may be why many of you are reading this book right now!

When you read research reports and journal articles, you will find that social scientists usually judge the strength of association on the basis of more statistics than just a cross-tabulation table. A **measure of association** is a descriptive statistic used to summarize the strength of an association. One measure of association in

Exhibit 8.14 Voting in 2008 by Gender

	Gender	
Voting	**Male**	**Female**
Voted	68.6%	75.3%
Did not vote	31.4	24.7
Total	100%	100%
(n)	(829)	(1,091)

Exhibit 8.15 Voting in 2008 by Education and Income by Education

Voting by Education				
	Education			
Voting	Grade School	High School	Some College	College Graduate
Voted	45.0%	68.2%	76.0%	87.6%
Did not vote	55.0	31.8	24.0	12.4
Total	100%	100%	100%	100%
(n)	(307)	(529)	(499)	(582)
Family Income by Education				
	Education			
Family Income	Grade School	High School	Some College	College Graduate
<$20,000	52.7%	26.6%	23.6%	10.1%
$20,000–$39,999	30.7	30.1	22.3	12.8
$40,000–$74,999	12.4	27.8	30.1	26.8
$75,000+	4.2	15.5	24.0	50.3
Total	100%	100%	100%	100%
(n)	(283)	(489)	(475)	(555)

cross-tabular analyses with ordinal variables is called **gamma**. The value of gamma ranges from −1 to +1. The closer a gamma value is to −1 or +1, the stronger the relationship between the two variables; a gamma of zero indicates that there is no relationship between the variables. Inferential statistics go further, addressing whether an association exists in the larger population from which the (random) sample was drawn. Even when the empirical association between two variables supports the researcher's hypothesis, it is possible that the association was just due to the vagaries of random sampling. In a crosstab, estimation of this probability can be based on the inferential statistic, **chi-square**. 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%)."

When the analyst feels reasonably confident (at least 95% confident, or $p < .05$) that an association was not due to chance, it is said that the association is statistically significant. **Statistical significance** basically means we conclude that the relationship is actually there; it's not a chance occurrence. Convention (and the desire to avoid concluding that an association exists in the population when it doesn't) dictates that the criterion be a probability of less than 5%. Statistical significance, though, doesn't equal substantive significance. That is, while the relationship is really occurring, not just happening accidentally, it may still not matter very much. It may be a minor part of what's happening.

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

Gamma: A measure of association that is sometimes used in cross-tabular analysis.

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

Statistical significance: The mathematical likelihood that an association is not due to chance, judged by a criterion the analyst sets (often that the probability is less than 5 out of 100, or $p < .05$).

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.

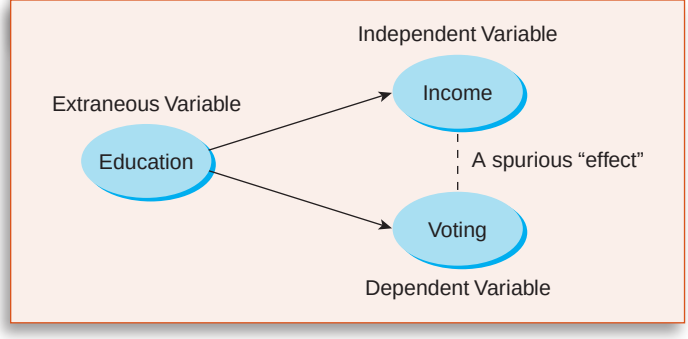
Elaboration analysis: The process of introducing a third variable into an analysis to better understand—to elaborate—the bivariate (two-variable) relationship under consideration. Additional control variables also can be introduced.

Controlling for a Third Variable

Cross-tabulation also can be used to study the relationship between three or more variables. The single most important reason for introducing a third variable is to see whether a bivariate relationship is spurious. A third, **extraneous variable**, for instance, may influence both the independent and dependent variables, creating an association between them that disappears when the extraneous variable is controlled. Ruling out possible extraneous variables helps to strengthen considerably the conclusion that the relationship between the independent and dependent variables is causal—that it is nonspurious. In general, adding variables is termed **elaboration analysis**: the process of introducing control or intervening variables into a bivariate relationship to better understand the relationship (Davis 1985; Rosenberg 1968).

For example, we have seen a positive association between incomes and the likelihood of voting; people with higher incomes are more likely to vote. But perhaps that association only exists because education influences both income and likelihood of voting; maybe when we control for education—that is, when we hold the value of education constant—we will find that there is no longer an association between income and voting. This possibility is represented by the hypothetical three-variable causal model in Exhibit 8.16, in which the arrows show that education influences both income and voting, thereby creating a relationship between the two. To test whether there is such an effect of education, we create the trivariate table in Exhibit 8.17, showing the bivariate crosstabs for various levels of education separately. This allows us to see if the income/voting relationship still exists after we hold education constant.

Exhibit 8.16 A Causal Model of a Spurious Effect



The trivariate cross-tabulation in Exhibit 8.17 shows that the relationship between voting and income is *not* spurious due to the effect of education. The association between voting and income occurs in all three subtables. So our original hypothesis—that income as a social status indicator has an effect on voting—is not weakened.

Our goal in introducing you to cross-tabulation has been to help you think about the association 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, at least to include the technique of *regression* or *correlation analysis*. These statistics have many advantages over cross-tabulation—as well as some disadvantages. You will need to take a course in social statistics to become proficient in the use of statistics based on regression and correlation.

Journal Link 8.1
Read an article where authors control for guilt.

Journal Link 8.2
Read about work assignment and career outcomes when controlling for work environment.

2 Analyzing Data Ethically: How Not to Lie With Statistics

Using statistics ethically means first and foremost being honest and open. Findings should be reported honestly, and the researcher should be open about the thinking that guided the decision to use particular

Exhibit 8.17 Voting in 2008 by Income and Education

Voting	Family Income			
	<$20,000	$20,000–$39,999	$40,000–$74,999	$75,000+
Education = Grade school				
Voted	38.0%	40.5%	55.2%	55.6%
Did not vote	62.0	59.5	44.8	44.4
Total	100%	100%	100%	100%
(n)	(137)	(79)	(29)	(9)
Education = High school				
Voted	58.4%	67.4%	65.6%	85.1%
Did not vote	41.6	32.6	34.4	14.9
Total	100%	100%	100%	100%
(n)	(125)	(138)	(131)	(74)
Education = Some college				
Voted	63.5%	80.6%	78.3%	81.5%
Did not vote	36.5	19.4	21.7	18.5
Total	100%	100%	100%	100%
(n)	(104)	(103)	(138)	(108)
Education = College graduate				
Voted	75.5%	82.6%	86.1%	92.2%
Did not vote	24.5	17.4	13.9	7.8
Total	100%	100%	100%	100%
(n)	(49)	(69)	(144)	(269)

statistics. Although this section has a mildly humorous title (after Darrell Huff's 1954 little classic, *How to Lie With Statistics)*, make no mistake about the intent: It is possible to distort social reality with statistics, and it is unethical to do so knowingly, even when the error is due more to carelessness than to deceptive intent. There are a few basic rules to keep in mind:

- Inspect the shape of any distribution for which you report summary statistics to ensure that the statistics do not mislead your readers because of an unusual degree of skewness.
- When you create graphs, be sure to consider how the axes you choose may change the distribution's apparent shape; don't deceive your readers. You have already seen that it is possible to distort the shape of a distribution by manipulating the scale of axes, clustering categories inappropriately, and the like.

Journal Link 8.3
Read about research on alcohol and partner violence.

- Whenever you need to group data in a frequency distribution or graph, inspect the ungrouped distribution and then use a grouping procedure that does not distort the distribution's basic shape.
- Test hypotheses formulated in advance of data collection as they were originally stated. When evaluating associations between variables, it becomes very tempting to search around in the data until something interesting emerges. Social scientists sometimes call this a "fishing expedition." Although it's not wrong to examine data for unanticipated relationships, inevitably some relationships between variables will appear just on the basis of chance association alone. Exploratory analyses must be labeled in research reports as such.
- Be honest about the limitations of using survey data to test causal hypotheses. Finding that a hypothesized relationship is not altered by controlling for some other variables does not establish that the relationship is causal. There is always 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 have to think about the possibilities and be cautious in our causal conclusions.

2 Conclusion

With some simple statistics (means, standard deviations, and the like), a researcher can describe social phenomena, identify relationships among them, explore the reasons for these relationships (especially through elaboration), and test hypotheses about them. Statistics—carefully constructed numbers that describe an entire population of data—are amazingly helpful in giving a simple summation of complex situations. Statistics provide a remarkably useful tool for developing our understanding of the social world, a tool that we can use both to test our ideas and to generate new ones.

Audio Link 8.2
Listen to more information about quantitative studies.

Unfortunately, to the uninitiated, the use of statistics can seem to end debate right there—one can't argue with the numbers. But you now know better. Numbers are worthless if the methods used to generate the data are not valid, and numbers can be misleading if they are not used appropriately, taking into account the type of data to which they are applied. In a very poor town with one wealthy family, the mean income may be fairly high—but grossly misleading. And even assuming valid methods and proper use of statistics, there's one more critical step, because the numbers do not speak for themselves. Ultimately, how we interpret and report statistics determines their usefulness.

Key Terms

Bar chart 158
Base number (*N*) 161
Bimodal 164
Central tendency 158
Chi-square 171
Cross-tabulation (crosstab) 168
Data cleaning 156
Descriptive statistics 155
Elaboration analysis 172
Extraneous variable 172
Frequency distribution 161
Frequency polygon 159
Gamma 171
Histogram 159
Inferential statistics 155
Interquartile range 167
Mean 164
Measure of association 170
Median 164
Mode (probability average) 164
Normal distribution 168
Outlier 167
Percentage 161
Quantitative data analysis 154
Quartiles 167

Highlights

- Data entry options include direct collection of data through a computer, use of scannable data entry forms, and use of data entry software. 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 the mode, mean, and median vary 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.
- Cell frequencies in cross-tabulation should normally be converted to 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 relationships.
- 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.
- Honesty and openness are the key ethical principles that should guide data summaries.

STUDENT STUDY SITE

The Student Study Site, available at **www.sagepub.com/chambliss4e,** includes useful study materials including web exercises with accompanying links, eFlashcards, videos, audio resources, journal articles, and encyclopedia articles, many of which are represented by the media links throughout the text. The site also features Interactive Exercises—represented by the green icon here—to help you understand the concepts in this book.

Exercises

Discussing Research

1. We presented in this chapter several examples of bivariate and trivariate cross-tabulations involving voting in the 2008 presidential election. What additional influences would you recommend examining to explain voting in elections? Suggest some additional independent variables for bivariate analyses with voting, as well as several additional control variables to be used in three-variable crosstabs.
2. When should we control just to be honest? Should social researchers be expected to investigate alternative explanations for their findings? Should they be expected to check to see if the associations they find occur for different subgroups in their samples? Justify your answers.

Finding Research

1. Do a web search for information on a social science subject in which you are interested. How much of the information 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 on which you relied.

2. The National Bureau of Economic Research provides many graphs and numeric tables about current economic conditions (www.nber.org/). Review some of these presentations. Which displays are most effective in conveying information? Summarize what you can learn from this site about economic conditions.

Critiquing Research

1. Become a media critic. For the next week, scan a newspaper or some magazines for statistics. How many articles can you find that use 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?

Doing Research

1. Create frequency distributions from lists in U.S. Census Bureau reports on the characteristics of states, cities, or counties or any similar listing of data for at least 100 cases (http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml0). You will have to decide on a grouping scheme for the distribution of variables, such as average age and population size; how to deal with outliers in the frequency distribution; and how to categorize qualitative variables, such as the predominant occupation. Decide what summary statistics to use for each variable. How well were the features of each distribution represented by the summary statistics? Describe the shape of each distribution. Propose a hypothesis involving two of these variables and develop a crosstab to evaluate the support for this hypothesis. Describe each relationship in terms of the four aspects of an association after converting cell frequencies to percentages in each table within the categories of the independent variable. Does the hypothesis appear to have been supported?

2. Exhibit 8.18 is a three-variable table created with survey data from 355 employees hired during the previous year at a large telecommunications company. Employees were asked if the presence of on-site child care at the company's offices was important in their decision to join the company.

Reading the table:

   a. Does gender affect attitudes?
   b. Does marital status affect attitudes?
   c. Which of the preceding two variables matters more?
   d. Does being married affect men's attitudes more than women's?

3. If you have access to the SPSS statistical program, you can analyze data contained in the 2010 General Social Survey (GSS) file on the Study Site for this text. See Appendix C for instructions on using SPSS.

   Develop a description of the basic social and demographic characteristics of the U.S. population in 2010. Examine each characteristic with three statistical techniques: a graph, a frequency distribution, and a measure of central tendency (and a measure of variation, if appropriate).

   a. From the menu, select Graphs and then Legacy Dialogs and Bar. Select Simple Define [Marital—Category Axis]. Bars represent % of cases. Select Options (do not display groups defined by missing values). Finally, select Histogram for each of the variables [EDUC, EARNRS, TVHOURS, ATTEND].
   b. Describe the distribution of each variable.
   c. Generate frequency distributions and descriptive statistics for these variables. From the menu, select Analyze/Descriptive Statistics/Frequencies. From the Frequencies window, set MARITAL, EDUC, EARNRS, TVHOURS, ATTEND. For the Statistics, choose the mean, median, range, and standard deviation.
   d. Collapse the categories for each distribution. Be sure to adhere to the guidelines given in the section "Grouped Data." Does the general shape of any of the distributions change as a result of changing the categories?

Exhibit 8.18 Is Child Care Important? By Gender and Marital Status

	MEN		WOMEN	
	Single	Married	Single	Married
Not important	54%	48%	33%	12%
Somewhat important	24%	30%	45%	31%
Very important	22%	22%	22%	57%
	100%	100%	100%	100%
n =	(125)	(218)	(51)	(161)

e. Which statistics are appropriate to summarize the central tendency and variation of each variable? Do the values of any of these statistics surprise you?

4. Try describing relationships with support for capital punishment by using graphs. Select two relationships you identified in previous exercises and represent them in graphic form. Try drawing the graphs on lined paper (graph paper is preferable).

Ethics Questions

1. Review the frequency distributions and graphs in this chapter. Change one of these data displays so that you are "lying with statistics." (You might consider using the graphic technique discussed by Orcutt & Turner, 1993.)
2. Consider the relationship between voting and income that is presented in Exhibit 8.13. What third variables do you think should be controlled in the analysis to understand better the basis for this relationship? How might social policies be affected by finding out that this relationship was due to differences in neighborhood of residence rather than to income itself?