CHAPTER 3

Descriptive Statistics

We've seen that statistical methods are *descriptive* or *inferential*. The purpose of descriptive statistics is to summarize data, to make it easier to assimilate the information. This chapter presents basic methods of descriptive statistics.

We first present tables and graphs that describe the data by showing the number of times various outcomes occurred. Quantitative variables also have two key features to describe numerically:

- The ***center*** of the data—a typical observation
- The ***variability*** of the data—the spread around the center

We'll learn how to describe quantitative data with statistics that summarize the center, statistics that summarize the variability, and finally with statistics that specify certain positions in the data set that summarize both center and variability.

3.1 DESCRIBING DATA WITH TABLES AND GRAPHS

Tables and graphs are useful for all types of data. We'll begin with categorical variables.

Relative Frequencies: Categorical Data

For categorical data, we list the categories and show the frequency (the number of observations) in each category. To make is easier to compare different categories, we also report proportions or percentages, also called ***relative frequencies***.

Relative Frequency

The ***relative frequency*** for a category is the ***proportion*** or ***percentage*** of the observations that fall in that category.

The *proportion* equals the number of observations in a category divided by the total number of observations. It is a number between 0 and 1 that expresses the share of the observations in that category. The *percentage* is the proportion multiplied by 100.

EXAMPLE 3.1 Household Structure in the U.S.

Table 3.1 lists the different types of households in the United States in 2005. Of 111.1 million households, for example, 24.1 million were a married couple with children. The proportion $24.1/111.1 = 0.22$ were a married couple with children.

TABLE 3.1: U.S. Household Structure, 2005

Type of Family	Number (millions)	Proportion	Percentage
Married couple with children	24.1	0.22	22
Married couple, no children	31.1	0.28	28
Single householder, no spouse	19.1	0.17	17
Living alone	30.1	0.27	27
Other households	6.7	0.06	6
Total	111.1	1.00	100

Source: U.S. Census Bureau, *2005 American Community Survey*, Tables B11001, C11003.

A percentage is the proportion multiplied by 100. That is, the decimal place is moved two positions to the right. For example, since 0.22 is the proportion of families that are married couples with children, the percentage is $100(0.22) = 22\%$. Table 3.1 shows the proportions and percentages for all the categories. ■

The sum of the proportions equals 1.00. The sum of the percentages equals 100. (In practice, the values may sum to a slightly different number, such as 99.9 or 100.1, because of rounding.)

It is sufficient in such a table to report the percentages (or proportions) and the total sample size, since each frequency equals the corresponding proportion multiplied by the total sample size. For instance, the frequency of married couples with children equals $0.22(111.1) = 24$ million. When presenting the percentages but not the frequencies, always also include the total sample size.

Frequency Distributions and Bar Graphs: Categorical Data

Table 3.1 lists the categories for household structure and the number of households of each type. Such a listing is called a *frequency distribution*.

Frequency Distribution

A ***frequency distribution*** is a listing of possible values for a variable, together with the number of observations at each value. A corresponding ***relative frequency distribution*** lists the possible values together with their proportions or percentages.

To construct a frequency distribution for a categorical variable, list the categories and count the number of observations in each.

To more easily get a feel for the data, it's helpful to look at a graph of the relative frequency distribution. A ***bar graph*** has a rectangular bar drawn over each category. The height of the bar shows the relative frequency in that category. Figure 3.1 is a bar graph for the data in Table 3.1. The bars are separated to emphasize that the variable is categorical rather than quantitative. Since household structure is a nominal variable, there is no particular natural order for the bars. The order of presentation for an ordinal variable is the natural ordering of the categories.

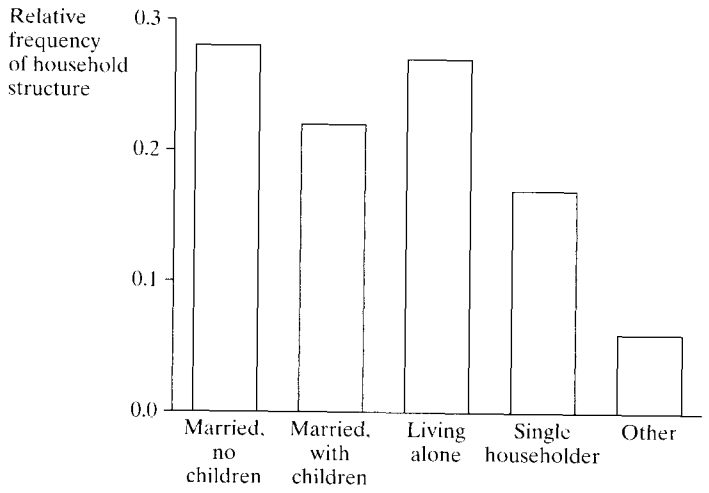

FIGURE 3.1: Relative Frequency of U.S. Household Structure Types, 2005

Another type of graph, the *pie chart*, is a circle having a "slice of the pie" for each category. The size of a slice represents the percentage of observations in the category. The bar graph is more precise than the pie chart for visual comparison of categories with similar relative frequencies.

Frequency Distributions: Quantitative Data

Frequency distributions and graphs also are useful for quantitative variables. The next example illustrates.

EXAMPLE 3.2 Statewide Violent Crime Rates

Table 3.2 lists all 50 states in the United States and their 2005 violent crime rates. This rate measures the number of violent crimes in that state in 2005 per 10,000 population. For instance, if a state had 12,000 violent crimes and a population size of 2,300,000, its violent crime rate was $(12{,}000/2{,}300{,}000) \times 10{,}000 = 52$. It is difficult to learn much by simply reading through the violent crime rates. Tables, graphs, and numerical measures help us more fully absorb the information in these data.

First, we can summarize the data with a frequency distribution. To do this, we divide the measurement scale for violent crime rate into a set of intervals and count the number of observations in each interval. Here, we use the intervals {0–11, 12–23, 24–35, 36–47, 48–59, 60–71, 72–83}. The values Table 3.2 reports were rounded, so for example the interval 12-23 represents values between 11.5 and 23.5. Counting the number of states with violent crime rates in each interval, we get the frequency distribution shown in Table 3.3. We see that considerable variability exists in the violent crime rates.

Table 3.3 also shows the relative frequencies, using proportions and percentages. For example, $3/50 = 0.06$ is the proportion for the interval 0–11, and $100(0.06) = 6$ is the percentage. As with any summary method, we lose some information as the cost of achieving some clarity. The frequency distribution does not identify which states have low or high violent crime rates, nor are the exact violent crime rates known. ■

The intervals of values in frequency distributions are usually of equal width. The width equals 12 in Table 3.3. The intervals should include all possible values of the

TABLE 3.2: List of States with Violent Crime Rates Measured as Number of Violent Crimes per 10,000 Population

Alabama	43	Louisiana	65	Ohio	33
Alaska	59	Maine	11	Oklahoma	51
Arizona	51	Maryland	70	Oregon	30
Arkansas	46	Massachusetts	47	Pennsylvania	40
California	58	Michigan	51	Rhode Island	29
Colorado	34	Minnesota	26	South Carolina	79
Connecticut	31	Mississippi	33	South Dakota	17
Delaware	66	Missouri	47	Tennessee	69
Florida	73	Montana	36	Texas	55
Georgia	45	Nebraska	29	Utah	25
Hawaii	27	Nevada	61	Vermont	11
Idaho	24	New Hampshire	15	Virginia	28
Illinois	56	New Jersey	37	Washington	35
Indiana	35	New Mexico	66	West Virginia	26
Iowa	27	New York	46	Wisconsin	22
Kansas	40	North Carolina	46	Wyoming	26
Kentucky	26	North Dakota	8		

TABLE 3.3: Frequency Distribution and Relative Frequency Distribution for Violent Crime Rates

Violent Crime Rate	Frequency	Relative Frequency	Percentage
0–11	3	0.06	6
12–23	3	0.06	6
24–35	18	0.36	36
36–47	11	0.22	22
48–59	7	0.14	14
60–71	6	0.12	12
72–83	2	0.04	4
Total	50	1.00	100.0

variable. In addition, any possible value must fit into one and only one interval; that is, they should be ***mutually exclusive***.

Histograms

A graph of a relative frequency distribution for a quantitative variable is called a ***histogram***. Each interval has a bar over it, with height representing the number of observations in that interval. Figure 3.2 is a histogram for the violent crime rates.

Choosing intervals for frequency distributions and histograms is primarily a matter of common sense. If too few intervals are used, too much information is lost. For example, Figure 3.3 is a histogram of violent crime rates using the intervals 0–29, 30–59, 60–89. This is too crude to be very informative. If too many intervals are used, they are so narrow that the information presented is difficult to digest, and the histogram may be irregular and the overall pattern of the results may be obscured. Ideally, two observations in the same interval should be similar in a practical sense. To summarize annual income, for example, if a difference of $5000 in income is not considered practically important, but a difference of $15,000 is notable, we might

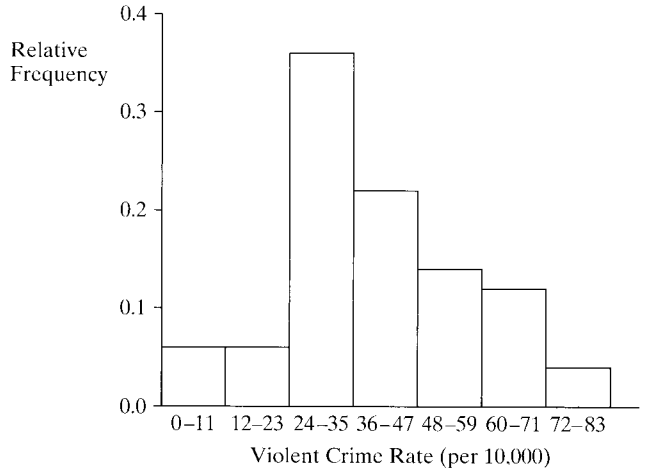

FIGURE 3.2: Histogram of Relative Frequencies for Statewide Violent Crime Rates

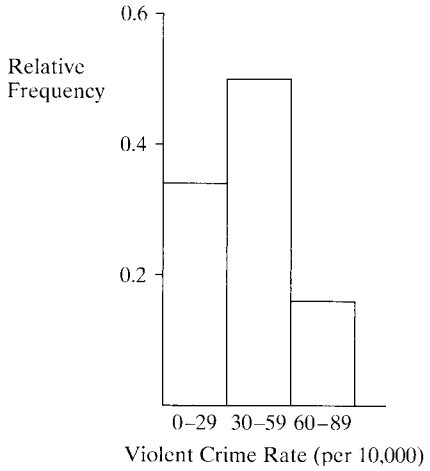

FIGURE 3.3: Histogram of Relative Frequencies for Violent Crime Rates, Using Too Few Intervals

choose intervals of width less than \$15,000, such as \$0–\$9999, \$10,000–\$19,999, \$20,000–\$29,999, and so forth. Statistical software can automatically choose intervals for us and construct frequency distributions and histograms.

For a discrete variable with relatively few values, a histogram has a separate bar for each possible value. For a continuous variable or a discrete variable with many possible values, you need to divide the possible values into intervals, as we did with the violent crime rates.

Stem-and-Leaf Plots

Figure 3.4 shows an alternative graphical representation of the violent crime rate data. This figure, called a ***stem-and-leaf plot***, represents each observation by its leading digit(s) (the *stem*) and by its final digit (the *leaf*). Each stem is a number to the left of the vertical bar and a leaf is a number to the right of it. For instance, on the second line, the stem of 1 and the leaves of 1, 1, 5, and 7 represent the violent crime rates 11, 11, 15, 17. The plot arranges the leaves in order on each line, from smallest to largest.

Stem	Leaf
0	8
1	1 1 5 7
2	2 4 5 6 6 6 6 7 7 8 9 9
3	0 1 3 3 4 5 5 6 7
4	0 0 3 5 6 6 6 7 7
5	1 1 1 5 6 8 9
6	1 5 6 6 9
7	0 3 9

FIGURE 3.4: Stem-and-Leaf Plot for Violent Crime Rate Data in Table 3.2

A stem-and-leaf plot conveys similar information as a histogram. Turned on its side, it has the same shape as the histogram. In fact, since the stem-and-leaf plot shows each observation, it displays information that is lost with a histogram. From Figure 3.4, the largest violent crime rate was 79 and the smallest was 8 (shown as 08 with a stem of 0 and leaf of 8). It is not possible to determine these exact values from the histogram in Figure 3.2.

Stem-and-leaf plots are useful for quick portrayals of small data sets. As the sample size increases, you can accommodate the increase in leaves by splitting the stems. For instance, you can list each stem twice, putting leaves of 0 to 4 on one line and leaves of 5 to 9 on another. When a number has several digits, it is simplest for graphical portrayal to drop the last digit or two. For instance, for a stem-and-leaf plot of annual income in thousands of dollars, a value of $27.1 thousand has a stem of 2 and a leaf of 7 and a value of $106.4 thousand has a stem of 10 and leaf of 6.

Comparing Groups

Many studies compare different groups on some variable. Relative frequency distributions, histograms, and stem-and-leaf plots are useful for making comparisons.

EXAMPLE 3.3 Comparing Canadian and U.S. Murder Rates

Stem-and-leaf plots can provide visual comparisons of two small samples on a quantitative variable. For ease of comparison, the results are plotted "back to back." Each plot uses the same stem, with leaves for one sample to its left and leaves for the other sample to its right. To illustrate, Figure 3.5 shows back-to-back stem and leaf plots of recent murder rates (measured as the number of murders per 100,000 population) for the 50 states in the U.S. and for the provinces of Canada. From this figure, it is clear that the murder rates tended to be much lower in Canada, varying between 0.7 (Prince Edward Island) and 2.9 (Manitoba) whereas those in the U.S. varied between 1.6 (Maine) and 20.3 (Louisiana). ■

Population Distribution and Sample Data Distribution

Frequency distributions and histograms apply both to a population and to samples from that population. The first type is called the ***population distribution***, and the second type is called a ***sample data distribution***. In a sense, the sample data distribution is a blurry photo of the population distribution. As the sample size increases, the sample proportion in any interval gets closer to the true population proportion. Thus, the sample data distribution looks more like the population distribution.

Canada	Stem	United States
7	0	
3 2 1	1	6 7
9 7 6 3 2 0	2	0 3 9
	3	0 1 4 4 4 6 8 9 9 9
	4	4 6
	5	0 2 3 8
	6	0 3 4 6 8 9
	7	5
	8	0 3 4 6 9
	9	0 8
	10	2 2 3 4
	11	3 3 4 4 6 9
	12	7
	13	1 3 5
	14	
	15	
	16	
	17	
	18	
	19	
	20	3

FIGURE 3.5: Back-to-Back Stem-and-Leaf Plots of Murder Rates from U.S. and Canada. Both share the same stems, with Canada leafs to the left and U.S. leafs to the right.

For a continuous variable, imagine the sample size increasing indefinitely, with the number of intervals simultaneously increasing, so their width narrows. Then, the shape of the sample histogram gradually approaches a smooth curve. This text uses such curves to represent population distributions. Figure 3.6 shows two sample histograms, one based on a sample of size 100 and the second based on a sample of size 500, and also a smooth curve representing the population distribution. Even if a variable is discrete, a smooth curve often approximates well the population distribution, especially when the number of possible values of the variable is large.

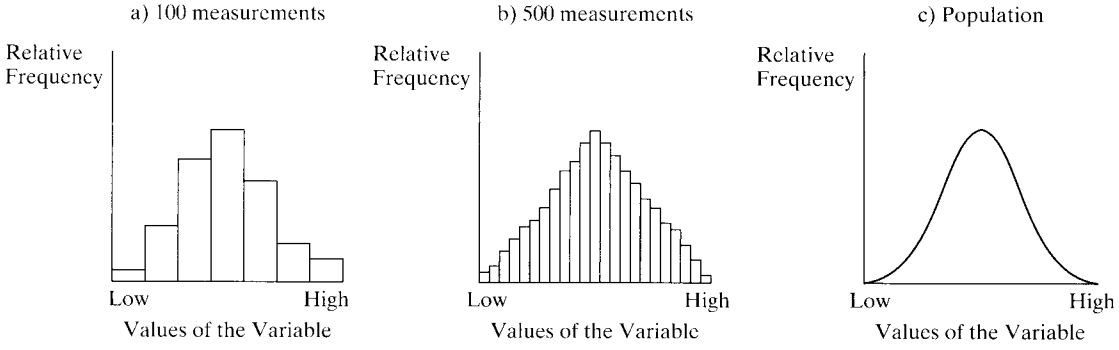

FIGURE 3.6: Histograms for a Continuous Variable. We use smooth curves to represent population distributions for continuous variables.

The Shape of a Distribution

One way to summarize a sample or a population distribution is to describe its shape. A group for which the distribution is bell-shaped is fundamentally different from

a group for which the distribution is U-shaped, for example. See Figure 3.7. In the U-shaped distribution, the highest points (representing the largest frequencies) are at the lowest and highest scores, whereas in the bell-shaped distribution, the highest point is near the middle value. A U-shaped distribution indicates a polarization on the variable between two sets of subjects. A bell-shaped distribution indicates that most subjects tend to fall near a central value.

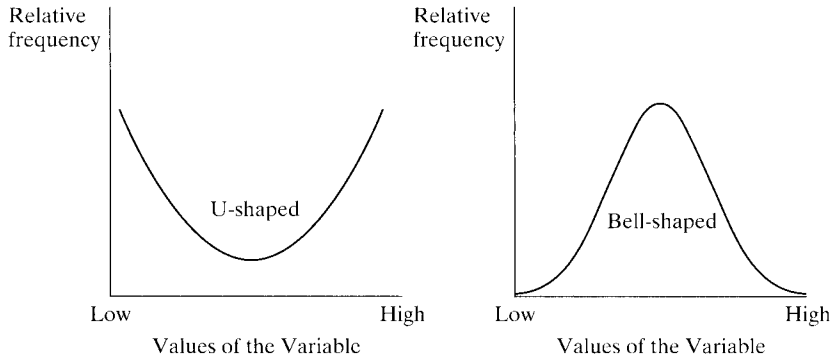

FIGURE 3.7: U-Shaped and Bell-Shaped Frequency Distributions

The distributions in Figure 3.7 are ***symmetric***: The side of the distribution below a central value is a mirror image of the side above that central value. Most distributions encountered in the social sciences are not symmetric. Figure 3.8 illustrates. The parts of the curve for the lowest values and the highest values are called the ***tails*** of the distribution. Often, as in Figure 3.8, one tail is much longer than the other. A distribution is said to be ***skewed to the right*** or ***skewed to the left***, according to which tail is longer.

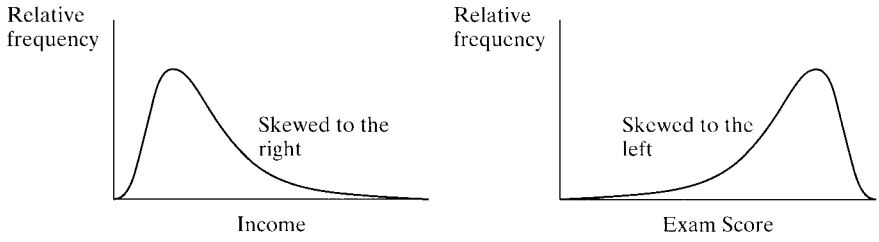

FIGURE 3.8: Skewed Frequency Distributions. The longer tail indicates the direction of skew.

To compare frequency distributions or histograms for two groups, you can give verbal descriptions using characteristics such as skew. It is also helpful to make numerical comparisons such as, "On the average, the murder rate for U.S. states is 5.4 higher than the murder rate for Canadian provinces." We now turn our attention to numerical descriptive statistics.

3.2 DESCRIBING THE CENTER OF THE DATA

This section presents statistics that describe the center of a frequency distribution for a quantitative variable. The statistics show what a *typical* observation is like.