

Unit 3

More Summarizing Data

Objectives:

- To construct and interpret a bar chart and a pie chart for qualitative data
- To construct and interpret a frequency distribution and a histogram for quantitative data

Thus far, our discussion of tabular and graphical displays for one variable has focused exclusively on quantitative variables. With a qualitative variable that does not involve any meaningful values, a five-number is not possible, nor are a dot plot, stem-and-leaf display and box plot. However, we can construct a frequency distribution. A *frequency distribution* is a list of distinct values or categories together with corresponding frequencies. A *raw frequency* is a count of the number of observations corresponding to a particular value or category; we can convert a raw frequency to a *relative frequency* by dividing by the total number of observations to obtain a fraction, proportion, or percentage.

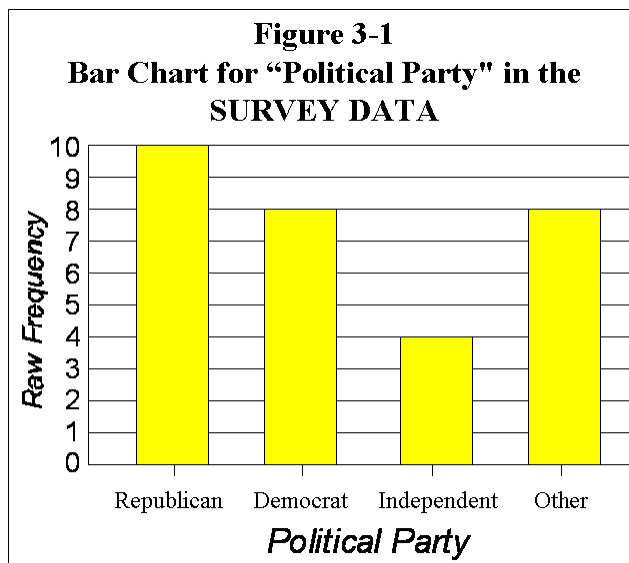
The variable "Political Party" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, to is a qualitative-nominal variable measured by using four categories: Republican, Democrat, Other, and Independent. From the SURVEY DATA, complete the second column of Table 3-1 by entering the raw frequencies, and verify that these raw frequencies sum to 30, as they should. Then, complete the third column of Table 3-1 by entering the relative frequencies as percentages, and verify that these percentages sum to 100%, as they always must (plus or minus some possible rounding error). This frequency distribution should be the same as that in the corresponding table displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

The distribution for a qualitative variable can be displayed with a *bar chart*, which is constructed by labeling a horizontal axis with the possible categories of a qualitative variable, and using the heights of bars extending from the horizontal axis to represent corresponding frequencies. Figure 3-1 is a bar chart for "Political Party," constructed from the relative frequencies in Table 3-1. An alternative to a bar chart is the *pie chart*, where a circle representing a "pie" is cut into slices with the size of each slice corresponding to a relative frequency. Figure 3-2 is a pie chart for the variable "Political Party." The raw and relative frequencies displayed beside the slices of the pie are taken from Table 3-1.

We can also construct a frequency distribution for a quantitative variable. Suppose we want a frequency distribution for the variable "Number of Children" in the SURVEY DATA. From the dot plot constructed for "Number of Children" in Figure 2-1, it is very easy to construct a frequency distribution for "Number of Children." Complete the columns for raw and relative frequency in Table 3-2. Note that the table contains two additional columns for cumulative frequencies, which are sometimes useful with a quantitative

Table 3-1
Frequency Distribution for
“Political Party” in the
SURVEY DATA

<i>Political Party</i>	<i>Raw Frequency</i>	<i>Relative Frequency</i>
Republican		
Democrat		
Independent		
Other		
	30	100.0%



variable. A *cumulative raw frequency* is a count of the number of all observations up to and including a given value; similarly, a *cumulative relative frequency* is the fraction, proportion, or percentage of observations up to and including a given value. Complete the fourth and fifth columns of Table 3-2. This frequency distribution should be the same as that in the corresponding table displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

When constructing a frequency distribution for a quantitative variable, it is not always feasible to list all of the individual observed values. Suppose we want a frequency distribution for the variable "Yearly Income" in the SURVEY DATA. From the dot plot constructed for "Yearly Income" in Figure 2-3, we see that there are many distinct values with relatively little repetition. Listing individual values, most of which occur only once or twice, would result in a very large, but not very useful, table. Instead, we shall define categories, often called *classes*. While we can give no universal rules for defining classes, we can state a few generally accepted guidelines to follow unless there is some strong reason not to do so. Usually, the number of classes should be between 5 and 15, and each class should be of equal length. The number of classes should be selected so that the data is well represented without losing too much information. A dot plot, such as the one for "Yearly Income," can be a very helpful in deciding how to define classes. We shall choose 6 classes for "Yearly Income," which is measured in thousands of dollars: "Above 20 to 30," "Above 30 to 40," ..., "Above 70 to 80."

Having defined our classes, we could construct the frequency distribution from the raw data, but it is generally easier to work from a dot plot or stem-and-leaf display. Make use of either the dot plot in Figure 2-3 or the stem-and-leaf display in Figure 2-4 to complete Table 3-3. This frequency distribution should be the same as that in the corresponding table displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

A *histogram* provides a graphical display of a frequency distribution. It is constructed by labeling a horizontal axis with possible values of a quantitative variable, and labeling a vertical axis with raw or relative frequencies. The heights of bars extending from the horizontal axis are proportional to the corresponding frequencies. Figure 3-3 is a histogram for the frequency distribution of "Yearly Income," and was constructed from the

relative frequencies in Table 3-3. The bars touch each other in Figure 3-3 to emphasize that each bar represents an entire range of values. Figure 3-4 is a histogram a histogram for the frequency distribution for "Number of Children" in the SURVEY DATA, and was constructed from the raw frequencies in Table 3-2. The bars do not touch each other in Figure 3-4 to emphasize that each bar represents only a single integer; however, it would not

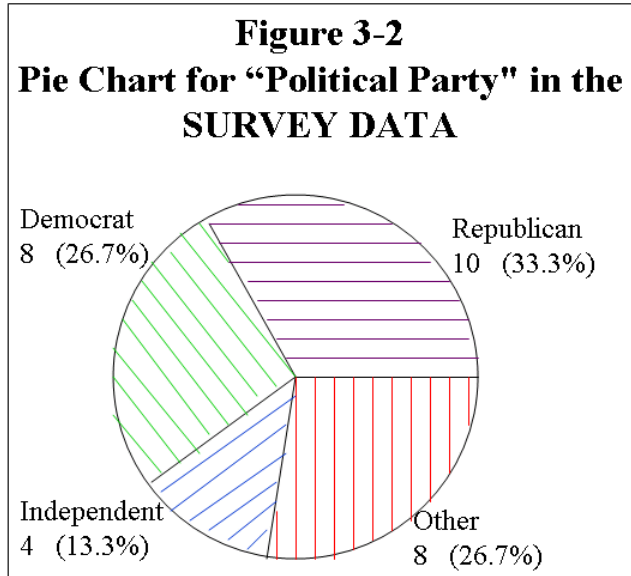


Table 3-2
Frequency Distribution for "Number of Children" in the SURVEY DATA

<i>Number of Children</i>	<i>Raw Frequency</i>	<i>Relative Frequency</i>	<i>Cumulative Raw Frequency</i>	<i>Cumulative Relative Frequency</i>
0	2	6.67%	2	6.67%
1				
2				
3				
4				
5				
6				
7				
8				
9				
<i>Totals</i>				

be incorrect to have the bars touch each other by making the width of each bar extend to one-half unit on either side of the integer it represents. The histogram in Figure 3-4 gives us a picture of the distribution of "Number of Children" virtually identical to the picture provided by the dot plot of Figure 2-1. However, the histogram in Figure 3-3 provides us with a much better picture of the distribution of "Yearly Income" than does the dot plot of Figure 2-3.

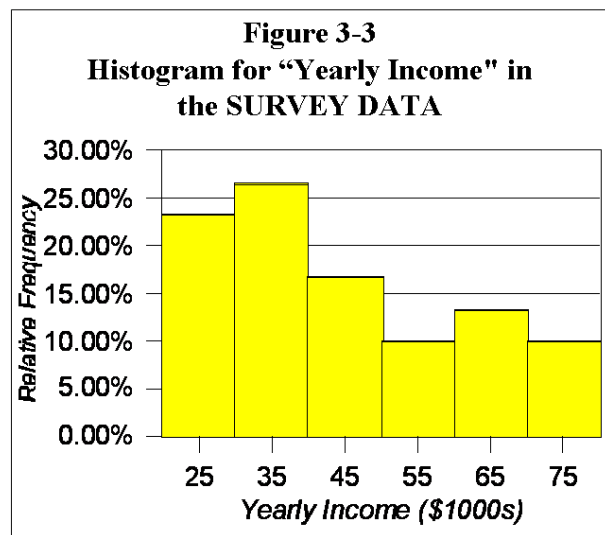
A histogram can give us the same information about the distribution of a quantitative variable as a box plot can. By looking at a histogram, we can get a sense about where the center of the distribution is, and we can also see how the data are dispersed. From the histogram in Figure 3-4, we can see that 13 out of the 30 values of the variable "Number of Children" are less than or equal to 2, which gives us the idea that the distribution is centered between 2 and 3; we also observe that 2 children, 3 children, and 4 children each occur more than any other value, and the other values are as small as 0 and as large as 9. From the histogram in Figure 3-3, we can see that roughly half of the values of the variable "Yearly Income" are less than or equal to 40 thousand dollars, which gives us the idea that the distribution is centered close to 40 thousand dollars; we also observe that incomes appear more frequently in the classes of 20 to 30 thousand dollars, 30 to 40 thousand dollars, and 40 to 50 thousand dollars than in the classes of 50 to 60 thousand dollars, 60 to 70 thousand dollars, and 70 to 80 thousand dollars.

We can also obtain some information about the shape of the distribution from a histogram. Figures 3-5a and 3-5b are histograms displaying two different symmetric shapes that a distribution might have. The histogram of Figure 3-5a displays a distribution with the same characteristics as the distribution displayed by the box plot of Type A orange weights in Figure 2-8; this type of symmetric distribution was called a *uniform* distribution, since the observations are uniformly distributed across a range. The histogram of Figure 3-5b displays a distribution with the same characteristics as the distribution displayed by the box plot of Type B orange weights in Figure 2-8; this type of symmetric distribution can be called a *bell-shaped* distribution, since values in the middle of the range occur with high frequency while the values toward either end of the range occur with low frequency.

Figures 3-5c and 3-5d are histograms displaying two different skewed shapes that a distribution might have. The histogram of Figure 3-5c displays a distribution with the same characteristics as the distribution displayed by the box plot of Type C orange weights in Figure 2-8; this type of distribution was called a *positively skewed* distribution, since the upper (positive) half of the observations are more widely dispersed than the lower half of observations. The histogram of Figure 3-5d displays a distribution with the same characteristics as the distribution displayed by the box plot of Type D orange weights in Figure 2-8; this type of

Table 3-3
Frequency Distribution for "Yearly Income" in the SURVEY DATA

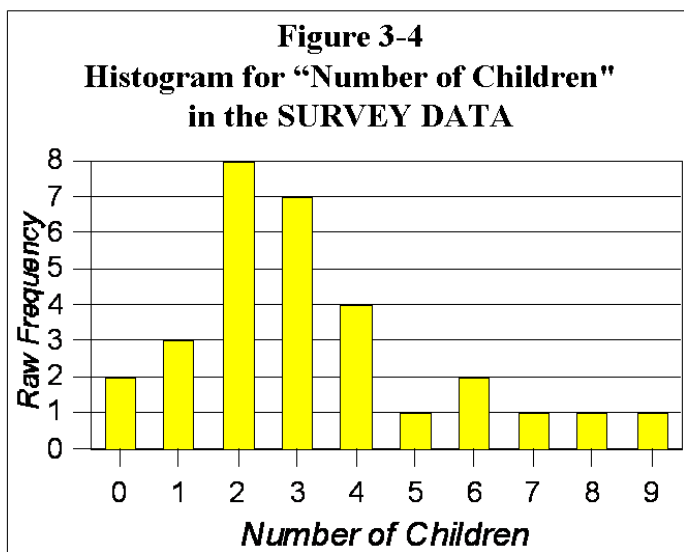
<i>Yearly Income (\$1000s)</i>	<i>Raw Frequency</i>	<i>Relative Frequency</i>	<i>Cumulative Raw Frequency</i>	<i>Cumulative Relative Frequency</i>
Above 20 to 30	7	23.33%	7	23.33%
Above 30 to 40				
Above 40 to 50				
Above 50 to 60				
Above 60 to 70				
Above 70 to 80				
<i>Totals</i>				



distribution was called a *negatively skewed* distribution, since the lower (negative) half of the observations are more widely dispersed than the upper half of observations.

From the histogram in Figure 3-4, we can see that the distribution of the variable "Number of Children" in the SURVEY appears to be positively skewed. We also see from the histogram in Figure 3-3 that the distribution of the variable "Yearly Income" in the SURVEY appears to be positively skewed.

The histograms of Figures 3-5a to 3-5d illustrate some of the types of distribution which we often encounter in practice. These are by no means the only distributions we might encounter, though. Figure 3-5e illustrates a distribution where values occur with highest frequencies at more than one point in the range; such a distribution is called a *multi-modal* distribution. Figure 3-5f illustrates a positively skewed distribution which appears significantly more positively skewed than the distribution of Figure 3-5c.



The number of different types of graphical displays possible is limited only by the imagination. We have introduced only a few ways to display the distribution for a variable. We shall briefly mention two other graphical displays which can be used to display the distribution for a quantitative variable.

A *frequency polygon* is constructed as a histogram is, except that dots connected by straight lines are used in place of bars. A dot is placed either above each value represented on the horizontal axis, or above the midpoint of each class represented on the horizontal axis. Figure 3-6 is a frequency polygon displaying exactly the same data as Figure 3-3. Notice that the horizontal and vertical axes in Figures 3-3 and 3-6 are scaled the same, but Figure 3-6 was constructed by placing dots above the midpoints of each class, including an extra class at each end of the range, and then connecting these dots with lines. The extra classes at each end are included merely so that the frequency polygon does not appear to be suspended in mid air, and are not absolutely essential.

An *ogive* is constructed in the same way that a frequency polygon is, except that cumulative frequencies are labeled on the vertical axis, and if classes are used, the dots are placed above the endpoints of each class, instead of the midpoints. Figure 3-7 is an ogive displaying exactly the same data as Figures 3-3 and 3-6.

Self-Test Problem 3-1. Use the data for the variable "Number of Children" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, to do each of the following:

- Complete the construction of a dot plot for the rural area, a dot plot for the suburban area, and a dot plot for the urban area, by completing the graph titled "Dot Plots for Self Test Problem 3-1".
- Does the distribution of "Number of Children" appear to be centered at different values for the three areas of residence (rural, suburban, urban)? If yes, describe the difference(s).
- Does the distribution of "Number of Children" appear to have a different amount of dispersion for the three areas of residence (rural, suburban, urban)? If yes, describe the difference(s).
- Suppose a histogram was constructed for each of the three areas of residence (rural, suburban, urban). Indicate which of Figures 3-5a to 3-5f each of the three histogram shapes would most resemble.
- Construct three frequency polygons, one for each area of residence, on the same graph.

Dot Plots for Self Test Problem 3-1

Rural Area

Suburban Area

Urban Area

0 1 2 3 4 5 6 7 8 9
Number of Children

0 1 2 3 4 5 6 7 8 9
Number of Children

0 1 2 3 4 5 6 7 8 9
Number of Children

Figure 3-5a

Uniform Distribution

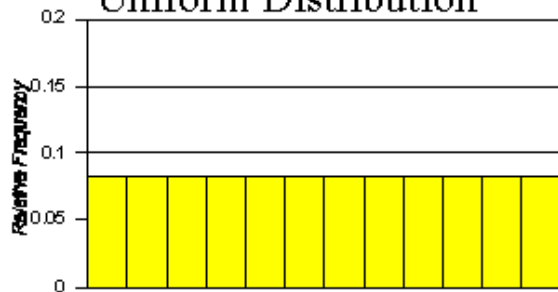


Figure 3-5b

Bell-Shaped Distribution

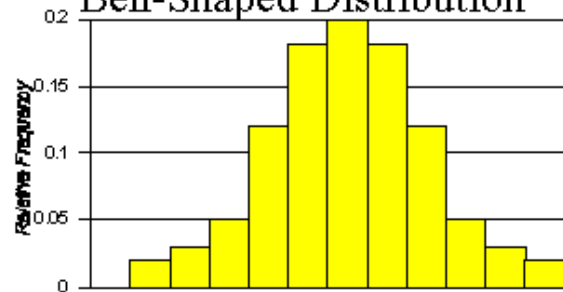


Figure 3-5c

Positively Skewed Distribution

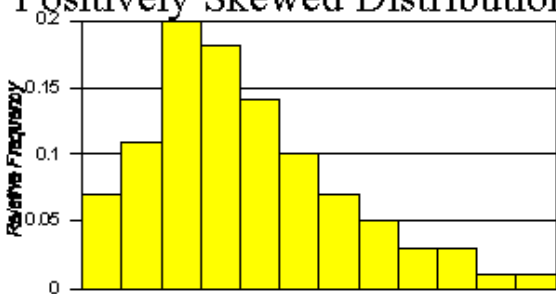


Figure 3-5d

Negatively Skewed Distribution

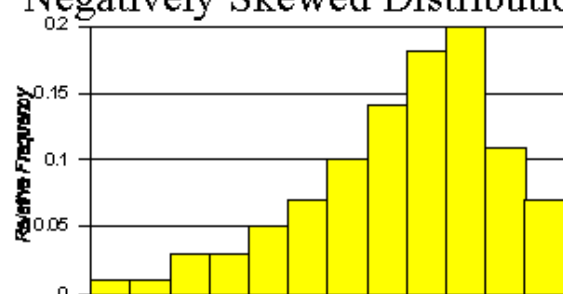


Figure 3-5e

Multimodal Distribution

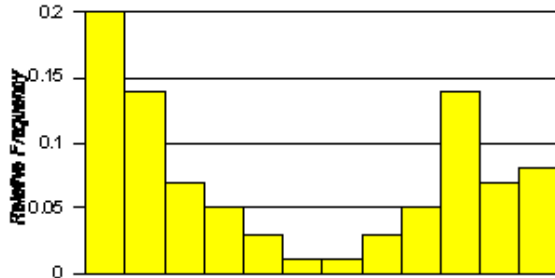
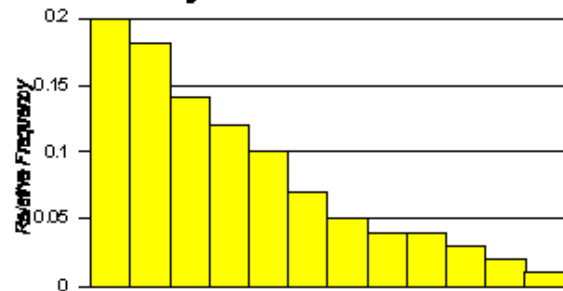
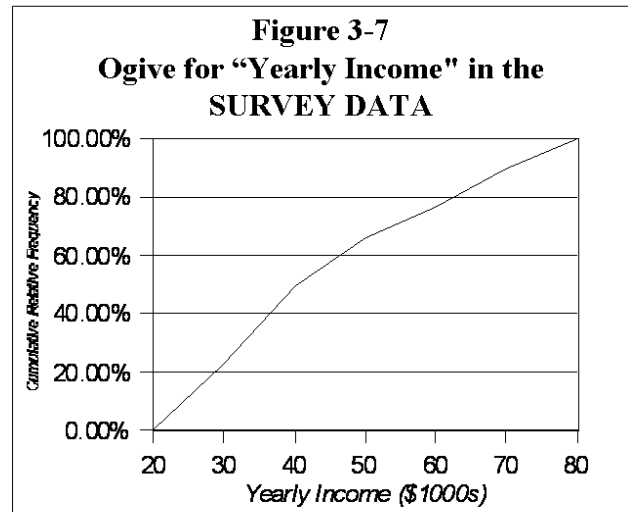
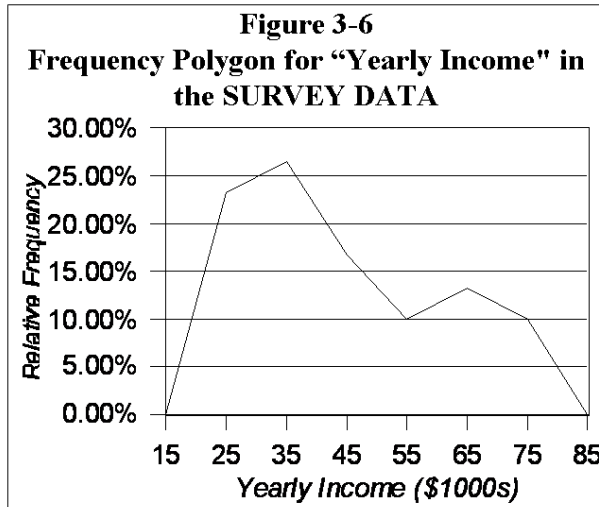


Figure 3-5f

Positively Skewed Distribution





Self-Test Problem 3-2. Figure 2-7 displays a box plot of yearly salaries for each of three mythical corporations. Suppose a histogram of the yearly salaries was to be constructed for each of the three corporations. For each of the box plots in Figure 2-7, indicate which of Figures 3-5a to 3-5f the corresponding histogram shape would most resemble.

Self-Test Problem 3-3. Decide whether it would be better to list individual values or to define classes in the construction of a frequency distribution, and explain your answer, for each of the following variables:

- The number of times a person rides one of the city busses in a given week is recorded for each of 200 residents.
- The number of times a person rides one of the city busses in a given month is recorded for each of 200 residents.
- The exact total dollars of income last year for a household is recorded for each of 200 households in a city.
- The number of household members who earned income last year is recorded for each of 200 households in a city.

Self-Test Problem 3-4. For each the following frequency distributions, indicate whether or not cumulative frequencies would be appropriate, and state why or why not:

- The exact total dollars of income last year for a household is recorded for each of 200 households in a city and organized into a frequency distribution.
- The county in which a person is employed is recorded for each of 200 residents of a state and organized into a frequency distribution.
- The number of times a person rides one of the city busses in a given week is recorded for each of 200 residents and organized into a frequency distribution.
- Each of 200 residents of a city are classified into one of five economic status categories (poverty, low income, middle class, high income, wealthy), and the data is organized into a frequency distribution.
- Which of the 12 busses that run each weekday a person rides most often is recorded for each of 200 city bus riders and organized into a frequency distribution.

Answers to Self-Test Problems

- 3-1** (a) See the completed version of the dot plots in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit. (b) "Number of Children" appears to be centered at a considerably higher value for the rural area than for the suburban and urban areas. (c) The amount of variation in "Number of Children" appears to be greatest for the rural area and least for the suburban area. (d) The shape of the histogram for the rural area would most resemble Figure 3-5c; the shape of the histogram for the suburban area would most resemble Figure 3-5b; the shape of the histogram for the urban area would most resemble Figure 3-5d (although since the distribution appears to be only slightly negatively skewed, one might say the shape is close to that of Figure 3-5b). (e) See graph titled "Frequency Polygons for Number of Children" in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.
- 3-2** The shape of each of the three histograms would most resemble Figure 3-5b.
- 3-3** (a) Since the observations are likely to consist of a few distinct values with much repetition, individual values could be listed in the construction of a histogram. (b) Since the observations are likely to consist of many different values with little repetition, it would be necessary to define classes in the construction of a frequency distribution. (c) Since the observations are likely to consist of many different values with little repetition, it would be necessary to define classes in the construction of a frequency distribution. (d) Since the observations are likely to consist of a few distinct values with much repetition, individual values could be listed in the construction of a histogram.
- 3-4** (a) Since the variable is quantitative, cumulative frequencies would be appropriate. (b) Since the categories have no natural ordering, cumulative frequencies would not be appropriate. (c) Since the variable is quantitative, cumulative frequencies would be appropriate. (d) Since the categories have a natural ordering, cumulative frequencies would be appropriate. (e) Since the categories have no natural ordering, cumulative frequencies would not be appropriate.

Summary

A *frequency distribution* is a list of distinct values or categories together with corresponding *raw frequencies* and *relative frequencies*; with a quantitative variable, we can also include *cumulative raw frequencies* and *cumulative relative frequencies*. The *bar chart* and *pie chart* are possible graphical for the distribution of a qualitative variable. The *histogram*, *frequency polygon*, and *ogive* are possible graphical displays for the distribution of a quantitative variable; such graphical displays can provide us with information about center, dispersion, and shape for the distribution. The number of different types of graphical displays possible for the distribution of a variable is limited only by the imagination.

SURVEY DATA Summaries for Data Set 1-1

Table 3-1		
Frequency Distribution for "Political Party" in the SURVEY DATA		
<i>Political Party</i>	<i>Raw Frequency</i>	<i>Relative Frequency</i>
Republican	10	33.3%
Democrat	8	26.7%
Independent	4	13.3%
Other	8	26.7%
	30	100.0%

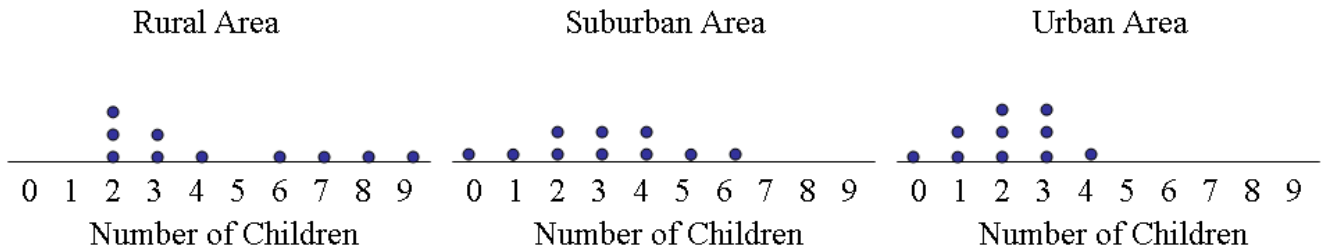
Table 3-2				
Frequency Distribution for "Number of Children" in the SURVEY DATA				
<i>Number of Children</i>	<i>Raw Frequency</i>	<i>Relative Frequency</i>	<i>Cumulative Raw Frequency</i>	<i>Cumulative Relative Frequency</i>
0	2	6.67%	2	6.67%
1	3	10.00%	5	16.67%
2	8	10.00%	13	43.33%
3	7	23.33%	20	66.67%
4	4	13.33%	24	80.00%
5	1	3.33%	25	83.33%
6	2	3.33%	27	90.00%
7	1	3.33%	28	93.33%
8	1	3.33%	29	96.67%
9	1	3.33%	30	100.00%
<i>Totals</i>	30	100.00%		

continued next page

SURVEY DATA Summaries for Data Set 1-1 - *continued*

Table 3-3				
Frequency Distribution for “Yearly Income” in the SURVEY DATA				
<i>Yearly Income (\$1000s)</i>	<i>Raw Frequency</i>	<i>Relative Frequency</i>	<i>Cumulative Raw Frequency</i>	<i>Cumulative Relative Frequency</i>
Above 20 to 30	7	23.33%	7	23.33%
Above 30 to 40	8	26.67%	15	50.00%
Above 40 to 50	5	16.67%	20	66.67%
Above 50 to 60	3	10.00%	23	76.67%
Above 60 to 70	4	13.33%	27	90.00%
Above 70 to 80	3	10.00%	30	100.00%
<i>Totals</i>	30	100.00%		

Dot Plots for Self Test Problem 3-1



Frequency Polygons for Number of Children

