

# Unit 2

## Summarizing Data

Objectives:

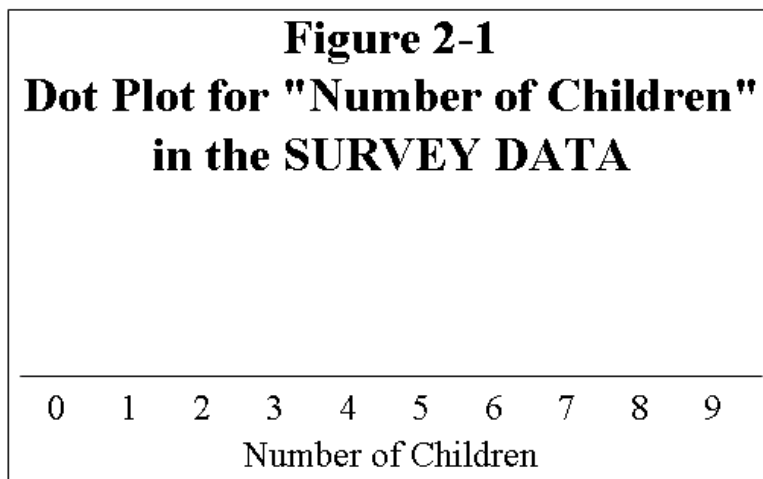
- To construct a dot plot and a stem-and-leaf display for quantitative data
- To obtain the five-number summary for quantitative data
- To construct a box plot for quantitative data
- To understand center, dispersion, and shape for the distribution of a quantitative variable

Information about one or more variables is called *data*. This information could be in the form of one or more tables, graphs, or numerical values. In order to construct a table or graph or to obtain numerical values which describe certain characteristics, we must first obtain and record observations of the variable(s) of interest. The recorded observations that we make to collect data are called *raw data*. The SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, is raw data, since it consists of the observations of nine variables (excluding the ID number assigned to each individual) made on several individuals.

We rarely obtain any important information by looking only at raw data, especially if the number of observations is large. For instance, as you scan through the SURVEY DATA there is not much information about the variables that is easy to extract from the sea of numbers you face. It is often necessary and useful to organize the raw data into some form of table or graph, or to obtain numerical values which describe informative characteristics about the data. We shall now consider how to accomplish this for data observed on a quantitative variable.

Two graphical displays which are generally easy to construct and often serve as a starting point for analyzing data are a *dot plot* and a *stem-and-leaf display*. After scaling a horizontal axis with values covering the range for a given variable, a dot plot can be constructed by simply representing each observed value with a dot placed above the corresponding value on the axis. A stem-and-leaf display is constructed by first placing integers called *stems* on the left side of a vertical line and placing integers called *leaves* on the right side of the vertical line.

Look at the values of the variable "Number of Children" in the SURVEY DATA. You should quickly see that the values range from 0 to 9. We can now construct a dot plot by first labeling a horizontal axis to contain the range from 0 to 9, then representing each observed number of children with a dot placed above the corresponding value on the axis. Use the SURVEY DATA to complete the dot plot in Figure 2-1 by placing dots in the appropriate places above the horizontal axis. The result should look like the corresponding figure displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

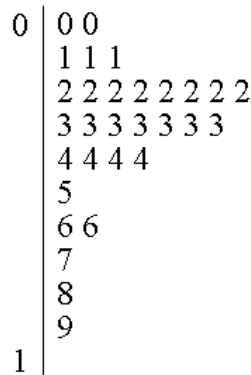


To construct a stem-and-leaf display for the variable "Number of Children," we must first place stems on the left side of a vertical line. Since values only range from 0 to 9, we decide to let each row represent one value. In Figure 2-2, we placed a zero at the top immediately on the left of the vertical line, and we placed a one at the bottom immediately on the left of the vertical line, leaving room for ten rows in between the zero and one. We then record the appropriate digit in the appropriate row for each observed value of number of children. No digits appear to the right of the one at the bottom, since number of children was always smaller than 10. The dot

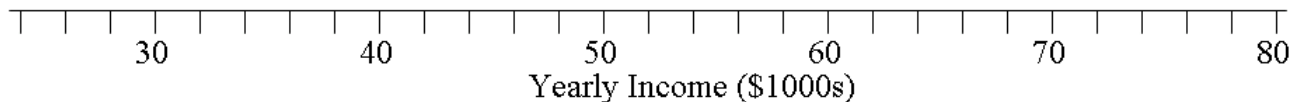
plot and the stem-and-leaf display each provide us with a picture of the data for the variable "Number of Children."

Now look at the values of the variable "Yearly Income" in the SURVEY DATA. You should see that there are no values less than 20 thousand dollars and no values greater than 80 thousand dollars. We can now construct a dot plot by first labeling a horizontal axis from 20 thousand to 80 thousand dollars, then representing each observed yearly income with a dot placed above the corresponding value on the axis. Use the SURVEY DATA to complete the dot plot in Figure 2-3 by placing dots in the appropriate places above the horizontal axis. The result should look like the corresponding figure displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

**Figure 2-2  
Stem-and-Leaf Display for  
"Number of Children" in the  
SURVEY DATA**



**Figure 2-3  
Dot Plot for "Yearly Income" in  
the SURVEY DATA**



To construct a stem-and-leaf display for the variable "Yearly Income," we must first place stems on the left side of a vertical line. Since values are all between 20 and 80 thousand dollars, we can first place the digits 2, 3, 4, 5, 6, and 7 on the left side of a vertical line, as has been done in Figure 2-4. Complete this stem-and-leaf display by recording the appropriate second digit in the corresponding row for each observed yearly income. The result should look like the corresponding figure displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit. The dot plot and the stem-and-leaf display each give us a picture of the data for the variable "Yearly Income."

Notice how different variations are possible in constructing stem-and-leaf displays, depending on the number of digits used in the data, the number of significant digits desired, etc. One very convenient use of a stem-and-leaf display is in the construction of an *ordered array*, which is a list of the observations ordered by value. The completed version of Figure 2-4 in the

**Figure 2-4  
Stem-and-Leaf Display for  
"Yearly Income" in the  
SURVEY DATA**



section titled **SURVEY DATA Summaries for Data Set 1-1** actually contains two stem-and-leaf displays. The ordered stem-and-leaf display on the right results from placing the digits in each row of the stem-and-leaf display on the left in ascending order. From the ordered stem-and-leaf display, it is easy to construct the following ordered array for yearly income:

25 26 27 28 29 30 30 33 34 34 35 39 39 39 40 41 44 45 45 49 53 55 60 61 64 65 68 71 75 78

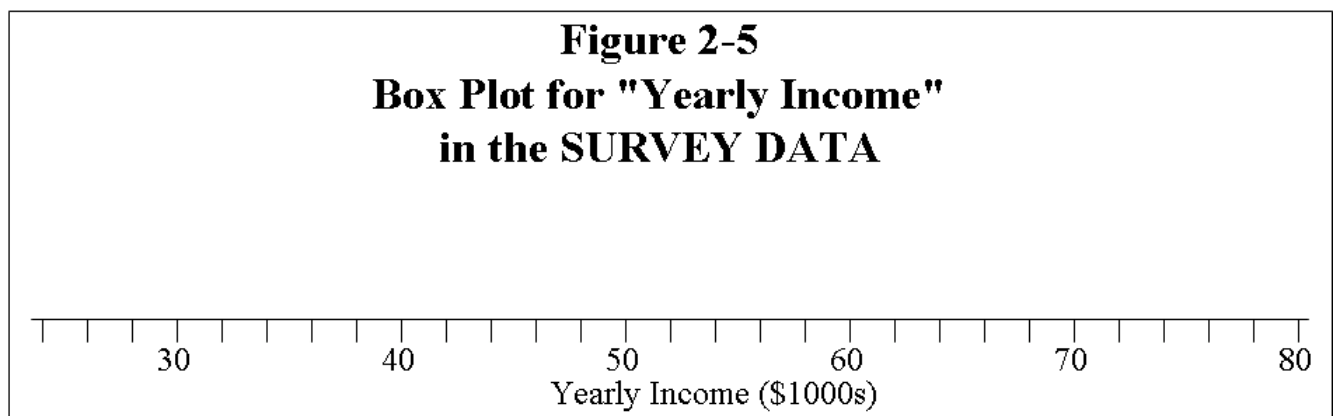
Generally, a data set contains too many numbers for us to extract any useful information just by looking at the ordered array. Consequently, we can summarize the data by obtaining the *five-number summary*. These five numbers are the *minimum*, the *first quartile*, the *second quartile*, the *third quartile*, and the *maximum*. The minimum and maximum are of course the smallest and largest values in the ordered array; the three quartiles are values that divide the ordered array into four sections, each containing the same number of observations.

From the ordered array of yearly incomes, the minimum and maximum are obviously 25 and 78. The second quartile is either the observation which divides the ordered array in half (when there are an odd number of observations) or the average of the two middle observations of the ordered array (when there are an even number of observations). Since there are 30 yearly incomes, the 15 smallest incomes make up the lower half, and the 15 largest incomes make up the upper half. To illustrate, we redisplay the ordered array as follows, with the observations which determine the five-number summary in boldface:

<b>25</b>	<b>40 41</b>	<b>78</b>
26	39 44	75
27	39 45	71
28	39 45	68
29	35 49	65
30	34 53	64
30	34 55	61
<b>33</b>	<b>60</b>	

The second quartile is  $(40+41)/2 = 40.5$ . The first and third quartiles are found by obtaining either the middle observation or the average of the two middle observations, for each of the lower half and upper half of the ordered array, respectively. The first quartile is 33, and the third quartile is 60. We have found then that the five-number summary for "Yearly Income" in the SURVEY DATA is 25, 33, 40.5, 60, 78.

We shall find it convenient to refer to the five-number summary in general as *Min*,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , *Max*. Of course, *Min* and *Max* refer respectively to the minimum and maximum. The first quartile, second quartile, and third quartile are denoted respectively as  $Q_1$ ,  $Q_2$ , and  $Q_3$ . Using either Figure 2-1 or Figure 2-2, find the five-number summary for "Number of Children" in the SURVEY DATA. (You should find the five-number summary to be 0, 2, 3, 4, 9.)



We shall see that we can obtain a great deal of information about the data from the five-number summary. Since it is almost always easier to look at a graph than to look at a list of numbers, we shall find it very useful to create a "picture" of the five-number summary. A *box plot* is a graphical display of the five-number summary and is among the most popular graphical displays of data. A box plot for a data set can be

constructed by first labeling an axis with a range of values that includes  $Min$  and  $Max$ . Next, a rectangle is drawn with one end at  $Q_1$  and the other end at  $Q_3$ ; this rectangle is then divided into two parts, with a line drawn at the  $Q_2$ . Finally, two lines parallel to the axis are extended from each end of the rectangle, one to  $Min$  and the other to  $Max$ .

The five-number summary for "Yearly Income" in the SURVEY DATA was found to be 25, 33, 40.5, 60, 78. Complete the box plot in Figure 2-5. The result should look like the corresponding figure displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

Let us now consider the variable "Yearly Income" in the SURVEY DATA separately for males and females. First, verify that the stem-and-leaf displays for the yearly incomes of males and the yearly incomes of females, shown in Figure 2-6, are both correct. Then, verify that the ordered array for the 15 males is as follows:

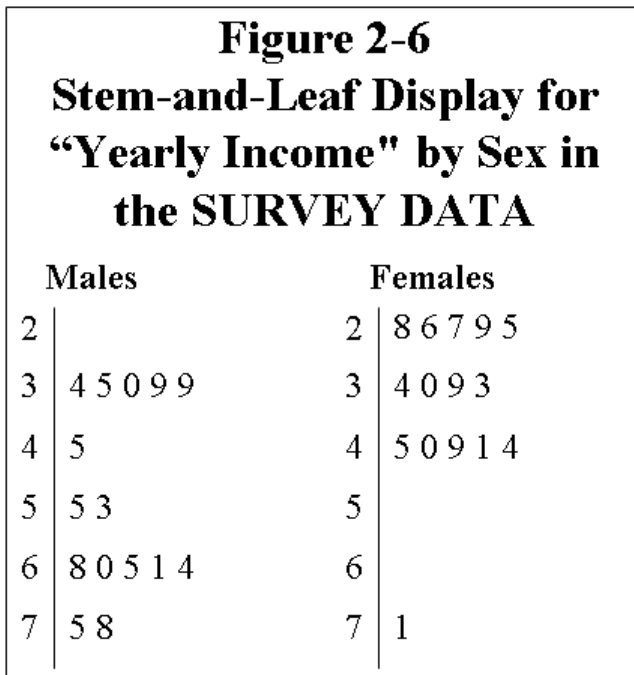
30 34 35 39 39 45 53 55 60 61 64 65 68 75 78

It should be easy to see that  $Min = 30$  thousand dollars and  $Max = 78$  thousand dollars. Since the number of observations is odd, there is exactly one observation which divides the ordered array in half, and this is  $Q_2 = 55$  thousand dollars. To obtain  $Q_1$ , we find the middle observation among the seven observations below  $Q_2 = 55$ ; to obtain  $Q_3$ , we find the middle observation among the seven observations above  $Q_2 = 55$ . Notice that when we split the ordered array into two halves in order to find  $Q_1$  and  $Q_3$ , we did not include the middle observation as part of either half of the ordered array. We find then that the five-number summary for males is 30, 39, 55, 65, 78. Now, you find the five-number summary for the females. (You should find that the five-number summary for females is 25, 28, 34, 44, 71.)

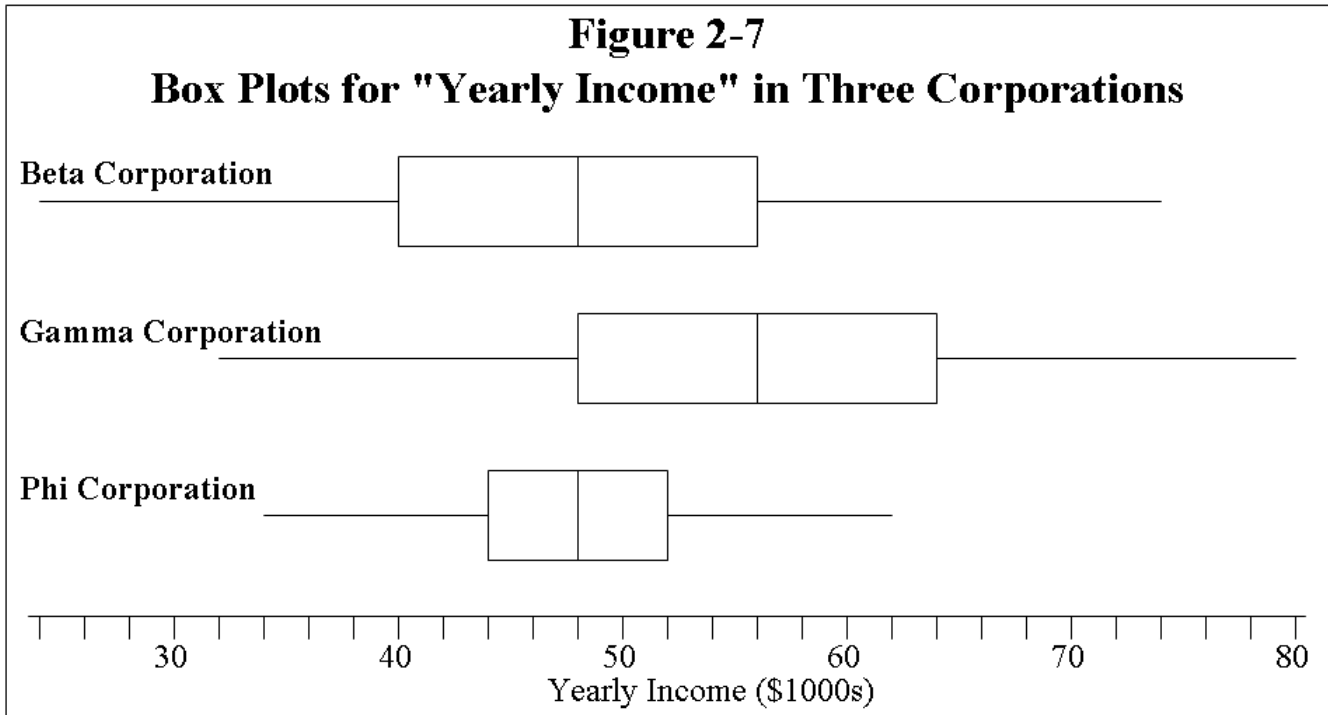
We have now introduced three different graphical displays of data (the dot plot, stem-and-leaf display, and box plot) and one way to summarize a data set with just a few numbers (the five-number summary). You may now be asking yourself, however, just exactly what it is you are suppose to be looking for in a graphical display or a five-number summary. We shall now begin to illustrate the kind of valuable information you can obtain from a graph and how to interpret graphs intelligently. Three main characteristics about the distribution of a quantitative variable are most often of interest. The *center value* in a distribution refers to a middle or average value for the distribution. The *dispersion* in a distribution refers to the amount of variation in the distribution. The *shape* of a distribution involves the type and amount of symmetry or non-symmetry present in the distribution.

Let us first consider center value for a distribution. Figure 2-7 displays a box plot of yearly salaries for each of three mythical corporations. The second quartile, which is the middle value in the five-number summary and is represented by the line which divides the box in the box plot into two parts, can be used as a measure of the center value of a distribution (since we know that about 50% of the observations lie on either side of the second quartile). Putting the three box plots on the same graph makes it is easy to compare the center value of salary for the three corporations. From Figure 2-7, we see that the distribution of salaries is centered at the same value (\$48,000) for the Beta and Phi corporations, and we see that the distribution of salaries appears to be centered considerably higher (\$56,000) for the Gamma corporation.

Let us now consider dispersion in a distribution. From Figure 2-7, we see that the distribution of salaries for the Beta and Gamma corporations appears to span across a wider range of salaries than for the Phi corporation. In other words, the distribution of salaries Beta and Gamma corporations show more variation or more dispersion than for the Phi corporation. By subtracting the minimum value from the maximum value, we can obtain what we formally define in statistics to the *range*. Since the range for the Beta corporation salaries is  $72 - 24 = 48$  thousand dollars, and the range for the Gamma corporation salaries is  $80 - 32 = 48$  thousand



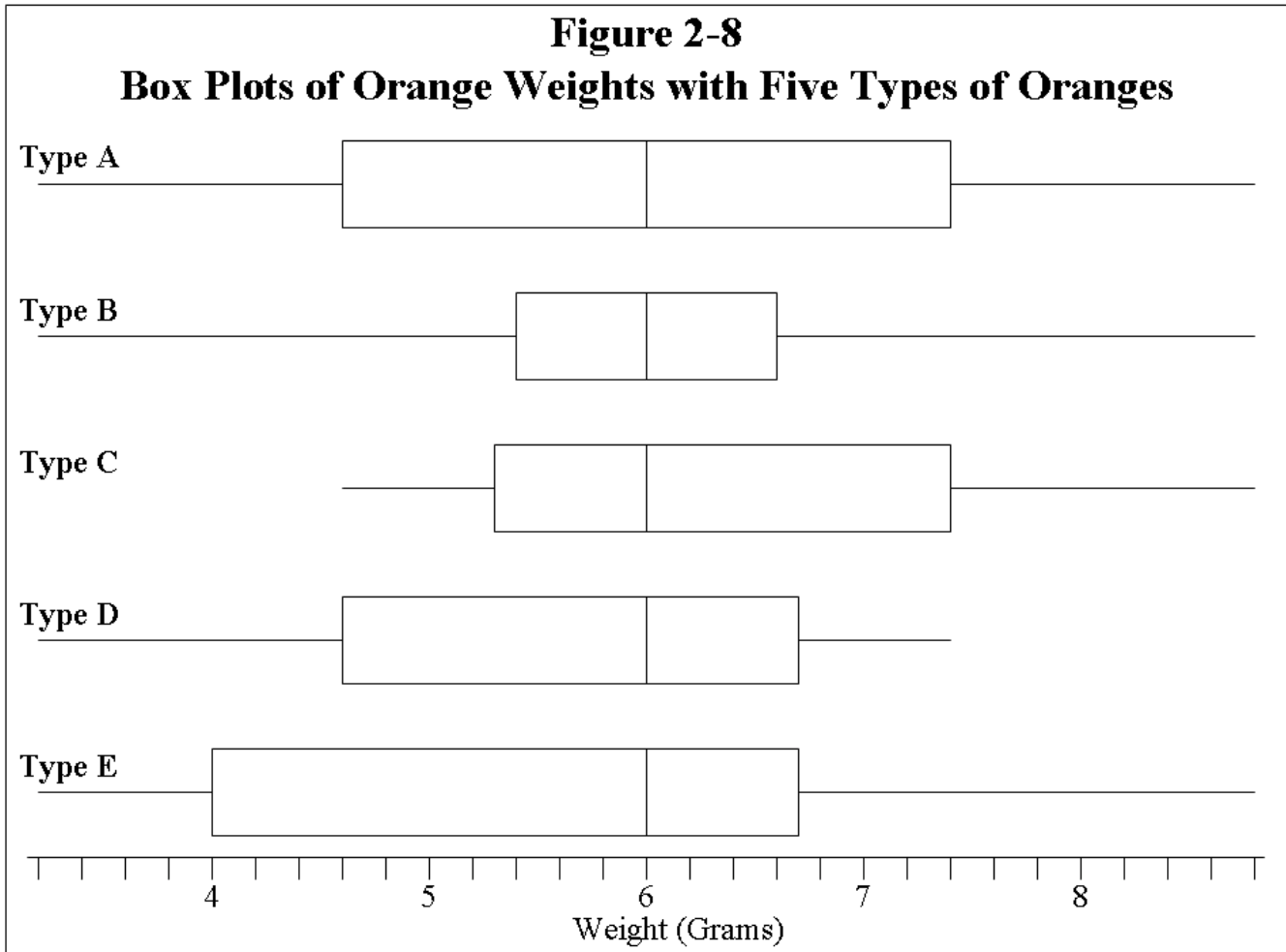
dollars, we find that the dispersion of salaries is the same for the Beta and Gamma corporations (even though Gamma corporation salaries are centered at a higher value). Since the range for the Phi corporation salaries is  $62 - 34 = 28$  thousand dollars, we find that the dispersion of salaries for the Phi corporation appears to be considerably less than for the Beta and Gamma corporations (even though Beta and Phi corporation salaries are centered at the same value).



Finally, let us consider the shape of a distribution. A distribution is called *symmetric* when the shape of the lower half and the shape of the upper half are mirror images of each other; otherwise the distribution is called skewed. Each of the distributions displayed in Figure 2-7 is symmetric. Figure 2-8 displays a box plot of orange weights for each of five mythical types of oranges. Note that the distribution of weights is symmetric for Type A and for Type B oranges. Note also that the range of weights is exactly the same for Type A and Type B oranges, that is, from 3.2 to 8.8 grams. However, the distribution of weights for Type B oranges shows less dispersion than that for Type A oranges. From the box plot for Type A, we observe that 25% of the weights are between 3.2 and 4.6 grams, 25% of the weights are between 4.6 and 6.0 grams, 25% of the weights are between 6.0 and 7.4 grams, and 25% of the weights are between 7.4 and 8.8 grams. This suggests that the Type A weights are rather uniformly distributed across the range from 3.2 to 8.8 grams. It will probably not come as a great shock that a distribution where observations are uniformly distributed across a range is called a *uniform* distribution. Although the distribution for Type A weights and Type B weights are both symmetric, we see that the Type B weights do not have a uniform distribution, since half (50%) of the weights are between 5.4 and 6.6 grams, which is considerably less than half of the entire range. The fact that such a large percentage of the weights is concentrated in such a small portion of the range is the reason why the Type B weights do not have a uniform distribution and is also the reason why the Type B weights show less dispersion than the Type A weights.

The distribution of weights is not symmetric for Type C, Type D, and Type E oranges, although we can think of each of these distributions as centered at 6.0 grams. From the box plot for Type C, we observe that the half of the weights above 6.0 grams are spread out over an interval which is almost three grams in length, while the half of the weights below 6.0 grams are all concentrated inside an interval just a little over one gram in length. We said earlier that a distribution where the lower half and the upper half are not mirror images of each other is called skewed. The Type C weights have a skewed distribution; more specifically, however, we can say that the Type C weights have a positively skewed distribution. A distribution is called *positively skewed* (or *skewed to the right*) when the observations in the upper half of the distribution show considerably more

dispersion than the observations in the lower half of the distribution; a distribution is called *negatively skewed* (or *skewed to the left*) when the observations in the lower half of the distribution show considerably more dispersion than the observations in the upper half of the distribution. Note that the Type D weights have a negatively skewed distribution, since the half of the weights below 6.0 grams are spread out over an interval which is almost three grams in length, while the half of the weights above 6.0 grams are all concentrated inside an interval just a little over one gram in length.



We would call the distribution of Type E weights skewed, since this distribution is obviously not symmetric; however, we can't really say the distribution is positively skewed, nor can we say the distribution is negatively skewed. There are many different ways in which a distribution can be skewed, and there are many ways in which a distribution can be symmetric. Our goal here is not to discuss every possible shape for a distribution, but instead merely to summarize a few basic ideas. Generally speaking, we will be interested in whether or not a distribution is approximately symmetric. If a distribution is symmetric, then we are interested in whether it is a uniform distribution or some other type of symmetric distribution; if a distribution is not symmetric, then we will be interested in whether it is skewed in a positive or negative direction.

Let us return to the variable "Yearly Income" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1. Earlier, you found that the five-number summary for males is 30, 39, 55, 65, 78, and that the five-number summary for females is 25, 28, 34, 44, 71. Once you have the five-number summaries for males and females, it is easy to construct a box plot for each sex on the same graph. Do this, and you will see that the distribution of "Yearly Income" appears centered at a higher value for males than females, but the amount of variation in incomes does not appear to be drastically different for the two sexes. You should also notice that the distribution of incomes for males is closer to looking symmetric than the distribution for females; the distribution for females looks very positively skewed.

**Figure 2-9**  
**Stem-and-Leaf Display for**  
**“Age” by Sex in the**  
**SURVEY DATA**

Males	Females
2	2
3	3
4	4
5	5
6	6

**Self-Test Problem 2-1.** Use the data for the variable "Age" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, to do each of the following:

- (a) Complete the construction of Figure 2-9 to obtain a stem-and-leaf display for males and a stem-and-leaf display for females.
- (b) Obtain an ordered array for males and an ordered array for females.
- (c) Find the five-number summary for males and the five-number summary for females.
- (d) Construct a box plot for males and a box plot for females.
- (e) Does the distribution of "Age" appear to be centered at a different value for males and females? If yes, for which sex does the center of the distribution appear to be greater?
- (f) Does the distribution of "Age" appear to have a different amount of dispersion for males and females? If yes, for which sex does the dispersion appear to be greater?
- (g) Does the distribution of "Age" appear to have a different shape for males and females? If yes, how does the shape appear to differ between the two sexes?
- (h) When comparing the distribution of "Age" for the two sexes as displayed from part (a), how might the difference in where the distributions are centered at least partially explain the difference in where the distributions of "Yearly Income" for the two sexes are centered in the stem-and-leaf displays constructed in Figure 2-6?

**Self-Test Problem 2-2.** Figure 2-8 displays a box plot of orange weights for each of five mythical types of oranges. Suppose the range of weights for a sixth type of orange, labeled Type F, was from 3.2 to 8.8 grams with less than half of the oranges weighing between 4 and 8 grams.

- (a) Give an example of what the five-number summary and corresponding box plot might be for Type F orange weights.
- (b) Would the box plot in part (a) for Type F orange weights show more dispersion or less dispersion than the Type A orange weights of Figure 2-8?

### Answers to Self-Test Problems

- 2-1** (a) See the completed version of Figure 2-9 in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit. (c) The five-number summary for males is 35, 44, 50, 59, 62, and the five-number summary for females is 20, 32, 40, 44, 59. (e) The ages appear to be centered at a higher value for males than for females. (f) The amount of variation in ages appears to be roughly the same for males and females. (g) The distribution of ages does not appear to be very skewed for either sex. (h) The fact that ages tend to be higher for males than females most likely contributes to the fact that yearly incomes tend to be higher for males than females.
- 2-2** (a) One possibility for the five-number summary would be 3.2, 3.9, 6.0, 8.1, 8.8. The corresponding box plot would look like one where the box extends over a very large range while the lines extending from either side of the box would each be very small. (b) The Type F orange weights would show considerably more dispersion than the Type A orange weights.

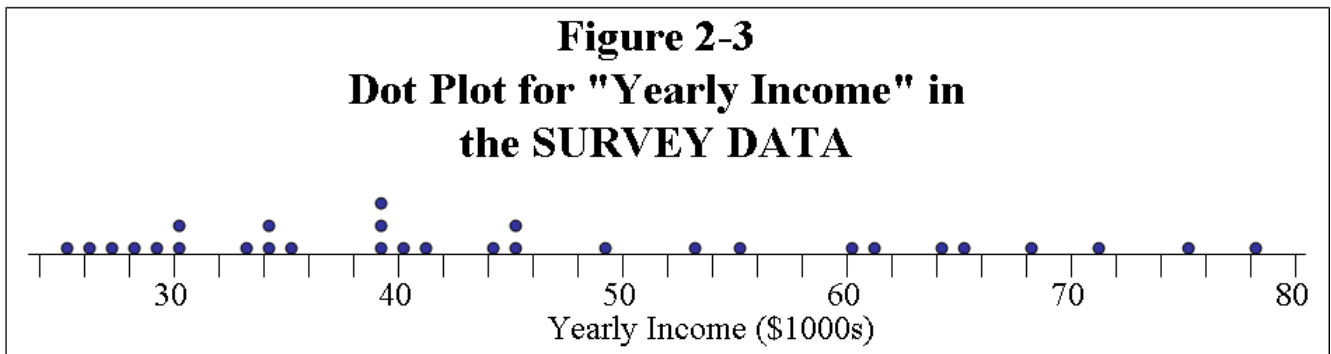
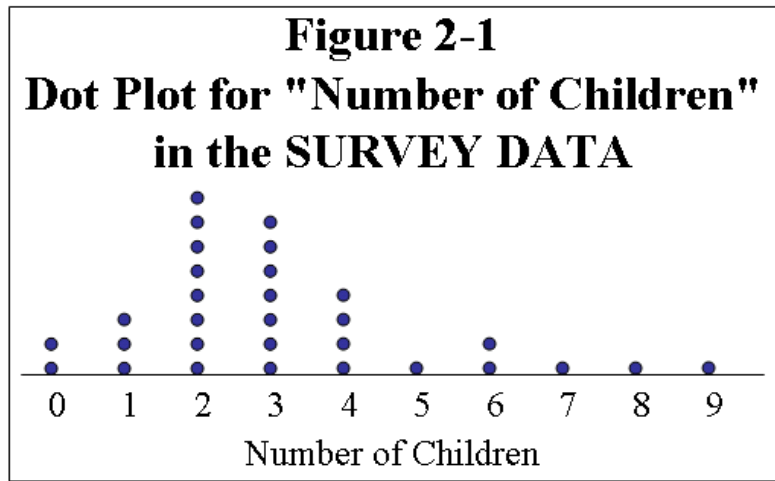
### Summary

It is often useful for us to organize *raw data* into some form of tabular and/or graphical display. Two graphical displays which are generally easy to construct and often serve as a starting point for analyzing data are a *dot plot* and a *stem-and-leaf display*. These visuals display of data make it easy to obtain an ordered array. From an ordered array, the *five-number summary* is easily obtained: the minimum *Min*, the first quartile  $Q_1$ , the second quartile  $Q_2$ , the third quartile  $Q_3$ , and the maximum *Max*. A *box plot* is a graphical display of the five-number summary.

Three main characteristics about the distribution of a quantitative variable are most often of interest. The *center value* in a distribution refers to a middle or average value for the distribution. The *dispersion* in a distribution refers to the amount of variation in the distribution. The *shape* of a distribution involves the type and amount of symmetry or non-symmetry present in the distribution. A distribution is called *symmetric* when the shape of the lower half and the shape of the upper half are mirror images of each other; otherwise the distribution is called *skewed*.

There are many different ways in which a distribution can be skewed, and there are many ways in which a distribution can be symmetric. A distribution where observations are uniformly distributed across a range is symmetric and is called a *uniform* distribution. A distribution is called *positively skewed* (or *skewed to the right*) when the observations in the upper half of the distribution show considerably more dispersion than the observations in the lower half of the distribution; a distribution is called *negatively skewed* (or *skewed to the left*) when the observations in the lower half of the distribution show considerably more dispersion than the observations in the upper half of the distribution.

## SURVEY DATA Summaries for Data Set 1-1



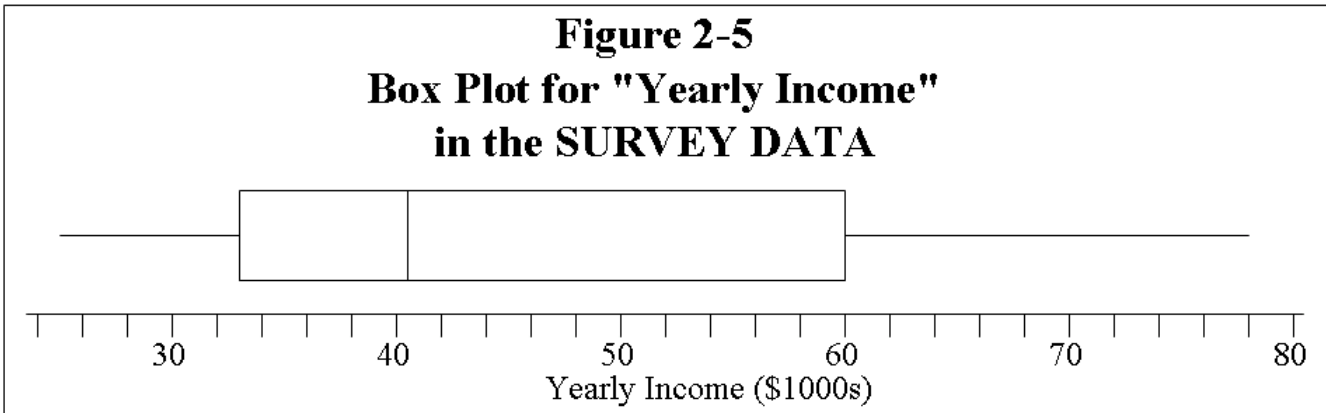
**Figure 2-4**  
**Stem-and-Leaf Display for**  
**"Yearly Income" in the**  
**SURVEY DATA**

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

*continued next page*

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

**Figure 2-5**  
**Box Plot for "Yearly Income"**  
**in the SURVEY DATA**



**Figure 2-9**  
**Stem-and-Leaf Display for**  
**"Age" by Sex in the**  
**SURVEY DATA**

Males		Females	
2		2	0 0 4
3	5 9	3	5 4 8 2
4	1 6 4 7 4	4	4 0 4 5 1 4
5	9 2 4 0 3 9	5	6 9
6	2 2	6	