

Unit 7

Comparisons and Relationships

Objectives:

- To understand the distinction between making a comparison and describing a relationship
- To select appropriate graphical displays for making comparisons
- To select appropriate graphical displays for describing relationships

Up to this point, our focus has been primarily on characteristics of the distribution for a single variable. We now want to consider situations involving at least two variables, where we can study the relationship between two variables. In particular we shall be discussing two concepts: making a comparison and describing a relationship. Making a *comparison* generally implies that different observations of the same numerical value(s) or commensurate numerical values (i.e., values measured on the same scale) are being compared. Describing a *relationship* implies describing how changes in one variable are influenced by changes in a second variable. Depending on the situation, making a comparison may involve only one variable or more than one variable; however, describing a relationship must always involve (at least) two variables.

As an illustration, let us suppose we would like to compare the distribution of weights for five different types of oranges (labeled A, B, C, D, and E). It seems clear that in this situation our interest is in making comparisons. In fact, Figure 2-8 displays a box plot of weights for each of the five types of oranges. Having the five box plots on the same graph makes it easy to compare the center of the distribution of weights, the dispersion of the distribution of weights, and the shape of the distribution of weights for the five types of oranges, which we have done previously.

While it seems clear that Figure 2-8 designed to make comparisons, it may surprise you to find out that in this particular situation making these comparisons is actually the same as describing a relationship between two variables! The two variables are “type of orange” (which is qualitative-ordinal with five categories) and “weight” (which is quantitative). If there were no relationship between “type of orange” and “weight,” then we would expect that the distribution of weights would look practically identical for the five types of oranges; however, if there were a relationship between “type of orange” and “weight,” then this relationship would have to be described in terms of how the distribution of weights differs for the five types of oranges, or in other words how changes in “weight” are influenced by changes in “type of orange.” Consequently, making a comparison of the distribution of weights among the five types of orange and describing the relationship between “type of orange” and “weight” are really just two different ways of thinking about the same thing!

We see then that the distinction between making a comparison and describing a relationship is not always a sharp one. Describing a relationship can sometimes be done by making comparisons, but making comparisons does not always imply that a relationship is being described. If we are making comparisons in order to describe how changes in one variable are influenced by changes in a second variable, then we are really describing the relationship between two variables. On the other hand, if we are making comparisons which do not involve how the values of one variable influence a second variable, then we are not describing a relationship.

For instance, consider the frequency distribution for political party constructed in Table 3-1 or the corresponding bar chart and pie chart displayed respectively as Figures 3-1 and 3-2 There is essentially only one variable involved here: “political party affiliation.” The frequency table and graphical displays enable us to make a comparison of the proportions of each category of the variable “political party affiliation,” but since we are only considering one variable, we cannot say that we are describing any relationship.

As another illustration, let us consider the box plot for yearly income constructed in Figure 2-5. There is essentially only one variable involved here: “yearly income.” Since we are only considering one variable, there is no way to refer to the relationship between two variables. However, the box plot might possibly be used to compare of the proportion of incomes in one range with the proportion of incomes in another range, or to compare the median income with some particular value.

It would be proper to say that Figure 2-8 can be used to compare the distribution of weights among orange types A, B, C, D, and E. It would also be proper to say that Figure 2-8 can be used to describe the

relationship between type of orange and weight. However, it would not be proper to say that Figure 2-8 can be used to describe the relationship between orange types A, B, C, D, and E, since A, B, C, D, and E are categories, not variables. Also, it would not be proper to say that Figure 2-8 can be used to compare type of orange and weight, since "type of orange" and "weight" are not commensurate variables.

Three different situations involving two variables are possible: one variable is qualitative and the other is quantitative, both variables are qualitative, or both variables are quantitative. Figure 2-8 provided us an example of the situation where one variable is qualitative and the other is quantitative; contiguous box plots were used to display the relationship between the two variables. However, we could alternatively have chosen contiguous stem-and-leaf displays, contiguous histograms, contiguous frequency polygons, etc. Recall how the contiguous dot plots constructed in Self-Test Problem 3-1(a) were used to compare the distribution of "Number of Children" in the SURVEY DATA for the rural, suburban, and urban areas; this is essentially what describes the relationship between "Residence" and "Number of Children." An alternative graphical display is the separate frequency polygon for each of the three areas of residence (all placed on the same graph) in Self-Test Problem 3-1(e). Having now considered examples of the situation where one variable is qualitative and the other is quantitative, we next want to consider examples of the other two possible situations.

Self-Test Problem 7-1. For each of several residents in a city, data is recorded which includes the variable type of job (white collar or blue collar), treated as qualitative-dichotomous, and the variable diastolic blood pressure, treated as quantitative.

- (a) Suppose the data are to be used to study a possible relationship between type of job and diastolic blood pressure. Is this goal stated properly? If not, why not?
- (b) Suppose the data are to be used to study the difference between type of job and diastolic blood pressure. Is this goal stated properly? If not, why not?
- (c) Suppose the data are to be used to study a possible relationship between diastolic blood pressure and blue collar workers. Is this goal stated properly? If not, why not?
- (d) Suppose the data are to be used to study the difference in diastolic blood pressure between white collar and blue collar workers. Is this goal stated properly? If not, why not?
- (e) How can the distribution of diastolic blood pressure be compared graphically for white collar workers and blue collar workers?

Self-Test Problem 7-2. The major source of income is recorded for each of many undergraduate college students.

- (a) Suppose the data are to be used to study the relationship between sources of income. Is this goal stated properly? If not, why not?
- (b) Suppose the data are to be used to compare the frequency of the different sources of income. Is this goal stated properly? If not, why not?
- (c) What graphical display can be used to compare the frequency of the different sources of income?

We shall now consider situations involving two variables where both variables are qualitative. Recall that Table 3-1 is a frequency distribution for the "Political Party" in the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1. Figure 3-1 displays this data in a bar chart. However, let us suppose that we want to simultaneously consider the variables "Residence" and "Political Party" in the SURVEY DATA. Both of these variables are qualitative. A contingency table can be used to display the frequencies for two qualitative variables.

Table 7-1 is a *contingency table* consisting of three rows representing "Residence" and four columns representing "Political Party." Each cell entry of this contingency table will contain a raw frequency; this raw frequency is the number of people who are from the area of residence represented by the corresponding row and who have a political party affiliation represented by the corresponding column. Use the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, to enter the appropriate raw frequency in each cell of Table 7-1.

Frequencies displayed along the right and lower edges just outside a contingency table are row and column totals. In Table 7-1, each row total is the total number of people in the corresponding area of residence, and each column total is the total number of people in the corresponding political party affiliation. The total number of people is displayed in the lower right corner just outside the contingency table and is called the grand

total. Enter the row totals, column totals, and grand total in Table 7-1. This table should now 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.

One way to display graphically the data of Table 7-1 is to modify the bar graph displayed as Figure 7-1a (taken from Figure 3-1). The height of each bar in Figure 7-1a represents the number of people in the corresponding political party category. For each political party category, we can divide the corresponding bar into pieces whose lengths are proportional to the raw frequencies for the three

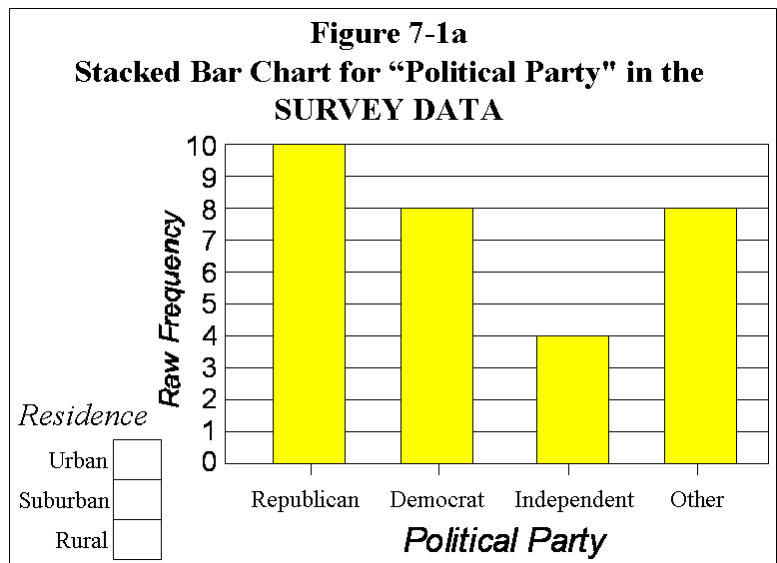
areas of residence: rural, suburban, and urban. A bar chart where each bar is divided into pieces whose lengths are proportional to frequencies corresponding to a second variable is called a *stacked bar chart*. For instance, we see that the height of the bar corresponding to the Republican party is 10, but this bar can be divided into three pieces. If the bottom piece were to represent Republicans from the rural area, it should have a length equal to 2. Similarly, if the middle piece were to represent Republicans from the suburban area, it should have a length equal to 4. This would leave a length of 4 for the top piece, indicating that 4 of the 10 Republicans are from the urban area. Each of the other bars can be divided in a similar manner. In each case, three different shades (or colors or design) can be used for respective pieces, one for the rural area, one for the suburban area, and one for the urban area. Legends could then be used to indicate which shading (or color or design) corresponds to which area of residence. Complete the stacked bar chart displayed as Figure 7-1a by appropriately dividing each bar into sections and completing the legend to distinguish between the different sections of the bars; you will find that two of the bars need only be divided into two pieces, since there are no residents from the urban area in each of these two cases. This stacked bar chart should be the same as that in the corresponding figure displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

Instead of using "Political Party" to label the horizontal axis of our stacked bar chart, we could have chosen to label the horizontal axis with "Residence" instead. With "Residence" labeled on the horizontal axis, each bar would be exactly the same height because the number of people from each area of residence is the same, 10. Complete the stacked bar chart displayed as Figure 7-1b by appropriately dividing each bar into sections and completing the legend to distinguish between the different sections of the bars. This stacked bar chart should be the same as that in the corresponding figure displayed in the section titled **SURVEY DATA Summaries for Data Set 1-1** at the end of this unit.

Figure 7-1a and Figure 7-1b each

Table 7-1
Contingency Table for "Residence" and "Political Party" in the SURVEY DATA

		Political Party			
		<i>Republican</i>	<i>Democrat</i>	<i>Independent</i>	<i>Other</i>
Residence	<i>Rural</i>				
	<i>Suburban</i>				
	<i>Urban</i>				



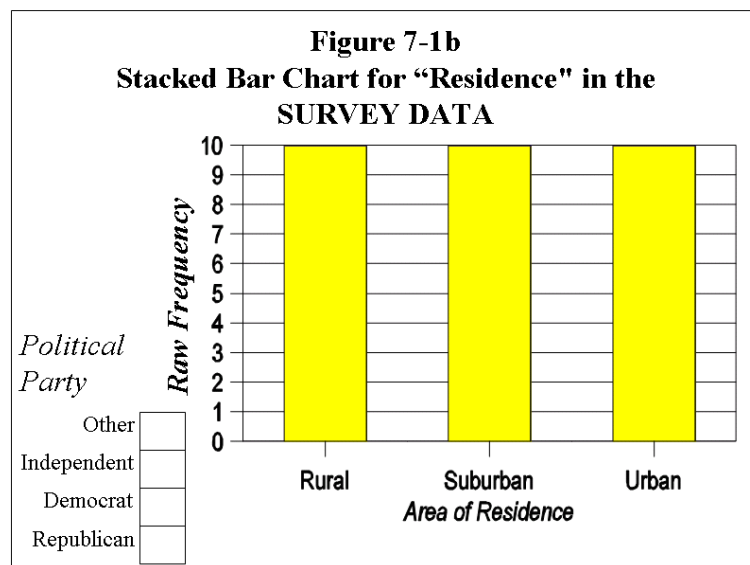
provide us with an example of a graphical display involving the two qualitative variables "Residence" and "Political Party." Let us consider Figure 7-1b for a moment. We can think of using this stacked bar chart to make a comparison of the three different areas of residence with regard to political party affiliation, or we can think of using this stacked bar chart to describe the relationship between the two qualitative variables "Residence" and "Political Party." These are just two different ways of thinking about the same thing. Figure 7-1b allows us to make a comparison of the proportion of a particular political party in one area of residence to the proportion in another area of residence. However, focusing on how the proportion of a particular political party differs from one area of residence to another is precisely what we need to do in order to describe the relationship between "Residence" and "Political Party," we must. Consequently, making a comparison of the areas of residence with regard to the distribution of political parties can be the same thing as describing the relationship between "Residence" and "Political Party." A similar comment can be made about Figure 7-1a; that is, making a comparison of political parties with regard to the distribution of areas of residence can essentially be considered the same thing as describing the relationship between "Residence" and "Political Party."

The way in which the data is for the two qualitative variables are collected often influences whether one thinks in terms of a making comparison or in terms of describing a relationship. If the data for Figure 7-1b were collected by selecting from each area of residence a fixed number of individuals, one might be more inclined to think in terms of making a comparison of the areas of residence with regard to political parties. On the other hand, if individuals were selected without regard to either of the variables, one might be more inclined to think in terms of describing a relationship between area and party affiliation.

When looking at Figure 7-1b, it is proper to refer to the difference between the rural and suburban areas, but it is not proper

to talk about a relationship between the rural and suburban areas; this is because the rural area and the suburban area are not two different variables, they are two different categories of the same variable. On the other hand, it is proper to refer to the relationship between area of residence and political party affiliation, but it is not proper to talk about a difference between area of residence and political party affiliation; this is because residence and political party affiliation are two distinct, but not commensurate, variables.

If we were going to make a choice of using Figure 7-1a or Figure 7-1b to describe the possible relationship between area and party affiliation, we might be more inclined to choose Figure 7-1b, since the bars all have the same height, making it easier to compare the three areas of residence. From this stacked bar chart, we see that we might describe the relationship by saying that the urban area has the highest proportion of democrats, the urban and suburban areas each have a higher proportion of republicans than the rural area, and the rural area has the highest proportion of independents. Whether or not the apparent relationship we might see in a stacked bar chart is significant is a subject to be addressed in a future unit.



Self-Test Problem 7-3. For a large number of people, data is recorded which includes location of residence (rural, urban) and whether the soft drink Napple was ever purchased (yes, no), both variables treated as qualitative-nominal.

- (a) What kind of table can be used to summarize and display the data?
- (b) What graphical display can be used to study a possible relationship between residence and whether Napple was ever purchased?

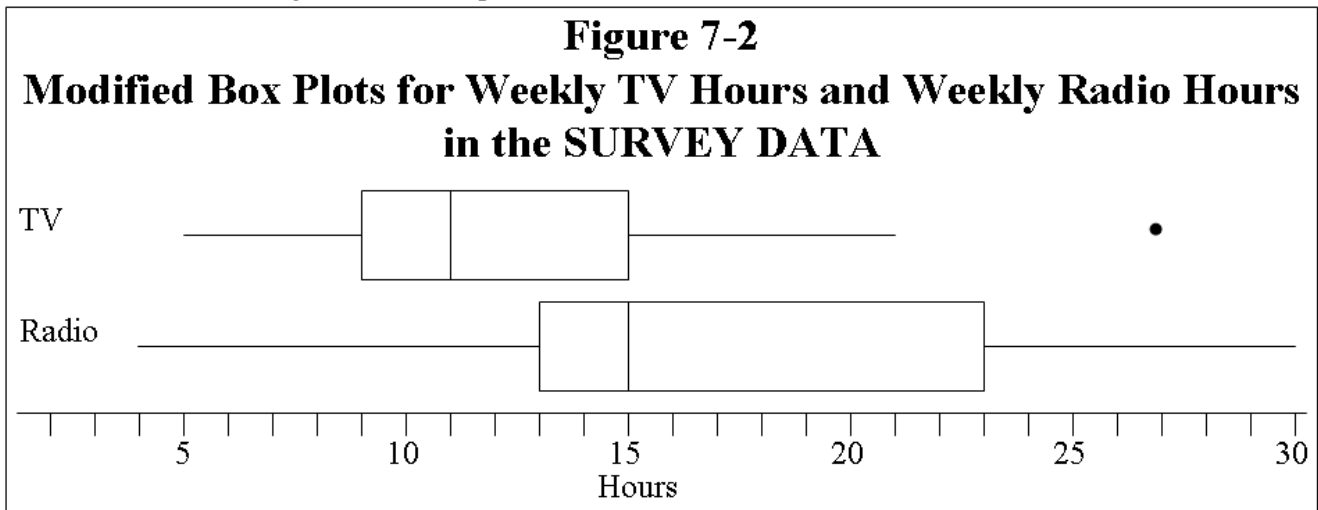
Self-Test Problem 7-4. Monthly income is recorded for each of several high school students.

- (a) Suppose the data are to be used to study the distribution monthly incomes among high school students. Is this goal stated properly? If not, why not?
- (b) Suppose the data are to be used to study a possible relationship between incomes among high school students. Is this goal stated properly? If not, why not?
- (c) What graphical display can be used to study the distribution of monthly incomes among high school students?

Self-Test Problem 7-5. For each of several residents in a city, data is recorded which includes the variable type of job (white collar or blue collar) and the variable area of residence (rural or urban), both treated as qualitative-dichotomous.

- (a) Suppose the data are to be used to study a possible relationship between white collar workers and blue collar workers. Is this goal stated properly? If not, why not?
- (b) Suppose the data are to be used to study a possible relationship between type of job and area of residence. Is this goal stated properly? If not, why not?
- (c) Suppose the data are to be used to study the difference between type of job and area of residence. Is this goal stated properly? If not, why not?
- (d) Suppose the data are to be used to study the difference in distribution of job types between the rural and urban areas. Is this goal stated properly? If not, why not?
- (e) How can the possible relationship between type of job and area of residence be displayed graphically?

Finally, we consider situations involving two variables where both variables are quantitative. Making a comparison of the distribution of the two quantitative variables is not the same as describing the relationship between the two quantitative variables. To illustrate, let us consider the two quantitative variables “weekly TV hours” and “weekly radio hours” with the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1. We may possibly be interested in making a comparing of the distribution of these two variables, but this would not be the same as describing the relationship between these two variables.



One way to compare the distribution of these two variables is with contiguous modified box plots, such as those displayed in Figure 7-2. (Of course, we may alternatively choose contiguous stem-and-leaf displays, contiguous histograms, or contiguous frequency polygons.) From Figure 7-2, it appears that the distribution for “weekly radio hours” is centered at a higher value than that for “weekly TV hours,” the distribution for “weekly radio hours” has somewhat more dispersion than that for “weekly TV hours,” and the distribution for “weekly radio hours” is a little less skewed than that for “weekly TV hours.”

Figure 7-2 does not enable us to describe the relationship between “weekly TV hours” and “weekly radio hours,” because we cannot see how the values of one variable changes with the values of the other variable. In order to study a possible relationship between “weekly TV hours” and “weekly radio hours,” we

need a graphical display which shows how changes in one variable are influenced by changes in the other variable. One way to visualize the relationship between two quantitative variables is to look at a *scatter plot*. A scatter plot is constructed by labeling a horizontal axis with possible values of one quantitative variable and labeling a vertical axis with possible values of the other quantitative variable. Dots are then used to represent each pair of observations. If we think of one variable as being predicted from the other, then it is customary to label the vertical axis with the variable being predicted and label the horizontal axis with the variable from which predictions are made; otherwise, which axis is labeled with which variable is just a matter of personal preference.

Figure 7-3 is a scatter plot for the variables "weekly TV hours" and "weekly radio hours". Since we do not choose to think of one variable as being predicted from the other, the choice of which variable to label on which axis was arbitrary. The 30 dots in the scatter plot represent the 30 people from which the data was obtained. For each person, a dot has been placed on the scatter plot above the value on the horizontal axis corresponding to that person's weekly radio hours and to the right of the value on the vertical axis corresponding to that person's weekly television hours. From this scatter plot, we might describe the relationship by saying that weekly TV hours appears to decrease as weekly radio hours increases.

Returning momentarily to Figure 7-2, note that the only reason that it made sense to compare the distribution of the two quantitative variables "weekly TV hours" and "weekly radio hours" with the contiguous box plots is because these two variables are commensurate, that is, they are both measured with hours. It would make no sense to talk about comparing the distribution of the two quantitative variables "age" and "yearly income" or to construct contiguous box plots of these two variables, because these two variables are not commensurate; that is, they are measured on completely different scales, with "age" measured in years and "yearly income" measured in dollars. However, we can certainly study a possible relationship between "age" and "yearly income."

Figure 7-4 is a scatter plot for the variables "age" and "yearly income." The 30 dots in the scatter plot represent the 30 people from which the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, was obtained. For each person, a dot has been placed on the scatter plot above the value on the horizontal axis corresponding to that person's age and to the right of the value on the vertical axis corresponding to that person's yearly income. Not only may we be interested in studying the relationship between "age" and "yearly income," but we might specifically be interested, say, in predicting "yearly income" from "age." This being the case, we decided to label the vertical axis with income (the variable being predicted) and label the horizontal axis with

Figure 7-3

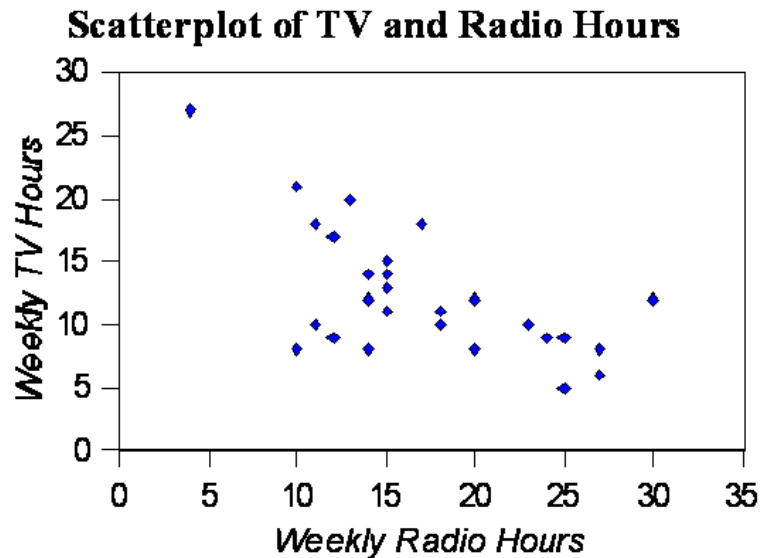
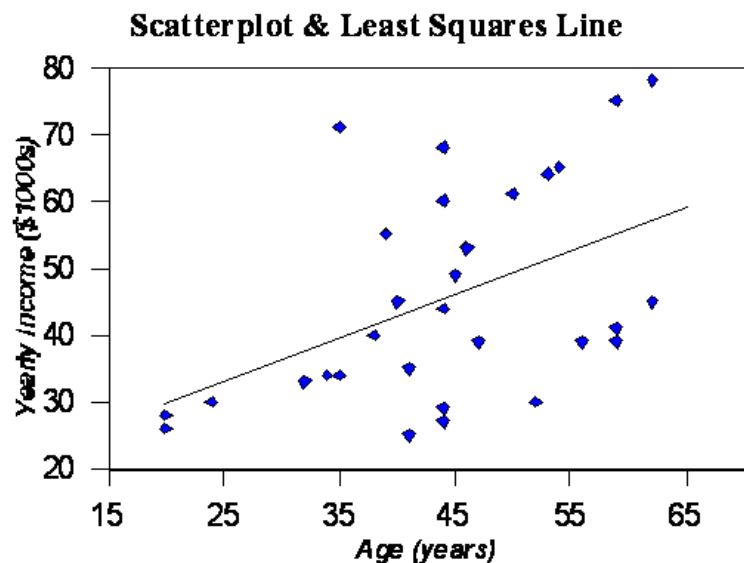


Figure 7-4



age (the variable from which predictions are made). You will notice that Figure 7-4 includes something called a least squares line with the scatter plot. This least squares line enables us to make predictions and will be discussed in a future unit. From this scatter plot, we might describe the relationship by saying that yearly income appears to increase as age increases.

Self-Test Problem 7-6. Data is recorded for several residents in a state, including weekly hours spent listening to the radio, weekly hours spent watching television, and yearly income. All three variables are treated as quantitative. Decide whether the goal in each situation is stated properly.

- (a) Suppose the data are to be used to study a possible relationship between time spent listening to the radio and time spent watching television. Is this goal stated properly? If not, why not?
- (b) Suppose the data are to be used to study the difference in distribution between time spent listening to the radio and time spent watching television. Is this goal stated properly? If not, why not?
- (c) Suppose the data are to be used to study a possible relationship between time spent listening to the radio and yearly income. Is this goal stated properly? If not, why not?
- (d) Suppose the data are to be used to study the difference in distribution between time spent listening to the radio and yearly income. Is this goal stated properly? If not, why not?
- (e) Suppose the data are to be used to study a possible relationship between time spent watching television and yearly income. Is this goal stated properly? If not, why not?
- (f) Suppose the data are to be used to study the difference in distribution between time spent watching television and yearly income. Is this goal stated properly? If not, why not?
- (g) Indicate how each of the following can be displayed graphically:
 - (i) the possible relationship between time spent listening to the radio and time spent watching television,
 - (ii) the difference in distribution between time spent listening to the radio and time spent watching television,
 - (iii) the possible relationship between time spent listening to the radio and yearly income,
 - (iv) the possible relationship between time spent watching television and yearly income.

Answers to Self-Test Problems

- 7-1** (a) This is stated properly. (b) This is not stated properly, since type of job and diastolic blood pressure are not commensurate. (c) This is not stated properly, since “blue collar worker” is a category, not a variable. (d) This is stated properly. (e) Two box plots displaying diastolic blood pressure, one for white collar and one for blue collar, would be appropriate (as would two stem-and-leaf displays, or two histograms, etc.).
- 7-2** (a) This is not stated properly, since only one variable “major source of income” is involved. (b) This is stated properly. (c) A bar chart with bars representing major source of income would be appropriate, since “source of income” is qualitative. (A pie chart could also be used.)
- 7-3** (a) A contingency table would be appropriate, since residence and purchase of Napple are both qualitative. (b) A stacked bar chart would be appropriate to display the relationship between two qualitative variables.
- 7-4** (a) This is stated properly. (b) This is not stated properly, since only one variable “monthly income” is involved. (c) A box plot (or a stem-and-leaf display, or a histogram, etc.) of the monthly incomes would be appropriate.
- 7-5** (a) This is not stated properly, since “white collar worker” and “blue collar worker” are categories, not variables. (b) This is stated properly. (c) This is not stated properly, since type of job and area of residence are not commensurate. (d) This is stated properly. (e) A stacked bar chart would be appropriate to display the relationship between two qualitative variables.
- 7-6** (a) This is stated properly. (b) This is stated properly. (c) This is stated properly. (d) This is not stated properly, since time spent listening to the radio and yearly income are not commensurate. (e) This is stated properly. (f) This is not stated properly, since time spent watching television and yearly income are not commensurate. (g) For (i), (iii), and (iv), a scatter plot would be appropriate; for (ii), contiguous box plots (or stem-and-leaf displays, or histograms, etc.) would be appropriate.

Summary

Making a *comparison* generally implies that different observations of the same numerical value(s) or commensurate numerical values (i.e., values measured on the same scale) are being compared. Describing a relationship implies describing how changes in one variable are influenced by changes in a second variable. Depending on the situation, making a comparison may involve only one variable or more than one variable; however, describing a relationship must always involve (at least) two variables. The distinction between making a comparison and describing a relationship is not always sharp. Describing a relationship can sometimes be done by making comparisons, but making comparisons does not always imply that a relationship is being described. If we are making comparisons in order to describe how changes in one variable are influenced by changes in a second variable, then we are really describing the relationship between two variables. On the other hand, if we are making comparisons that do not involve how the values of one variable influence a second variable, then we are not describing a relationship.

The relationship between a qualitative variable and a quantitative variable can be visually displayed with contiguous box plots, contiguous stem-and-leaf displays, contiguous histograms, etc.; this relationship is defined by how the distribution of the quantitative variable is different for each category of the qualitative variable. Data observed on two qualitative variables can be organized into a *contingency table*, and a *stacked bar chart* can be used to display visually the relationship between the two qualitative variables; this relationship is defined by how the distribution of categories for one qualitative variable changes across the categories of the other qualitative variable. A *scatter plot* can be used to display visually the relationship between two quantitative variables; this relationship is defined by how changes in the value of one variable are influenced by changes in the value of the other variable. If two (or more) quantitative variables are commensurate, then their distributions can be compared with contiguous box plots, contiguous stem-and-leaf displays, contiguous histograms, etc.; however, these graphical displays provide no information about a possible relationship.

SURVEY DATA Summaries for Data Set 1-1

Table 7-1
Contingency Table for "Residence" and
"Political Party" in the SURVEY DATA

		Political Party				
		<i>Republican</i>	<i>Democrat</i>	<i>Independent</i>	<i>Other</i>	
Residence	<i>Rural</i>	2	1	3	4	10
	<i>Suburban</i>	4	1	1	4	10
	<i>Urban</i>	4	6	0	0	10
		10	8	4	8	30

