

# Unit 9

## Describing Relationships in Scatter Plots and Line Graphs

Objectives:

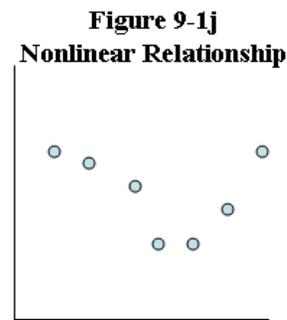
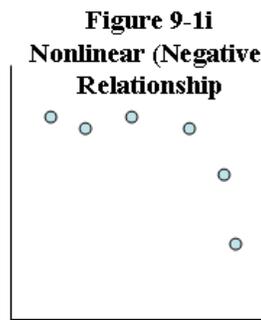
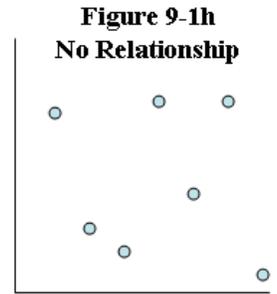
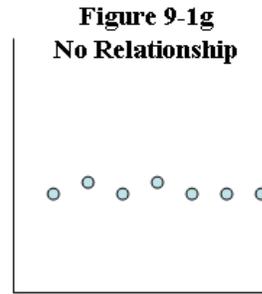
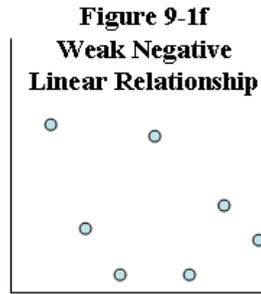
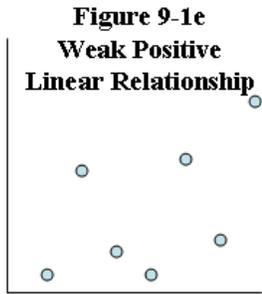
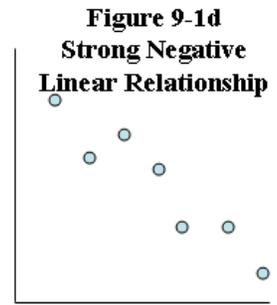
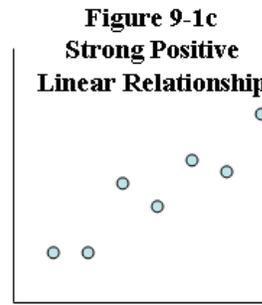
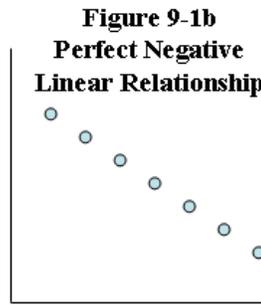
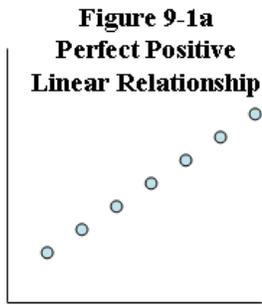
- To construct and interpret a scatter plot or line graph for two quantitative variables
- To recognize linear relationships, non-linear relationships, or independence between two quantitative variables
- To decide whether a linear relationship is negative or positive and whether the linear relationship appears strong or weak
- To construct and interpret a line graph for displaying changes over time of a quantitative variable
- To describe changes over time of a quantitative variable with an index number

We have considered how a stacked bar chart displays the relationship between two qualitative variables; we now want to consider how a scatter plot displays the relationship between two quantitative variables. If we think of one variable as being predicted from another, 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. Recall that Figure 7-3 is a scatter plot for the variables "Weekly TV Hours" and "Weekly Radio Hours". Since we did 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. When we looked at this scatter plot previously, we described the relationship by saying that weekly TV hours appears to decrease as weekly radio hours increases.

When a relationship exists between two quantitative variables, one of our first goals is to decide whether the relationship is *linear* or *nonlinear*. Roughly speaking, a linear relationship is said to exist between two quantitative variables when a straight line on a graph can be used with at least some reasonable degree of accuracy to predict the values of one quantitative variable from the values of the other quantitative variable. A nonlinear relationship is said to exist between two quantitative variables when a curve can be used to predict the values of one quantitative variable from the values of the other quantitative variable with considerably more accuracy than a straight line. To illustrate some of the many different aspects of a relationship between two quantitative variables, we shall consider Figures 9-1a to 9-1j.

Figures 9-1a and 9-1b are each a scatter plot illustrating a perfect linear relationship between two quantitative variables. In both figures, the data points all lie exactly on a straight line; that is, we can predict perfectly the value of one variable from the other. In Figure 9-1a, we observe that as one variable increases in value, the other variable increases in value; in Figure 9-1b, we observe that as one variable increases in value, the other variable decreases in value. When two variables tend to increase together, we say the variables have a *positive* relationship; when one variable tends to decrease as another variable increases, we say the variables have a *negative* relationship. Consequently, Figure 9-1a has been labeled as illustrating a perfect positive linear relationship, and Figure 9-1b has been labeled as illustrating a perfect negative linear relationship. Very rarely, if ever, do we observe in real data a relationship where one variable can perfectly be predicted from another. With real data, some component of random variation is always present.

Figures 9-1c and 9-1d illustrate scatter plots of a type more likely to be observed with real data than scatter plots of the type in Figures 9-1a and 9-1b. In Figures 9-1c and 9-1d, the data points lie relatively close to, but not exactly on, a straight line. We see that one variable cannot be predicted perfectly from the other variable, but it does look as if one variable can be predicted from the other with a relatively strong degree of accuracy. In Figure 9-1c, we observe that as one variable increases in value, the other variable tends to increase in value; in Figure 9-1d, we observe that as one variable increases in value, the other variable tends to decrease in value. Consequently, Figure 9-1c has been labeled as illustrating a strong, positive, linear relationship, and Figure 9-1d has been labeled as illustrating a strong, negative, linear relationship.



In Figures 9-1e and 9-1f, we find that the data points seem widely scattered around a straight line. In both scatter plots, it looks as if one variable can be predicted from the other with some degree of accuracy, but not with as much accuracy as in the case of Figures 9-1c and 9-1d. In Figure 9-1e, we observe that as one variable increases in value, the other variable tends to increase in value; in Figure 9-1f, we observe that as one variable increases in value, the other variable tends to decrease in value. Consequently, Figure 9-1e has been labeled as illustrating a weak, positive, linear relationship, and Figure 9-1f has been labeled as illustrating a weak, negative, linear relationship.

The scatter plots of Figures 9-1g and 9-1h illustrate data where as one variable increases there does not seem to be any discernable change in the behavior of the other variable. In Figure 9-1g, it appears that as the variable on the horizontal scale changes, the variable on the vertical scale seems to vary randomly within a relatively small range without tending to increase or decrease significantly. In Figure 9-1h, it appears that as the variable on the horizontal scale changes, the variable on the vertical scale seems to vary randomly within a relatively large range without tending to increase or decrease significantly. In both scatter plots, we find no evidence that the value of one variable is significantly influenced by changes in the value of the other variable; in other words, there appears to be no relationship between the two variables in each of these scatter plots.

We have previously stated that when two variables show no relationship, we say that the variables are *independent*. We have already seen that when two qualitative variables are independent, the distribution for one qualitative variable is the same for each of the categories of the other qualitative variable; this can also be said for two quantitative variables or for one qualitative variable and one quantitative variable. Consequently, we can say that Figures 9-1g and 9-1h are scatter plots illustrating independence between two quantitative variables.

In Figures 9-1a to 9-1h, we have seen scatter plots illustrating linear relationships of varying degrees of strength as well as scatter plots illustrating independence. Figures 9-1i and 9-1j are scatter plots which illustrate

nonlinear relationships. The scatter plot of Figure 9-1i illustrates a situation where as one variable increases in value, the other variable tends to decrease; this can also be said about Figure 9-1d, but the relationship we see in Figure 9-1d can be described with a straight line, whereas the relationship we see in Figure 9-1i is better described by a curve than by a straight line. Consequently, we can call the relationship we see in Figure 9-1d linear, but we would have to call the relationship we see in Figure 9-1i nonlinear. Linear relationships are of only two types: positive and negative. However, many different types of nonlinear relationships are possible, but there is no easy way to classify all of them. Figure 9-1j illustrates a nonlinear relationship which is quite different from the nonlinear relationship illustrated by Figure 9-1i. In Figure 9-1j, we see that as the variable on the horizontal axis increases in value, the other variable will sometimes tend to increase and sometimes tend to decrease.

The scatter plot of Figure 7-3 seems to suggest that there is a roughly negative linear relationship between weekly TV hours and weekly radio hours. Making decisions about whether or not the relationship we see in any graphical display should be considered significant is a subject we shall address in a future unit. We are not yet prepared to discuss exactly how to make this decision; for now, we simply use our best judgment in deciding whether or not a relationship appears to exist.

**Self-Test Problem 9-1.** For each pair of quantitative variables, describe the type of scatter plot likely to be observed if data were taken.

- (a) For each of several school children in grades 1 through 6, the variable "height" is measured in inches, and the variable "spelling ability" is measured as a score from 0 to 100 on a particular spelling test.
- (b) For each of several undergraduate college students, the variable "height" is measured in inches, and the variable "spelling ability" is measured as a score from 0 to 100 on a particular spelling test.
- (c) For each of several adults, the variable "time to go through a particular maze" is measured in minutes, and the variable "practice time" is measured in hours ranging from 0 to 10.
- (d) For each of several adults, the variable "time to go through a particular maze" is measured in minutes, and the variable "practice time" is measured in hours ranging from 0 to 90.
- (e) Each day at 3:00 pm in a certain city, the variable "temperature" is measured in degrees Fahrenheit, and the variable "temperature" is measured in degrees Centigrade; this is done for several days.

One noteworthy situation when describing and displaying the relationship between two quantitative variables is when one of the two quantitative variables is time. The study of how one or more variables change with time occurs often. The data of Table 9-1 provide an illustration.

The three leftmost columns of Table 9-1 contain (imaginary) prices and quantities sold for an Econo color printer over a five-year period. Let us consider how we might graphically display this data. If we wanted to focus a possible relationship between price and quantity sold, a scatter plot would be appropriate, since the variable "price" and the variable "quantity" are each quantitative.

**Table 9-1**  
**Prices and Quantities Sold for**  
**an Econo Color Printer**

<u>Year</u>	<u>Price</u> <u>(\$)</u>	<u>Millions</u> <u>Sold</u>	<u>Price</u> <u>Index</u>	<u>Quantity</u> <u>Index</u>
2007	1400	0.50	100.0	100.0
2008	1000	0.75	71.4	150.0
2009	500	1.20	35.7	240.0
2010	300	1.50	21.4	300.0
2011	200	2.30	14.3	460.0

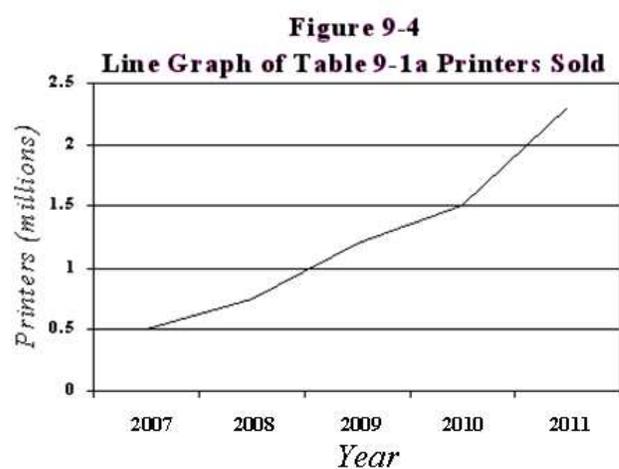
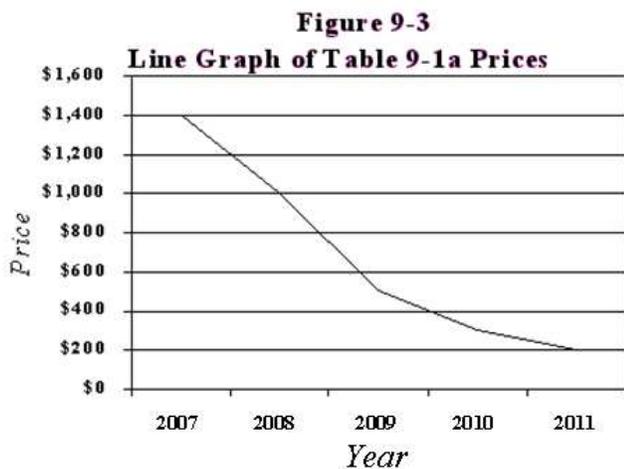
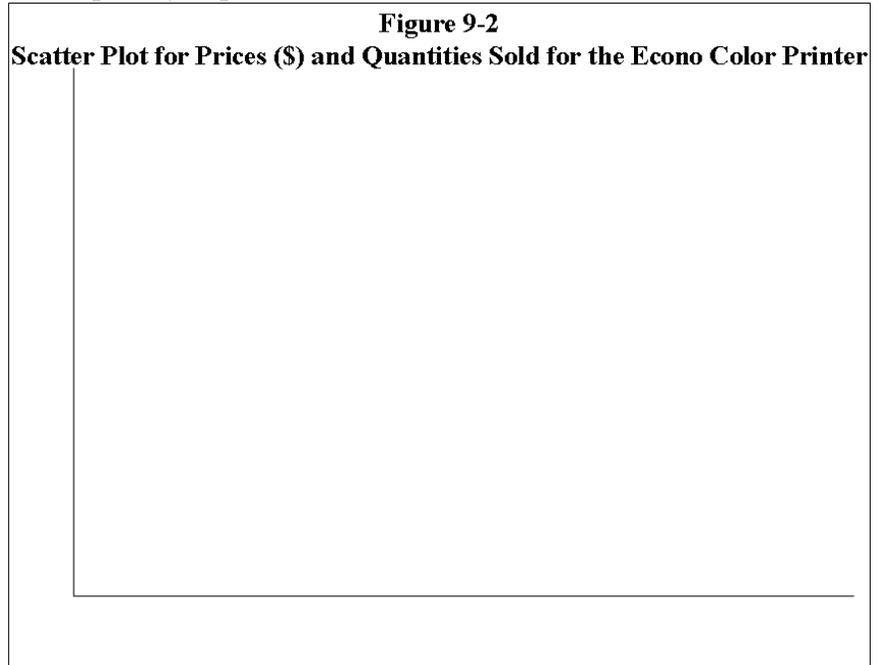
Complete the scatter plot in Figure 9-2 and underneath the scatter plot describe the type of relationship, if any, that appears to exist between price and quantity; you may choose either variable for the horizontal axis and the other variable for the vertical axis, but be sure to label each axis completely. (You should see a negative relationship which looks nonlinear.)

Suppose we wanted to focus on how the prices change over time. We might consider a scatter plot to display how the prices have changed over time, since the variable "time" and the variable "price" are each

quantitative; also, it is natural to think of predicting how price might change over time, which implies we should put time on the horizontal axis. In order to help visualize how time influences price, we could connect consecutive points over time on the scatter plot with line segments. In general, whenever we want to display how a quantitative variable changes over time, time is always scaled on the horizontal axis, and with only one measurement for each time period, we can sequentially connect the points on a scatter plot to create a *line graph*. The line graph is popular as a visual display to trace how the value of a variable changes through time. Figure 9-3 is a line graph displaying the change over time in the prices of Table 9-1, and Figure 9-4 is a line graph displaying the change over time in the quantity of printers sold of Table 9-1.

Often when studying how a variable changes over time, the percentage of change is of interest. In the past, we have used percentage as a way to express relative frequency (i.e., percentage found from the ratio of a raw frequency to a total frequency). A percentage representing a relative frequency must always be between 0% and 100%, but a percentage representing change over time may exceed 100%. A *percentage representing change* is the ratio of one value of a quantitative variable to another value of the same variable at a different time.

Let us return to the prices of Table 9-1. We see that the price for the printer in 2010 is \$300 and that the price for the printer in 2011 is \$200. Since  $200/300 = 2/3 = 66.7\%$ , one way to look at this change is to say that the 2011 price is two thirds of the 2010 price, or using a percentage, we would say that the 2011 price is 66.7% of the 2010 price.



Looking at the quantities of printers sold Table 9-1, we see that the quantity sold in 2010 is 1.50 million and that the quantity sold in 2011 is 2.30 million. Since  $2.30/1.50 = 1.535 = 153.3\%$ , one way to look at this change is to say that the 2011 quantity is 153.3% of the 2010 quantity.

A percentage representing change is called an *index number*; the time period corresponding to the denominator of the ratio is called the *base time period*. Since it is understood that an index number is a percentage, often the percent sign (%) is not included. For instance, we could say that the price index number for 2011 with 2010 as the base year is 66.7; we could also say that the quantity index number for 2011 with 2010 as the base year is 153.3.

When an index number is smaller than 100, then the change is a decrease; when an index number is greater than 100, then the change is an increase. Of course, the index number 100 implies no change at all. Since the 2011 price index with 2010 as the base year is 66.7, the price decreased from 2010 to 2011. Since the 2011 quantity index with 2010 as the base year is 153.3, the quantity increased from 2010 to 2011. The percentage of an increase or a decrease is the ratio of the amount of change to the value in the base time period. For instance, the price decreased by \$100 from 2010 to 2011, or in other words, we would say that the price decreased by  $100/300 = 0.333 = 33.3\%$  from 2010 to 2011; note that the percent of decrease is the amount that the index number 66.7 is below 100. Similarly, the quantity increased by 0.8 million from 2010 to 2011, or in other words, we would say that the quantity increased by  $0.8/1.50 = 0.533 = 53.3\%$  from 2010 to 2011; note that the percent of increase is the amount that the index number 153.3 is above 100.

By selecting one specific base time period, we can generate a sequence of index numbers for future time periods, which allows us to observe how a variable changes over time relative to a common standard; this is useful when we want to compare changes over

<u>Year</u>	<u>Alloy A</u>		<u>Alloy B</u>		<u>Alloy C</u>	
	<u>Price per Ton</u>	<u>Millions of Tons Sold</u>	<u>Price per Ton</u>	<u>Millions of Tons Sold</u>	<u>Price per Ton</u>	<u>Millions of Tons Sold</u>
2007	\$10.51	5.5	\$5.50	2.8	\$7.55	3.0
2008	\$11.10	6.0	\$5.53	3.0	\$7.75	3.0
2009	\$12.00	6.0	\$5.60	3.1	\$8.10	3.1

time for more than one quantitative variable not measured on the same scales. In Table 9-1, find each price index with 2007 as the base year, and find each quantity index with 2007 as the base year. Check that your results are exactly as those displayed in the two rightmost columns of Table 9-1.

**Self-Test Problem 9-2.** Use the prices and quantities sold for the color printer of Table 9-1 to

- find and interpret the price index number for 2011, with 2009 as the base year.
- find and interpret the quantity index number for 2011, with 2009 as the base year.

**Self-Test Problem 9-3.** Table 9-2 contains the prices and quantities sold for three different alloys over a period of three years.

- With alloy C, find and interpret the price index number for 2009, with 2007 as the base year.
- With alloy B, find and interpret the quantity index number for 2009, with 2008 as the base year.

### Answers to Self-Test Problems

- 9-1** (a) Since the older children in the higher grades are more likely to be better at spelling and are also more likely to be taller, then we would expect to observe a scatter plot similar to that of either Figure 9-1c or Figure 9-1e. (b) Since we believe there is no relationship between a college student's spelling ability and height, then we would expect to observe a scatter plot similar to that of either Figure 9-1g or Figure 9-1h. (c) Since it is likely that the time to go through the maze will decrease as practice time increases, then we would expect to observe a scatter plot similar to that of either Figure 9-1d or Figure 9-1f. (d) Although it is likely that the time to go through the maze will decrease as practice time increases, there will be for each person a point at which more practice will not decrease the time to go through the maze; as a result, we would expect to observe a scatter plot exhibiting decreasing curve which levels off at some point. (e) Since degrees Fahrenheit and degrees Centigrade are just different units of measurement for exactly the same quantity, we would expect to observe a scatterplot similar to that of Figure 9-1a.
- 9-2** (a) The 2011 price is 40.0% of the 2009 price; in other words, the price decreased by 60.0% from 2009 to 2011. (b) The 2011 quantity is 191.7% of the 2009 quantity; in other words, the quantity increased by 91.7% from 2009 to 2011.
- 9-3** (a) The 2009 price for alloy C is 107.3% of the 2007 price; in other words, the price for alloy C increased by 7.3% from 2007 to 2009. (b) The 2009 quantity for alloy B is 103.3% of the 2008 quantity; in other words, the quantity for alloy B increased by 3.3% from 2008 to 2009.

### Summary

A scatter plot is a natural way to display visually the relationship between two quantitative variables and is constructed by first 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 to 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.

A *linear* relationship is said to exist when a straight line on a graph can be used with at least some reasonable degree of accuracy to predict the values of one quantitative variable from another quantitative variable. A *nonlinear* relationship is said to exist when a curve can be used with considerably more accuracy than a straight line to predict the values of one quantitative variable from another quantitative variable. When two variables tend to increase together, we say the variables have a *positive* relationship; when one variable tends to decrease as another variable increases, we say the variables have a *negative* relationship.

When considering the changes in a quantitative variable over time, we can sequentially connect the points on a scatter plot to create a *line graph*, with time always scaled on the horizontal axis. A *percentage representing change* is the ratio of one value of a quantitative variable to another value of the same variable at a different time; such a percentage is called an *index number*; the time period corresponding to the denominator is called the *base time period*. Since it is understood that an index number is a percentage, we often omit the percent sign (%), and the percentage of increase or decrease can be determined by how much the index number lies above or below 100.