CHAPTER 4: CORRELATION AND SCATTER PLOTS

- To study relationships between variables, we must measure the variables on the same group of individuals.
- If we think that a variable x may explain or even cause changes in another variable y, we call x an **explanatory variable** and y a **response variable.**
- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph. Always plot the explanatory variable, if there is one, on the x axis of a scatterplot.
- Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.
- In examining a scatterplot, look for an overall pattern showing the **direction, form,** and **strength** of the relationship, and then for **outliers** or other deviations from this pattern.
- **Direction:** If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).
- **Form: Linear relationships,** where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and **clusters** are other forms to watch for.
- **Strength:** The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.
- The **correlation r** measures the direction and strength of the linear association between two quantitative variables x and y. Although you can calculate a correlation for any scatterplot, r measures only straight-line relationships.
- Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to −1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.
- Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r.

EXERCISES

1. Because elderly people may have difficulty standing to have their heights measured, a study looked at predicting overall height from height to the knee. Here are data

Knee height, X	57.7	47.4	43.5	44.8	55.2
Height, Y	192.1	153.3	146.4	162.7	169.1

   - Make a scatterplot of the data and describe the pattern of association.
   - Compute the correlation coefficient and interpret it.

2. A student wonders if tall women tend to date taller men than do short women. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then she measures the next man each woman dates. Here are the data (heights in inches):

Women, X	66	64	66	65	70	65
Men, Y	72	68	70	68	71	65

a. Make a scatterplot of these data. Based on the scatterplot, do you expect the correlation to be positive or negative? Near ±1 or not?

b. Find the correlation r between the heights of the men and women. Do the data show that taller women tend to date taller men?

C H A P T E R 5: REGRESSION

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. You can use a regression line to **predict** the value of y for any value of x by substituting this x into the equation of the line.
- The **slope** b of a regression line ˆ$y = a + bx$ is the rate at which the predicted response ˆy changes along the line as the explanatory variable x changes. Specifically, b is the change in ˆy when x increases by 1.
- The **intercept** a of a regression line ˆ$y = a + bx$ is the predicted response ˆy when the explanatory variable $x = 0$. This prediction is of no statistical interest unless x can actually take values near 0.
- The most common method of fitting a line to a scatter plot is least squares. The **least-squares regression line** is the straight line ˆ$y = a + bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.
- The least-squares regression line of y on x is the line with slope $b = r\ sy/sx$ and intercept $a = y - bx$. This line always passes through the point (x, y).
- **Correlation and regression** are closely connected. The correlation r is the slope of the least-squares regression line when we measure both x and y in standardized units. The **square of the correlation** r^2 is the fraction of the variation in one variable that is explained by least-squares regression on the other variable.
- Correlation and regression must be **interpreted with caution. Plot the data** to be sure the relationship is roughly linear and to detect outliers and influential observations. A plot of the **residuals** makes these effects easier to see.
- Look for **influential observations,** individual points that substantially change the correlation or the regression line. Outliers in the x direction are often influential for the regression line.
- Avoid **extrapolation,** the use of a regression line for prediction for values of the explanatory variable far outside the range of the data from which the line was calculated.
- **Lurking variables** may explain the relationship between the explanatory and response variables. Correlation and regression can be misleading if you ignore important lurking variables.
- Most of all be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated. **High correlation does not imply causation.** The best evidence that an association is due to causation comes from an **experiment** in which the explanatory variable is directly changed and other influences on the response are controlled.

REVIEW QUESTIONS

ONE

How strongly do physical characteristics of sisters and brothers correlate? Here are data on the heights (in inches) of 11 adult pairs

(a) Use your calculator or software to find the correlation and the equation of the least-squares line for predicting sister's height from brother's height. Make a scatter plot of the data and add the regression line to your plot.

(b) Damien is 70 inches tall. Predict the height of his sister Tonya. Based on the scatter plot and the correlation *r*, do you expect your prediction to be very accurate? Why?

TWO

In the business statistics course the correlation between the students' total scores prior to the final examination and their final examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final-exam scores have mean 75 and standard deviation 8. The lecturer has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.

(a) What is the slope of the least-squares regression line of final-exam scores on pre-exam total scores in this course? What is the intercept?

(b) Use the regression line to predict Julie's final-exam score.

(c) Julie doesn't think this method accurately predicts how well she did on the final exam. Use *r* 2 to argue that her actual score could have been much higher (or much lower) than the predicted value.