Chapter 6 – Scatterplots, Association, and Correlation

   a) Either weight in grams or weight in ounces could be the explanatory or response variable. Greater weights in grams correspond with greater weights in ounces. The association between weight of apples in grams and weight of apples in ounces would be positive, straight, and perfect. Each apple’s weight would simply be measured in two different scales. The points would line up perfectly.

   b) At first, it appears that there should be no association between ice cream sales and air conditioner sales. When the lurking variable of temperature is considered, the association becomes more apparent. When the temperature is high, ice cream sales tend to increase. Also, when the temperature is high, air conditioner sales tend to increase. Therefore, there is likely to be an increase in the sales of air conditioners whenever there is an increase in the sales of ice cream. The association is positive, straight, and moderate. Either one of the variables could be used as the explanatory variable.

   c) Either shoe size or GPA could be the explanatory or response variable. There would be no association between shoe size and GPA of college freshmen, so the scatterplot would show no form, no direction, and no strength.

   d) Number of miles driven is the explanatory variable, and gallons remaining in the tank is the response variable. The greater the number of miles driven, the less gasoline there is in the tank. If a sample of different cars is used, the association is negative, straight, and moderate. If the data is gathered on different trips with the same car, the association would be strong.

   a) File size is the explanatory variable, and cost of the data plan is the response variable. The larger the file size, the greater the cost. The association is positive, straight, and moderately strong, since some companies charge more than others.

   b) Distance from lightning is the explanatory variable, and time delay of the thunder is the response variable. The farther away you are from the strike, the longer it takes the thunder to reach your ears. The association is positive, straight, and fairly strong, since the speed of sound is not a constant. Sound travels at a rate of around 770 miles per hour, depending on the temperature.

   c) Distance from the streetlight is the explanatory variable, and brightness is the response variable. The further away from the light you are, the less bright it appears. The association is negative, curved, and strong. Distance and light intensity follow an inverse square relationship. Doubling the distance to the light source reduces the intensity by a factor of four.
There is likely very little association between the weight of the car and the age of the owner. However, some might say that older drivers tend to drive larger cars. If that is the case, there may be a positive, straight, and very weak association between weight of a car and the age of its owner.

3. Scatterplots.
   a) None of the scatterplots show little or no association, although # 4 is very weak.
   b) #3 and #4 show negative association. Increases in one variable are generally related to decreases in the other variable.
   c) #2, #3, and #4 each show a linear association.
   d) #2 shows a moderately strong association.
   e) #1 and #3 each show a very strong association. #1 shows a curved association and #3 shows a straight association.

4. Scatterplots.
   a) #1 shows little or no association.
   b) #4 shows a negative association.
   c) #2 and #4 each show a straight association.
   d) #3 shows a moderately strong, curved association.
   e) #2 and #4 each show a very strong association, although some might classify the association as merely “strong”.

5. Bookstore sales.
   a) The scatterplot is to the right.
   b) There is a positive association between bookstore sales and the number of sales people working.
   c) There is a linear association between bookstore sales and the number of sales people working.
   d) There is a strong association between bookstore sales and the number of sales people working.
   e) The relationship between bookstore sales and the number of sales people working has no outliers.

Copyright © 2015 Pearson Education, Inc.
   a) The scatterplot is to the right.
   b) There is a positive association between price and capacity of disk drives.
   c) With the exception of the outlier, there is a linear association between price and capacity of disk drives.
   d) With the exception of the outlier, there is a strong association between price and capacity of disk drives.

7. Performance IQ scores vs. brain size.
   The scatterplot of IQ scores vs. Brain Sizes is scattered, with no apparent pattern. There appears to be little or no association between the IQ scores and brain sizes displayed in this scatterplot.

   Winning speeds in the Kentucky Derby have generally increased over time. The association between year and speed is moderately strong, and seems slightly curved, with a greater rate of increase in winning speed before 1970 and a smaller rate of increase after 1970, suggesting that winning speeds have leveled off over time.

   a) True.
   b) False. The correlation will remain the same.
   c) False. Correlation has no units.

10. Correlation facts II.
   a) False. This is a very weak association.
   b) False. Standardizing does not change the correlation.
   c) True.
11. Bookstore sales again.

This conclusion is not justified. Correlation does not demonstrate causation. The analyst argues that the number of sales staff working causes sales to be higher. It is possible (perhaps more plausible) that the store hired more people as sales increased. The causation may run in the opposite direction of the analyst’s argument.


The director’s conclusion is not justified. The lurking variable is the severity of the blizzard. Particularly severe blizzards require more snowplows, and they also prevent people from leaving home, where they are more likely to make online purchases, especially since they have to leave home to go to a store.

13. Firing pottery.

a) A histogram of the number of broken pieces is at the right.

b) The distribution of the number broken pieces per batch of pottery is skewed right, centered around 1 broken piece per batch. Batches had from 0 and 6 broken pieces. The scatterplot does not show the center or skewness of the distribution.

c) The scatterplot shows that the number of broken pieces increases as the batch number increases. If the 8 daily batches are numbered sequentially, this indicates that batches fired later in the day generally have more broken pieces. This information is not visible in the histogram.


a) A histogram of daily sales is at the right.

b) The scatterplot shows that, in general, the sales have been increasing over time. The histogram does not show this.

c) The histogram shows that the mean of the daily sales for the coffee shop was between $300 and $400, and that this happened on 8 days. The scatterplot does not show this.

15. Matching.

a) 0.006   b) 0.777   c) -0.923   d) -0.487


a) -0.977   b) 0.736   c) 0.951   d) -0.021
17. Politics.

The candidate might mean that there is an association between television watching and crime. The term correlation is reserved for describing linear associations between quantitative variables. We don’t know what type of variables “television watching” and “crime” are, but they seem categorical. Even if the variables are quantitative (hours of TV watched per week, and number of crimes committed, for example), we aren’t sure that the relationship is linear. The politician also seems to be implying a cause-and-effect relationship between television watching and crime. Association of any kind does not imply causation.

18. Car thefts.

It might be reasonable to say that there is an association between the type of car you own and the risk that it will be stolen. The term correlation is reserved for describing linear associations between quantitative variables. Type of car is a categorical variable.


a) It is appropriate to calculate correlation. Both height of the drop and speed are quantitative variables, the scatterplot shows an association that is straight enough, and there are not outliers.

b) There is a strong, positive, straight association between drop and speed; the greater the height of the initial drop, the higher the top speed.

20. Antidepressants.

a) It is appropriate to calculate correlation. Both placebo improvement and treated improvement are quantitative variables, the scatterplot shows an association that is straight enough, and there are not outliers.

b) There is a strong, positive, straight association between placebo and treated improvement. Experiments that showed a greater placebo effect also showed a greater mean improvement among patients who took an antidepressant.


It is not appropriate to summarize the strength of the association between water hardness and pH with a correlation, since the association is curved, not Straight Enough.

22. Traffic headaches.

It is not appropriate to summarize the strength of the association between highway speed and total delay with a correlation. The scatterplot shows evidence of outliers, and the main cluster of data is not Straight Enough.
23. Cold nights.

The correlation is between the number of days since January 1 and temperature is likely to be near zero. We expect the temperature to be low in January, increase through the spring and summer, then decrease again. The relationship is not Straigh Enough, so correlation is not an appropriate measure of strength.


The researcher should have plotted the data first. A strong, curved relationship may have a very low correlation. In fact, correlation is only a useful measure of the strength of a linear relationship.

25. Prediction units.

The correlation between prediction error and year would not change, since the correlation is based on z-scores. The z-scores are the same whether the prediction errors are measured in nautical miles or miles.


The correlation between prediction error and year would not change, since the correlation is based on z-scores. The z-scores of the prediction errors are not changed by adding or subtracting a constant.

27. Correlation errors.

a) If the association between GDP and infant mortality is linear, a correlation of –0.772 shows a moderate, negative association. Generally, as GDP increases, infant mortality rate decreases.

b) Continent is a categorical variable. Correlation measures the strength of linear associations between quantitative variables.


a) Correlation must be between –1 and 1, inclusive. Correlation can never be 1.22.

b) A correlation, no matter how strong, cannot prove a cause-and-effect relationship.

29. Height and reading.

a) Actually, this does mean that taller children in elementary school are better readers. However, this does not mean that height causes good reading ability.

b) Older children are generally both taller and are better readers. Age is the lurking variable.


a) No. It simply means that in countries where cell phone use is high, the life expectancy tends to be high as well.
b) General economic conditions of the country could affect both cell phone use and life expectancy. Richer countries generally have more cell phone use and better health care. The economy is a lurking variable.

31. Correlations conclusions I.
   a) No. We don’t know this from correlation alone. The relationship between age and income may be non-linear, or the relationship may contain outliers.
   b) No. We can’t tell the form of the relationship between age and income. We need to look at the scatterplot.
   c) No. The correlation between age and income doesn’t tell us anything about outliers.
   d) Yes. Correlation is based on z-scores, and is unaffected by changes in units.

32. Correlation conclusions II.
   a) No. We don’t know this from correlation alone. The relationship between fuel efficiency and price may be non-linear, or the relationship may contain outliers.
   b) No. We can’t tell the form of the relationship between fuel efficiency and price. We need to look at the scatterplot.
   c) No. The correlation between fuel efficiency and price doesn’t tell us anything about outliers.
   d) No. Correlation is based on z-scores, and is unaffected by changes in units.

33. Baldness and heart disease.
   Even though the variables baldness and heart disease were assigned numerical values, they are categorical. Correlation is only an appropriate measure of the strength of linear association between quantitative variables. Their conclusion is meaningless.

34. Sample survey.
   Even though zipcodes are numbers, they are categorical variables representing different geographic areas. Likewise, even thought the variable `datasource` has numerical values, it is also categorical, representing the source from which the data were acquired. Correlation is only an appropriate measure of the strength of linear association between quantitative variables.

35. Income and housing.
   a) There is a positive, moderately strong, linear relationship between `Housing Cost Index` and `Median Family Income`, with several states whose `Housing Cost Index` seems high for their `Median Family Income`, and one state whose `Housing Cost Index` seems low for their `Median Family Income`.
   b) Correlation is based on z-scores. The correlation would still be 0.65.
c) Correlation is based on $z$-scores, and is unaffected by changes in units. The correlation would still be 0.65.

d) Washington D.C. would be a moderately high outlier, with Housing Cost Index high for its Median Family Income. Since it doesn’t fit the pattern, the correlation would decrease slightly if Washington D.C. were included.

e) No. We can only say that higher Housing Cost Index scores are associated with higher Median Family Income, but we don’t know why. There may be other variables at work.

36. Interest rates and mortgages.

a) There is a negative, strong, linear relationship between Total Mortgages and Interest Rate. There are no outliers in the relationship.

b) Correlation is based on $z$-scores. The correlation would still be $-0.84$.

c) Correlation is based on $z$-scores, and is unaffected by changes in units. The correlation would still be $-0.84$.

d) The given year has a very high mortgage rate for an interest rate that is that high. It doesn’t fit the overall pattern, so the correlation would weaken (get closer to zero).

e) No. We can only say that lower interest rates are associated with larger mortgage amounts, but we don’t know why. There may be other economic variables at work.


a) A scatterplot of expected fuel economy vs. horsepower is at the right.

b) There is a strong, negative, straight association between horsepower and mileage of the selected vehicles. There don’t appear to be any outliers. All of the cars seem to fit the same pattern. Cars with more horsepower tend to have lower mileage.

c) Since the relationship is linear, with no outliers, correlation is an appropriate measure of strength. The correlation between horsepower and mileage of the selected vehicles is $r = -0.909$. 
d) There is a strong linear relationship in the negative direction between horsepower and highway gas mileage. Lower fuel efficiency is associated with higher horsepower.

38. Drug abuse.

a) A scatterplot of percentage of teens who have used other drugs vs. percentage who have used marijuana in the U.S. and 10 Western European countries is at the right.

b) The correlation between the percent of teens who have used marijuana and the percent of teens who have used other drugs is $r = 0.934$.

c) The association between the percent of teens who have used marijuana and the percent of teens who have used other drugs is positive, strong, and straight. Countries with higher percentages of teens who have used marijuana tend to have higher percentages of teens that have used other drugs.

d) These results do not confirm that marijuana is a “gateway drug”. An association exists between the percent of teens that have used marijuana and the percent of teens that have used other drugs. This does not mean that one caused the other.


There is no apparent association between the number of grams of fat and the number of milligrams of sodium in several brands of fast food burgers. The correlation is only $r = 0.199$, which is close to zero, an indication of no association. One burger had a much lower fat content than the other burgers, at 19 grams of fat, with 920 milligrams of sodium. Without this (comparatively) low fat burger, the correlation would have been even lower.
40. Burgers II.

The correlation between the number of calories and the number of grams of fat in several fast food burgers is $r = 0.961$. The association between the number of calories and the number of grams of fat in several fast food burgers is positive, straight, and strong. Typically, burgers with higher fat content have more calories. Even if the outlier at 410 calories and 19 grams of fat is set aside, the correlation is still quite strong at 0.837.

41. Attendance 2010.

a) Number of runs scored and attendance are quantitative variables, the relationship between them appears to be straight, and there are no outliers, so calculating a correlation is appropriate.

b) The association between attendance and runs scored is positive, straight, and moderate in strength. Generally, as the number of runs scored increases, so does attendance.

c) There is evidence of an association between attendance and runs scored, but a cause-and-effect relationship between the two is not implied. There may be lurking variables that can account for the increases in each. For example, perhaps winning teams score more runs and also have higher attendance. We don’t have any basis to make a claim of causation.

42. Second inning 2010.

a) Winning teams generally enjoy greater attendance at their home games. The association between home attendance and number of wins is positive, somewhat straight, and moderately strong.

b) Neither. The correlations, $r = 0.533$ for wins and attendance, and $r = 0.538$ for runs and attendance, are so close that one cannot say that one is more strongly correlated than the other.

c) The correlation between number of runs scored and number of wins is $r = 0.919$, indicating a possible strong association. However, since there is no scatterplot of wins vs. runs provided, we can’t be sure the relationship is straight. Correlation may not be an appropriate measure of the strength of the association.
43. Thrills.

The scatterplot at the right shows that the association between duration and length is straight, positive, and moderate, with no outliers. Generally, rides on coasters with a greater length tend to last longer. The correlation between length and duration is 0.698, indicating a moderate association.

44. Thrills II.

a) With a correlation of 0.980, there is a very strong, positive, and linear relationship between the initial drop of a roller coaster and its maximum speed. It appears that the maximum speed of a roller coaster is directly related to the height of the first drop.

b) Similarly to part (a), the height and initial drop are directly related. The relationship is strong, positive, and reasonably linear and has a correlation of 0.966.

c) The initial drop of a coaster clearly affects the height and speed of the coaster. This is apparent, as most coasters start with a tall ascent that is the largest and the fastest. The initial drop is also moderately correlated with the steepness of the angle \((r = 0.603)\) and somewhat strongly correlated with the length of the coaster \((r = 0.903)\).

45. Thrills III.

a) We would expect that as one variable (say length of ride) increases, the rank will improve, which means it will decrease.
b) Drop has the strongest correlation \((r = -0.193)\), but even that correlation is very weak. The scatterplot shows no apparent association. The number one ranked coaster, Bizarro, has a fairly typical drop. There appear to be other factors that influence the rank of coaster more than any of the ones measured in this data set.

c) There may be other variables that account for the ranking. For example, other quantitative variables such as number of loops and number of corkscrews, categorical variables such as whether the coaster is made of wood or steel and whether or not there are tunnels may all have an affect on the rank.

46. Vehicle weights.

a) A scatterplot of the Static Weight vs. Weight-in-Motion of the test truck is at the right.

b) The association between static weight and weight-in-motion is positive, strong, and roughly straight. There may be a hint of a curve in the scatterplot.

c) As the static weight of the test truck increased, so did the weight-in-motion, but the relationship appears weaker for heavier trucks.

d) The correlation between static weight and weight-in-motion is \(r = 0.965\).

e) Weighing the trucks in kilograms instead of pounds would not change the correlation. Correlation, like \(z\)-score, has no units. It is a numerical measure of the degree of linear association between two variables.

f) When the test truck weighed approximately 35,500 pounds, it weighed higher in motion. The scale may need to be recalibrated. If the scale were calibrated exactly, we would expect the points to line up perfectly, with no curve, and no deviations from the pattern.
47. Planets (more or less).

a) The association between Position Number of each planet and its distance from the sun (in millions of miles) is very strong, positive and curved. The scatterplot is at the right.

b) The relationship between Position Number and distance from the sun is not linear. Correlation is a measure of the degree of linear association between two variables.

c) The scatterplot of the logarithm of distance versus Position Number (shown at the right) still shows a strong, positive relationship, but it is straighter than the previous scatterplot. It still shows a curve in the scatterplot, but it is straight enough that correlation may now be used as an appropriate measure of the strength of the relationship between logarithm of distance and Position Number, which will in turn give an indication of the strength of the association.


a) The correlation between the year and the number of flights is 0.441.

b) There is a positive, nonlinear association between the year and the number of flights. From 2000 to 2005, the number of flights tends to increase. From 2005 to 2010, the number of flights tends to decrease.

c) Correlation is not appropriate since the relationship is not linear.