Unit 12: Correlation

SUMMARY OF VIDEO

How much are identical twins alike – and are these similarities due to genetics or to the environment in which the twins were raised? The Minnesota Twin Study, a classic correlation study on genes versus environment done in the 1980s, studied subjects like Jerry Levey and Mark Newman, identical twins raised apart. The two looked alike, were both involved with volunteer fire departments, and even had the same beer preference. Given they were separated at birth, the measure of correlation or similarity between them should be attributed to genetics. Contrast this with correlations between identical twins raised together. Here the similarities ought to be due to common family environment in addition to common genes. The difference between the size of the correlation between these two groups of twins tells researchers about the influence of the common family environment.

So how do you assess the size of the correlation? We can often get a pretty good idea simply by looking at a scatterplot of the data. Take, for example, Figure 12.1, a scatterplot of heights of pairs of twins who have been raised apart.

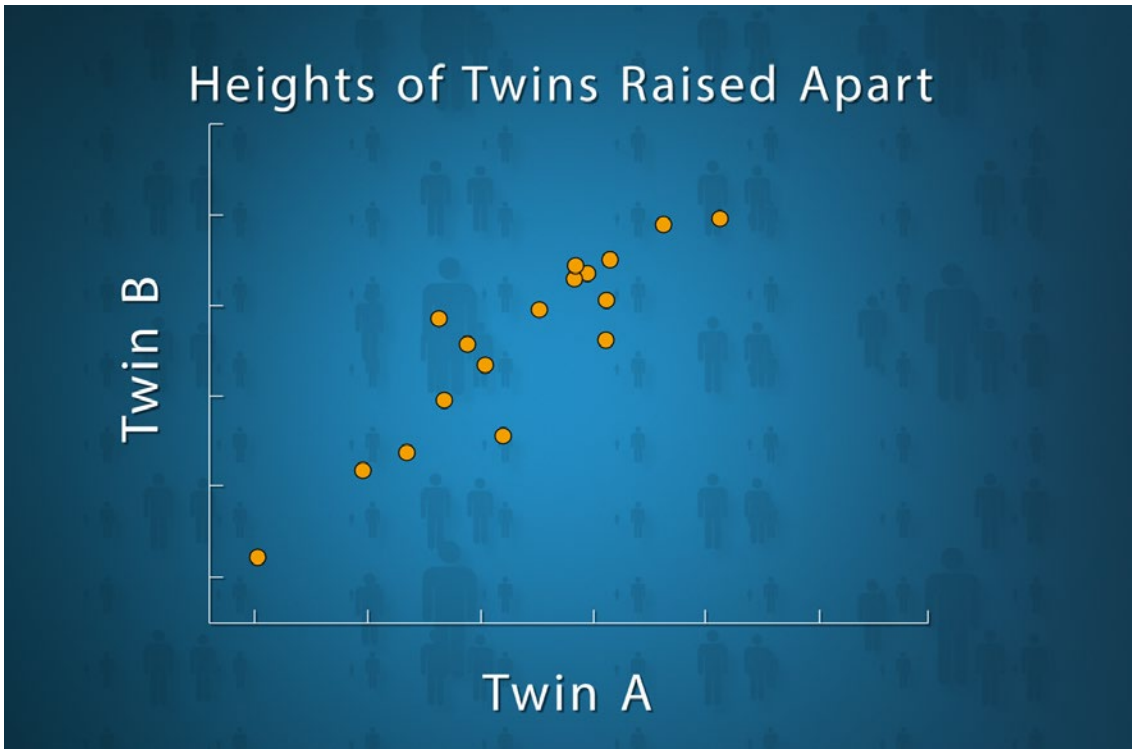

Figure 12.1. Scatterplot of heights.

We can quickly see that the taller one twin of a pair is, the taller is the other; there is a positive correlation between the two. The pattern appears quite strong, which is not surprising for a physical trait. But would we also find correlations between behavioral traits? Figure 12.2 shows a scatterplot of a personality inventory study given to pairs of identical twins raised apart.

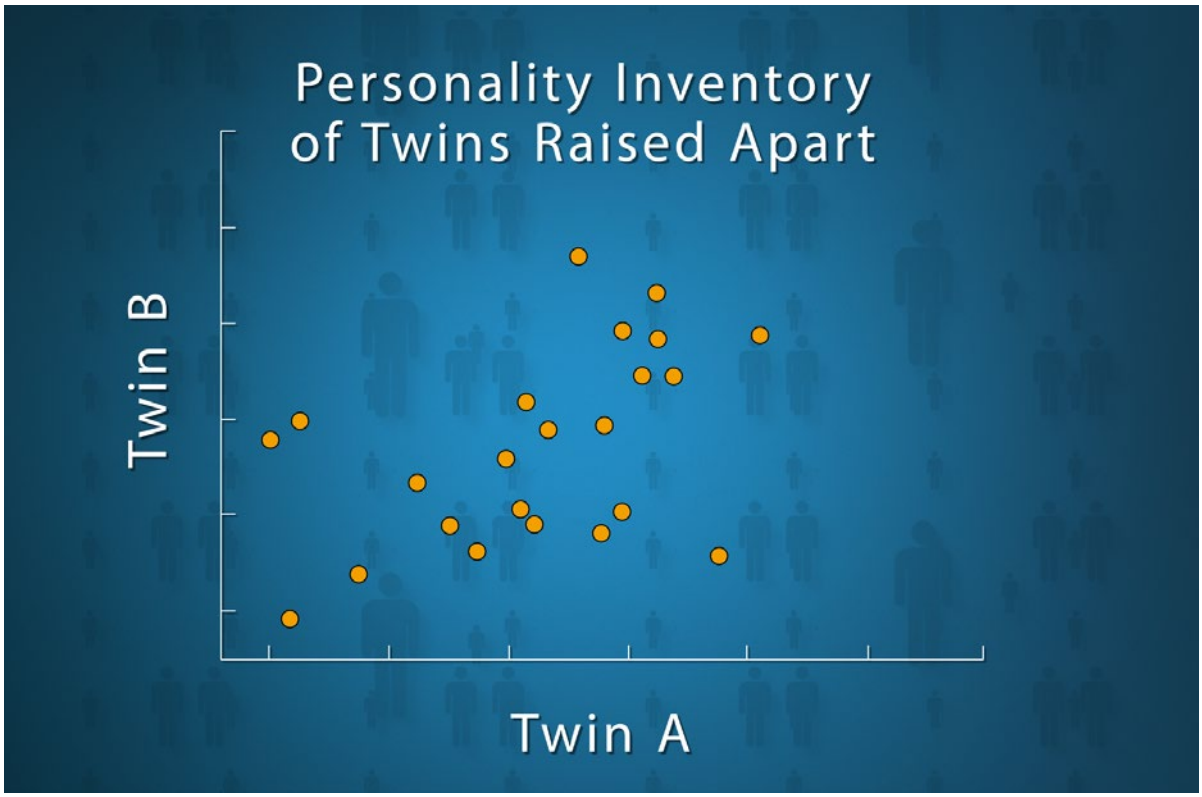

Figure 12.2. Scatterplot of personality inventory.

While the relationship is not as clear as it was for height, the points do tend to increase together. Remember, the twins were raised in different families so the fact that a correlation exists at all can only be attributed to their common genes. We can compare these two scatterplots more objectively with a direct measure of correlation denoted as *r*. The formula for calculating *r* is given below.

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

However, in practice, most people use software or a calculator that finds *r* from the keyed in data on *x* and *y*.

The value of *r* is always a number between –1 and 1; positive *r* means positive association, and the closer *r* is to 1, the closer to a straight line the scatterplot is; *r* = +1 is perfect positive linear association, in which case all the points lie exactly on a straight line that has

positive slope; negative *r* similarly measures negative linear association. Some examples of scatterplots together with their corresponding values of *r* are shown in Figure 12.3.

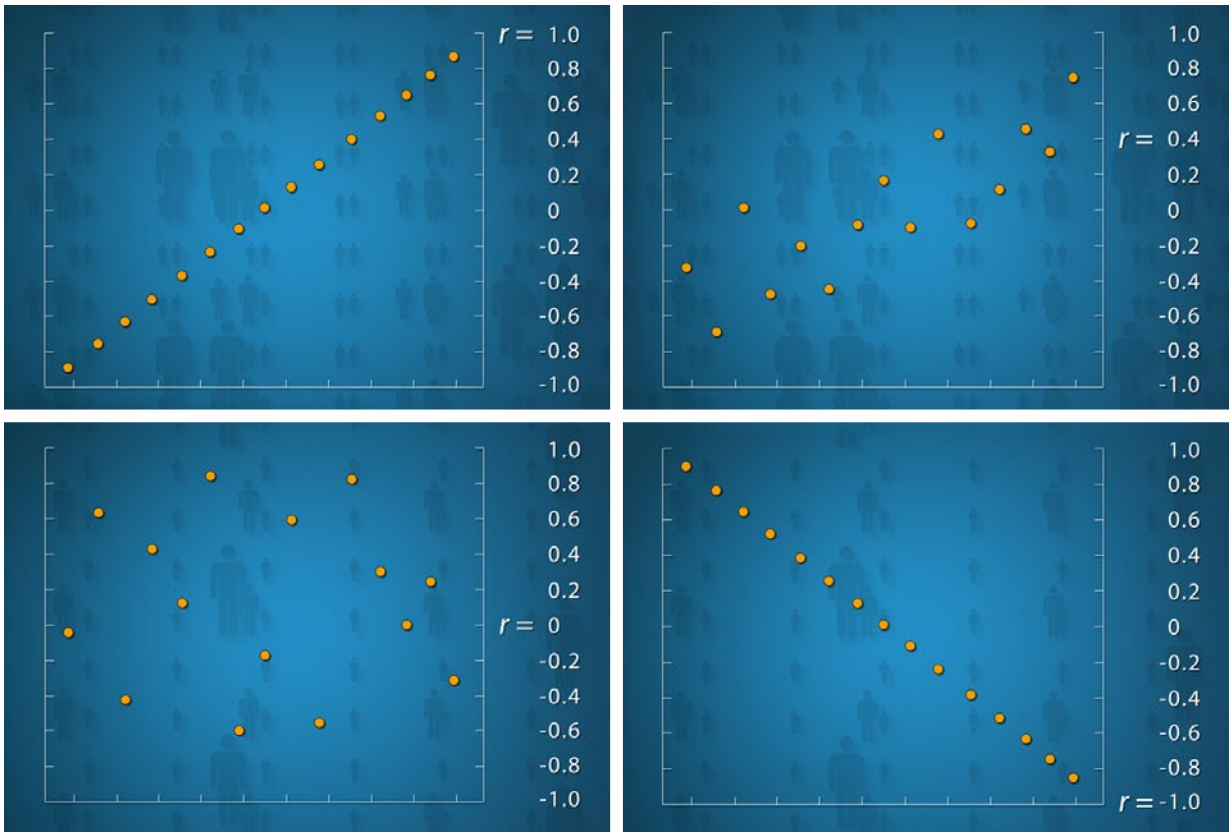

Figure 12.3. Values of r for four scatterplots.

The scatterplot of twins' heights in Figure 12.1 has $r = 0.92$, which is very close to 1 indicating a strong, positive, linear association. These twins were separated shortly after birth and raised apart, so the high correlation suggests that inheritance has a lot to do with determining height. For the personality study, the correlation is $r = 0.49$. Twins have somewhat similar personalities, but the relation is not as strong as for height but is still suggestive of a strong genetic influence.

Studies like the Minnesota Twin Study were only possible back when it had been common for identical twins to be separated when placed up for adoption. Nowadays we don't like to separate twins. So, in her study of the role of genes and environment on personality traits, Kim Saudino takes a different approach – comparing fraternal twins, who share approximately half their genes, with identical twins, who share all their genes. She records activity levels of twins by placing motion detectors on the twins. Two scatterplots in Figure 12.4 show the relationship between the activity level of the twins; identical twins are on the left and fraternal twins on the right.

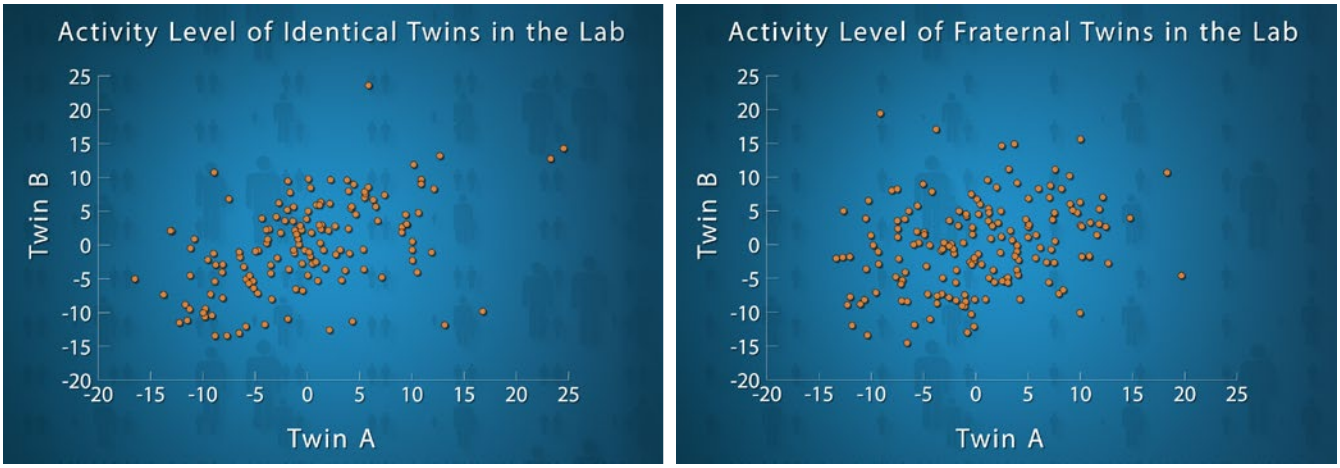

Figure 12.4. Activity levels of identical and fraternal twins in laboratory setting.

As expected, the correlation for the identical twins, $r = 0.48$, is higher than the correlation for fraternal twins, $r = 0.26$. Since the environments these toddlers were raised in were the same, the difference in correlations can only be accounted for by the genes they inherited. But that's not the end of the story. These data were collected in a laboratory environment. Next, the researcher collected the same type of data on the twins in their home environment. In the home setting, the difference in the correlations largely disappeared – for the identical twins, $r = 0.87$, and for the fraternal twins, $r = 0.70$. The conclusion: it looks as if twins' behavioral patterns are governed both by genes and by environment.

STUDENT LEARNING OBJECTIVES

A. Recognize the correlation coefficient *r* as a measure of the strength and direction of a linear relationship between two quantitative variables.

B. Be aware of the basic properties of *r*:

- The sign of *r* shows positive or negative association.
- The value of *r* always satisfies $-1 \le r \le 1$.
- The value of *r* remains the same when the two variables are interchanged and also when the units of the variables are changed.
- The value of *r* moves away from 0 toward -1 or 1 as the scatterplot points show a closer straight-line pattern; $r = \pm 1$ means a perfect straight-line relation.

C. Be able to use the formula to calculate *r* from small data sets, say 5 observations; be able to use technology to calculate *r* for larger data sets.

D. Understand that a strong correlation can have various interpretations and that correlation does not imply causation.

E. Understand the importance of looking at a scatterplot of the data when using *r* to interpret the strength of a linear relationship. Know that a single outlier can have a dramatic effect on the value of *r*.

CONTENT OVERVIEW

Correlation is the usual measure of association between two quantitative variables. More specifically, Pearson's product moment correlation coefficient *r*, or the correlation coefficient for short, measures the strength and direction of linear (straight-line) relationships. Given a linear relationship exists between two quantitative variables, *r* is positive if the data fall about a line that has a positive slope and r is negative if the data fall about a line that has a negative slope. The value of *r* is always between -1 and 1. If *r* = -1, the data fall exactly on a line with negative slope and if *r* = 1, the data fall exactly on a line with positive slope. The correlation measures both the strength and direction of a linear relationship. Figure 12.5 provides some guidelines:

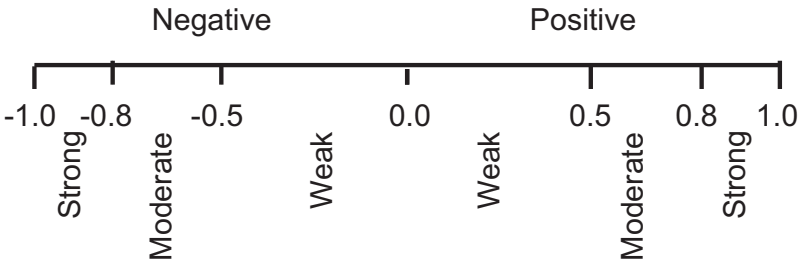

Figure 12.5. Using r to measure the strength and direction of a linear relationship.

There are a variety of formulas that can be used to compute *r*, but all are algebraically equivalent to the one given below.

Formula for Calculating *r*

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Graphing calculators, spreadsheets, and statistical computing packages compute *r* very efficiently. So, unless the size of the data set is quite small, it is best to use technology to compute the value of *r.* However, the formula in the form presented above does provide the following insight:

- Notice that *r* consists of the product of z-scores for the *x* and *y* values. Therefore, the correlation coefficient is unit-free because units associated with the *x* and *y*

values cancel out. If we change the units of our data, for example change inches to centimeters, the value of *r* will remain the same.

- Interchanging *x* and *y* does not affect *r*.
- If we add a constant to all data values, either the *x*'s or *y*'s, the value of *r* does not change. If we multiple all data values, either the *x*'s or *y*'s, by a constant, the value of *r* does not change.

Always make a scatterplot of the data before interpreting correlation. A single extreme outlier added to data that otherwise has a positive association can result in a negative correlation. A strong relationship that happens to be curved can produce a value of *r* close to 0. So, always check to see that a scatterplot of the data has linear form and is free of extreme outliers before using correlation to measure the strength of a relationship between two quantitative variables. In addition, it should be noted that a strong correlation does not always mean that there is a direct cause-and-effect link between the variables. Unit 14, The Question of Causation, looks at causation in detail.

KEY TERMS

Correlation, denoted by *r*, measures the direction and strength of a linear relationship between two quantitative variables. The formula for computing Pearson's correlation coefficient is:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. If it were true that two identical twins always had the same height, what would the scatterplot of the heights of several pairs of identical twins look like? What would be the correlation *r* between the heights?

2. What are all the possible values of the correlation coefficient *r*?

3. If heredity plays a strong role in determining personality, will the correlation between twins raised together be about the same as, or much larger than, the correlation between twins raised apart?

4. Is it easy to guess how large the correlation is by looking at a scatterplot? Explain.

UNIT ACTIVITY:
SCATTERPLOTS AND CORRELATION

Pearson's product moment correlation coefficient, *r,* measures the strength of a linear relationship. So, before computing *r*, make a scatterplot to check that the relationship between two variables has linear form. The value of *r* always lies between -1 and 1 and provides information both on a relationship's direction (positive or negative association) and strength (closeness of data points to a line). In the questions that follow, use technology (graphing calculator, spreadsheet, statistical computing software) to compute *r*. You can either make the scatterplots by hand or use technology.

1. Enter the following data into two columns, one for *x* and the other for *y*.

x	8	11	5	2	4
y	17	23	11	5	9

Table 12.1. Data set A.

a. Make a scatterplot of y versus x. Does the pattern appear to be linear or nonlinear? Is the association between x and y positive or negative?

b. Calculate the correlation, *r*. What does this tell you about the pattern of dots in your scatterplot?

2. Repeat question 1 using Data set B from Table 12.2.

x	10	3	5	1	6
y	-30	-2	-10	6	-14

Table 12.2. Data set B.

3. In the next data sets, the scatter of the points is increased. Your task will be to see how the increase of scatter affects correlation.

x	2	4	5	8	11
y_1	11.5	13.2	14	22.3	19.8
y_2	10.7	11.2	33	36.7	22.7

Table 12.3. Data with more scatter.

a. Using the data in Table 12.3, make scatterplots for y_1 versus x and y_2 versus x. Draw two separate scatterplots using the same scaling on the axes for both plots.

b. In which plot does the relationship between x and y appear stronger? In other words, in which scatterplot do the dots appear to lie closer to a line?

c. Since both scatterplots show a positive relationship between the two variables, the correlations should be positive. Calculate the correlation between x and y_1 and between x and y_2. Based on the value of the correlation coefficient r, classify the relationships between the variables as strong, moderate, or weak. Use the guidelines which can be found in Figure 12.5 to make that classification.

4. Return to data set A from Table 12.1.

a. Change y-value associated with $x = 11$ from $y = 23$ to $y = 10$. Make a scatterplot of the data and compute the correlation.

b. Change the y-value associated with $x = 11$ to $y = 0$. Make a scatterplot of the data and compute the correlation.

5. Summarize what you have observed about correlation from this activity.

EXERCISES

1. Archaeopteryx is an extinct beast having flight feathers like a bird but teeth and a long bony tail like a reptile. Only five complete fossil specimens are known. These specimens differ greatly in size, so some experts think they are different species. Others think they are individuals from the same species but of different ages. Correlation can help decide the question. If the specimens belong to the same species and differ in size because they are at different stages of growth, there should be a strong straight-line relationship between the lengths of a pair of bones from all individuals. Outliers from this relationship would suggest a different species. Table 12.4 gives the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five specimens.

Femur (cm)	38	56	59	64	74
Humerus (cm)	41	63	70	72	84

Table 12.4. Bone lengths from fossils.

a. Make a scatterplot of the data in Table 12.4. Does the relationship appear linear? Is the association between femur length and humerus length positive or negative?

b. Calculate the correlation coefficient, r , using the formula.

c. Based on the value of r , do you conclude that these specimens are all from the same species? Explain.

2. Each of the following statements contains a blunder. Explain in each case what is wrong.

a. There is correlation $r = 0.6$ between the gender of students and their scores on a mathematics exam.

b. We found a high correlation ($r = 1.09$) between students' scores on the math part of the SAT and their scores on the verbal part of the SAT.

c. The correlation between amount of fertilizer and yield of corn was found to be $r = 0.23$ bushel.

3. Foal weight at birth is one indicator of the newborn's health. Is a mare's (mother's) weight related to the weight of her foal? Data on the weights of 15 mares and their foals appears in Table 12.5.

Mare's Weight (kg)	Foal's Weight (kg)
556	129.0
638	119.0
588	132.0
550	123.5
580	112.0
642	113.5
568	95.0
642	104.0
556	104.0
616	93.5
549	108.5
504	95.0
515	117.5
551	128.0
594	127.5

Table 12.5. Weight of mares and their newborn foals.
Source: "Suckling behavior does not measure milk intake in horses, Equus caballus," Careron, E. et al. (Animal Behavior [1999]: p673 – 678).

a. Make a scatterplot of the data in Table 12.5. Which variable did you put on the horizontal axis. Explain your choice.

b. Based on your scatterplot, does the association between foal weight and mare weight appear to be positive, negative, or neither? Explain.

c. Based on your scatterplot, would you expect the correlation to be closer to -1, 0, or 1? Justify your choice.

d. Calculate the value of the correlation coefficient *r.* Does your result confirm or refute your answer to (c)?

4. Table 12.6 gives the average times by age (ages 18 – 50) for female runners in the 2012 Boston Marathon.

Age	Average Time (min)
18	288
19	286
20	260
21	274
22	273
23	265
24	271
25	272
26	272
27	265
28	264
29	270
30	265
31	268
32	261
33	265
34	268
35	260
36	261
37	264
38	271
39	264
40	268
41	269
42	271
43	266
44	267
45	273
46	280
47	276
48	279
49	282
50	281

Table 12.6. Average time by age of female runners in 2012 Boston Marathon.

a. What is the correlation between runners' average time and age? What does this tell you about the relationship between age and average time to run the race?

b. Make a scatterplot of average time versus age. Describe the relationship between these variables.

c. Explain why it is important to make a scatterplot of data before trying to interpret the value of the correlation coefficient *r.* Refer to your solutions to parts (a) and (b) as part of your answer.

REVIEW QUESTIONS

1. A student wonders if people of similar heights tend to date each other. She measures herself and several of her friends. Then she measures the next man each woman dates. Table 12.7 contains the data collected by the student.

Female Height (in)	66	64	66	65	70	65
Male Height (in)	72	68	70	68	71	65

Table 12.7. Heights of women and their dates.

a. Make a scatterplot of these data. Based on the scatterplot, do you expect the correlation to be positive or negative? Near +1 or -1, or neither?

b. Find the correlation *r* between the heights of the men and women. (Unless instructed otherwise, feel free to use technology.) Based on this correlation would you classify the strength of the linear relationship as strong, moderate, or weak? Explain.

2. Return to the data from Table 12.7.

a. If every woman in the study dated a man exactly 3 inches taller than she is, what would be the correlation between male and female heights? Explain.

b. How would *r* change if all the men were 6 inches shorter than the heights given in Table 12.7? Does the correlation help answer the question of whether women tend to date men taller than themselves? Explain.

c. Change all the heights in Table 10.1 from inches to centimeters. (Recall 1 inch = 2.54 centimeters.) Recalculate the correlation using the heights data measured in centimeters. How did this conversion from inches to centimeters affect the value of *r*?

3. Some students are good in mathematics and others are better at reading or writing. The question is whether there is any relationship between a student's ability in math and his/her ability in reading or writing. The SAT, a standardized test for college admissions that is widely used in the United States, has three sections, Math, Critical Reading, and Writing. Table 12.8 contains SAT Math, Writing, and Critical Reading test scores for 20 randomly chosen students accepted by a university.

Math	Writing	Critical Reading
440	410	410
550	570	520
520	520	540
420	470	410
550	620	530
650	560	560
610	620	550
610	520	600
340	470	400
600	540	620
680	580	580
440	430	470
440	450	370
390	430	390
460	600	600
460	520	500
520	570	580
540	530	570
420	430	470
550	480	530

Table 12.8. SAT test scores from 20 students.

a. We are interested in the relationship between students' scores on the SAT Math and their scores on the SAT Critical Reading and SAT Writing. Make two scatterplots, one of SAT Math versus SAT Critical Reading and the other of SAT Math versus SAT Writing. (In both scatterplots, SAT Math is being treated as the response variable.) Use the same scaling for both scatterplots. Based on your scatterplots, which variable has a stronger correlation with the SAT Math, the SAT Critical Reading or the SAT Writing? Explain.

b. Calculate the correlation between SAT Math scores and SAT Critical Reading scores. Then do the same for SAT Math scores and SAT Writing scores. Which variable, SAT Critical Reading or SAT Writing, is more highly correlated with SAT Math? Would you classify the strength of this relationship as strong, moderate, or weak?