Unit 12: Correlation

PREREQUISITES

Unit 10: Scatterplots is a prerequisite for this unit. However, this unit does not discuss the close relationship between correlation and regression, so Unit 11: Regression, is not a prerequisite. This also allows flexibility in presentation order between Units 11 and 12. The formula for computing the correlation coefficient makes use of both the mean and the standard deviation of the *x* and *y* values. Therefore, students should be familiar with those measures, which were covered in Unit 4: Measures of Center, and Unit 6: Standard Deviation. In addition, students should have some familiarity with summation notation.

ADDITIONAL TOPIC COVERAGE

Additional coverage on correlation can be found in *The Basic Practice of Statistics*, Chapter 4, Scatterplots and Correlation.

ACTIVITY DESCRIPTION

In the unit activity on correlation, students observe how the scatter of points about a line affects the value of *r*. Of particular importance is the discovery that even a single outlier can have a huge effect on correlation, even as extreme as turning an otherwise strong positive correlation into a negative correlation. Students should use technology – graphing calculators, spreadsheet or statistical computing software – for computation of correlation. The data sets in this activity are small. So, students can either draw the scatterplots by hand or use technology.

THE VIDEO SOLUTIONS

1. It would form a straight line. In fact, the points would all fall on the line $y = x$.
In this case, $r = 1$.

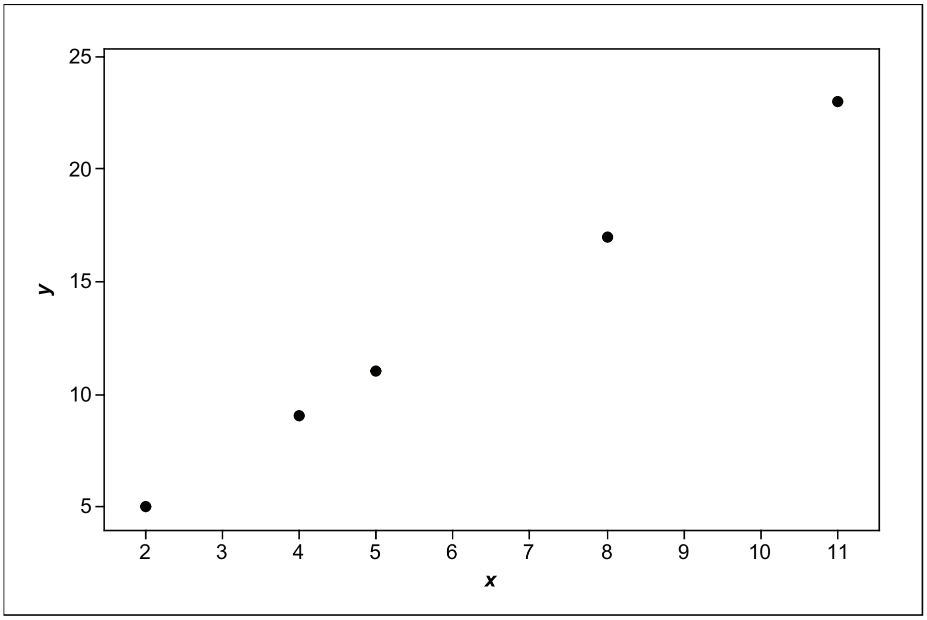
2. $-1 \leq r \leq 1$

3. The correlation between twins raised apart should be close to the correlation between twins raised together.

4. No – scaling can make a fairly strong correlation appear to be fairly weak and a weak correlation appear relatively strong. If two scatterplots are drawn on the same scale, then the data which produced the plot that is more scattered will have the lower correlation.

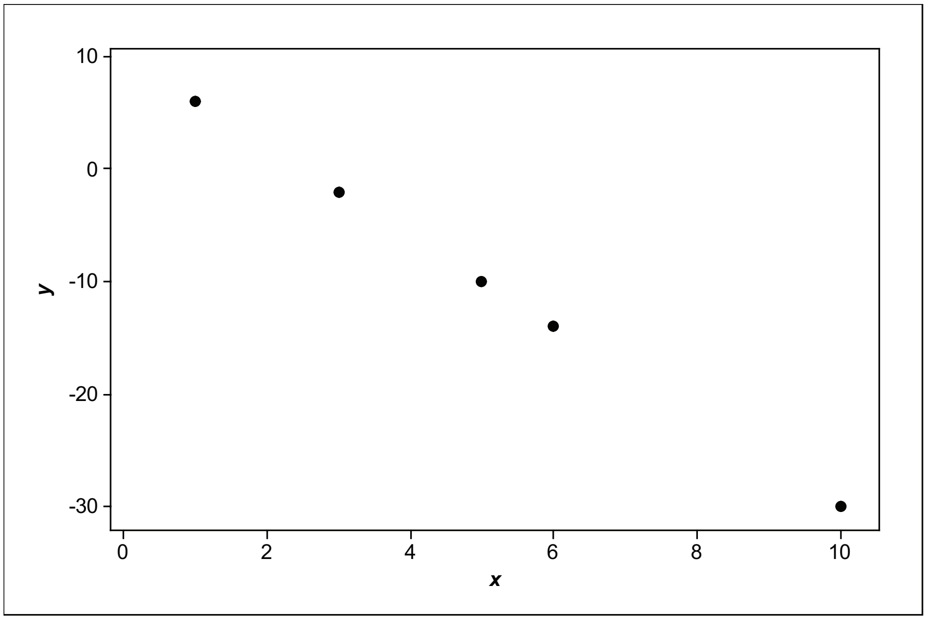
UNIT ACTIVITY SOLUTIONS

1. a.



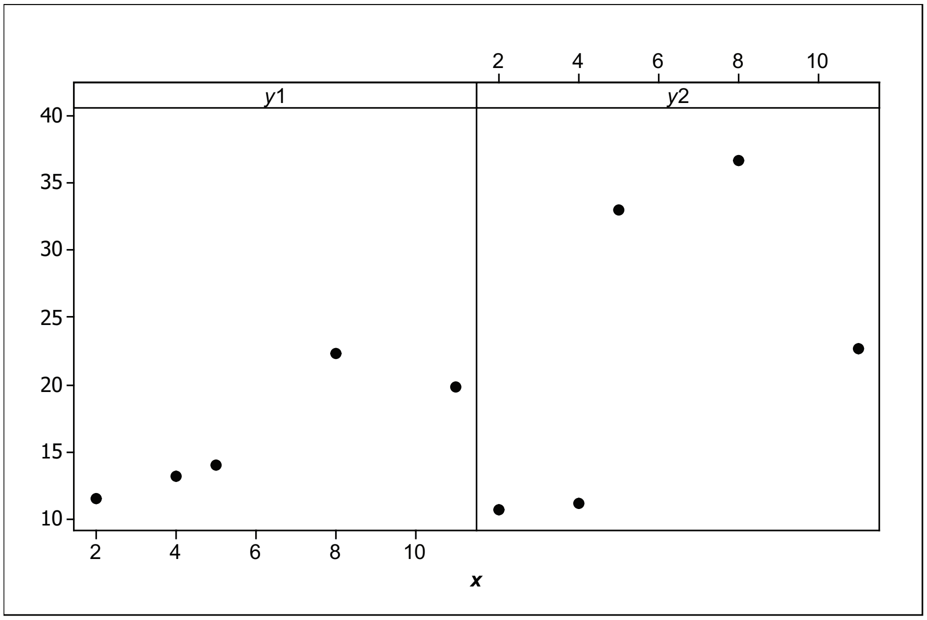
b. *r* = 1; the dots in the scatterplot fall exactly on a line with positive slope.

2. a.



b. The correlation coefficient: *r* = -1. The dots in the scatterplot fall exactly on a line with negative slope.

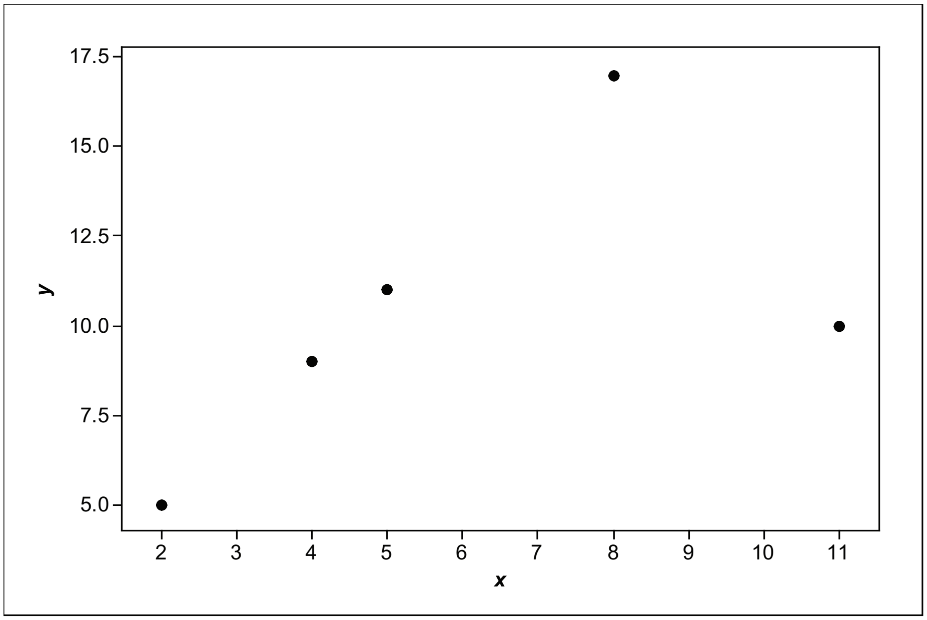
3. a.



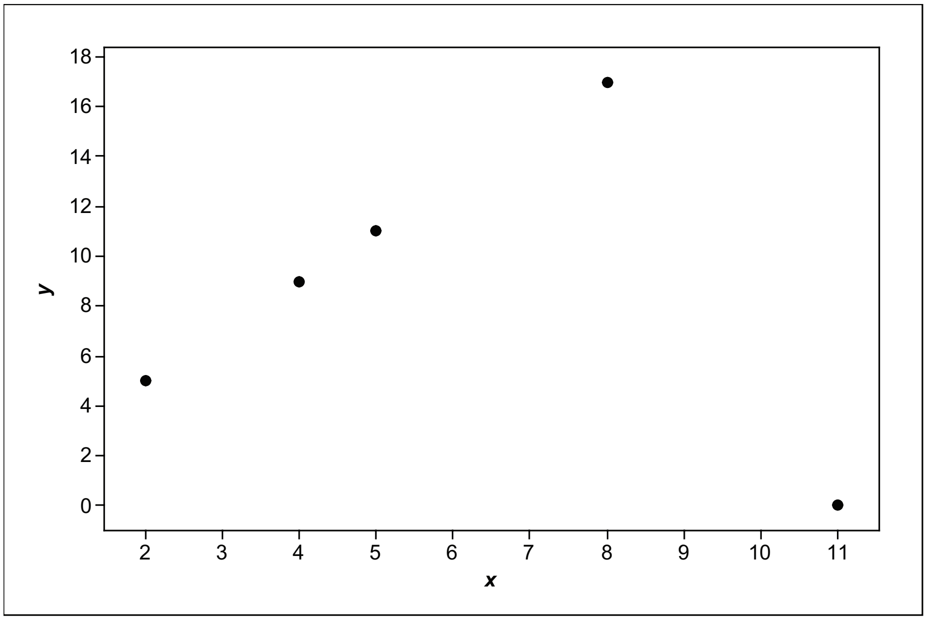
b. The plot of y_1 versus x (or the first plot) appears to show a stronger relationship.

c. For x and y_1: $r = 0.872$; strong relationship.
For x and y_2: $r = 0.522$; moderate relationship.

4. a. $r = 0.571$



b. *r* = - 0.166



5. Sample answer: When data points fall exactly on a line, *r* = ±1; if the slope of the line is positive, then *r* = +1 but if the slope of the line is negative, *r* = -1. The more scattered the data points are about a line, the closer *r* is to 0. A single outlier can have a huge effect on correlation. Even when the remaining data fall perfectly on a line with positive slope, a single outlier can dramatically reduce the value of *r* or even change it from positive to negative.

EXERCISE SOLUTIONS

1. a. Data from Marilyn A. Houck, et al. 1990. Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica. Science* 247: 195 – 198. The authors conclude from a variety of evidence that all specimens represent the same species.

There is no explanatory/response relation, so either variable can serve as *x*. In the sample solution, we have used femur length as *x* and humerus length as *y*.

The scatterplot shows a strong positive linear relationship between femur length and humerus length.

b. Femur length: $\bar{x} = 58.2$ cm and $s_x = 13.198$ cm

Humerus length: $\bar{y} = 66.0$ cm and $s_y = 15.890$ cm

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

$$r = \frac{1}{4}\left[\left(\frac{38-58.2}{13.198}\right)\left(\frac{41-66}{15.890}\right)+\left(\frac{56-58.2}{13.198}\right)\left(\frac{63-66}{15.890}\right)+\left(\frac{59-58.2}{13.198}\right)\left(\frac{70-66}{15.890}\right)+\right.$$
$$\left.\left(\frac{64-58.2}{13.198}\right)\left(\frac{72-66}{15.890}\right)+\left(\frac{74-58.2}{13.198}\right)\left(\frac{84-66}{15.890}\right)\right]$$

$$r \approx 0.994$$

Note: Expect some round off error in student answers, especially if students chose to round the standard deviations to fewer than three decimals.

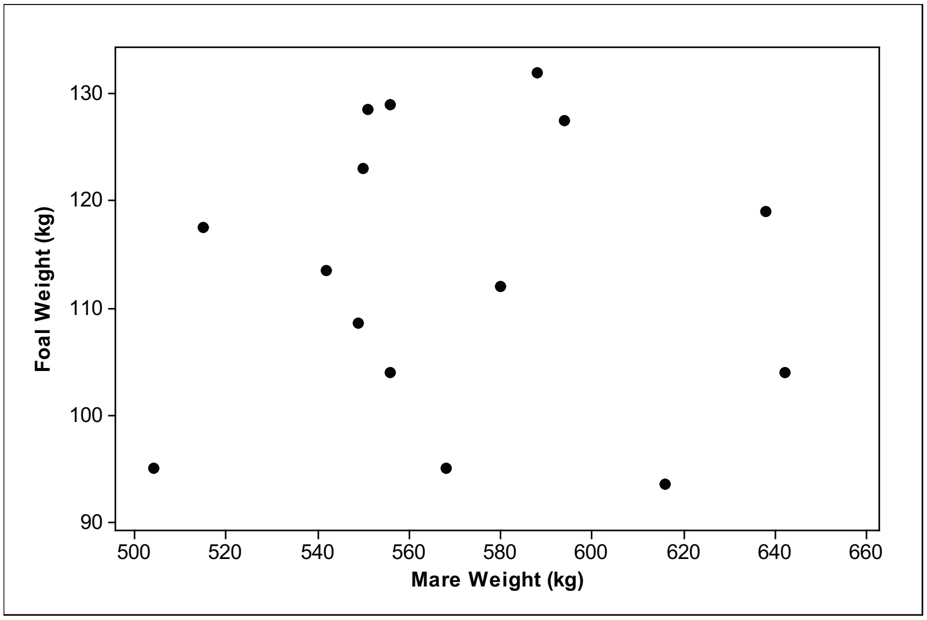
c. The high positive correlation supports the hypothesis that all five specimens appear to belong to the same species.

2. a. Gender is not a quantitative variable. You cannot calculate a correlation between gender (male or female) and anything. Correlation only makes sense for quantitative variables. (Note: “Correlation” is sometimes used to mean any kind of association, including an association between two categorical variables. However, it is better to restrict its usage to quantitative variables and *r*.)

b. Since $-1 \le r \le 1$, it is impossible for r = 1.09 > 1

c. *r* has no units – it's just a number between -1 and 1. So, *r* can't be measured in bushels.

3. a.



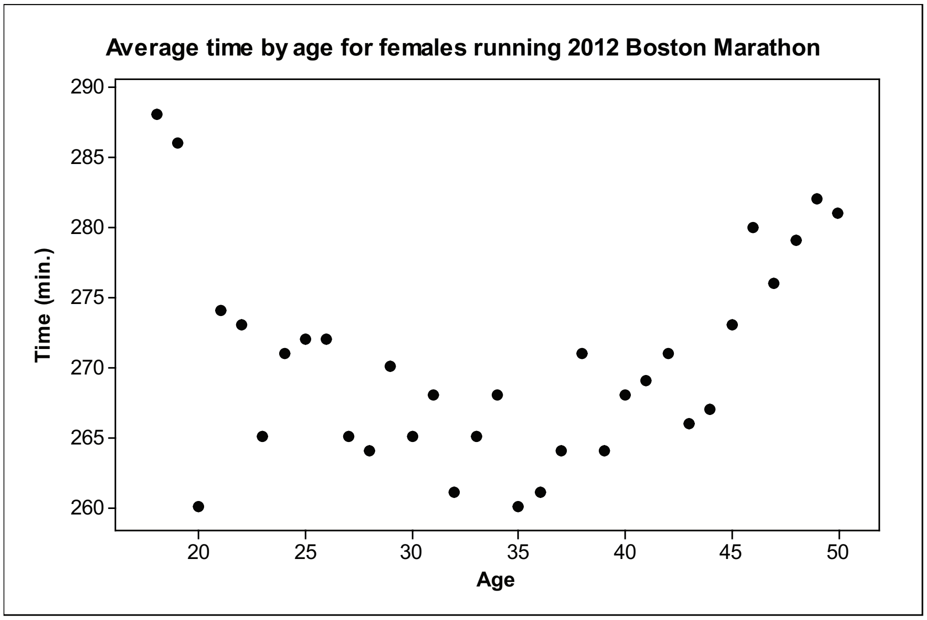
b. There doesn't appear to be any association between foal weight and mare weight. Some of the heaviest mares had the lightest foals while one of the lightest mares had a fairly heavy foal.

c. Expect the correlation to be closer to 0 than to 1 or -1.

d. *r* = 0.001. This confirms the answer to (c).

4. a. *r* = 0.097. If the relationship between average time and age were linear, it would mean there was a very weak, or practically no, relationship between these two variables.

b.

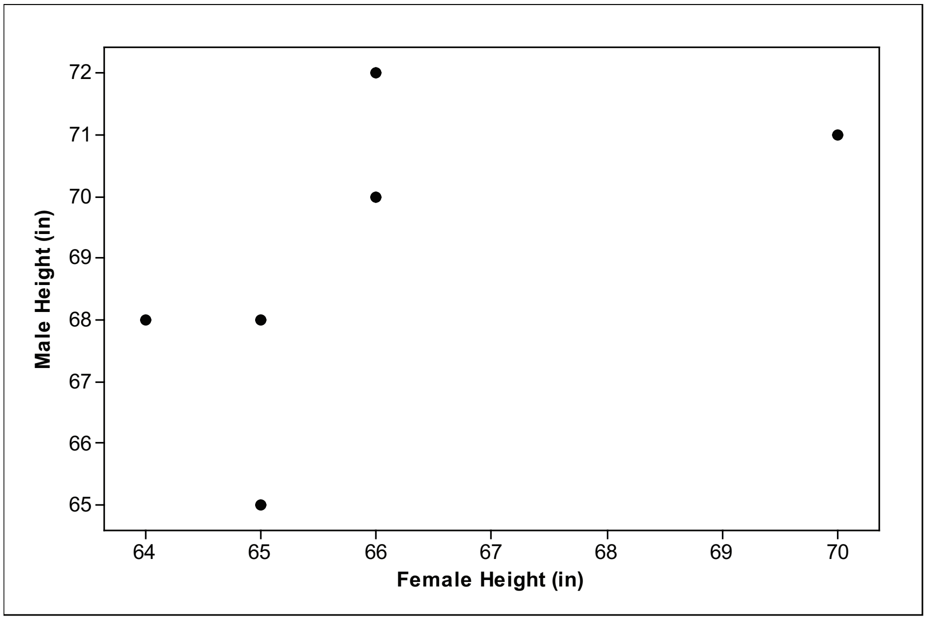


The relationship shows a curved pattern. Younger runners tend to have higher times but older runners also tend to have higher times. The best runners appear to be between 30 and 40.

c. Sample answer: The correlation coefficient measures the strength of a *linear* relationship. In part (a), the correlation was close to 0, indicating little relationship between the two variables. However, in (b) a curved relationship was noted – so there was a relationship between these two variables, but it didn't happen to be linear. Correlation should not be used to measure the strength of a relationship unless a scatterplot indicates that the form of that relationship is linear.

REVIEW QUESTIONS SOLUTIONS

1. a. In the scatterplot below, female height is on the horizontal axis since we expect that the height of the woman may influence whom she is willing to date. From the scatterplot the association between female and male heights appears to be positive. Hence, the correlation coefficient *r* should be positive. The data appear to be fairly spread out, so although the value of *r* should be positive, it should be less than 1 and not terribly close to 1.



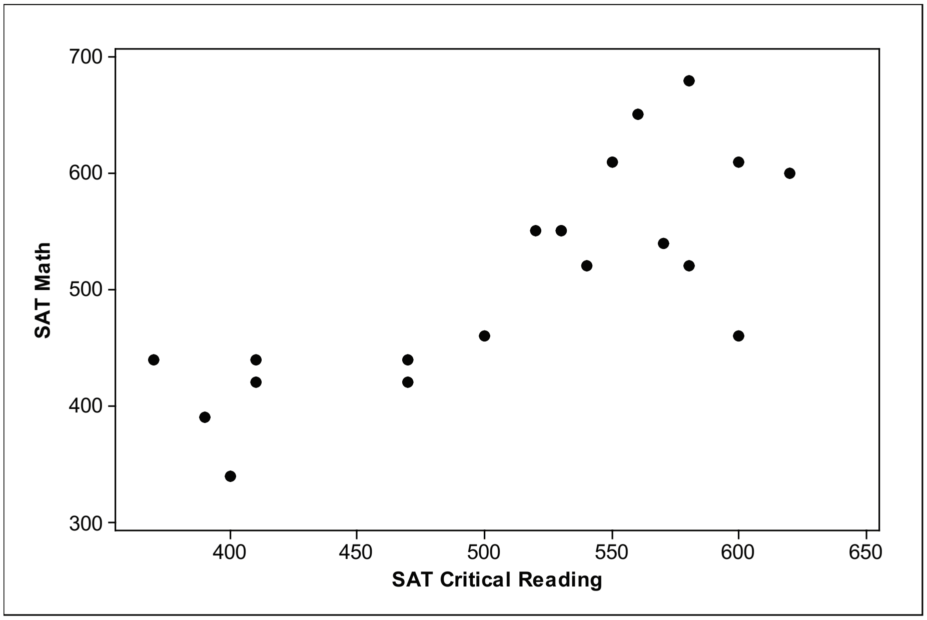
b. *r* = 0.565. The strength of the relationship between female heights and the heights of their male dates is in the low end of the moderate range.

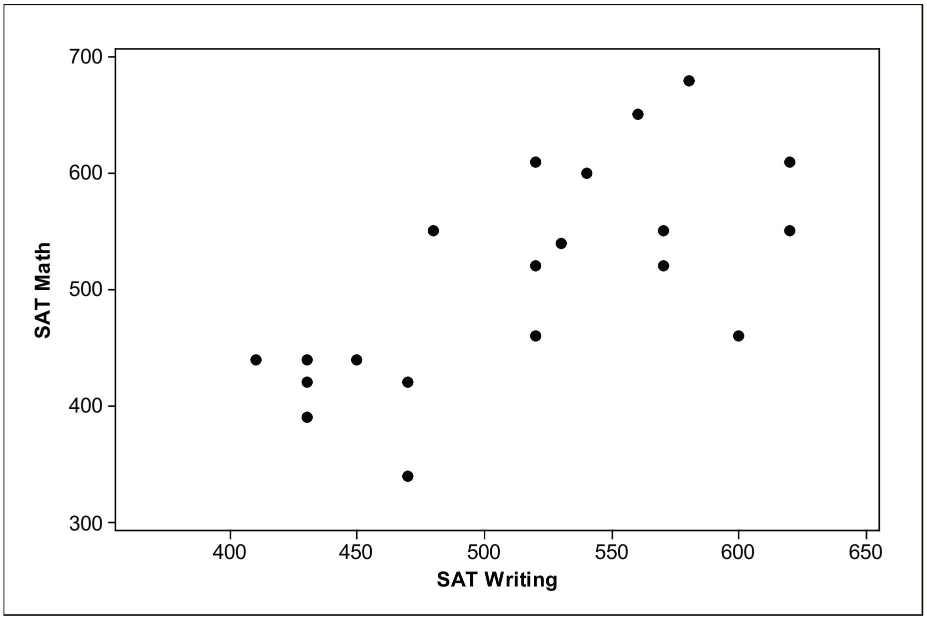
2. a. It would be *r* = 1. In this case the points fall exactly on the line *y* = *x* + 3, so there is a perfect positive linear association between the variables.

b. Changing all male heights by 6 inches does not change the correlation *r*. A positive correlation indicates that taller women tend to date taller men than shorter women date. It doesn't say anything about whether the women are dating men that are taller than they are.

c. Changing the height units from inches to centimeters did not affect the correlation; *r* = 0.565.

3. a.





The SAT Critical Reading exam appears to be more highly correlated with the SAT Math exam than the SAT Writing exam. The dots in the scatterplot of SAT Math versus SAT Critical Reading appear less spread out than the dots in the scatterplot of SAT Math versus SAT Writing.

b. The correlation between SAT Math and SAT Critical Reading is $r = 0.784$; the correlation between SAT Math and SAT Writing is $r = 0.680$. So the correlation between the SAT Math exam and the SAT Critical Reading exam is higher. The strength of that relationship should be classified as moderate (but it's at the upper end of moderate).