Lesson 2.2.1

2-42. a. It might give us an idea into which of these is a better predictor of average test score.

b. Both have fairly strong linear associations with no obvious outliers. Dataset A has a positive association and dataset B has a negative association. It is hard to tell which is stronger.

2-43. a. $\frac{1.2-4.02}{4.00}$ is about –0.7. Her absence percentage is about 0.7 standard deviations below the average.

b. $\frac{91-76}{11.49} = 1.3$. Her average test score is about 1.3 standard deviations from the average test score.

c. Since the values are negatively associated, numbers below average in one variable should generally be above average in the other.

d. She is further from average in the test-score variable, 1.3 standard deviations from the mean, than in the absence variable where she is only 0.7 standard deviations from the mean.

e. She would be equally out of the ordinary in both cases, though one could still be negative while the other positive.

2-44. a. When you standardize a data set, you are giving it a mean of 0 and a standard deviation of 1.

b. They have exactly the same association shape, form, direction, and strength. Only the scales have changed.

c. Since the data is standardized, the means of both data variables are now 0. It makes sense for the intersection of the means to be in the middle.

d. Students have already seen that LSRLs pass through the "average point." In standardized data sets, the mean point is always (0, 0).

2-45. a. Dataset A will be positive, because you are almost always multiplying two positives or two negatives. Dataset B will be negative, because you are almost always multiplying one of each.

b. A graph with a negative association will have a negative r. A graph with a positive association will have a positive r.

c. r for dataset A = 0.85. r for dataset B = –0.88

d. r is the same for the non-standardized and standardized data.

e. r is the slope of the standardized best fit lines!

2-46. Possible observations: r gets closer to 1 as the points get closer to the line of best fit. The largest r that is possible is 1. The smallest is –1. It can be zero if the points are completely scattered (no obvious trend or direction). The closer r is to 1 or –1 (the further from zero), the stronger the linear association is. It is not resistant to outliers (or "influential points"). One point can have a drastic effect on r.

2-47. a. $r = 1$

b. $r = -1$

c. All the points lie exactly on the LSRL.

d. r becomes negative, between 0 and –1.

e. The third point should be chosen so that the LSRL is as horizontal as possible.

f. The largest possible r is 1. The smallest is –1. A scatterplot with $r = 1$ has a positive slope with all points *on* the line (no scatter). A scatterplot with $r = -1$ has a negative slope with all points on the line. A scatterplot with $r = 0$ has no apparent trend (completely scattered).

2-48. –0.88 is closer to –1 than 0.85 is to 1, so dataset A has a slightly stronger linear association.

2-49. a. –0.9

b. 0.1

c. 0.5

d. –0.6

2-50. a. Strong. $r = 0.997$

b. Linear, positive, strong, with no apparent outliers. The strength is measured by $r = 0.997$; this could also be measured by S, but that would need to be compared to the actual values to be meaningful. The direction is positive because r is positive and so is the slope of the best fit line.

c. The LSRL is $f = 1.66 + 0.13d$, so the predicted value for 600 inches is about 80 inches. This is probably meaningful, but it is a large extrapolation and should not be trusted without further information.

2-51. a. $\hat{h} = 103.80 - 48.60g$, 65.89 ft

b. $r = -0.579$. It is negative because the association is negative.

c. S is the standard deviation of the residuals, and it means the typical tree's height is 4.018 feet away from the height predicted by its specific gravity.

d. It is a moderate, negative association.

2-52. By standardizing the data, we can compare the data set to other data sets with different units or values. It is also how we calculated r.

2-53. The LSRL, S, and r are identical in all four of these sets, but the forms are wildly different: only the first one fits our normal instinct for what a set with $r = 0.82$ might look like. r only tells us the strength of the association *when the form is treated as linear*—if the form is not linear, r is meaningless as a measure of strength. Certainly, high r does not make a data set linear. Be very careful to only interpret r as representing the strength of a linear association, which can only be seen in a plot.

2-54. a. 13

b.

0	9
1	0, 2, 4, 7, 7, 8
2	0, 1, 4, 5
3	0, 1

4|1 means 41

c. Because the stem-and-leaf plot plot shows all of the data points, Jerome can calculate all of the measures of central tendency.

d. It is concentrated at values between 10 and 25 and not spread too widely. The range is 15.

e. No, there are too many data points that do not repeat so it would end up just being a number line with dots at each value.

