Chapter 5: Summarizing Bivariate Data

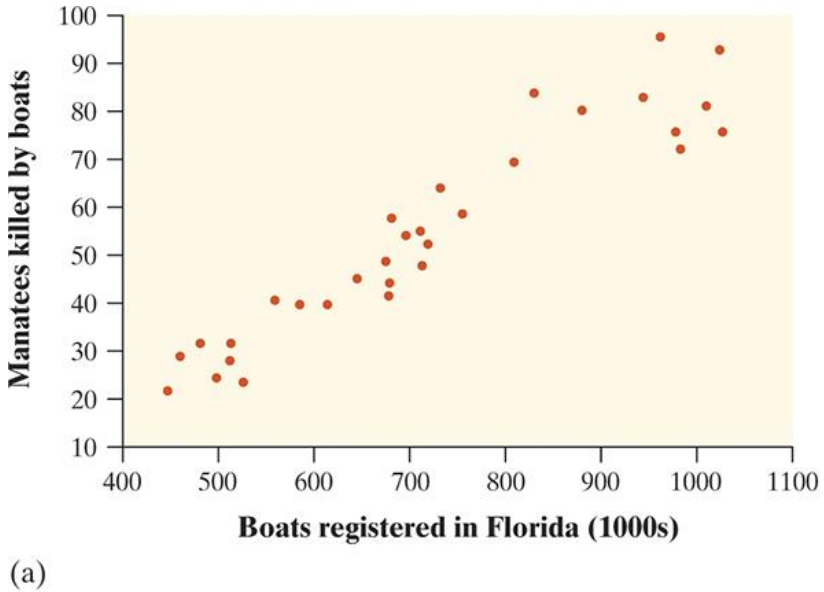
Section 5.1 & 3.4

Scatterplots and Correlation

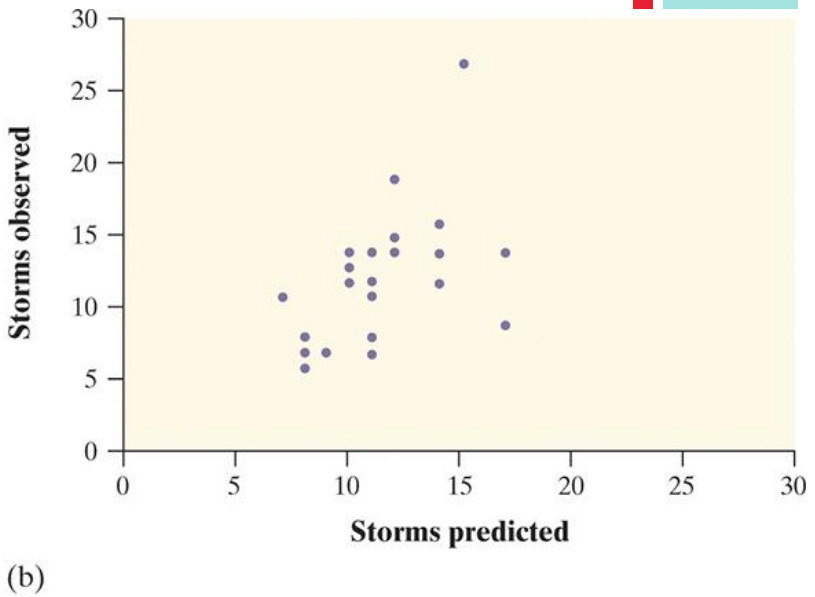
Statistics & Data Analysis, 5th edition – For AP*

PECK, OLSEN, DEVORE

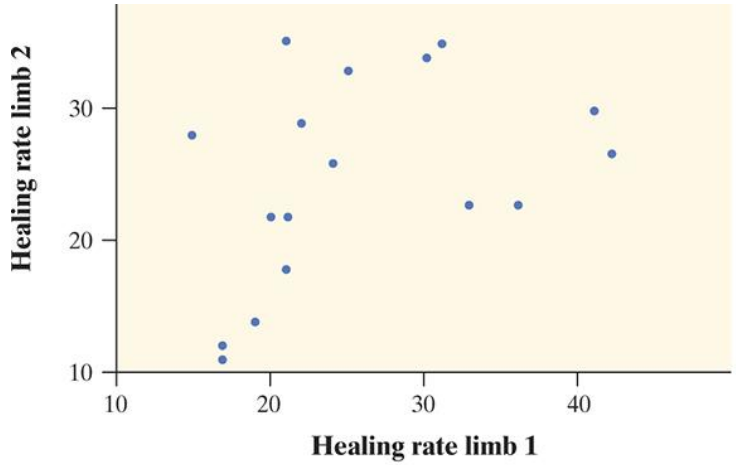
Scatterplots: like "dot plots" for two variables of interest *(bivariate data)*



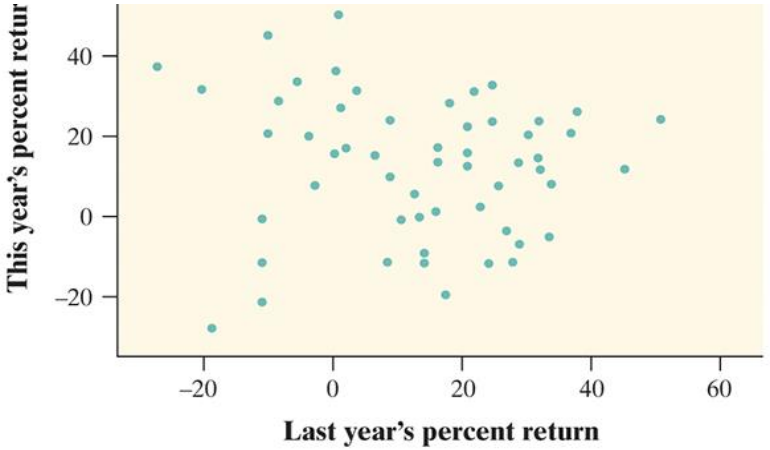
(a)



(b)



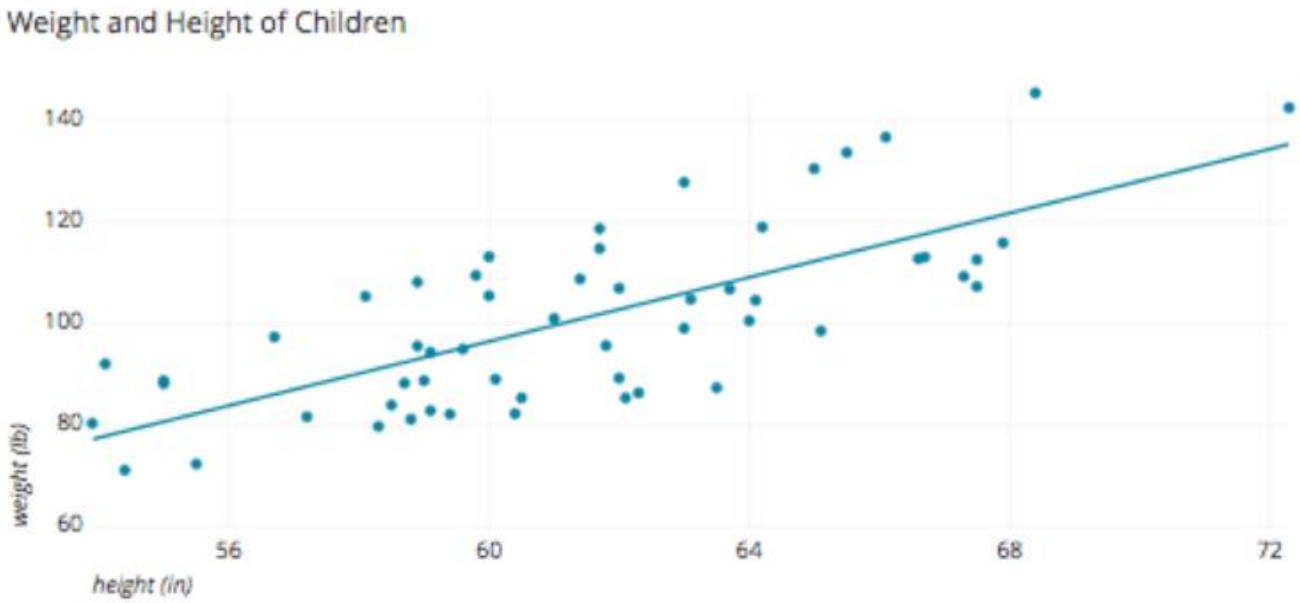
(c)



(d)

Scatterplots: like “dot plots” for two variables of interest *(bivariate data)*

- **A scatter plot, also known as a scatter graph or a scatter chart, is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis.**



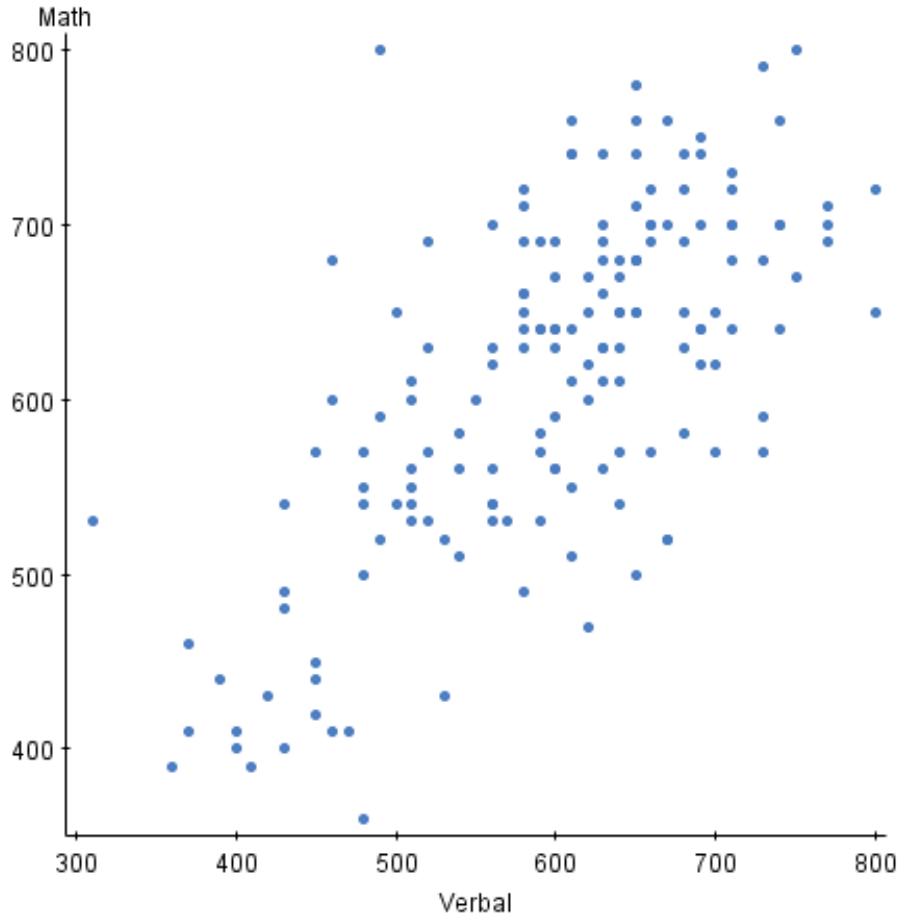
Example: this scatter plot shows the height and weight of a set of children.

Scatter plots or scatterplots? (to-may-to or *to-ma-to)*

- Scatter plots are used when you want to show the *relationship between two variables*. Scatter plots are sometimes called ***correlation plots*** because they show how two variables are correlated. In the height and weight example, the chart wasn't just a simple list of the height and weight of a set of children, but it also visualized the relationship between height and weight - namely that weight increases as height increases.

- Notice that the relationship isn't perfect (because the scatter plot **does not** form a *perfect line*), some taller children weight less than some shorter children, but the general trend is pretty strong and we can see that weight is correlated with height.

How do SAT Math and SAT Verbal scores correlate?



Measuring Linear Association: Correlation

A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables.

Linear relationships are important because a straight line is a simple pattern that is quite common. Unfortunately, our eyes are not good judges of how strong a linear relationship is.

Definition:

The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

- *r* is always a number between -1 and 1
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 towards -1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.

Correlation

The formula for *r* is a bit complex. It helps us to see what correlation is, but in practice, you should use your calculator or software to find *r*.

How to Calculate the Correlation *r*

Suppose that we have data on variables *x* and *y* for *n* individuals.

The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on.

The means and standard deviations of the two variables are *x-bar* and s_x for the *x*-values and *y-bar* and s_y for the *y*-values.

The correlation *r* between *x* and *y* is:

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right)+\ldots+\left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

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

Facts about Correlation

How correlation behaves is more important than the details of the formula. Here are some important facts about *r*.

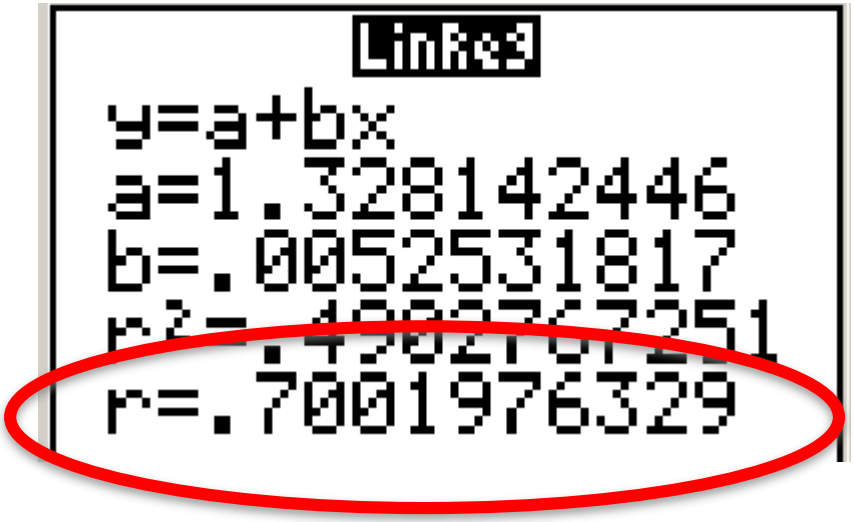
1. **Correlation makes no distinction between explanatory and response variables.**
2. ***r* does not change when we change the units of measurement of *x*, *y*, or both.**
3. **The correlation *r* itself has no unit of measurement.**

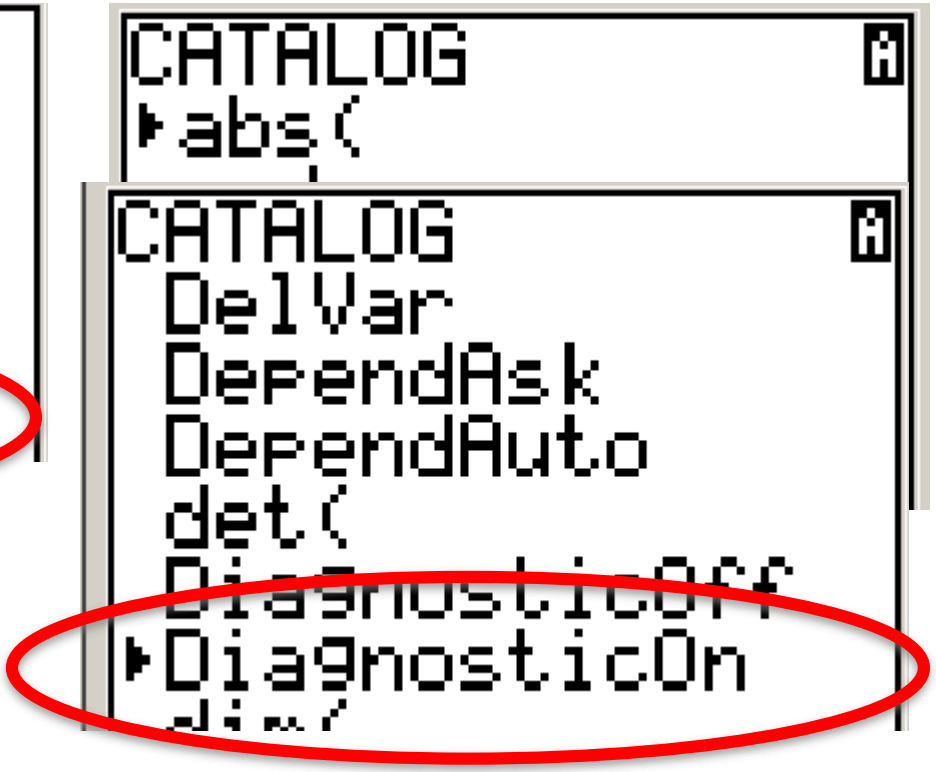
Cautions:

- Correlation requires that both variables be **quantitative.**
- Correlation does not describe curved relationships between variables, no matter how strong the relationship is.
- Correlation is not resistant. *r* is strongly affected by a few outlying observations (Beware of outliers!).
- Correlation is not a complete summary of two-variable data.

Correlation Coefficient (r) on your calculator

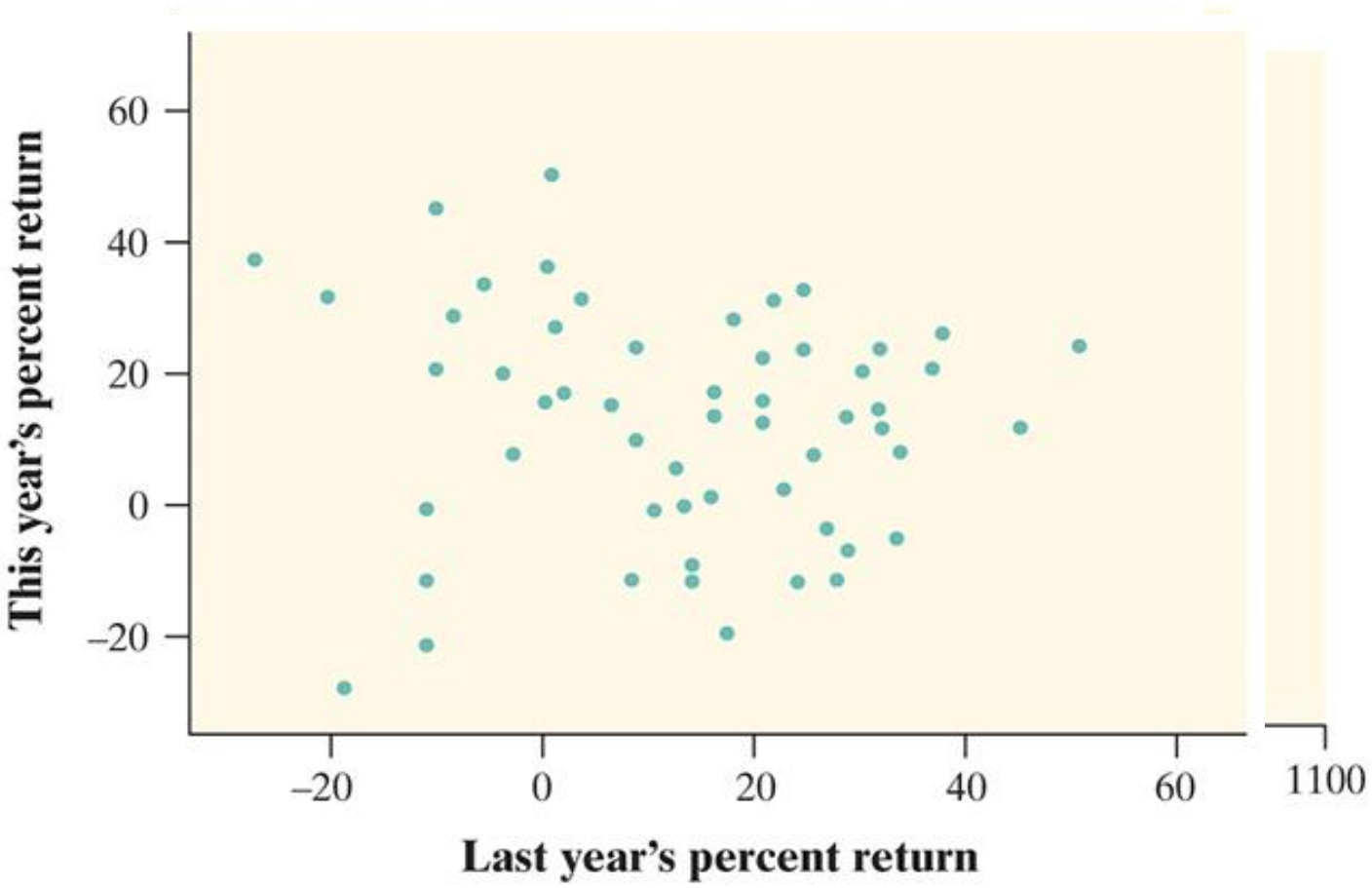
- Turn Diagnostics ON for your calculator to show these





Correlation Practice

For each graph, estimate the correlation *r* and interpret it in context.



(d)