3.1

Scatterplots and Correlation

Most statistical studies examine data on more than one variable. In many of these settings, the two variables play different roles.

Explanatory variable (independent) – "predicts" changes in response variable

Response variable (dependent) – varies depending on the explanatory variable

One way to display the relationship between two variables is a scatterplot.

Scatterplot:

1. Shows the relationship between TWO *quantitative* variables measured on the same individuals.
2. The only choice for displaying the relationship between two quantitative variables.
3. Explanatory variable always gets plotted on the x-axis (x).
4. Response variable gets plotted on the y-axis (y).
5. Axes usually do not start at 0 and often not on the same scale.
6. The scale on each axis must be consistent.
7. Clearly label the variable on each axis.

Interpreting scatterplots:

1. **Direction:** Does the pattern have a positive association, negative association, or neither?
2. **Form:** straight, curved, something exotic, or no pattern? Are there any clusters?
3. **Strength:** Do the points appear tightly clustered or are they widely scattered?
    a. Numerical methods help us determine strength (Correlation *r*)
4. **Outliers:** individual value that falls outside the overall pattern of the relationship
    a. If covering up one value makes the form go from nonlinear to linear, you should call it a linear relationship with an outlier.
5. If asked to describe a scatterplot, you are expected to discuss the above *in context* as well as possible outliers. Use the acronym **DOFS** to help.
    a. Association does not imply causation!

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.

Example #1

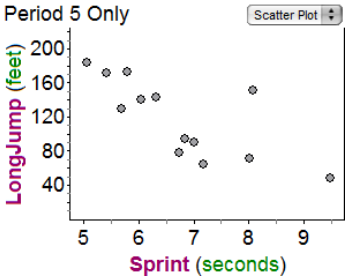
Julie wants to know if she can predict a student's weight from his or her height. Information about height is easier to obtain than information about weight. Jim wants to know if there is a relationship between height and weight. Identify the explanatory and response variable, if possible.

Julie is treating a student's height as the explanatory variable and the student's weight as the response variable. Jim is just interested in exploring the relationship between the two variables, so there is no clear explanatory or response variable.

Example #2

The table below shows data for 13 students in a statistics class. Each member of the class ran a 40-yard sprint and then did a long jump (with a running start). Make a scatterplot of the relationship between sprint time (in seconds) and long-jump distances (in inches). Interpret the scatterplot.

Sprint time (s)	5.41	5.05	9.49	8.09	7.01	7.17	6.83	6.73	8.01	5.68	5.78	6.31	6.04
Distance (in)	171	184	48	151	90	65	94	78	71	130	173	143	141



Direction: Students who take longer to run the sprint typically have shorter jumps. This means there is a negative association between sprint time and distance jumped.

Form: There is a somewhat linear pattern in the scatterplot.

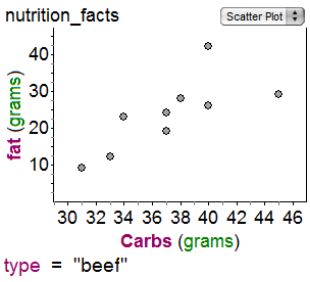
Strength: Since the points do not closely conform to a linear pattern, the association is not strong.

Outliers: There is one possible outlier—the student who took 8.09 seconds for the sprint but jumped 151 inches.

Example #3

Over the past few years, many people have gone on "low-carb" diets while others have tried "low-fat" diets. Here are the carbohydrate contents in grams (g) and fat contents in grams for nine different types of hamburgers at McDonalds. What is the relationship between the amount of carbs and the amount of fat in McDonald's hamburgers?

Type	Carbs(g)	Fat(g)
Hamburger	31	9
Cheeseburger	33	12
Double Cheeseburger	34	23
Quarter Pounder	37	19
Quarter Pounder w/ Cheese	40	42
Double Quarter Pounder w/ Cheese	40	42
Big Mac	45	29
Big N' Tasty	37	24
Big N' Tasty w/ Cheese	38	28



The scatterplot shows a positive association—hamburgers with more carbs tend to have more fat. The form is linear and fairly strong. There are no obvious outliers.

As the amount of carbs increases, the amount of fat increases linearly.

Correlation (correlation coefficient *r*)

Measures the direction and strength of the linear relationship between two quantitative variables.

Properties:

1. Always a number between -1 and 1
2. r > 0 for positive association and r < 0 for negative association
3. Values of r near 0 indicate a very weak linear relationship
4. r = 1 or r = -1 *only* when the points lie exactly along a straight line
5. A correlation near 1 doesn't mean that the relationship is linear. It is easy to create curved data with a correlation close to 1

Facts About Correlation:

1. *r* does not change when you change the units of x, y, or both

2. *r* does not have a unit of measurement. It is just a number

3. Both variables must be quantitative

4, Measures the strength of *only* the linear relationship between two variables. Variables can be strongly associated but still have a small correlation if the association isn't linear.

5. Correlation is not resistant

Interpreting Strength:

.8 to 1.0 (very strong relationship)

.6 to .8 (strong relationship)

.4 to .6 (moderate relationship)

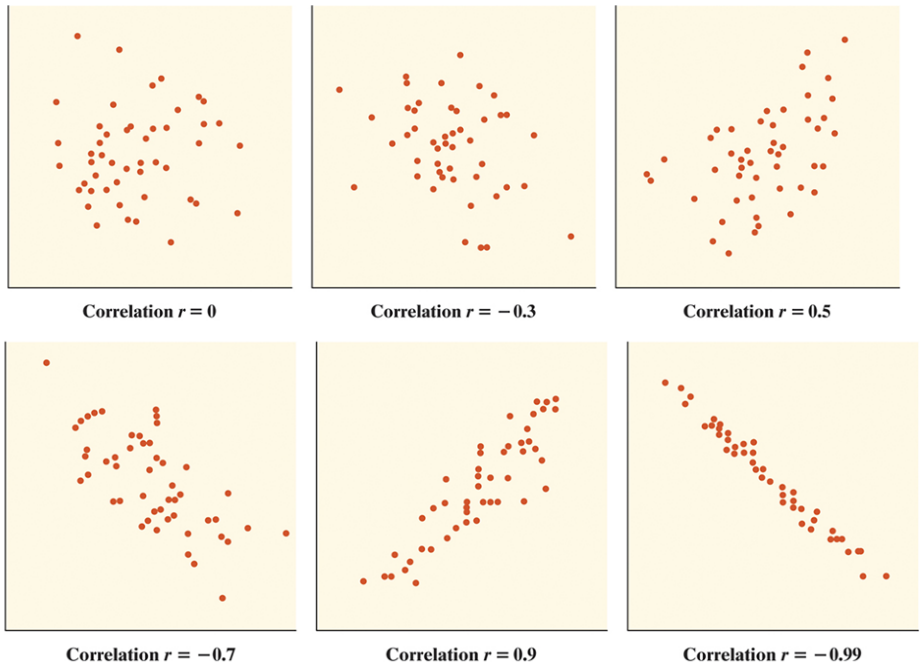
.2 to .4 (weak relationship)

0 to .2 (weak or no relationship)

Association vs. Correlation:

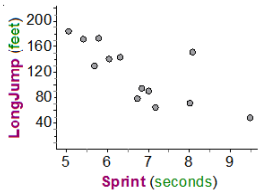
1. Association is a deliberately vague term describing the relationship (not necessarily linear) between variables. Association is NOT causation! There could be other things affecting this relationship that the researchers don't know about.
2. Correlation is a *precise* term describing the strength and direction of the *linear* relationship between two quantitative variables.

Linear Association ⟷ Correlation



Example #5

Here is the scatterplot of the sprint time and long-jump distance data from Example #2. The correlation coefficient r = -0.75.



a) Explain what this value means?

There is a strong, negative linear association between sprint times and long jump distances.

b) What effect would removing the student at (8.09, 151) have on the correlation? At (9.49, 48)?

It would be closer to -1 since this point is outside the pattern of the rest of the data. The new correlation is r = -0.88

It would make the correlation closer to 0 since this point was in line with the rest of the data and particularly influential since its x-value is larger than any of the others. The new correlation is r = -0.67