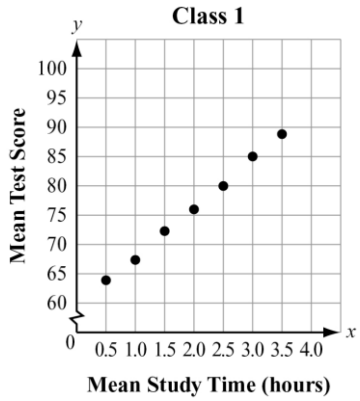
Correlation

MCC9-12.S.ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
MCC9-12.S.ID.9 Distinguish between correlation and causation.

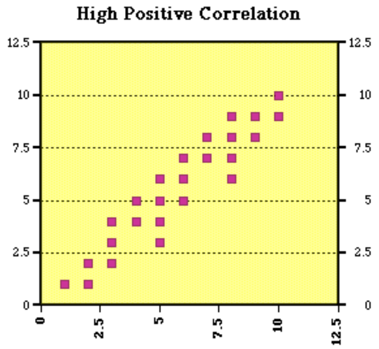
A **scatter plot** is often used to present bivariate **quantitative** data. Each variable is represented on an axis and the axes are labeled accordingly.

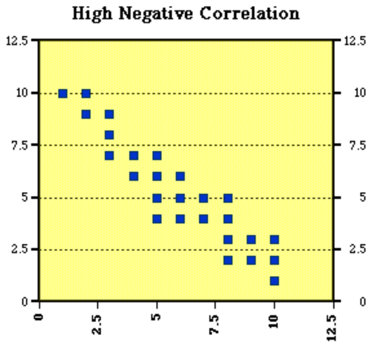


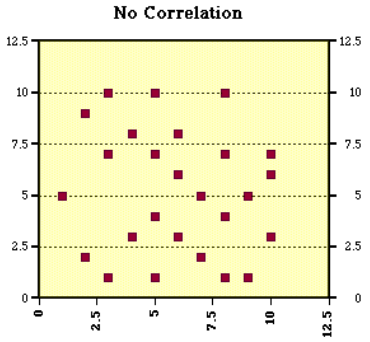
A scatter plot displays data as points on a grid using the associated numbers as coordinates or ordered pairs (x, y). The way the points are arranged by themselves in a scatter plot may or may not suggest a relationship between the two variables. For instance, by reading the graph below, do you think there is a relationship between the hours spent studying and exam grades?

If y tends to increase as x increases, then the data have **positive** correlation.

If y tends to decrease as x increases, then the data have **negative** correlation.





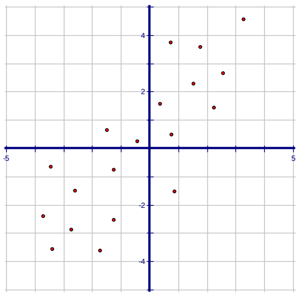


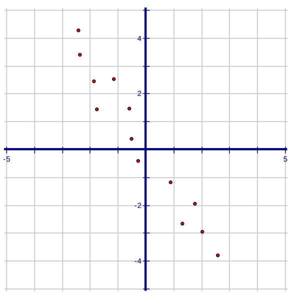
A correlation coefficient, denoted by r, is a number from -1 to 1 that measures how well a line fits a set of data pairs (x, y). If r is near 1, the points lie close to a line with a positive slope. If r is near -1, the points lie close to a line with a negative slope. If r is near 0, the points do not lie close to any line.

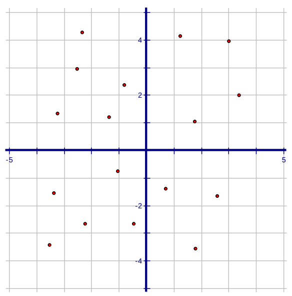
<u>**Practice Problems**</u>:

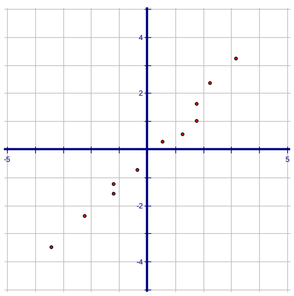
For each scatter plot, tell whether the data have a positive correlation, a negative correlation, or no correlation. Then, match the correlation coefficient to the appropriate scatter plot.

1. _______________ 2. _______________ 3. _______________ 4. _______________









A. $r = 0.05$ B. $r = -0.67$ C. $r = 0.89$ D. $r = 0.42$

3. Positive, negative, or no correlation?

a. Amount of exercise and percent of body fat ____________________

b. A person's age and the number of medical conditions they have ___________________

c. Temperature and number of ice cream cones sold ___________________

d. The number of students at Hillgrove and the number of dogs in Atlanta ___________________

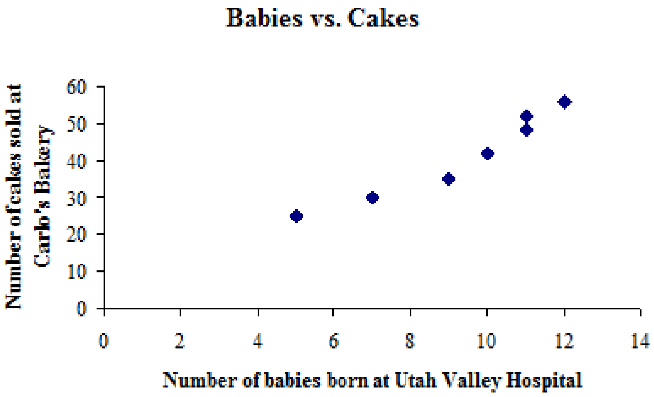
e. Age of a tadpole and the length of its tail ___________________

Correlation vs. Causation

When a scatter plot shows a correlation between two variables, even if it's a strong one, there is *not necessarily a cause-and-effect relationship*. Both variables could be related to some third variable that actually causes the apparent correlation. Also, an apparent correlation simply could be the result of chance.

Example 1: During the month of June the number of new babies born at the Utah Valley Hospital was recorded for a week. Over the same time period, the number of cakes sold at Carlo's Bakery in Hoboken, New Jersey was also recorded. What can be said about the correlation? Is there causation? Why or why not?

Number of babies born	Number of cakes sold
5	25
7	30
9	35
10	42
11	48
11	52
12	56



Example 2: Mr. Jones gave a math test to all the students in his school. He made the startling discovery that the taller students did better than the short ones. His Causation Statement: *As your height increases, so does your math ability.*
What can be said about the correlation? Is there causation? Why or why not?

Example 3: In this present economy families are trying to find ways to save money. Families might be thinking about not eating out to spend less money. Causation Statement: *The more you eat out, the more money you spend at restaurants.*
What can be said about the correlation? Is there causation? Why or why not?

Scatter Plots and Line of Best Fit

MCC9-12.S.ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
MCC9-12.S.ID.6a Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use the given functions or choose a function suggested by the context. Emphasize linear and exponential models.
MCC9-12.S.ID.6c Fit a linear function for a scatter plot that suggests a linear association.

The **best fitting line or curve** is the line that lies as close as possible to all the data points.

Regression is a method used to find the equation of the best fitting line or curve.

Line of Best Fit using the calculator

1) Use the table below to answer the questions about the population p (in millions) in Florida.

Year, t	**2002**	**2003**	**2004**	**2005**
Population (millions)	16.4	17.0	17.4	17.8

a) Find the best-fitting line for the data and the correlation coefficient.

b) Using this model, what will be the population in 2020?

2) Use the table below to answer the questions about the U.S. residential carbon dioxide emissions from 1993 to 2002. Emissions are measured in million metric tons.

Year, t	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**
Emissions	1027.6	1020.9	1026.5	1086.1	1077.5	1083.3	1107.1	1170.4	1163.3	1193.9

a) Find the best-fitting line for the data and the correlation coefficient.

b) Using this model, how many residential tons were emitted in 1990? In 2010?