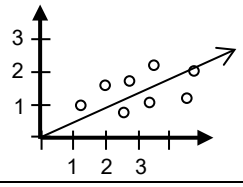
STUDENT RESOURCES

Word or Phrase	Definition
association	In statistics, an association between two variables is a relationship between the variables, so that the variables are statistically dependent. In the case of numerical variables, if the relationship is linear, we refer to a linear association between the variables.
bivariate data	Bivariate data is data that has two variables. Bivariate data can be represented by ordered pairs. A list of country of origin and batting average for each baseball player is a bivariate data set with one categorical variable and one numerical variable.
bivariate numerical data	Bivariate numerical data is data that has two numerical variables. Bivariate numerical data can be represented by a scatter plot, so that the relationship (if any) between the variables is more easily seen. A list of heights and weights for each player on a football team is a bivariate numerical data set.
categorical data	Categorical data is data sorted into categories, such as colors, ranges of measurements, or other attributes of the data. Generally, there are only finitely many categories.
data set	A data set is a collection of pieces of information about a population, often numbers, obtained from observation, questioning, or measuring.
frequency table	A frequency table is a table that lists items and the number of times they occur in a data set.
line of best fit	A line of best fit for a scatter plot is a straight line that best represents (in some sense) the data points in the scatter plot. 
measurement data	Measurement data is numerical data that comes from making measurements. Measurement data can be obtained by measuring such things as heights, weights, temperatures, lengths, areas, and volumes.
numerical data	Numerical data is data consisting of numbers. The numbers allow for statistical calculations, such as finding the mean or median.
outlier	An outlier of a data set is a data value that is unusually small or unusually large relative to the overall pattern of values in the data set. For the data set {1, 1, 1, 3, 5, 6, 6, 7, 23}, the data value 23 is a potential outlier.

Word or Phrase	Definition
population	In statistics, the population refers to the source of a data set. If we wish to make statistical inferences about the students at a school, we may take a random sample of the students, or we may gather data from all the students. In either case, the population refers to the students in the school.
relative frequency table	A relative frequency table is a frequency table that lists items and the proportion (or percent) of times they occur.
statistical question	A statistical question is a question where numerical data that has potential for variability can be collected and analyzed for the purpose of answering the question. A statistical question: “How much TV do middle school students watch on average?” NOT a statistical question: “How many hours of TV did you watch last week?”
two-way table	A two-way table is a table that displays bivariate categorical data, in which the rows correspond to the categories of one variable, and the columns correspond to the categories of the other. A two-way table that includes the number of data observations is called a "two-way frequency table". A two-way table that includes the percentage of the number of data observations relative to the total number of observations is called a "two-way relative frequency table".

Numerical Data

Numerical data is data consisting of numbers. Measurement data is numerical data that comes from making measurements.

Numerical survey questions are used to collect numerical data. Numerical data typically come from counting or measurements. Examples of numerical survey questions include:

- How many dogs do you own? (a counting question)
- How many minutes did you exercise last week? (a measurement question)

Some ways to report one-variable (or univariate) numerical data include:

- Measures of center such as mean, median, mode
- Measures of spread such as range, mean absolute deviation (MAD), and 5-number summary
- Data displays such as tables, line plots, histograms, and box plots

Some ways to report two-variable (or bivariate data) numerical data include:

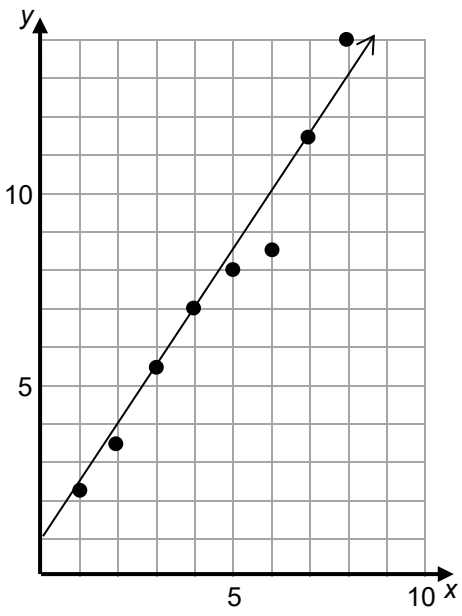
- Tables
- Graphs
- Equations

Lines of Best Fit

A line of best fit for a scatter plot is a straight line that best represents (in some sense) the data points in the scatter plot.

Example: When the data in the table below is graphed in a scatter plot, the data points cluster along a straight line. We conclude that there is likely a linear association between x and y. One possible such line may be estimated by the equation graphed below, $y = \frac{3}{2}x + 1$. Using a graphing calculator, another estimated equation is given as $y = 1.6x + 0.3$ (not graphed).

x	1	2	3	4	5	6	7	8
y	2.2	3.5	5.5	7	8	8.5	11.5	14

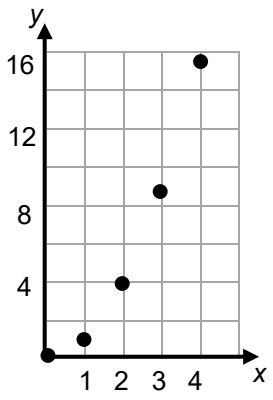


Nonlinear Associations

Not all associations are linear. Here is an example of a scatter plot of bivariate data that appears to have a nonlinear association.

Example: For this data set, the graphed points do not fall in a linear pattern. They increase at an increasing rate.

x	0	1	2	3	4
y	0	1	4	9	16

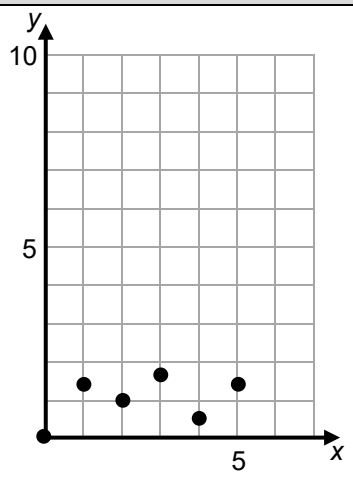


Outliers

An outlier of a data set is a data value that is unusually small or unusually large relative to the overall pattern of values in the data set.

Outliers can create the illusion that an association exists when one does not. They can also distract us from seeing an association when there clearly is one.

Example 1: In the scatter plot to the right, the data point (6, 10) is a potential outlier. Its *y*-coordinate 10 appears to be unusually large compared to the other *y*-coordinates.



Example 2: In a 6th grade classroom, students were asked how many pets they had. All students but one replied with numbers of pets that ranged from 0 to 8. That one pet owner said she had 40 fish. This number of fish appears to be an outlier, because it is unusually large compared to the other numbers of pets.

Categorical Data

Categorical data is data sorted into categories, such as colors, ranges of measurements, or other attributes of the data. Generally, there are only finitely many categories.

Categorical survey questions are used to collect categorical data. Responses to these questions are usually in words. Examples of categorical survey questions include:

- What types of pets do you own? (Answers include dog, cat, bird, no pets, etc.)
- Do you have a curfew? (A yes-no answer)

Some ways to report one-variable categorical data include

- Frequency tables
- Relative frequency tables
- Pie charts (circle graphs)
- Bar graphs

Some ways to report two-variable categorical data include:

- Two-way frequency tables
- Two-way relative frequency tables