STUDENT RESOURCES

Word or Phrase	Definition
experimental probability	In a repeated probability experiment, the experimental probability of an event is the number of times the event occurs divided by the number of trials. This is also called empirical probability. If, in 25 rolls of a number cube, we obtain an even number 11 times, we say that the experimental probability of rolling an even number is $\frac{11}{25}$ = 0.44 = 44%.
measure of center	A measure of center is a statistic describing the middle of a numerical data set. The mean, the median, and the mode are three commonly used measures of center. For the data set {3, 3, 5, 6, 6}, the mean (average) is $\frac{(3+3+5+6+6)}{5}$ = 4.6, and the median is 5. There are two modes, 3 and 6. Each of these numbers can be viewed as the "center" of the data set in some way.
measure of spread	A measure of spread (or a measure of variability) is a statistic describing the variability of a numerical data set. It describes how far the values in a data set are from the mean or median. The standard deviation (SD or σ), the mean absolute deviation (MAD), and the interquartile range (IQR) are three measures of spread.
population	The population is the entire group of individuals (objects or people) to which a statistical question refers. If a survey is taken to investigate how many pets the students at Seaside School own, the population under study is the entire student body of Seaside School.
sample	A sample is a subset of the population that is examined in order to make inferences about the entire population. The sample size is the number of elements in the sample. In order to estimate how many phones coming off the production line were defective, the plant manager randomly selected a sample of 50 phones and tested them to see if they worked properly.
simulation	Simulation is the imitation of one process by means of another process. We may simulate rolling a number cube by drawing a card blindfold from a group of six identical cards labeled one through six. We may simulate the weather by means of computer models.
theoretical probability	The theoretical probability of an event is a measure of the likelihood of the event occurring. In the probability experiment of rolling a (fair) number cube, there are six equally likely outcomes, each with probability $\frac{1}{6}$. Since the event of rolling an even number corresponds to 3 of the outcomes, the theoretical probability of rolling an even number is 3 out of 6, or $3 \bullet \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$.

Dot Plots (Line Plots)

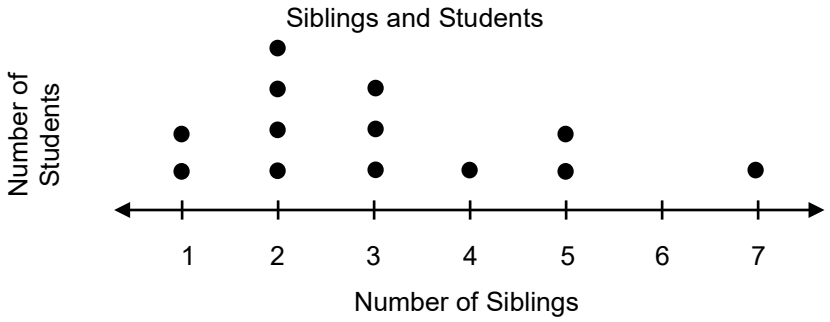
A dot plot (also called a line plot) displays data on a number line with a dot (•) or an X to show the frequency of data values.

Here are the number of siblings (brothers and sisters) for 13 different students:

3, 4, 5, 2, 2, 3, 3, 2, 2, 5, 7, 1, 1

To make a dot plot of this data set:

- Make a number line that extends from the minimum data value to the maximum data value
- Mark a dot or an X for every data value
- Write a title and add vertical and horizontal labels.



Measures of Center

Here are the number of siblings for 13 different students:

3, 4, 5, 2, 2, 3, 3, 2, 2, 5, 7, 1, 1

To find the mean (average) of a data set, add all the values in the data set and divide the total by the number of values (number of observations, *n*).

Number of observations: $n = 13$
To find the mean: $3 + 4 + 5 + 2 + 2 + 3 + 3 + 2 + 2 + 5 + 7 + 1 + 1 = 40$
$40 \div 13 \approx 3.08$

To find the median (*M*), order the values from least to greatest and find the middle number. If there is an even number of values in the data set, the median is the mean (average) of the two middle numbers.

For the siblings data set: {1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 5, 7}
↑
median

To find the mode, find the value that occurs most often. (Some data sets may have more than one mode.)

For the siblings data set, the mode is 2. It is the value 2 occurs most often.

The Range, the Quartiles, and the Five-Number Summary

Here are the number of siblings for 13 different students:

3, 4, 5, 2, 2, 3, 3, 2, 2, 5, 7, 1, 1

To find the range of a data set, find the difference between the greatest value and the least value in the data set.

For the siblings data set, the range is 6, since 7 – 1 = 6.

To find quartiles, first put the numbers in numerical order. Then locate the points that divide the data set into four equal parts.

1	1	2	2	2	2	3	3	3	4	5	5	7

min Q_1 $Q_2 = M$ median Q_3 *max*

For the siblings data set:
$Q_1 = 2$ (the 1st quartile)
$Q_2 = 3$ (the 2nd quartile; also the median)
$Q_3 = 4.5$ (the 3rd quartile)

Q_1 is the median of the first half of the data set, and Q_3 is the median of the second half.

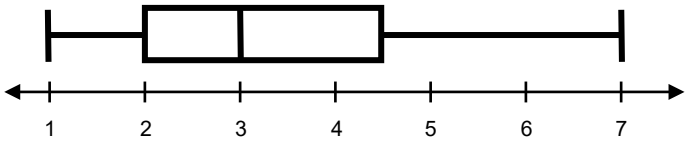
The five-number summary is (*min*, Q_1, Q_2, Q_3, *max*) = (1 , 2 , 3 , 4.5 , 7).

Box Plots (Box-and-Whisker Plots)

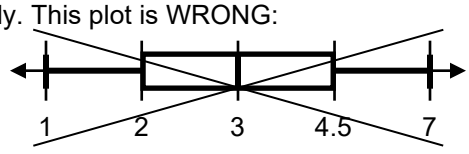
A box plot (or box-and-whisker plot) provides a visual representation of the center and spread of a data set. The display is based on the five-number summary.

For the sibling data, the five-number summary is (1 , 2 , 3 , 4.5 , 7).

To make a box plot:
- Locate the five-number summary values on a number line, and indicate each value with a vertical segment
- Create a "box" to highlight the interval from the first to the third quartile, and draw "whiskers" that extend to the minimum and maximum

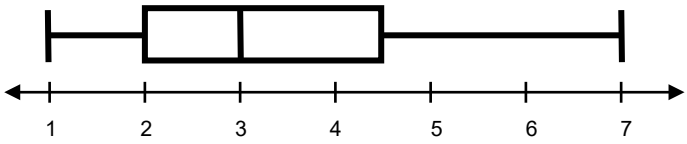


Be sure to scale the box plot properly. This plot is WRONG:



Interpreting a Box Plot

This is a box plot.



- Each of the four "sections" (the two whiskers and the two rectangular parts of the box) contains (close to) one-fourth of the data points. Be careful: If one section appears larger than another, we cannot say it has more data points, but only that the data points are spread out over a wider range.

- Sometimes we use the word "quartile" to refer to specific data points. Sometimes the word "quartile" is also used to refer to one of the four quarters, or sections, of the data set. For example, data points that lie within the farthest left section may be referred to as "in the first quartile."

Mean Absolute Deviation

The mean absolute deviation (MAD) is a measure of spread of a numerical data set. It is the arithmetic average of the distance (absolute value) of each data point to the mean. To calculate the MAD statistic:

For the sibling data, there are 13 data points:

3, 4, 5, 2, 2, 3, 3, 2, 2, 5, 7, 1, 1

To find the MAD statistic:

- Find the mean of the sample.

  The mean is 3.08.

- Find the distance (absolute value) from each data point to the mean.

  See the table entries to the right.

- Find the sum of the distances.

  See the bottom row of the table.

- Divide the sum of the distances by the number of data points to find the average distance from the mean.

  See the calculation below.

Sibling Data	Distance from data point to mean
3	$\lvert 3.08 - 3 \rvert = 0.08$
4	$\lvert 3.08 - 4 \rvert = 0.92$
5	$\lvert 3.08 - 5 \rvert = 1.92$
2	$\lvert 3.08 - 2 \rvert = 1.08$
2	$\lvert 3.08 - 2 \rvert = 1.08$
3	$\lvert 3.08 - 3 \rvert = 0.08$
3	$\lvert 3.08 - 3 \rvert = 0.08$
2	$\lvert 3.08 - 2 \rvert = 1.08$
2	$\lvert 3.08 - 2 \rvert = 1.08$
5	$\lvert 3.08 - 5 \rvert = 1.92$
7	$\lvert 3.08 - 7 \rvert = 3.92$
1	$\lvert 3.08 - 1 \rvert = 2.08$
1	$\lvert 3.08 - 1 \rvert = 2.08$
Sum of distances from mean	17.4

$$\text{MAD} = \frac{\text{sum of distances from mean}}{\text{number of data points}} = \frac{17.4}{13} = 1.34$$

Sampling

Sampling refers to selecting a subset of a population to be examined for the purpose of drawing statistical inferences about the entire population. If the sample is representative of the entire population, we may make valid inferences about the entire population based on properties of the sample.

Suppose you want to know how many hours per week students in our school spend watching television. From the population of all students, you select a sample and you ask the students in the sample how many hours they watch television. You would like to infer that the average time spent watching TV for all students is about the same as for students in the sample.

An easy way to select a sample might be to ask your friends how many hours they watch TV. Such a sample is called a convenience sample. However, your friends may not be representative of all students.

To select a more representative sample, you might assign a number to the name of each student in the school, and then use a computerized "random number generator" to select a certain number of the students' numbers. This sort of sample is referred to as a random sample. Its mathematical properties allow us to draw inferences about the population.

The Percent Error Formula

A percent error computation is used to quantify the difference between an estimated value and an actual value as a percentage of the actual value. The formula for percent error (as a percent) is:

$$\text{Percent Error}_{\text{(as a percent)}} = \frac{|\text{actual} - \text{estimate}|}{\text{actual}}\text{, written as a percent}$$

If the estimated number of fish in a lake is 45, and the actual number is 50, then the percent error is

$$\frac{|50 - 45|}{50} = \frac{|5|}{50} = \frac{1}{10} = 10\%$$

If the estimated number of fish in a lake is 50, and the actual number is 45, then the percent error is

$$\frac{|45 - 50|}{50} = \frac{|-5|}{50} = \frac{1}{10} = 10\%$$