Handout #3: Summarizing data

Collected data is summarized using numerical quantities by calculating:

a. Measure of central tendency: a value that represent central entry of a data.
Includes: the mean, the median and the mode.

b. Measure of dispersion (variability): is a value that quantify the variability in the data.
Includes: the range, interquartile range, standard deviation and variance

c. Measure of position: describe the position of a particular data value within a given data set.
Includes: percentiles, quartiles, and standard scores (aka, z-scores)

In Handout #3, you will learn about measures of central tendencies. In Handout #4, you will learn about measures of dispersion & measures of position.

Measures of Central Tendency

a. Mean (aka average) = $\frac{\text{Sum of the data values}}{\text{\# of data values}}$

Population mean	Sample mean
$\mu =$	$\bar{x} =$

Here

$\sum$:	N :
x :	n :

Typically population mean is unknown, the sample mean is used to estimate it.

Example: Mark operates Technology Titans, a Web site service that employs 8 people. The age of his employees are: 55, 63, 34, 59, 29, 46, 51, 41

i. Is this a sample or population data?

ii. Find the mean age.

Example: Consider the sample data values: 12, 8, 11, 10, 7, 10, 15, 13, 14 and 9. Calculate an average. How do you use it to describe your data?

b. <u>Median</u> (M): the "middle value" of the ordered data set. It divides the data into two equal parts.

For odd # of values – median is the middle data value.

For even # of values – median is the mean of two middle data values.

Example: Find the median of the following data: 12, 2, 16, 8, 14, 10, 6

Example: Find the median of the following data: 7, 9, 3, 4, 11, 1, 8, 6, 1, 4

c. <u>Mode</u> (M_0): Value(s) that occurs most frequently in a data set.

Example: Consider the data: 5, 6, 7, 7, 7, 7, 7, 7, 8, 9, 9, 10, 11, 12, 13, 13, 15.

A data set can have one mode, more than one mode, or no mode.

If no entry is repeated, the data set has no mode.

If two entries occur with the same greatest frequency, each entry is a mode and the data set is called bimodal.

i. 21, 29, 24, 31, 27, 23, 32, 33, 19.

ii. 23, 21, 29, 24, 31, 21, 27, 23, 24, 32, 33, 19.

Mode is the only measure of central tendency for qualitative data.

Example: You begin to observe to the color of clothing your employees wear. Your goal is to find what color is worn most frequently so that you can offer company shirts to your employees.

Monday: Red, Blue, Black, Pink, Green, and Blue
Tuesday: Green, Blue, Pink, White, Blue, and Blue
Wednesday: Orange, White, White, Blue, Blue, and Red
Thursday: Brown, Black, Brown, Blue, White, and Blue
Friday: Blue, Black, Blue, Red, Red, and Pink

What is the mode of the colors above?

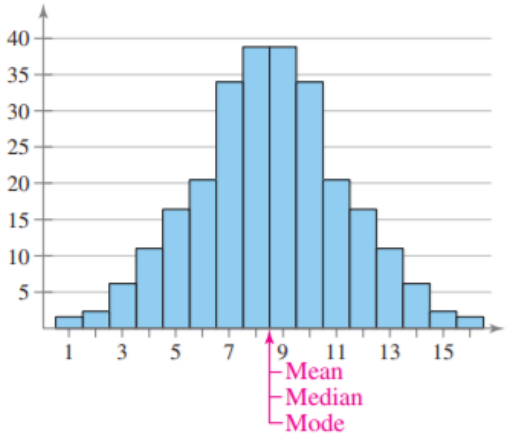
Comparing the Mean, Median, and Mode:

Example: Find the mean, median, and mode of the sample ages of a class shown. Which measure of central tendency best describes a typical entry of this data set? Are there any outliers?

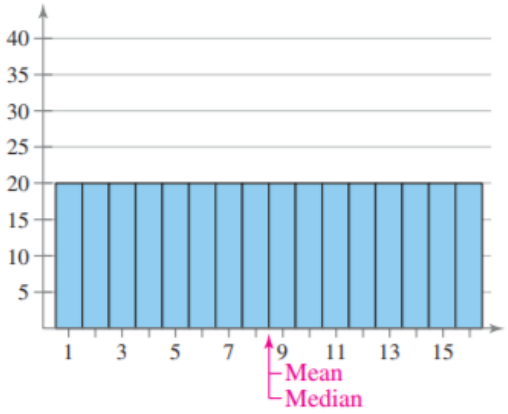
Ages in a class:	20	20	20	20	20	20	21	21	21	21
	22	22	22	23	23	23	23	24	24	65

Which measure was more affected by the outlier?

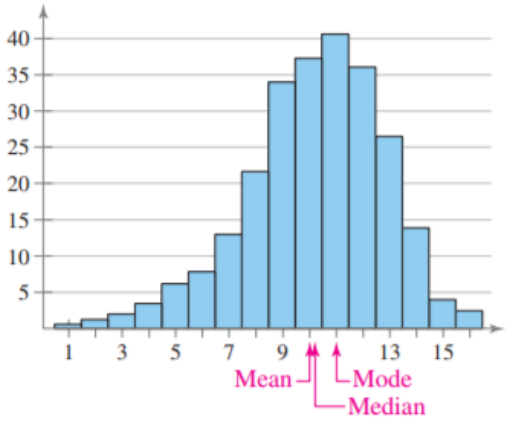
Shapes of Distributions and location of mean, median and mode



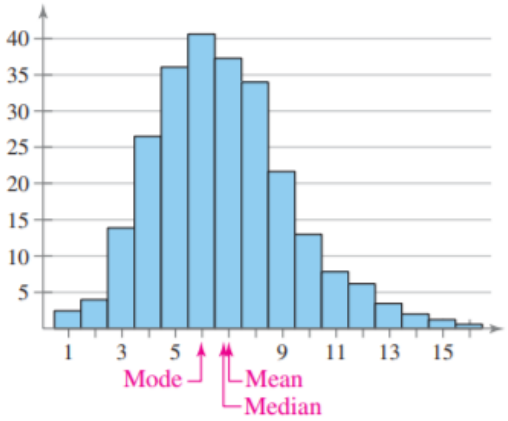
Symmetric Distribution



Uniform Distribution



Skewed Left Distribution



Skewed Right Distribution

The most appropriate measure of central tendency depends on the data set:

For skewed data?

For symmetric data?

For categorical data?

Weighted Mean:

Is a mean of a set of numbers in which some elements of the set carry more importance (weight) than others.

Data value: $x_1,\ x_2, \ldots\ldots\ldots, x_n$

Weight: $w_1, w_2, \ldots\ldots\ldots, w_n$

Then, weighted mean ($\bar{x}_w$) =

Example: The scores and their percent of the final grade for an archeology student are shown below. What is the student's mean score?

	Score	Percent of final grade
Articles reviews	95	10%
Quizzes	100	10%
Midterm exam	89	30%
Student lecture	100	10%
Final exam	92	40%

Example: A student receives the following grades, with an A worth 4 points, a B worth 3 points, a C worth 2 points, and a D worth 1 point. What is the student's mean grade point score?

A in 2 three-credit classes, B in 1 two-credit class, D in 1 three-credit class, and an C in 1 four- credit class.

Mean of Grouped Data:

Mean $(\bar{x}) = \frac{\sum(f.\ x)}{n}$, where $n = \sum f$, $x =$ midpoint

Class	Frequency
51-55	2
56-60	7
61-65	8
66-70	4

Trimmed Mean

To find the 10% **trimmed mean** of a data set, order the data, delete the lowest 10% of the entries and the highest 10% of the entries, and find the mean of the remaining entries.

Consider the data:
44 51 11 90 76 36 64 37 43 72 53 62 36 74 51 72 37 28 38 61 47 63 36 41 22 37 51 46 85 13

(a) Find the mean using all the data entries.
(b) Find the 10% trimmed mean for the data.
(c) Compare the means from (a) and (b).
(d) What is the benefit of using a trimmed mean versus using a mean found using all data entries? Explain your reasoning.