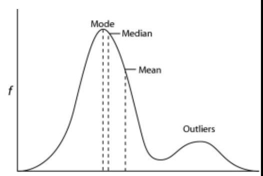
MDM4U: Mathematics of Data Management

Measures of Central Tendency (Part 1)

Mean, Median and Mode

MDM4U: Data Management



Mean, Median and Mode

The *mean* of a data set (commonly called the *average*) is the sum of all values, divided by the number of values.

For a population, the mean is represented by the Greek letter μ ("mu"), while for a sample it is represented by $\overline{x}$.

For a population, $\mu = \dfrac{x_1 + x_2 + \ldots + x_N}{N} = \dfrac{\sum x}{N}$.

For a sample, $\overline{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n} = \dfrac{\sum x}{n}$.

For example, the mean of the data set $\{3, 4, 5\}$ is $\dfrac{3+4+5}{3} = 4$.

Mean, Median and Mode

The *median* is the central datum when they are ordered (either from highest to lowest or vice versa).

If there is an even number of values, the median is the mean of the two central values.

For example, the median of the data set $\{1, 2, 3, 4, 5\}$ is 3, whereas the median of the data set $\{4, 3, 2, 1\}$ is 2.5.

Mean, Median and Mode

The *mode* is the most frequently occurring value in a data set.

A data set may have no mode if all values occur with the same frequency, or several modes if more than one value occurs with the highest frequency.

Consider the data sets $A = \{1, 2, 2, 1, 2, 3\}$, $B = \{4, 5, 6, 7\}$ and $C = \{8, 9, 8, 9, 10\}$.

The mode of A is 2. B has no mode. The modes of C are 8 and 9.

Mean, Median and Mode

Example

Five randomly-selected employees at a company earn annual salaries of \$32 000, \$42 000, \$32 000, \$40 000 and \$150 000. Determine the mean, median and mode of the data.

There are five values to use in the calculation of the mean.

$\overline{x} = \dfrac{32\,000 + 42\,000 + 32\,000 + 40\,000 + 150\,000}{5}$ or \$59 200

To find the median, order the data from highest to lowest: \$150 000, \$42 000, \$40 000, \$32 000, \$32 000. The median is the central value, or \$40 000.

The mode is \$32 000 since it occurs twice in the data set.

Appropriateness of Measures of Central Tendency

Depending on the situation, some measures of central tendency are more appropriate than others.

The mode, for example, uses the lowest value in the data. Is this the best measure of the centre of the data?

The mean was inflated by the highest salary, \$150 000. Is this the best measure?

The median represents the "middle" salary. Is the median a useful measure for salaries?

Appropriateness of Measures of Central Tendency

Some general rules of thumb that may or may not apply to a situation:

- If there are extreme values that do not represent the general nature of the data (called *outliers*), the mean may not be appropriate, as it may be skewed.
- If the frequency of the data is important, the mode might be a better measure of central tendency.
- If further mathematical analysis of the data is required, the mean is generally the best measure for use.

Note that these rules are not set in stone, and the "best" measure of central tendency is situation-specific.

Appropriateness of Measures of Central Tendency

Example

Which measure of central tendency is most appropriate for each scenario?

- Mr. Garvin examines the scores from the last unit test. Median.
- An insurance company examines the ages of 1000 individuals involved in car accidents. Mode.
- The Ministry of Transportation examines the model years of 5000 cars on a highway. Mode.

The Effect of Outliers

What effect do outliers have on the mean, median and mode of a data set?

Consider the data sets $A = \{1, 2, 2, 4, 6\}$ and $B = \{1, 2, 2, 4, 51\}$. Calculate the mean, median and mode of each data set.

For set A, the mean is 3, the median is 2 and the mode is 2.

For set B, the mean is 12, the median is 2 and the mode is 2.

In general, outliers can greatly affect the mean, but often have little or no effect on the median or mode.

The Effect of Outliers

Example

A fisherman records the length, in centimetres, of 10 bass caught in a stream.

15	22	19	18	15	45	27	18	18	51

- Determine the mean, median and mode of the data.
- Identify any outliers.
- Determine the mean, median and mode of the data with any outliers removed.
- Discuss the appropriateness of each measure of central tendency.

The Effect of Outliers

Example (Continued)

For the original data set, the mean is 24.8 cm, the median is 18.5 cm and the mode is 18 cm.
There are two outliers, at 45 cm and 51 cm.
With the two outliers removed, the mean is 19 cm. The median is 18 cm, and the mode remains at 18 cm.

The Effect of Outliers

Example (Continued)

Any of the measures of central tendency may be appropriate for this data set, depending on the situation.
If the fisherman wants to know the most frequent size of fish caught, then the mode is the best measure.
If he wishes to know the "average" length he can expect to catch, the mean is the best measure.
To complete the phrase "half of my catch is at least X centimetres," the median is the best measure.