Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.

As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

❖ **Mean:** The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data. There are three types of mean:

1. **Arithmetic Mean:** The AM is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have 'n' values in a data set and they have values x_1, x_2, x_3,…,x_n the sample AM, usually denoted by $\bar{x}$ (pronounced "x bar"), is:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. **Geometric Mean:** The Geometric mean of set of 'n' positive observations equals the n^{th} positive root of the product of these observations. For n positive observations x_1, x_2, x_3,…,x_n the GM is denoted by G is define as:

$$G = \sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

For group of data which has size N:

Observation	x_1	x_2	x_3	… … …	x_n
Frequency	f_1	f_2	f_3	… … …	f_n

$$G = \sqrt[N]{(x_1)^{f_1} \times (x_2)^{f_2} \times (x_3)^{f_3} \times \ldots \times (x_n)^{f_n}}$$

$$\Rightarrow \log(G) = \log\left(\sqrt[N]{(x_1)^{f_1} \times (x_2)^{f_2} \times (x_3)^{f_3} \times \ldots \times (x_n)^{f_n}}\right)$$

$$\Rightarrow \log(G) = \log(((x_1)^{f_1} \times (x_2)^{f_2} \times (x_3)^{f_3} \times \ldots \times (x_n)^{f_n})^{\frac{1}{n}})$$

$$\Rightarrow \log(G) = \frac{1}{N} \times (\log(\ (x_1)^{f_1} \times (x_2)^{f_2} \times (x_3)^{f_3} \times \ldots \times (x_n)^{f_n}))$$

$$\Rightarrow \log(G) = \frac{1}{N} \times (\log((x_1)^{f_1}) + \log((x_2)^{f_2}) + \log((x_3)^{f_3}) + \cdots + \log((x_n)^{f_n}))$$

$$\Rightarrow \log(G) = \frac{1}{N} \times (\ f_1 \log x_1 + f_2 \log x_2 + f_3 \log x_3 + \cdots + f_n \log x_n\ )$$

$$\Rightarrow \log(G) = \frac{1}{N} \times \sum_{i=1}^{n} f_i \log x_i$$

$$\Rightarrow G = antilog\left(\frac{1}{N} \times \sum_{i=1}^{n} f_i \log x_i\right) \ where, N = \sum_{i=1}^{n} f_i$$

$$\therefore G = 10^{\left(\frac{1}{N} \times \sum_{i=1}^{n} f_i \log x_i\right)} \text{ [As the log has base 10]}$$

3. **Harmonic Mean:** The Harmonic mean of set of 'n' positive observations equals the reciprocal of the arithmetic mean of the individual values. For n observations x_1, x_2, x_3,…,x_n the HM is denoted by H is define as:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

When not to use the mean

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	2	3	4	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is $30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the $12k to 18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide

the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values. This is explained in more detail in the skewed distribution section later in this guide.

- **Median:** The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. For n positive observations $x_1, x_2, x_3, \ldots, x_n$ the Median is denoted by M_e is define as:

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & when\ n\ is\ odd \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}}\right) & when\ n\ is\ even \end{cases}$$

- **Mode:** The mode is value of distribution for which the frequency is maximum. In other words, mode is the value of a variable which occur with highest frequency. Mode can be null or multiple. There could be three case in getting mode.

    a. **Case-**1: S = {2, 5, 2, 7, 3, 2, 4, 5} here, mode = 2
    b. **Case-**2: S = {1, 2, 3, 4, 5, 6, 7, 8} here, mode=null as the set contain unique elements.
    c. **Case-**3: S = {2, 5, 7, 4, 3, 5, 7, 1} here, mode=5 and 7.