Statistic

Statistics is the science of collecting, organizing and interpreting numerical facts which we often call **data**. Synonyms for data are scores, measurements and observations. The study and collection of data involves classifying data in various heads. The process involves lot of representations of a characteristic by numbers and it is termed as **measurement**. In other words **Data** are measurements of a situation under consideration.

Example: The measurements of heights of all creatures in world is a data. All numerical characteristics are called **variables**. A large number of observations on a single variable can be summarized in a table of frequencies. Any particular pattern of variation is termed as distribution.

1 MEASURES OF CENTRAL TENDENCY

The most commonly used measures of central tendency are

- The mode
- The median
- The arithmetic mean

1.1 THE MODE

The mode is the most frequently occurring value in a distribution of a variable.

1.2 THE MEDIAN

The median is the middlemost point in a rank ordered set of measures.

If the number of observations is odd, then the median is $\left(\frac{n+1}{2}\right)$ th observation. If the number of observations is even, then median is the mean of $\left(\frac{n}{2}\right)$ th and $\left(\frac{n}{2}+1\right)$ th observations.

1.3 THE ARITHMETIC MEAN

The arithmetic mean is defined as the sum of the values divided by the total number of values. $X=\left(\frac{\sum X_i}{N}\right)$. The mean is not necessarily the middle of the distribution, as is the median. The mean is the point in a distribution about which deviations from it sum to zero. A deviation score (*x*) is defined as the distance between a value and its mean. They can be either positive (+) or negative (-).In this sense the mean is a *centroid*.

It is observed that the measures of central tendency are not sufficient to give complete information about the given data. “Variability” is another factor which is required to be studied under statistics. The single number which describes variability, is known as ‘Measures of dispersion’.

2 MEASURES OF DISPERSION

Dispersion means **‘Scatteredness’**. Dispersion measures the degree of scatteredness of the variable about a central value.

There are following measures of dispersion:

- Range
- Quartile deviation
- Mean deviation
- Variance

- Standard deviation.

  In this chapter, we shall study all of these measures of dispersion, except the Quartile deviation.

- **RANGE**

  The range is the simplest measure of variation to find. The usual definition of range is the difference between the maximum and minimum values of a population.

  RANGE = MAXIMUM VALUE – MINIMUM VALUE

  For example, consider the following series

60	60	60	60	60	60	60	60	60	60	60	Range = 0
0	2	3	15	20	60	89	91	95	99	100	Range = 100
0	49	50	51	54	60	74	75	76	78	100	Range = 100

- Since the range only uses the largest and smallest values, it is greatly affected by extreme values.
- The range of data gives us a rough idea of variability or scatter.

3 MEAN DEVIATION

Mean deviation of a distribution is the arithmetic mean of the absolute deviations of the terms of the distribution from its statistical mean (arithmetic mean, median or mode).

- Mean deviation may be obtained from any measure of central tendency. However, mean deviation from mean and median are commonly used in statistical studies.
- Mean deviation about the median is least.

3.1 MEAN DEVIATION FOR UNGROUPED DATA

Let $x_1, x_2, x_3, ..., x_n$ are n values of a variable X and *k* be the statistical mean (A.M., median, mode) about which we have to find the mean deviation. The mean deviation (M.D.) about *k* is given by

$$\text{M.D.}(k) = \frac{|x_1 - k| + |x_2 - k| + |x_3 - k| + ... + |x_n - k|}{n} = \frac{\sum_{i=1}^{n} |x_i - k|}{n}$$

3.2 MEAN DEVIATION FOR GROUPED DATA

(a) Discrete Frequency Distribution:

Let $x_1, x_2, x_3, ..., x_n$ be n observations occurring with frequencies $f_1, f_2, f_3, ..., f_n$ respectively and *k* be the statistical mean (A.M., median, mode). The mean deviation (M.D.) about *k* is given by

$$\text{M.D.}(k) = \frac{|x_1 - k|f_1 + |x_2 - k|f_2 + |x_3 - k|f_3 + ... + |x_n - k|f_n}{f_1 + f_2 + f_3 + ... + f_n}$$

$$= \frac{\sum_{i=1}^{n} |x_i - k| f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} |d_i| f_i}{N},$$

Where $d_i = x_i - k$ and $N = \sum_{i=1}^{n} f_i$ = total frequency

- The mean of given discrete frequency distribution is given by

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_n x_n}{f_1 + f_2 + f_3 + ... + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

- To find the median of given discrete frequency distribution, observations are arranged in ascending order. After this, cumulative frequencies are obtained, then the observation is identified whose cumulative frequency is equal to or just greater than $\frac{N}{2}$, this value of the observation lies in the middle of the data, it is the required median.

(b) Continuous Frequency Distribution:

The mean of a continuous frequency distribution is calculated with the assumption that the frequency in each class is centered at its mid-point.

Let x_i be the mid-value of the i^{th} class, f_i be the frequency of the i^{th} class and k be the statistical mean (A.M., median, mode), then the mean deviation (M.D.) about k is given by

$$\text{M.D.}(k) = \frac{\sum_{i=1}^{n}|x_i - k|f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n}|d_i|f_i}{N}, \quad \text{Where } d_i = x_i - k \text{ and } N = \sum_{i=1}^{n} f_i = \text{total frequency}$$

- **SHORTCUT METHOD FOR CALCULATING MEAN DEVIATION ABOUT MEAN**

For the given continuous frequency distribution, arithmetic mean can be calculated by shortcut (step-deviation)method. Rest of the procedure is same.

In this method,

(i) Take an assumed mean (just middle data or close to it).

(ii) Calculate deviations of the observations (or mid-point of classes) from the assumed mean.

(iii) If there is a common factor of all the deviations, divide them by this common factor to simplify the deviations.

(iv) Now the arithmetic mean $(\bar{x})$ by step-deviation method is given by

$$\bar{x} = a + \frac{\sum_{i=1}^{n} f_i d_i}{N} \times h, \qquad \text{where } d_i = \frac{x_i - a}{h}$$

a = assumed mean, h = common factor and $N = \sum_{i=1}^{n} f_i$

- **TO CALCUATE THE MEDIAN FOR A CONTINUOUS FREQUENCY DISTRIBUTION**

(i) Calculate $\frac{N}{2}, \left(N = \sum_{i=1}^{n} f_i\right)$.

(ii) The class corresponding to cumulative frequency just more than $\frac{N}{2}$ is known as median class.

(iii) Median = $l + \frac{h}{f}\left(\frac{N}{2} - c\right)$,

where l = lower limit of median class
f = frequency of the median class
h = width of the median class
c = cumulative frequency of the class just preceding the median class

(i) The sum of the absolute deviations about the mean is greater than the sum of the absolute deviations from median, in fact mean deviation about median is least. Therefore mean deviation about mean is not very suitable.

(ii) In the series, where the degree of variability is very high, the median is not a representative of central tendency. The mean deviation about the mean is not a very good measure of dispersion.

(iii) Mean deviation is calculated on the basis of absolute values of the deviations and therefore can not be subjected to further algebraic treatment.

Important formulae/points

- *Mean deviation about the median is least.*
- *Mean deviation*

Ungrouped data: $M.D. = \dfrac{\sum_{i=1}^{n} | x_i - k |}{n}$ *; k be the statistical mean (A.M., median, mode)*

Discrete/ continuous frequency distribution: $M.D. = \dfrac{\sum_{i=1}^{n} | x_i - k | f_i}{\sum_{i=1}^{n} f_i}$

where x_i *= observations/ mid-value of* i^{th} class

f_i *=frequency of* i^{th} observation/ class

n = number of observations

k = statistical mean (A.M., Median, Mode)

4 VARIANCE AND STANDARD DEVIATION

While calculating the mean deviation, the absolute values of the deviations were taken to avoid the difficulty which arose due to the signs of deviation. The another way is to take squares of all the deviations.

4.1 VARIANCE

The variance of a variate is the arithmetic mean of the squares of all deviations from mean (A.M.) and is denoted by σ^2 or var(x).

Therefore, if $x_1, x_2, x_3, ..., x_n$ be n given values of a variate and $\bar{x}$ be their mean, then

$$\text{Variance } (\sigma^2) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

4.2 STANDARD DEVIATION

In the calculation of variance, the units of individual observations x_i and the unit of their mean $(\bar{x})$ are different from that of variance. The proper measure of dispersion about the mean of a set of observations is expressed as positive square root of variance and is known as standard deviation (σ).

Hence $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

4.3 VARIANCE AND STANDARD DEVIATION IN DIFFERENT CASES

(a) In case of individual series (ungrouped data):

Let $x_1, x_2, x_3, ...x_n$ are n values of a variable x, then by definition

Variance, $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$,

(where $\bar{x}$ is A.M. of $x_1, x_2, x_3, ...x_n$ i.e., $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$)

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}(\bar{x})^2\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2$$

and standard deviation,

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2}$$

(b) In case of discrete frequency distribution:

Let $x_1, x_2, x_3, \ldots\ldots\ldots x_n$ be n observations having frequency $f_1, f_2, f_3, \ldots\ldots, f_n$ respectively, then

Variance, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i$,

(where $\bar{x}$ is A.M. of $x_1, x_2, x_3, \ldots\ldots\ldots x_n$ i.e., $\bar{x} = \frac{\sum x_i f_i}{N}$ and $N = \sum_{i=1}^{n} f_i$)

$$= \frac{1}{N}\sum\left(x_i^2 - 2\bar{x}x_i + (\bar{x})^2\right)f_i = \frac{\sum x_i^2 f_i}{N} - 2\bar{x}\frac{\sum x_i f_i}{N} + (\bar{x})^2\frac{\sum f_i}{N}$$

$$= \frac{\sum x_i^2 f_i}{N} - 2\bar{x}\,\bar{x} + (\bar{x})^2 = \frac{\sum x_i^2 f_i}{N} - (\bar{x})^2$$

Hence standard deviation, $\sigma = \sqrt{\frac{\sum x_i^2 f_i}{N} - (\bar{x})^2}$

(c) In case of continuous frequency distribution:

Let x_i = mid-value of i th class

f_i = frequency of i th class

$N = \sum_{i=1}^{n} f_i$ (total frequency)

$\bar{x}$ = A.M. of given observations

then variance, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i = \frac{\sum x_i^2 f_i}{N} - (\bar{x})^2$

and standard deviation, $\sigma = \sqrt{\frac{\sum x_i^2 f_i}{N} - (\bar{x})^2}$

$$= \frac{1}{N}\sqrt{N\sum_{i=1}^{n} f_i x_i^2 - \left(\sum_{i=1}^{n} f_i x_i\right)^2} \text{ , as } \bar{x} = \frac{\sum f_i x_i}{N}$$

4.4 SHORTCUT METHOD TO FIND VARIANCE AND STANDARD DEVIATION: (When x_i are large)

Let the assumed mean be A and the width of class interval be h.

Let $u_i = \frac{x_i - A}{h} \Rightarrow x_i = A + hu_i$ …………….. (1)

The arithmetic mean $\bar{x} = \frac{\sum f_i x_i}{N}$

$$= \frac{\sum f_i (A + hu_i)}{N}$$

$$\Rightarrow \quad \bar{x} = A + h\frac{\sum f_i u_i}{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2)}$$

from (1) and (2),

$$x_i - \bar{x} = h\left(u_i - \frac{\sum f_i u_i}{N}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3)}$$

Now, variance(σ^2) = $\dfrac{\sum f_i (x_i - \bar{x})^2}{N} = \dfrac{h^2}{N}\sum\left(u_i - \dfrac{\sum f_i u_i}{N}\right)^2 f_i$ {using (3)}

$$= \frac{h^2}{N}\sum (u_i - \bar{u})^2 f_i \quad = \quad h^2 \times \text{(variance of variable } u_i\text{)}$$

$$\Rightarrow \sigma_x{}^2 = h^2 \sigma_u{}^2 \quad \Rightarrow \sigma_x \ = h\sigma_u \quad \ldots\ldots\ldots\ldots\text{(4)}$$

$$\Rightarrow \sigma_x = \frac{h}{N}\sqrt{N\sum f_i u_i{}^2 - \left(\sum f_i u_i\right)^2} \quad \text{as} \quad \sigma = \frac{1}{N}\sqrt{N\sum f_i x_i{}^2 - \left(\sum f_i x_i\right)^2}$$

5 ANALYSIS OF FREQUENCY DISTRIBUTION

In order to compare the variability of two series with same mean, which are measured in different units, merely calculating the measures of dispersion are not sufficient, but we require such measures which are independent of the units. The measure of variability which is independent of units is called coefficient of variation(C.V.) and defined as

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100, \quad \bar{x} \neq 0$$

Where σ and $\bar{x}$ are standard deviation and mean of data respectively.

The series having greater C.V. is said to be more variable than the other series. The series having lesser C.V. is said to be more consistent than the other.

- **COMPARISION OF TWO FREQUENCY DISTRIBUTION WITH SAME MEAN**

Let σ_1 and σ_2 be the standard deviation of two series with mean $\bar{x}$, then

$$\text{C.V. (I}^{\text{st}}\text{ distribution)} = \frac{\sigma_1}{\bar{x}} \times 100, \quad \bar{x} \neq 0$$

$$\text{C.V. (II}^{\text{nd}}\text{ distribution)} = \frac{\sigma_2}{\bar{x}} \times 100, \quad \bar{x} \neq 0$$

Hence, the above two C.V. can be compared on the basis of values of σ_1 and σ_2 only.

Therefore, for the two series with equal means, the series with greater standard deviation (or variance) is called more variable or dispersed than the other, Also the series with lesser value of standard deviation (or variance) is said to be more consistent than the other.