Measure of dispersion :

The measurement of the scatter of the values of a data set among themselves is called a measure of dispersion or variation.

Measures of central tendency serve to locate the 'center' of a distribution but they do not reveal how the items or the observations are spread out or scattered on each side of the center. Absence of dispersion in the data indicates perfect uniformity. This situation arises when all observations in the distribution are identical.

Properties of a good measure of variation :

A good measure of variation should possess, as far as possible, the following properties:

(i) It should be simple to understand.

(ii) It should be easy to compute.

(iii) It should be rigidly defined.

(iv) It should be based on each and every observation of the distribution.

(v) It should be amenable to further algebric treatment.

(vi) It should have sampling stability.

(vii) It should ~~be~~ not be unduly affected by extreme observations.

The frequently used measures of dispersion are:

a) Absolute measures

(i) The range

(ii) The quartile deviation

(iii) The mean deviation

~~(iv) The variance~~

(iv) The standard deviation

b) Relative measures

i) Co-efficient of QD (ii) Co-efficient of variation (iii) Co-efficient of MD

Range: Range is the simplest method of studying variation. It is defined as the difference between the value of the smallest observation and the value of the largest observation included in the distribution.

Symbolically, $Range = L - S$

L = Largest value and

S = Smallest value.

For example - for a set of values: 7, 4, 1, 10, 9, 2 and 8, the range is given by:

$R = 10 - 1 = 9$

Merits, Demerits, Uses : Rabindra Nath Shil - Subash Chandra Debnath

Quartile deviation:

A measure similar to the above measures is the inter-quartile range (Q). It is the difference between the third quartile (Q_3) and the first quartile (Q_1). Thus

$$Q = Q_3 - Q_1$$

The inter-quartile range is frequently reduced to the measure of semi-interquartile range, also known as the quartile deviation (QD), by dividing it by 2. Thus

$$QD = \frac{Q_3 - Q_1}{2}$$

same as before.

Scanned with CamScanner

Quartile divide a set of observations into four equal parts. Hence 25 percent of the observations will be less than the first quartile. Seventy-five percent of the observations will be less than the 3rd quartile.

To locate the first quartile, we use formula –

$$L_p = (n+1)\frac{p}{100} \quad ; \; p = 25 \text{ . let } n = 15$$

$= 4$ th position

To locate the 3rd quartile, we use formula –

$$L_p = (n+1)\frac{p}{100} \quad ; \; p = 75$$

$= 12$ th position .

$$Q_i = l_i + \frac{h}{f_R}\left(\frac{in}{4} - F_{-1}\right)$$

Box plots:

A box plot is a graphical display based on quartiles, that helps us to picture a set of data. To construct a box plot, we need only five statistics: the minimum value, Q_1, the median, Q_3 and the maximum value.

These quantities are known as the five-number summary of a distribution. The box extends from the Q_1 to Q_3 representing the inter-quartile range and so encloses the 50% of the values. The whiskers are the lines that extend from the box to the highest and lowest values and thus illustrate the range. A line across the box indicates the median (Q_2). The edges of the box are known as hinges which are approximated by Q_1 and Q_3 values. The points lying beyond 1.5 times the inter-quartile range (i.e. above Q_3 and below Q_1) are known as outliers.

let,

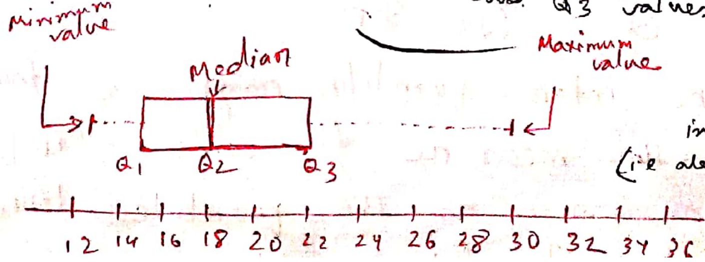
minimum value $= 13$ minutes

$Q_1 = 15$ minutes

Median $= 18$ minutes

$Q_3 = 22$ minutes

Maximum value $= 30$ minutes

edges =


Scanned with CamScanner

Mean deviation:

The mean deviation is an average of absolute deviations of individual observations from the central value of a series.

If $x_1, x_2, \ldots, x_n$ form a sample of observation the formula for computing the average or mean deviation from arithmatic mean is –

$$MD(\bar{x}) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^{n} |d_i|}{n}$$

where, $d_i = x_i - \bar{x}$, which stands for the deviations of the individual observations from the mean.

If a grouped data frequency distribution is constructed, as is usually done with large samples the average deviation is –

$$MD(\bar{x}) = \frac{\sum_{i=1}^{K} f_i |x_i - \bar{x}|}{n}$$

where, $MD(\bar{x})$ = average deviation about mean

K = number of classes.

x_i = mid point of the ith class

f_i = frequency of the ith class

$n = \sum_{i=1}^{K} f_i$.

Merits, Demerits, Uses.

II Co-efficient of variation:

The ratio of the standard deviation to the arithmetic mean, expressed as a percent.

In terms of a formula for a sample:

$$CV = \frac{S}{\bar{x}} \times 100$$

[multiplying by 100 converts the decimal to a percent]

It is a very useful measure when:

(i) The data are in different units (such as dollars and days absent).

(ii) The data are in the same units, but the means are far apart (such as the incomes of the top executives and the incomes of the unskilled employees).

Moments: A set of descriptive measures, which can provide a unique characterisation of a distribution and hence can determine the distribution uniquely, is called moments.

Standard Deviation :

The arithmatic mean of the squares of the deviations of the given observations from their arithmatic mean is known as variance. The positive square root of variance is the standard deviation.

If $x_1, x_2, \ldots, x_n$ be n observations of a variable, then standard deviation is defined by –

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

In case of frequency distribution or grouped data –

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

where $n = \sum_{i=1}^{n} f_i$ and $\bar{x}$ is the aritmatic mean of the distribution.

Merits, Demerits and Uses of S.D

Moments :

If $x_1, x_2, \ldots x_n$ be n observations of a variate then the rth raw moment is defined by –

$$\mu_r' = \frac{1}{n}\sum_{i=1}^{n}(x_i-A)^r$$; where A is any arbitrary value.

The rth central moment is defined by –

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r$$; where $\bar{x}$ is the AM.

If $x_1, x_2, \ldots, x_K$ occur with frequencies $f_1, f_2, \ldots, f_K$ respectively then the rth raw moment is –

$$\mu_r' = \frac{1}{n} \sum_{i=1}^{K} f_i (x_i - A)^r \text{ where } n = \sum_{i=1}^{K} f_i$$

The rth central moment is defined by –

$$\mu_r = \frac{1}{n} \sum_{i=1}^{K} f_i (x_i - \bar{x})^r, \text{ where } n = \sum_{i=1}^{K} f_i$$

Relation between raw and central moments:

we know, rth raw moment is –

$$\mu_r' = \frac{1}{n} \sum_{i=1}^{n} (x_i - A)^r$$

putting $r = 1, 2, 3, 4, \ldots$ etc

$$\mu_1' = \frac{1}{n} \Sigma (x_i - A)$$

$$\mu_2' = \frac{1}{n} \Sigma (x_i - A)^2$$

$$\mu_3' = \frac{1}{n} \Sigma (x_i - A)^3$$

$$\mu_4' = \frac{1}{n} \Sigma (x_i - A)^4$$

$\vdots$

rth central moment is –

$$\mu_r = \frac{1}{n} \Sigma (x_i - \bar{x})^r$$

putting $r = 1, 2, 3, \ldots$ etc

$$\mu_1 = \frac{1}{n} \Sigma (x_i - \bar{x})$$

$$\mu_2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2$$

$$\mu_3 = \frac{1}{n} \Sigma (x_i - \bar{x})^3$$

$$\mu_4 = \frac{1}{n} \Sigma (x_i - \bar{x})^4$$

$$\mu_1' = \frac{\Sigma(x_i - A)}{n} = \frac{\Sigma x_i - nA}{n} = \bar{x} - A$$

$$\mu_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = \bar{x} - \bar{x} = 0$$

$$\mu_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\Sigma\{(x_i - A) - (\bar{x} - A)\}^2}{n}$$

$$= \frac{\sum_{i=1}^{n}(x_i - A)^2}{n} - 2\frac{\Sigma(x_i - A)}{n}\cdot(\bar{x} - A) + \frac{\sum_{i=1}^{n}(\bar{x} - A)^2}{n}$$

$$= \mu_2' - 2\mu_1'\mu_1' + \mu_1'^2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n} = \frac{\sum_{i=1}^{n}\{(x_i - A) - (\bar{x} - A)\}^3}{n}$$

$$= \frac{\Sigma(x_i - A)^3}{n} - 3\frac{\Sigma(x_i - A)^2}{n}\times(\bar{x} - A) + 3\frac{\Sigma(x_i - A)}{n}\times(\bar{x} - A)^2 - (\bar{x} - A)^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 3\mu_1'\mu_1'^2 - \mu_1'^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

Shape characteristics of a Distribution:

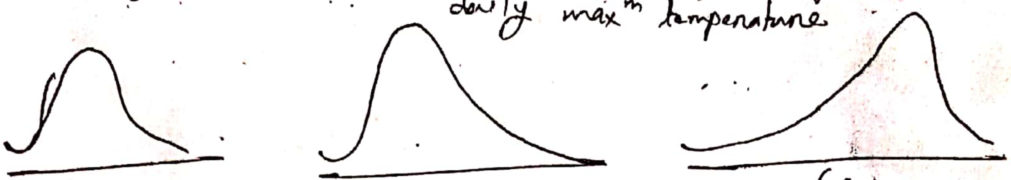
<u>Skewness:</u>

Skewness means "lack of symmetry" i.e. departure from symmetry of a distribution. The skewness may be either positive or negative. When the skewness is positive, the associated distribution is called positively skewed distribution. When the skewness is negative we call the distribution a negatively skewed distribution. Absence of skewness makes the distribution symmetrical.

(a) symmetrical distribution : height, weight, examination score

(b) positively skewed distribution : family size, female age at marriage, wages of the employees

(c) negatively skewed distribution : reaction times for an experiment, daily max^m temperature



(a) (b) (c)

Measures of skewness :

Pearson's coefficient of skewness

$$= \frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

If mean > mode, the skew is positive

If mean < mode, the skew is negative

If mean = mode, the skew is zero, in which case the distribution is symmetrical.

A relative measure of skewness denoted by β_1 is defined as follows :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\mu_2^{3/2}}$$

γ_1 measures the skewness more directly as compared to β_1.

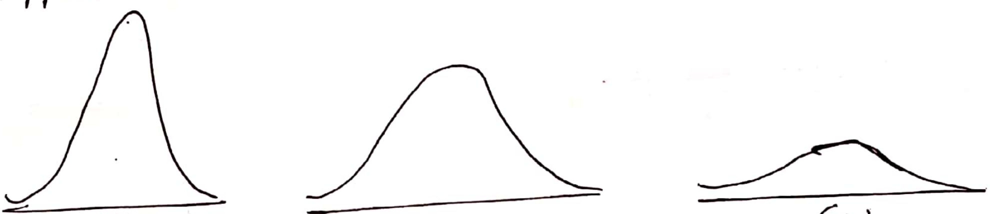
Kurtosis :

~~Kurtosis is defined~~ The degree of peakness or flatness of a distribution relative to a normal distribution is called Kurtosis.

A curve having relatively higher peak than the normal curve, is known as leptokurtic.

If the curve is more flat-topped than the normal curve, it is called platykurtic.

A normal curve itself is called mesokurtic, which is neither too peaked nor too flat-topped.



(a) leptokurtic (b) Mesokurtic (c) Platykurtic

Measures of kurtosis :

The most important measure of kurtosis based on second and fourth moments is β_2, defined as –

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

For normal distribution $\beta_2 = 3$. When the value of β_2 is greater than 3, the curve is more peaked than normal curve, in which case it is leptokurtic.
When the value of β_2 is less than 3, the curve is less peaked than the normal curve, in which case, it is platykurtic.

if $\beta_2 - 3 > 0$, the distribution is leptokurtic

if $\beta_2 - 3 < 0$, the " " platykurtic

if $\beta_2 - 3 = 0$, the " " mesokurtic.

Example: for a distribution, the four central moments were found to be as follows:

$\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.7$

find β_1 and β_2 and hence comment on the nature of the distribution.

Solution:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \quad = 0.031$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3$$

Based on the values of β_1 and β_2 we conclude that the distribution is slightly positively skewed and mesokurtic.

Scanned with CamScanner