Mean square deviation

Suppose $d = x - a$ is a deviation taken from an arbitrary reference point a. To get rid of algebraic sign of d, we either use $|d|$ or d^2. The measure of dispersion based on $|d|$ is mean deviation, we have already study. Using d^2, we can develop measure of dispersion which is better than mean deviation clearly

Mean square deviation $= \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2$ for individual Observation.

$$= \frac{\sum_{i=1}^{n} f_i (x_i - a)^2}{\sum_{i=1}^{n} f_i}$$ for freq^n distribution

However M.S.D. is affected by choice of a. Thus it creates difficulty in measuring dispersion clearly. We try to find a measure which will overcome this difficulty.

Property :-

(i) Sum of square of deviations taken from AM is minimum.

(ii) Minimality property of MSD.

MSD is least if the deviations are taken From AM.

Since, sum of square of deviation taken from

AM is minimum

ie $\sum (x_i - a)^2 \geq \sum (x_i - \bar{x})^2$

$\frac{1}{n}\sum (x_i - a)^2 \geq \frac{1}{n}\sum (x_i - \bar{x})^2$

MSD about a $\geq$ MSD about $\bar{x}$

Variance, standard deviation and coefficient of variation.

Variance :-

The A.M of squares of deviations taken from mean is called as variance. It is denoted as var(x) or σ^2.

$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ for individual observation & $\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}$ for freq^n distn.

Remark: The units of original items and that of variance are not same.

standard deviation (S.D)

It is positive square root of variance. It is denoted by σ.

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

$$= \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i}} \quad \text{for freq}^n \text{ distribution}$$

Simplified formula :→

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

$$= \frac{1}{n}\sum x_i^2 - 2x_i\bar{x} + \bar{x}^2$$

$$= \frac{1}{n}\sum x_i^2 - \frac{2\bar{x}}{n}\sum x_i + \bar{x}^2 \frac{1}{n}\cdot\sum 1$$

$$= \frac{1}{n}\sum x_i^2 - 2\bar{x}\cdot\bar{x} + \bar{x}^2$$

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

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

standard deviation is measure of dispersion which satisfies most of the requisites of good measure. It is free from drawbacks present in the other measure of dispersion.

Coeff^n of variance (C.V).

$$C.V = \frac{S.D}{|AM|} \times 100 = \frac{\sigma}{|\bar{x}|} \times 100$$

C.V is always expressed in percentage

Remarks :—

R.H.S of above includes multiplier 100 because $\frac{\sigma}{|\bar{x}|}$ is too small in many cases thus for convenience it is multiplied by 100

(ii) S.D of weights of a group of elephants will be larger than that of group of human beings suppose S.D of weights of group of elephants is 15kg and that of human beings is also 15 kg. In this case we cannot says both the groups have identical variations This is because average weight of group of elephants is larger than that of average weight of a group of persons Therefore for compairing variations bet^n two different data set, a measure based on ratio of σ & $\bar{x}$ would be appropriate This is achieved in CV

* Uses —

i) for compairing variability of two data sets.

ii) determining uniformity of data smaller CV higher is uniformity

iii) small CV data is more homogeneous

iv) for compairing consistency of data

Stability

Properties :-

(1) Mean square deviation ≥ variance

(2) Variance is invariant to the change of origin. In other words if a constant is added (or subtracted) to each item then variance (SD) remains same.

→ Let $x_1, x_2 \dots x_n$ is set of observations.

$y_i = x_i - a$ where a is constant,

we have to S.T $var(y) = var(x)$.

As

$$y_i = x_i - a \Rightarrow \bar{y} = \bar{x} - a$$

we have

$$var(y) = \frac{1}{n}\sum (y_i - \bar{y})^2$$

$$= \frac{1}{n}\sum [x_i - a - (\bar{x} - a)]^2$$

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

$$= var(x).$$

(3) If $u = \frac{x-a}{h}$, a, h being constant.

Then $var(u) = \frac{1}{h^2} var(x)$.

→ Let $u_i = \frac{x_i - a}{h} \Rightarrow \bar{u} = \frac{\bar{x} - a}{h}$

we have

$$var(u) = \frac{1}{n}\sum (u_i - \bar{u})^2$$

$$= \frac{1}{n}\sum \left[\frac{x_i - a}{h} - \frac{\bar{x} - a}{h}\right]^2$$

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

$$\text{var}(u) = \frac{1}{h^2} \text{var}(x).$$

(4) If $y = ax + b$ Then $\text{var}(y) = a^2 \text{var}(x)$.

Use of variance & SD.

Practically in almost all advanced statistical methods such as sampling, statistical quality control, statistical inference deal with variance. variance & SD are used in no. of situations some of them are as follows:

a) Precision of an instrument is inversely proportional to variance. Therefore

$$\text{precision} = \frac{K}{\text{variance}}$$

b) In portfolio analysis risk is described in terms of variance of prices of shares.

c] for the comparison of performance of two or more instruments, machines, coeff^n of variation is used.

* If all observations are equal, SD is { }
* If data contains only one observation, SD is ?

1) Compute S.D & C.V of marks scored by 10 candidates given below

54, 61, 64, 69, 58, 56, 49, 57, 55, 50

→

$u = x - 57$

x_i	54	61	64	69	58	56	49	57	55	50	Total
u_i	-3	4	7	12	1	-1	-8	0	-2	-7	3
u_i^2	9	16	49	144	1	1	64	0	4	49	337

$$\sigma = \sqrt{\frac{1}{n}\sum u_i^2 - \left(\frac{\sum u}{n}\right)^2}$$

$$= 5.7974.$$

To compute C.V we need $\bar{x}$

Here $\bar{u} = \bar{x} - 57 \Rightarrow \bar{x} = \bar{u} + 57$

$$= \frac{\sum u}{n} + 57 = 57.3$$

$$C.V = \frac{\sigma}{|\bar{x}|} \times 100 = \frac{5.7974}{57.3} \times 100 = 10.1176\ \%$$

(2) Calculate S.D & C.V for freq^n distribution of marks of 100 candidates given below

Marks	0-20	20-40	40-60	60-80	80-100
freq^n	5	12	32	40	11

→ Let $u = \frac{x - 50}{20}$

class	mid values x_i	f_i	$f_i u_i^2$	$f_i u_i$
0-20	10	5	100	
20-40	30	12		
40-60	50	32		
60-80	70	40		
80-100				

$u = \frac{x - 50}{20} \Rightarrow x = 50 + 20u$

$\bar{x} = 50 + 20\bar{u}$

$= 50 + 20 \frac{\sum f_i u_i}{\sum f_i} = 50 + 20 \times \frac{40}{100}$

$= 58$

$Var(x) = (20)^2 Var(u)$

$= 400 \left[\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum u}{n} \right)^2 \right]$

$= 400 \left[\frac{116}{100} - \left(\frac{40}{100} \right)^2 \right] = 400$

$S.D(x) = 20$

$C.V(x) = \frac{S.D}{Am} \times 100 = \frac{20}{58} \times 100$

$= 34.4828$

(3) The number of runs scored by cricker A & B in 10 test matches are shown below.

A	5	20	90	76	102	90	6	108	20	16
B	40	35	60	62	58	76	42	30	30	20

Find i) which cricketer is better in average
ii) which cricketer is more consistent.

→ mean of A = $\frac{533}{10} = 53.3$

SD A = $\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = 40.9293$

cv of A = 76.79%.

Mean of B = $\frac{\sum y}{n} = \frac{453}{10} = 45.3$

SD of B = $\sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = 16.8882$

cv of B = 37.28.

A gives better average run (∵ mean A > mean B)

ii) B is more consistent (c.v of B < c.v of A)

Moments:

Moments give tools for comparison of data set. Using moments we can study the four aspects of comparison viz average, dispersion, symmetry & kurtosis.

There are several aspects of studying freq^n distribution. In earlier type we have studied two of them. viz average & dispersion. In order to study few more aspects such as symmetry, shape of freq^n distr^r moments are useful.

Defⁿ :-

The quantity $\sum \frac{(x_i - a)^r}{n}$ is called r^{th} moment (or moment of order r) about a.

According to choice of a we make 3 categories of moments namely raw moments, central moments, moments about a.

* Raw moments :

Raw moments of order r (r^{th} raw moment) is denoted by μ'_r & is

$$\mu'_r = \frac{\sum_{i=1}^{n} x_i^r}{n} \quad \text{for individual observation}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i^r}{\sum_{i=1}^{n} f_i} \quad \text{for freqⁿ distribution.}$$

Note :- (i) $\mu'_1 = \frac{\sum x_i}{n} = \bar{x} = A.M.$

(ii) The raw moments are also called as moments about origin.

Central moments:

Central moment of order r, μ_r is given by

$$\mu_r = \frac{\sum (x_i - \bar{x})^r}{n} \quad \text{for individual observation}$$

$$= \frac{\sum f_i (x_i - \bar{x})^r}{\sum f_i} \quad \text{for freq}^n \text{ distribution}$$

Here

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

$$= \frac{1}{n}\sum x_i - \frac{\bar{x}}{n}\sum 1$$

$$= \bar{x} - \bar{x} = 0$$

$$\mu_2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} = \sigma^2 = \text{variance}.$$

Mainly we need central moments. However these are difficult to compute as compared to raw moments. Therefore we require to find relation betn them.

Relation betn Raw & central moment.

1) $\mu_1 = 0$

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

$= \frac{1}{n}\sum x_i^2 - \left(\frac{\sum x}{n}\right)^2$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

sim 14

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{n}\sum x_i^3 - 3x_i^2\bar{x} + 3x_i\bar{x}^2 - (\bar{x})^3$$

$$= \frac{1}{n}\sum x_i^3 - \frac{3\bar{x}}{n}\sum x_i^2 + 3\bar{x}^2\frac{1}{n}\sum x_i - \left(\frac{\sum x}{n}\right)^3$$

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

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

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

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^3$$

In general

$$\mu_r = \mu_r' - {}^rC_1\,\mu_{r-1}'\mu_1' + {}^rC_2\,\mu_{r-2}'(\mu_1')^2 + \cdots$$

Moments about a :-

Suppose a is any arbitrary constant then r^{th} moment about a is denoted by $\mu_r(a)$ if is

$$\mu_r(a) = \frac{\sum(x_i - a)^r}{n}$$ for individual observation

$$= \frac{\sum f_i(x_i - a)^r}{\sum f_i}$$

① If $a = 0$ then $\mu_r(a) = \mu_r'$

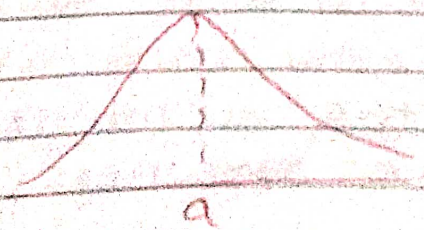
② If $a = \bar{x}$ then $\mu_r(a) = \mu_r$

Skewness & kurtosis:

In previous types we have studied two aspects ~~of freq^n dis~~ in the study of freq^n distribution viz average, dispersion. However in order to compare two freq^n distribution, average & dispersion are not adequate sometimes two freq^n distribution have same average & dispersion however they differ in symmetry. so symmetry is the third aspect in the study of freq^n distribution. The term skewness carries opposite meaning to symmetry ie lack of symmetry.

Symmetry :—

A freq^n distribution is symmetric about a value a if the corresponding freq^n curve is symmetric about a. In other words the ordinate at $x=a$ divides freq^n curve into two equal part For symmetric freq^n curve these two parts are mirror images of each other. The point a turns out to be Am, mode & median.



eg — class 0-10 10-20 20-30 30-40 40-50

~~freq^n dist~~ 5 12 20 12 5

Here freq^n of first is the same as that of last class, simly second & second last class have equal freq^n. & so on.

Properties

(1) In case of bell shaped unimodal symmetric freq^n distribution, Am, mode, median coincide

(2) The quarties of S.D are equispaced

$$Q_3 - Q_2 = Q_2 - Q_1$$

(3) ~~The odd Order~~

In day to day life we come across several distributions which are not symmetric

eg: i) distribution of income of individual,

ii) of agricultural land holdings,

iii) of no. of misprints per page

In these situation we need to measure the extent of departure from symmetry.

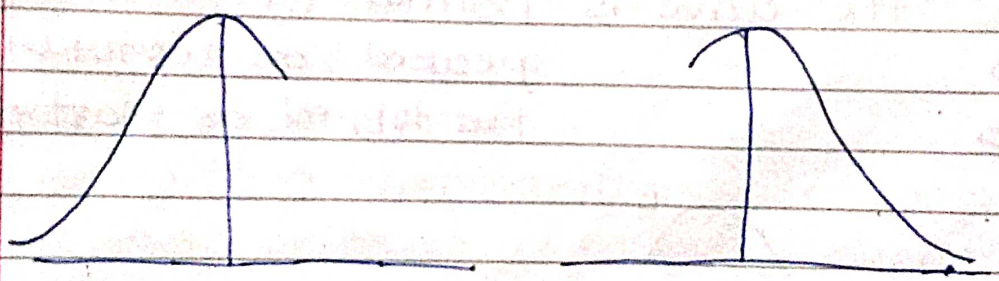
Skewness:-

is lack of symmetry or departure from symmetry. If the distribution is skew, the corresponding ~~distribution~~ curve is elongated on either side.

If the curve is elongated towards right side the distribution is said to positive skewness on other hand if it is elongated to left side the distribution is said to possess -ve skewness.

In other words in case of +ve skewness the freq^n increases rapidly to reach max^m and further decreases slowly. Exactly reverse process is observed in case of the distribution with -ve skewness.

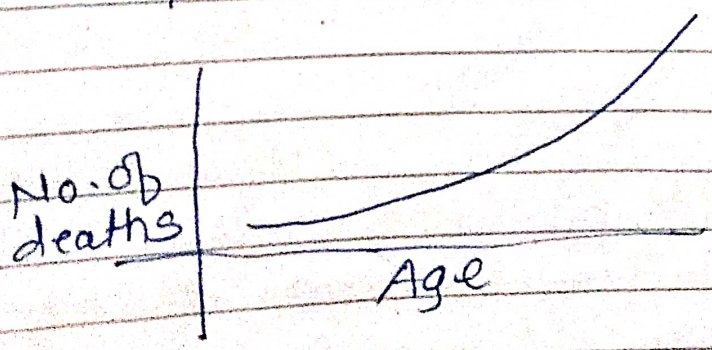
In case of +vely skew distribution we observe that Mode < median < AM.
whereas in case of -vely skew distribution we observe that
AM < median < mode.



-ve skewness +ve skew.

eg: The freq^n curve of annual income is +ve skew

The freq^n curv. of deaths among adults is -vely skew.



coeff^n of skewness = $\frac{\text{mean} - \text{mode}}{\text{S.D.}}$

Kurtosis :-

It measures the degree of peakedness of a distribution. It is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \quad \text{where} \quad \mu_4 = \frac{\sum (x - \bar{x})^4}{n}$$

$$\mu_2 = \frac{\sum (x - \bar{x})^2}{n}$$

$\beta_2 = 3$ the curve is normal or mesokurtic

$\beta_2 > 3$ peaked or leptokurtic

$\beta_2 < 3$ flat topped or platykurtic.