MAT2001-Statistics for Engineers:Module-1

Dr. Nalliah M

Assistant Professor

Department of Mathematics

School of Advanced Sciences

Vellore Institute of Technology

Vellore,Tamil Nadu,India.

nalliah.moviri@vit.ac.in

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

July 24, 2020

Mean deviation and Coefficient of Mean Deviation

The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average.

The mean deviation is measure of dispersion based on all items in a distribution.

Definition:Mean deviation

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored. According to Clark and Schekade.

We usually compute mean deviation about any one of the three averages mean, median or mode.

Some times mode may be ill defined and as such mean deviation is computed from mean and median.

Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

Mean deviation calculated by any measure of central tendency is an absolute measure.

For the purpose of comparing variation among different series, a relative mean deviation is required.

The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

$$\text{Coefficient of mean deviation} = \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}}$$

Mean deviation-Ungrouped data

- Calculate the average(mean, median or mode) of the series.
- Take the deviations of items from average ignoring signs and denote these deviations by $|D|$.
- Compute the total of these deviations, i.e., $\sum |D|$
- Divide this total obtained by the number of items.

$$M.D = \frac{\sum |D|}{n}.$$

Problem

Calculate mean deviation from mean and median for the following data: 100,150,200,250,360,490,500,600,671. Also,calculate coefficients of M.D.

Mean deviation-Grouped data:Discrete Series

- Calculate the average(mean, median or mode) of the series.
- Find out the deviation of the variable values from the average, ignoring signs and denote these deviations by $|D|$.
- Compute $\sum f|D|$
- Divide this total obtained by the total frequency N.

$$M.D = \frac{\sum f|D|}{N}.$$

Problem

Calculate mean deviation from mean and median for the following data:

Also,calculate coefficients of M.D.

Height in cms	Number of persons
158	15
159	20
160	32
161	35
162	33
163	22
164	20
165	10
166	8

Mean deviation-Grouped data:continuous Series

- Calculate the average (mean, median or mode) of the series.
- Compute $|D| = |Mid - average|$, where Mid is the mid point of class interval.
- Compute $\sum f|D|$
- Divide this total obtained by the total frequency N.

$$M.D = \frac{\sum f|D|}{N}.$$

Problem

Calculate mean deviation from mean and median for the following data:

Also,calculate coefficients of M.D.

Marks	Number of students
0-10	20
10-20	25
20-30	32
30-40	40
40-50	42
50-60	35
60-70	10
70-80	8

Solution

- To find Mean and Median.

  Mean $\bar{x} = \frac{\sum fMid}{N}$, where *Mid* is the mid point of the class interval and N is the total frequency.

  Therefore, $\bar{x} = 36.5$

- To find Median

  Now, $\frac{N}{2} = \frac{212}{2} = 106$

  Therefore, Median class is 30-40

  Median $Md = l + h\left(\frac{\frac{N}{2} - cf}{f}\right)$

  Here $l = 30, cf = 77, f = 40, h = 10.$

  Hence Median $Md = 37.25.$

Solution Cont...

Here $|D_1| = |Mid - \bar{x}|$ and $|D_2| = |Mid - Md|$,$\bar{x} = 36.5$ and $Md = 37.25$.

Class	Mid	f	fMid	c.f	$\|D_1\|$	$f\|D_1\|$	$\|D_2\|$	$f\|D_2\|$
0-10	5	20	100	20	31.5	630	32.5	645
10-20	15	25	375	45	21.5	537.5	22.25	556.25
20-30	25	32	800	77	11.5	368	12.25	392
30-40	35	40	1400	117	1.5	60	2.25	90
40-50	45	42	1890	159	8.5	357	7.75	325.5
50-60	55	35	1925	194	18.5	647.5	17.75	621.25
60-70	65	10	650	204	28.5	285	27.75	277.5
70-80	75	8	600	212	38.5	308	37.75	302
Total		212	7740			3192.5		3209.5

Mean deviation from mean $M.D_1 = \frac{\sum f|D_1|}{N} = \frac{3192.5}{212} = 15.06$

Coefficient of mean deviation from mean$=\dfrac{\text{Mean deviation}}{\text{Mean}} = \frac{15.06}{36.5} = 0.41$

Mean deviation from median $M.D_2 = \frac{\sum f|D_2|}{N} = \frac{3209.5}{212} = 15.14$

Coefficient of mean deviation from

Median$=\dfrac{\text{Mean deviation}}{\text{Median}} = \frac{15.14}{37.25} = 0.41$

Standard Deviation and Coefficient of variation

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean.

The standard deviation is denoted by σ.

Standard deviation for Ungrouped data

- $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$, where $\bar{x}$ is the mean.
- $\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$, where $d = x - A$, A is assumed value.

Problem

Calculate the standard deviation from the following data.

14, 22, 9, 15, 20, 17, 12, 11.

Coefficient of Variation

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

$$\text{Coefficient of variation(C.V)} = \frac{\text{standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100.$$

Coefficient of Variation

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous.

If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

Standard deviation for Grouped data-Discrete Series

- $\sigma = \sqrt{\frac{\sum fd^2}{N}}$, where $d = x - \bar{x}$ and N is total frequency.
- $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$, where $d = x - A$, A is assumed value.
- $\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times c$, where $d' = \frac{x-A}{c}$, A is assumed value and c is the interval between each value.

Problem

Calculate Standard deviation from the following data.

X	*20*	*22*	*25*	*31*	*35*	*40*	*42*	*45*
f	*5*	*12*	*15*	*20*	*25*	*14*	*10*	*6*

Standard deviation for Grouped data- Continuous Series

- $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times c$, where $d = \frac{x-A}{c}$, A is assumed value and c is the interval between each value. N is total frequency.

Problem

Calculate Standard Deviation for the following data.

Marks	Number of students
0-10	6
10-20	5
20-30	8
30-40	15
40-50	7
50-60	6
60-70	3

Solution

x	Mid=m	f	$d = \frac{m-35}{10}$	fd	fd^2
0-10	5	6	-3	-18	54
10-20	15	5	-2	-10	20
20-30	25	8	-1	-8	8
30-40	35	15	0	0	0
40-50	45	7	1	7	7
50-60	55	6	2	12	24
60-70	65	3	3	9	27
Total		N=50		-8	140

Solution Cont...

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times c$$

$$= \sqrt{\frac{140}{50} - \left(\frac{-8}{50}\right)^2} \times 10$$

$$= \sqrt{\frac{140}{50} + \frac{64}{2500}} \times 10$$

$$= \sqrt{\frac{7000}{2500} + \frac{64}{2500}} \times 10$$

$$= \sqrt{\frac{7064}{2500}} \times 10$$

$$= 1.68 \times 10 = 16.8$$

Problem

Find the coefficient of variation for the above problem.

Problem

Prices of a particular commodity in five years in two cities are given below:

Price in City A	Price in City B
20	10
22	20
19	18
23	12
16	15

Which city has more stable prices?

Moments

Moments can be defined as the arithmetic mean of various powers of deviations taken from the mean of a distribution. These moments are known as central moments.

Moments

The first four moments about arithmetic mean or central moments are defined below.

The firsr four Central moments	Individual series	Discrete series	Continuous series
μ_1	$\frac{\sum(x-\bar{x})}{n}=0$	$\frac{\sum f(x-\bar{x})}{N}=0$	$\frac{\sum f(mid-\bar{x})}{N}=0$
μ_2	$\frac{\sum(x-\bar{x})^2}{n}=\sigma^2$	$\frac{\sum f(x-\bar{x})^2}{N}$	$\frac{\sum f(mid-\bar{x})^2}{N}$
μ_3	$\frac{\sum(x-\bar{x})^3}{n}$	$\frac{\sum f(x-\bar{x})^3}{N}$	$\frac{\sum f(mid-\bar{x})^3}{N}$
μ_4	$\frac{\sum(x-\bar{x})^4}{n}$	$\frac{\sum f(x-\bar{x})^4}{N}$	$\frac{\sum f(mid-\bar{x})^4}{N}$

If the mean is a fractional value, then it becomes a difficult task to work out the moments. In such cases, we can calculate moments about a working origin and then change it into moments about the actual mean. The moments about an origin are known as raw moments.

The first four raw moments are defined below.

The first four Raw Moments	Individual series $d_1 = X - A$ A- any origin	Discrete series $d_2 = \frac{X-A}{C}$ C-common point	Continuous series $d_3 = \frac{mid-A}{C}$ c-class interval width
μ_1'	$\frac{\sum d_1}{n}$	$\frac{\sum fd_2}{N} \times C$	$\frac{\sum fd_3}{N} \times C$
μ_2'	$\frac{\sum d_1^2}{n}$	$\frac{\sum fd_2^2}{N} \times C^2$	$\frac{\sum fd_3^2}{N} \times C^2$
μ_3'	$\frac{\sum d_1^3}{n}$	$\frac{\sum fd_2^3}{N} \times C^3$	$\frac{\sum fd_3^3}{N} \times C^3$
μ_4'	$\frac{\sum d_1^4}{n}$	$\frac{\sum fd_2^4}{N} \times C^4$	$\frac{\sum fd_3^4}{N} \times C^4$

Relationship between Raw Moments and Central moments

Relation between moments about arithmetic mean and moments about an origin are given below.

$$\mu_1 = \mu_1' - \mu_1' = 0$$

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

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

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

Problem

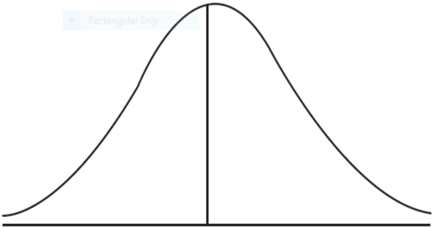
From the data given below, calculate the first four raw and central moments

X	f
30-33	2
33-36	4
36-39	26
39-42	47
42-45	15
45-48	6

Skewness

Skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data.If in a distribution mean$=$ median $=$ mode, then that distribution is known as symmetrical distribution. If in a distribution mean $\neq$ median $\neq$ mode, then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

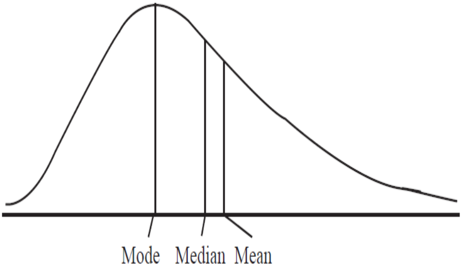
Symmetrical distribution



Mean = Median = Mode

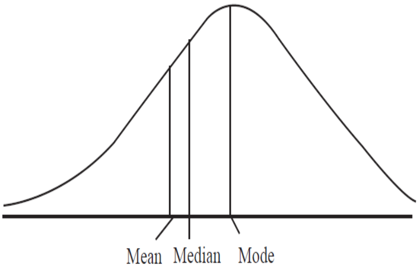
It is clear from the above diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the center point of the curve.

Positively skewed distribution



It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

Negatively skewed distribution



It is clear from the above diagram, in a negatively skewed distribution, the value of the mode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

Measures of skewness

The important measures of skewness are

- Karl-Pearason's coefficient of skewness
- Bowley's coefficient of skewness
- Measure of skewness based on moments

Karl Pearson's Coefficient of skewness

According to Karl Pearson, the absolute measure of skewness = mean mode. This measure is not suitable for making valid comparison of the skewness in two or more distributions because the unit of measurement may be different in different series. To avoid this difficulty use relative measure of skewness called Karl Pearson s coefficient of skewness given by:

Karl Pearson's Coefficient Skewness $= \dfrac{Mean - Mode}{\sigma}$,

where σ-Standard Deviation.

In case of mode is ill-defined, the coefficient can be determined by the formula:

Coefficient of skewness $= \dfrac{3(Mean - MMedian)}{\sigma}$,

where σ-Standard Deviation.

Bowley's Coefficient of skewness

In Karl Pearson's method of measuring skewness the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the median; ie., Median-Q_1=Q_3-Median. But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula:

Bowley's Coefficient of skewness$=\dfrac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$.

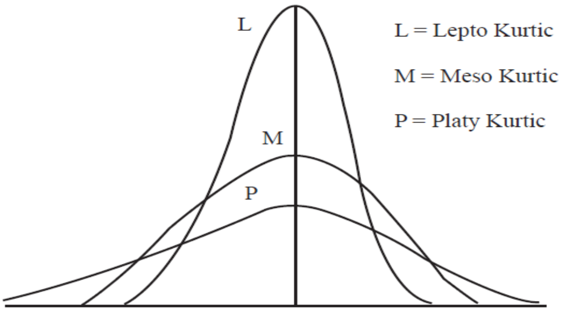
Measure of skewness based on moments

The measure of skewness based on moments is denoted by β_1 and is given by:

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

The expression 'Kurtosis' is used to describe the peakedness of a curve. The three measures central tendency, dispersion and skewness describe the characteristics of frequency distributions. But these studies will not give us a clear picture of the characteristics of a distribution.
As far as the measurement of shape is concerned, we have two characteristics skewness which refers to asymmetry of a series and kurtosis which measures the peakedness of a normal curve. All the frequency curves expose different degrees of flatness or peakedness. This characteristic of frequency curve is termed as kurtosis.

Measure of kurtosis denote the shape of top of a frequency curve. Measure of kurtosis tell us the extent to which a distribution is more peaked or more flat topped than the normal curve, which is symmetrical and bellshaped, is designated as Mesokurtic. If a curve is relatively more narrow and peaked at the top, it is designated as Leptokurtic. If the frequency curve is more flat than normal curve, it is designated as platykurtic.


L
M
P
L = Lepto Kurtic
M = Meso Kurtic
P = Platy Kurtic

Measure of Kurtosis

The measure of kurtosis of a frequency distribution based moments is denoted by β_2 and is given by

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

- If $\beta_2 = 3$, the distribution is said to be normal and the curve is mesokurtic.
- If $\beta_2 > 3$, the distribution is said to be more peaked and the curve is leptokurtic.
- If $\beta_2 < 3$, the distribution is said to be flat topped and the curve is platykurtic.

Thank you