Module 3 :

Statistics :

It is a branch of science dealing with collection of data, organising, summarising, presenting and analysing data & drawing valid conclusions & thereafter making reasonable decisions on the basis of such analysis.

Frequency distribution :

It is the arranged data, summarised by distributing it into classes or categories with their frequencies.

Graphical representation :

It is useful to represent $freq^n$ distribution by means of diagram. The different types of diagrams are

(i) Histogram (ii)
2) $freq^n$ polygon (3) $freq^n$ curve
(iv) cumulative $freq^n$ curve or ogive.
(v) Bar chart (vi) Pie diagrams

Average or measure of central tendancy :-

An average is a value which is representative of a set of data. Average value may be termed as measure of central tendancy.

eg:- Average marks, average profit, average run rate of team in one day

A single value is suitable for comparison so average is essential quantity. Average is a value around which most of the observations are clustered. hence this single value itself gives clear idea regarding phenomenon under study.

There are five types of average in common

i) A.M (ii) Median (iii) Mode (iv) G.M (v) H.M

Among the above stated averages A.M, G.M & H.M are called as mathematical averages & the median & mode are called as Positional averages.

A.M :-

A.M or mean is sum of observations divided by no: of observations.

$$A.M = \frac{\text{sum of observations}}{\text{No of observations.}}$$

case(i) :- If $x_1, x_2 \dots x_n$ are n numbers then

$$A.M = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\Sigma x_i}{n}$$

(ii) Discrete freq^n distribution :-

If the no. x_1 occurs f_1 times, x_2 occurs f_2 times & so on then

$$A.M = \frac{x_1 f_1 + x_2 f_2 + \dots + x_n f_n}{f_1 + f_2 + \dots + f_n}$$

$$= \frac{\Sigma x_i f_i}{\Sigma f_i}$$

Scanned by CamScanner

iii) Continuous freq^n distribution:

In this case freq^n is associated to the entire class & not to any specific single value. This creates difficulty in choosing $x_1, x_2 \ldots x_n$.

For calculation purpose we make assumption that freq^n is associated with mid point of class. Thus taking $x_1, x_2 \ldots x_n$ as mid values of class intervals so mean is

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

Median:—

We have seen that A.M can not be calculated for qualitative observations like debating skill, honesty, blindness. moreover if freq^n distribution includes open end class mean does not exist. In order to overcome these drawbacks other measures of central tendancy such as median or mode are used.

eg:— The A.M of 38, 41, 43, 39, 52, 48, 60, 167 is 61. This can't be said to be representative value. because among these 8 observations 7 are smaller than A.M. Thus in case extreme items are widely separated from most of observations. A.M does not suitable, median is suitable.

Page No.
Date

Def^n –

It is the value of middle most observation in the data when the observations are arranged in increasing (or decreasing) order of their values.

i) When the total no. of the items is odd & equal to say n then value of $\frac{(n+1)}{2}^{th}$ item gives median.

ii) When total no. of freq^n is even say n then there are two middle items & mean of the values of $\frac{n}{2}^{th}$ & $\left(\frac{n}{2}+1\right)^{th}$ item is median.

iii) Continuous freq^n distribution :-

Suppose N is total freq^n.

Regardless of N, whether it is even or odd in continuous freq^n distribution, we take median to be the value of $N/2^{th}$ observation.

Procedure :-

i) obtain class boundaries
ii) obtain less than c.f
iii) locate median class. median class is the class in which median ie $(N/2)^{th}$ observation falls. In other words it is the class where less than cf is equal to or exceeds $N/2$ for the first time
iv) $$\text{median} = l + \frac{N/2 - c.f}{f} \times h$$

where

Scanned by CamScanner

l – lower boundary of class
N – total freq^n
cf – less than c.f of class just preceeding median class
f – freq^n of median class
h – class width.

Ex:

① calculate median for following freq^n dst.

marks	below 20	21-40	41-60	61-80	81-100
No. of students	1	9	32	16	7

→

class boundaries	freq^n	less than cf	
0 - 20.5	1	1	$< N/2$
20.5 - 40.5	9	10	$< N/2$
40.5 - 60.5	32	42	$> N/2$
60.5 - 80.5	16	58	
80.5 - 100.5	7	65	
	65		

10
80.5
99.5

The value of $N/2^{th}$ observeation = 32.5^{th} obseravation
Median class = 40.5 – 60.5
Here $l = 40.5$, $N/2 = 32.5$, $cf = 10$
$f = 32$, $h = 20$

$$\text{median} = l + \frac{N/2 - cf}{f} \times h$$

$$= 54.5625$$

Scanned by CamScanner

② Find the median from following data.

No. of days for which absent (less than)	5	10	15	20	25	30	35	40	
No. of students	29	224	465	582	634	644	650	653	6

→

Class interval	cf	ordinary freq^n
0-5	29	29
5-10	224	224-29 = 195
10-15 =	465 >	465-224 = 241
15-20	582	582-465 = 117
20-25	634	634-582 = 52
25-30	644	644-634 = 10
30-35	650	650-644 = 6
35-40	653	3
40-45	655	2

Median = size of $\left(\frac{655}{2}\right)^{th}$ = 327.5^{th} obser.

median class = 10-15

cf = 224

median = $10 + \frac{327.5 - 224}{241} \times 5$

= 12.15

Page No.
Date

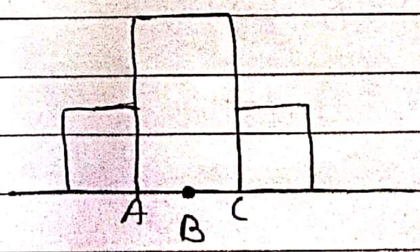
Mode

The most repeated observation is called mode.

Eg: In case of election results, a political party with largest votes (max^m $freq^n$) is considered as representative. Thus it is mode. sim^{ly} to estimate the crop yield too good quality or too poor quality crop is not considered. A quality of crop most commonly found is taken into account. In titration expt. out of 3 readings a repeated reading taken to be final reading. It is mode & not A.M. In number of situation mode is appropriate

For continuous $freq^n$ distribution :-

Mode lies in the class with max^m $freq^n$. The position of mode depends on premodal & postmodal $freq^n$. clearly if premodal & postmodal frequencies are equal mode occupies the centre B of modal class AC.



However if premodal $freq^n$ is larger than postmodal $freq^n$ mode shifts earlier to the centre proportionality. ie B shifts towards A. on other hand if postmodal $Freq^n$ is larger than premodal $freq^n$ then mode shifts towards C. It is like tug off war bet^n pre & postmodal class $freq^n$. In derivation we need to to find exactly by how much quantity mode shifts from centre.

suppose f_m = freq^n of modal class
f_1 - premodal class freq^n
f_2 - post modal ——"——

The shift in the mode from centre is in the proportion $(f_m - f_1)$ to $(f_m - f_2)$
In other words mode (pt B) divides line AB internally in the ratio $\frac{f_m - f_1}{f_m - f_2}$

computation:—
i) obtain class boundaries
ii) Locate modal class. modal class is the class in which mode lies or class with largest freq^n.
iii)
$$\text{mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$
where
l - lower boundary of modal class
f_m - freq^n of modal class
f_1 - freq^n of premodal class
f_2 - ——"—— postmodal class
h - width of modal class.

Mean - mode = 3(mean - median)

Find the mode from following data.

Age	0-6	6-12	12-18	18-24	24-30	30-36	36-42
freq^n	6	11	25	35	18	12	6

→

Age	freq^n
0-6	6
6-12	11
12-18	25
$\boxed{18-24}$	35
24-30	18
30-36	12
36-42	6

Here modal class = 18-24.

$$\text{mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

$$= 18 + \frac{35 - 25}{70 - 25 - 18} \times 6 = 20.22$$

Calculate AM & mode for following data.

Daily salary	No. of workers
Below 400	0
Below 600	4
Below 800	14
1000	33
1200	45
1400	49
1600	90

Page No.
Date

class	freq^n f	mid value x	xf
400-600	4	500	2000
600-800	10	700	7000
800-1000	19	900	17100
1000-1200	12	1100	13200
1200-1400	4	1300	5200
1400-1600	1	1500	1500
			46000

$$A \cdot M = \frac{46000}{50} = 920$$

Now,

modal class = 800 - 1000

$$\text{mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

$$= 800 + \frac{19-10}{38-10-12} \times 200$$

$$= 812.5$$

① Geometric mean (GM)

If $\{x_1, x_2 \dots x_n\}$ be n values of x then Geometric mean

$$G = (x_1 \cdot x_2 \dots x_n)^{1/n}$$ ie n^{th} root of their product.

eg: G.M. of 4, 8, 16 is $(4 \times 8 \times 16)^{1/3} = 8$

② $$G = \left(\prod x_i\right)^{1/n}$$
↑ product.

Sometimes $\prod x_i$ is too large hence it is difficult to compute. so we use logarithm & simplify it.

$$\log G = \frac{1}{n} \log\left(\prod x_i\right) \Rightarrow \log G = \frac{1}{n} \sum \log x_i$$

(← product)

$$G = \text{Antilog}\left(\frac{1}{n} \sum \log x_i\right)$$ $\left(\because \log(ab) = \log a + \log b\right)$

③ If $\{(x_i, f_i), i = 1, 2 \dots k\}$ is freq^n distribution

Then

$$G = \left(x_1^{f_1} \cdot x_2^{f_2} \dots x_k^{f_k}\right)^{1/N}$$ where $N = \sum f_i$

$$G = \left(\prod x_i^{f_i}\right)^{1/N}$$

Taking log

$$G = \text{Antilog}\left[\frac{1}{N} \sum_{i=1}^{k} f_i \log x_i\right]$$

Page No.
Date

Uses of GM :→

A.M is an important & widely used average. Whereas G.M is not much in use. However in certain situations G.M is more appropriate. The following are situations where G.M is preferred.

i) Average change in %.
ii) Average of bank interest rates
iii) Average of depreciation in the cost of certain machine
iv) Average of population growth
v) Average rate of returns on share

In general G.M is appropriate if the values are ratios or percentages.

Harmonic Mean (HM).

A H.M is defined as reciprocal of A.M of reciprocals of observations.

If $x_1, x_2 \ldots x_n$ are the observations then harmonic mean H is

H = Reciprocal of (AM of reciprocals)

$$= \frac{1}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right) \div n}$$

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

If $\{(x_i, f_i), i = 1, 2 \ldots k\}$ is freq^n distribution

then

$$H = \frac{N}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots \frac{f_k}{x_k}} = \frac{N}{\sum (f_i / x_i)} \quad \text{where } N = \sum f_i$$

* Uses of HM:

HM is appropriate to compute average speed, average rates e.g. (where the rates are specified in units per Re).

* $A.M \geq GM \geq HM$

* Measures of Dispersion:→

We have seen that average condenses information into single value. However, average alone is not sufficient to describe distribution completely. There may be two distributions with same mean but distributions may not be identical.

eg. – Marks of students A, B, C in 5 subjects are as follows.

Student	Marks	AM
A	51 52 50 48 49	50
B	30 35 50 65 70	50
C	0 15 45 95 95	50

Note that the average marks of all

students are same but they differ in variation. clearly we can see that A is more consistent than B & B is more consistent than C

For further study & analysis it becomes essential to measure the extent of variation. observations are scattered or dispersed from central value. This variation is called dispersion

Average remains good representative if dispersion is less (ie if the observations are close to it)

Measures of dispersion :-

In this chapter we study the following measures of dispersion.

i) range (ii) quartile deviation
iii) Mean deviation (iv) standard deviation.

These measures have the same units as that of the observation,
eg: Rs, cm, hrs etc.

Range and Coefficient of Range:-

Range is a crude measure of dispersion However it is the simplest measure and suitable if the extent of variation is small.

Def^n:-

If L is the largest observation & S is the smallest observation then

Range = L-S & corresponding relative measure is

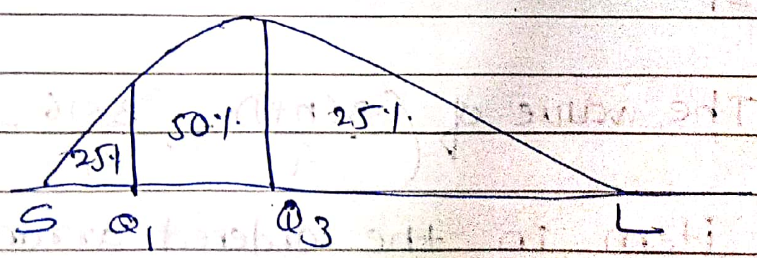
$$\text{Coeff}^n \text{ of range} = \frac{L-S}{L+S}$$

Scanned by CamScanner

In cace of freq^n distribution mid values of first and last class intervals are taken to be largest & smallest observation resp.

Application of Range:
Range is suitable measure of dispersion in case of small group with less variation.
(i) It is widly used in the branch of Statistics known as statistical Quality control.
(ii) The changes in prices of shares, the lowest and highest values are recorded
(iii) Temp^r at certain place is recorded using max^m and minimum value
(iv) Range is used in medical science to check whether BP, HB count etc is normal.

Quartile Deviation or semiinterquartile Range:
∴ The range use only two extreme items. Hence any change in between observations is not going to affect the range. This is main drawback of range



$$Q.D = \frac{Q_3 - Q_1}{2}$$

& corresponding relative measure is

$$\text{Coeff}^n \text{ of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Here Q_1 is the median of lower half of data &

Q3 is median of the upper half of the data

Ex:-

① Compute i) range & coeffⁿ of range

ii) Q.D & coeffⁿ of Q.D. for following data

100, 24, 14, 105, 21, 35, 106, 16, 100, 72, 68, 103, 61, 90, 20.

→ Here S = 14 & L = 106

Range = L − S = 92

Coeffⁿ of range = $\frac{L-S}{L+S}$ = 0.7667

ii) To find Q.D we arrange observations in ascending order as

14, 16, 20, 21, 24, 35, 61, 68, 72, 90, 100, 100, 103, 105, 106.

Q_1 = The value of $\left(\frac{n+1}{4} = \frac{15+1}{4} = 4\right)$th item in the ordered arrangement

= 21

Q_3 = The value of $\left(\frac{3(n+1)}{4} = \frac{3\times16}{4} = 12\right)$th item in the ordered arrangement

= 100

Q.D = $\frac{Q_3-Q_1}{2}$ = 39.5

Coeffⁿ of QD = $\frac{Q_3-Q_1}{Q_3+Q_1}$ = 0.6529 classmate

Date
Page

(2) The no. of vehicles sold by a major Toyota showroom in a day was recorded for 10 working days. The data is

Day	1	2	3	4	5	6	7	8	9	10
freq^n	20	15	18	5	10	17	21	19	25	28

Find Q.D & coeff^n of Q.D.

→

Ascending order data

5, 10, 15, 17, 18, 19, 20, 21, 25, 28

& $n = 10$

$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ item $= 2.75^{th}$ term

$= 2^{nd}$ term $+ 0.75 \times (3^{rd} - 2^{nd}$ term$)$

$= 10 + 0.75 \times (15 - 10) = 13.75$

$Q_3 = \frac{3(n+1)}{4}^{th}$ term $= 8.25^{th}$ term

$= 8^{th}$ term $+ 0.25 \times (9^{th} - 8^{th}$ term$)$

$= 21 + 0.25(25 - 21)$

$= 22$

$Q.D. = \frac{Q_3 - Q_1}{2} = 4.125$

coeff^n of Q.D $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = 23.08$

Mean deviation and coeffⁿ of M.D.

A prime requirement of a good measure is that it should be based on all the observation. It is not satisfied by both range & Q.D.

Naturally the use of deviations taken from a certain point of reference is appropriate. Preferably we take deviations from A.M, we require to combine all these ~~observations~~ deviations into single value. One of the appropriate techniques is to take A.M. However the sum of deviations taken from AM is zero. So A.M of deviations fails to serve the purpose. A.M behaves like a centre of gravity, it balances both +ve & -ve deviations giving total zero. Hence it is required to get rid of the algebraic signs of deviations. This can be done in two ways.

a) Taking absolute deviations
b) Taking squares of deviations.

Defn :-

The A.M of absolute deviations from any average (mean or median or mode) is called M.D about respective averages.

i) M.D about mean :-

$$\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n} = \frac{\sum_{i=1}^{n}|d_i|}{n} \quad \text{for individual observation}$$

$$= \frac{\sum_{i=1}^{n} f_i|d_i|}{N} \quad \text{for freq}^n \text{ distribution where } N=\sum f_i$$

coeffⁿ of M D about mean = $\frac{\text{M.D about mean}}{\text{mean}}$

ii) M.D about median

$= \frac{\sum_{i=1}^{n} |d_i|}{n}$ for individual observation

where $|d_i| = |x_i - \text{median}|$

$= \frac{\sum_{i=1}^{n} f_i |d_i|}{N}$ for freq^n distribution

coeff^n of M D about median = $\frac{\text{M.D about median}}{\text{median}}$

iii) M.D about mode: -

$= \frac{\sum_{i=1}^{n} |d_i|}{n}$ for individual observations

where $|d_i| = |x_i - \text{mode}|$

$= \frac{\sum_{i=1}^{n} f_i |d_i|}{N}$ for freq^n distribution.

coeff^n of M D about mode = $\frac{\text{M.D about mode}}{\text{mode}}$

Computational Procedure:

i) obtain required average (mean or mode or median)

ii) obtain $|d_i| = |x_i - \text{average}|$ for each observation,

iii) find $\sum|d_i|$ for individual observation

& $\sum f_i|d_i|$ for freq^n distribution

iv) Compute MD as

$\frac{\sum|d_i|}{n}$ for individual observation

$\frac{\sum f_i|d_i|}{n}$ for freq^n distribution.

Ex:—

Compute i) MD about mean & coeff^n of MD about mean.

ii) MD about median & coeff^n of M.D about median for the prices per 10 kg of sugar for 7 days in certain week

80, 82, 79, 78, 85, 80, 83.

→ $AM = \frac{\sum x}{n} = \frac{567}{7} = 81$

x_i	80	82	79	78	85	80	83	Total
$\lvert d_i\rvert = \lvert x_i - 81\rvert$	1	1	2	3	4	1	2	14

MD about mean = $\frac{\sum|Hi|}{n} = \frac{14}{7} = 2$

coeffn of MD about mean = $\frac{M.D}{mean} = \frac{2}{81} = 0.0247$

ii) To find median, we use ordered arrangement
78, 79, 80, 80, 82, 83, 85

Median = $\left(\frac{n+1}{2}\right)^{th}$ observation

= 4^{th} observation = 80

x_i	80	82	79	78	85	80	83	Total
$\lvert d_i \rvert = \lvert x_i - 80 \rvert$	0	2	1	2	5	0	3	13

MD about median = $\frac{\sum|d_i|}{n} = \frac{13}{7} = 1.8571$

coeffn of MD about median = $\frac{MD}{median} = 0.0232$

② Obtain MD about i) mean (ii) median & the absolute measure of dispersion in each case for the following freqn distribution:

class	2-4	4-6	6-8	8-10
freqn	3	4	2	1

class	mid values x_i	f_i	$x_i f_i$	C.f.
2-4	3	3	9	3
4-6	5	4	20	7 $> N/2$
6-8	7	2	14	9
8-10	9	1	9	10
		10	52	

$$\text{mean} = \frac{\sum x_i f_i}{\sum f_i} = \frac{52}{10} = 5.2$$

median = The size of $(N/2)^{th}$ observation

ie 5^{th} observation

median class = 4-6

$$\text{median} = l + \frac{N/2 - cf}{f} \times h$$

Here $l = 4$, $N/2 = 5$, $cf = 3$, $h = 2$

median = 5

x_i	f_i	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$	$\lvert x_i - \text{median} \rvert$	$f_i \lvert x_i - \text{med} \rvert$
3	3	2.2	6.6	2	6
5	4	0.2	0.4	0	0
7	2	1.8	3.6	2	4
9	1	3.8	3.8	4	4
			14.8		14

$$\text{M.D about mean} = \frac{\sum f_i \lvert x_i - \bar{x} \rvert}{N} = \frac{14.8}{10} = 1.48$$

Coeffn of M D about mean = $\frac{M \cdot D}{mean}$ = 0.2846

M D about median = $\frac{14}{10}$ = 1.4

Coeffn of M D about median = $\frac{1.4}{5}$ = 0.28

Among all mean deviations mean deviation about median is minimum.

Therefore in order to avoid the effect of choice of average, mean deviation about median is preferred.