PARUL UNIVERSITY - FACULTY OF IT & COMPUTER SCIENCE

Department of Applied Science & Humanities

SYLLABUS FOR 3rd SEM BCA/IMCA PROGRAMME
Statistics **(**05191206**)**

Unit 3. Measure of Central tendency and dispersion

- Measures of Central Tendency: There are three types
  (i)Mean
  (ii) Median
  (iii) Mode

Mean: The most important measure of location is the **mean** or average value, for a variable. The mean provides a measure of central location for the data. If the data are for sample, the mean is denoted by.If the data are for a population, the mean is denoted by the Greek letter *μ*.
In statistical formulas, it is customary to denote the value of variable *x* for the first observation by x_1, the value of variable *x* for the second observation by x_2, and so on. In general, the value of variable *x* for the i^{th} observation is denoted by . For a sample with *n* observations, the formula for the sample mean is as follows.

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

In the preceding formula, the numerator is the sum of the values of the *n* observations. That is,

The Greek letter ∑ is the summation sign.

Example: Find the mean of 10 students and their weights in kilograms are following:
$32, 26, 41, 35, 28, 42, 36, 40, 33, 42$

Solution: $\bar{x} = \frac{32+26+41+35+28+42+36+40+33+42}{10} = \frac{355}{10} = 35.5$ kilogram

Grouped Data
If the data for a **grouped** frequency then the formula for the mean is as follows,

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

Where = mid value of the class

Total frequency = $\sum f_i$

Example: Find the mean of the following:

x_i	0	1	2	3	4	5
f_i	3	20	15	8	3	1

Solution:

x_i	f_i	$x_i f_i$
0	3	0
1	20	20
2	15	30
3	8	24
4	3	12
5	1	5
	n **= 50**	$\sum x_i f_i$ **=91**

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{91}{50} = 1.82$$

Example: Find the mean of the following data:

Class	0-10	10-20	20-30	30-40	40-50
f_i	5	8	15	16	6

Solution:

Class	f_i	x_i	$x_i f_i$
0-10	5	5	25
10-20	8	15	120
20-30	15	25	375
30-40	16	35	560
40-50	6	45	270
	n **= 50**		$\sum x_i f_i$ **=1350**

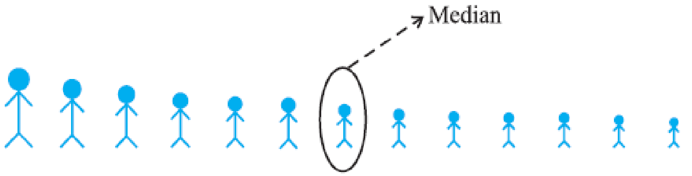
$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{1350}{50} = 27$$

Median: The **median** is another measure of central location. The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value) or descending order (largest value to smallest value) the median can be denoted as M.

If there is an odd number of an observation then median can be found as:

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

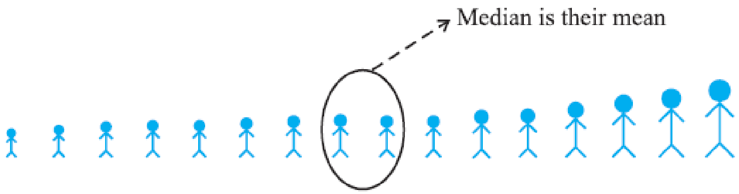
Where, n= number of observation
Example: 2,3,5,6,7,8,9 for given data median is 6.



If there is an even number of an observation then median can be found as:

Where, n= number of observation

Example: 3, 4, 5,6,7,8 for given data median is 5.5.



If the data for **grouped** frequency then the formula for the median is as follows:

$$M = L + \frac{\frac{n}{2} - cf}{f} * c$$

Where,
L =lower limit of median class
n = total number of frequency
cf=cumulative frequency of above median class
f = frequency of median class
c = class width

Example: Find the median of the following data:

x_i	0	1	2	3	4
f_i	4	1	6	11	3

Solution:

0	4	4
1	1	5
2	6	11
3	11	22
4	3	25
	n **= 25**	

Here, **= 25**

If we observed in Cumulative frequency (cf) 13^{th} *observation* lies between 11-22,

So the **median =3**

Example: Find the median of the following:

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f_i	4	8	12	20	24	15	7

Solution:

Class	f_i	cf
0-10	4	4
10-20	8	12
20-30	12	24
30-40	20	**44**
40-50	24	68
50-60	15	83
60-70	7	90
	n **= 90**	

Here, n **= 90**

$$M = \left(\frac{n}{2}\right)^{th} observation$$

$$= \left(\frac{90}{2}\right)^{th} observation$$

$$= 45^{th}\ observation$$

If we observed in cumulative frequency (cf) 45^{th} observation lies between 44-68,

So here
$Median\ \ class\ = 40 - 50,\ L = 40, n = 90, cf = 44, f = 24, class\ \ width\ \ c = 10$

$$Median,\ M = L + \frac{\frac{n}{2} - cf}{f} * c$$

$$= 40 + \left(\frac{\frac{90}{2} - 44}{24} \right) * 10$$

$$= 40.42$$

Mode: A third measure of location is the **mode**. The mode is defined as follows. The mode is the value that occurs with greatest frequency. Mode is denoted as Z.

Example: 1,2,3,4,5,6,4,5,2,4,1,2,2
In above given data 4 occur maximum time so our mode is 4.

Example: Find the mode of the following:

x_i	0	1	2	3	4
f_i	12	20	10	6	2

Solution: Here we can see that maximum frequency is **20**and corresponding value of maximum frequency is 1. So, the **mode** is **1.**

If the data for **grouped** frequency then the formula for the mode is as follows:

$$Z = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) * c$$

Where,
L=lower limit of modal class
f_0 =frequency of the class preceding the modal class
f_1=frequency of the modal class
f_2=frequency of the class succeeding the modal class
c= width of the class

Example: Find the mode of the following:

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f_i	5	9	11	13	10	7	2

Solution:
In the given data maximum frequency is 13, so that the **modal class** is **30-40**
$L = 30 \ \ , f_0 = 11, f_1 = 13, f_2 = 10, c = 10$

$$Mode, \ Z = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) * c$$
$$= 30 + \left(\frac{13 - 11}{(2*13) - 11 - 10} \right) * 10$$
$$= 30 + \left(\frac{2}{5} \right) * 10$$
$$= 34$$

Example: Find the mean, median and mode of the following data:

Class	10-19	20-29	30-39	40-49	50-59
f_i	2	9	15	14	10

Solution:

Class	Class	f_i	x_i	$f_i \ x_i$	cf
10-19	9.5-19.5	2	14.5	29	2
20-29	19.5-29.5	**9**	24.5	220.5	**11**
30-39	**29.5-39.5**	**15**	34.5	517.5	26
40-49	39.5-49.5	**14**	44.5	623	40
50-59	49.5-59.5	10	54.5	545	50
		n **= 50**		$\sum x_i f_i$ **=1935**	

Mean: $\bar{x} = \frac{\sum x_i f_i}{n} = \frac{1935}{50} = 38.7$

Median:

$$M = \left(\frac{n}{2}\right)^{th} observation$$

$$= \left(\frac{50}{2}\right)^{th} observation \quad = 25^{th} observation$$

In cumulative frequency (cf) 25^{th} *observation* lies between 11-26.So that **median class** is **29.5-39.5**

$L = 29.5, cf = 11, f = 15, n = 50, c = 10$

$$Median, M = L + \frac{\frac{n}{2} - cf}{f} * c$$

$$= 29.5 + \left(\frac{\frac{50}{2} - 11}{15}\right) * 10$$

$$= 29.5 + 9.3$$

$$= 38.8$$

Mode:

In the given data the maximum frequency is 15, so that the **modal class** is **29.5-39.5**

$L = 29.5\ , f_0 = 9, f_1 = 15, f_2 = 14, c = 10$

$$Mode, Z = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) * c$$

$$= 29.5 + \left(\frac{15 - 9}{(2 * 15) - 9 - 14}\right) * 10$$

$$= 29.5 + \left(\frac{6}{7}\right) * 10$$

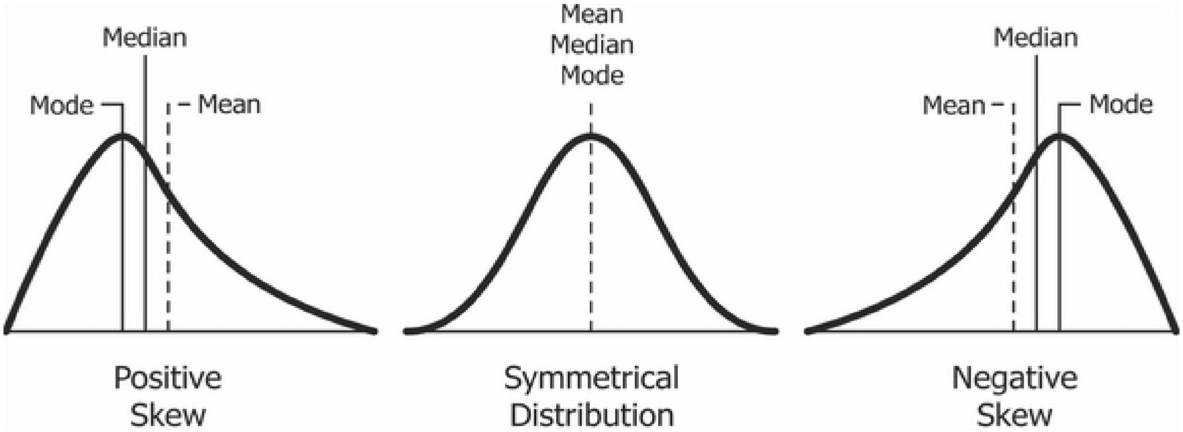
$$= 38.07$$

Relationship among the mean (), median () and mode () can be denoted as following:

Measures of variability: Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Negative Skewness: The **left tail is longer**; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be **left-skewed, left tailed or skewed to the left.**

Positive Skewness: The **right tail is longer**; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be **right-skewed, right tailed or skewed to the right.**



Measure of Deviation:

Variance: The **variance** is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation and the mean. The difference between each and the mean (for a sample, *μ* for a population) are called a deviation about the mean. For a sample, a deviation about the mean is written; for a population, it is written**.**

Population
Variance

Sample variance

NOTE: if in sample size is greater than 30 then n-1 is replaced by n because it does not affect numerical value of variance.

Standard Deviation: Standard deviation (s.d.) is the square root of variance. Standard deviation can be defined as follows:

OR

Where, n is the number of observation.

If the data for **grouped frequency** then the formula for the standard deviation is as follows:

OR

Example: Find the variance and the standard deviation.

X	6	7	8	9	10	11	12
F	3	6	9	13	8	5	4

Solution:

X	F	XF	$\left(X-\overline{X}\right)^2$	$F\left(X-\overline{X}\right)^2$					
6	3	18	9	27					
7	6	42	49	294					
8	9	72	64	576					
9	13	117	81	1053					
10	8	80	100	800					
11	5	55	121	605					
12	4	48	144	576					
63	48	432	568	3931					

$$\overline{X} = \frac{\sum FX}{\sum F} = \frac{432}{48} = 9$$

Variance $$\sigma^2 = \frac{\sum F\left(X-\overline{X}\right)^2}{\sum F} = \frac{3931}{48} = 81.8958$$

Std. Deviation $$\sigma = \sqrt{\frac{\sum F\left(X-\overline{X}\right)^2}{\sum F}} = \sqrt{\frac{3931}{48}} = 9.0496$$

Example: Find the variance and the standard deviation from the following table:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

Coefficient of Variation
In some situations we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the **coefficient ofvariation** and is usually expressed as a percentage.

*100
Coefficient of variation is better is said that more variable or less consistent.
Coefficient of variation is less is said that less variable or more consistent.

Correlation Analysis: we have studied problems relating to one variable only. In practice we come across a large number of problems involving the use of two or more variables. If two quantities vary in such a way that change in one variable are effects a change in the value of other. These quantities are correlated.

Types of correlation: There are three types of correlation.

(i) **Positive or Negative correlation:** If two variables are changing in the same direction, correlation is said to be **positive or direct correlation**. If two variables are changing in the opposite direction, correlation is said to be **negative or inverse correlation**.
For example: The correlation between heights and weights of group of people is positive and the correlation between pressure and volume of a gas is negative.

(i) **Simple, partial or multiple:** The difference between the simple, partial or multiple correlation is based on the number of variable studied. When only two variable are studied correlation is said to be **simple correlation**. When three or more variable are involved then the problem may be either **partial or multiple correlation.**

(ii) **Linear or Non-linear correlation:** If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then the correlation is said to be **linear correlation.**

For example: consider to variables X and Y

X	5	10	15	20	25	30
Y	50	100	150	200	250	300

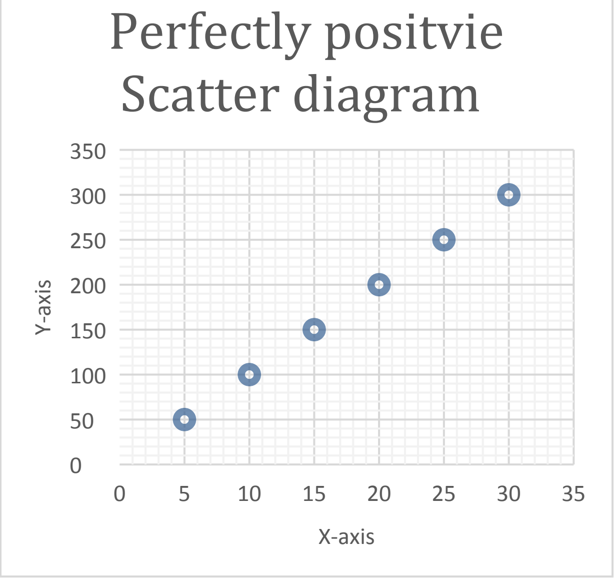
It is clear shows that the ratio of change in both the variables is same.

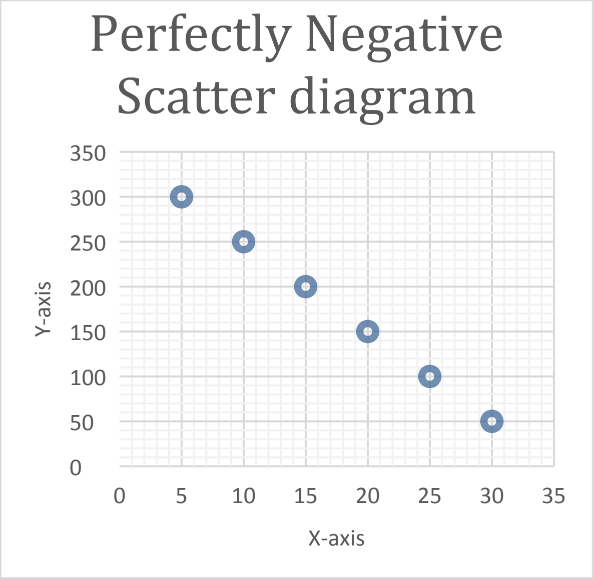
If the amount of change in one variable does not tend to bear a constant ratio to the amount of change in the other variable then the correlation is said to be **Non-linear correlation** or **curly linear correlation.**

Methods of studying correlation: There are mainly three types of methods.

(i) Scatter Diagram
(ii) Karl Pearson's method
(iii) Spearman's method of rank correlation

(i) **Scatter diagram:** This is a very simple method studying the relationship between two variables. In this method one variable is taken on X-axis and the other variable is taken on Y-axis and for each pair of values, points are plotted as follows:





(ii) **Karl Pearson's coefficient of correlation:** The several mathematical methods of measuring correlation the Karl Pearson's popularly known as Pearson's coefficient of correlation is most widely used. It is denoted by r. The formula for computing the coefficient of correlation is as follows:

Where,

This formula also can be written as follow:

Correlation coefficient for the grouped data the formula can be written as follows:

OR

Properties of the coefficient of correlation:

(1) The coefficient of correlation always lies between -1 and 1 including -1 and 1.
i.e.

(2) The correlation coefficient is independent of change of origin and scale.

(3) The correlation coefficient is an absolute number and it is independent of units of measurements.

Example: Find the Pearson's Correlation Coefficient of the following data:

x	100	101	102	102	100	99	97	98	96	95
y	98	99	99	97	95	92	95	94	90	91

Solution:

x	y	$x-\bar{x}$	$y-\bar{y}$	$\left(x-\bar{x}\right)^2$	$\left(y-\bar{y}\right)^2$	$\left(x-\bar{x}\right)\left(y-\bar{y}\right)$

100	98	1	3	1	9	3
101	99	2	4	4	16	8
102	99	3	4	9	16	12
102	97	3	2	9	4	6
100	95	1	0	1	0	0
99	92	0	-3	0	9	0
97	95	-2	0	4	0	0
98	94	-1	-1	1	1	1
96	90	-3	-5	9	25	15
95	91	-4	-4	16	16	16
$\sum x$ =990	$\sum y$ =950	$\sum\left(x-\bar{x}\right)$ =0	$\sum\left(y-\bar{y}\right)$ =0	$\sum\left(x-\bar{x}\right)^2$ =54	$\sum\left(y-\bar{y}\right)^2$ =96	$\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$ =61

$$\bar{x}=\frac{\sum x}{n}=\frac{990}{10}=99$$

$$\bar{y}=\frac{\sum y}{n}=\frac{950}{10}=95$$

$$Correlation\ Coefficient\ ,\ r=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\left(x-\bar{x}\right)^2}\sqrt{\left(y-\bar{y}\right)^2}}=\frac{61}{\sqrt{54}\sqrt{96}}=0.85$$

Calculated by following formula:

Where, n=number of pairs

In case finding out **rank correlation coefficient** when the observations are paired the above formula can be written as:

In $\sum d^2$, $\frac{m}{12}\left(m^2-1\right)$ is added where m is the number of times an item is repeated. The value of correlation coefficient by Spearman’s method also lies between -1 and +1. If the ranks are same for each pair of two series then each value of d=0. Hence =0 and the value of r=+1, which shows that perfect positive correlation between the two variables. If the ranks are exactly in reverse order for each pair of two series, then the value of r = which shown perfect negative correlation between the variables.

Example: Two judges have given ranks to 10 students for their honesty. Find the rank correlation coefficient of the following data:

1st Judge	3	5	8	4	7	10	2	1	6	9
2nd judge	6	4	9	8	1	2	3	10	5	7

Solution:

Rank given by 1st judge	Rank given by 2nd judge	Difference in ranks d	d^2
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
			$\sum d^2$ =214

$$Rank\,Correlation\ ,r = 1-\frac{6\sum d^2}{n\left(n^2-1\right)} = 1-\frac{6*214}{10\left(100-1\right)} = 1-\frac{1284}{990} = 1-1.30 = -0.30$$

Example: Find the Coefficient of rank correlation of the following data:

x	35	40	42	43	40	53	54	49	41	55
y	102	101	97	98	38	101	97	92	95	95

Solution:

x	y	Ranks in x	Ranks in y	Difference d	d^2
35	102	10	1	9	81
40	101	8.5	2.5	6	36
42	97	6	5.5	0.5	0.25
43	98	5	4	1	1

40	38	8.5	10	-1.5	2.25
53	101	3	2.5	0.5	0.25
54	97	2	5.5	-3.5	10.25
49	92	4	9	-5	25
41	95	7	7.5	-0.5	0.25
55	95	1	7.5	-6.5	42.25
					$\sum d^2$ =200.25

$$Rank\ Correlation\ ,r=1-\frac{6\left\{\sum d^2+\frac{m}{12}\left(m^2-1\right)+\frac{m}{12}\left(m^2-1\right)+\frac{m}{12}\left(m^2-1\right)+\frac{m}{12}\left(m^2-1\right)\right\}}{n\left(n^2-1\right)}$$

$$=1-\frac{6\{200.50+0.5+0.5+0.5+0.5\}}{990}$$

$$=-0.227$$

Unit 3: Measures of central tendency and Dispersion

MCQ s:

Which of the following statistics measures the most frequently occurring value in a set of data?

a) median b) mode c) mean d) None of above

2) The mean temperature for the past ten days was 22° Celsius. If the sum of the temperatures for the first nine days was 200, what was the temperature on day 10?

a) 22 b) 32 c) 10 d) 20

3) The method used to compute average or central value of the collected data is considered as

a) measures of positive variation b) measures of central tendency c) measures of negative skewness d) measures of negative variation

4) Given the following set of data, what is the variance? [2 6 8 3 7 9 1 4]

a) 40 b) 2.74 c) 5 d) 7.5

5) Let (100,-20),(103,-10),(107,-5),(109,1),(111,-30) be the ordered pairs of the variables (u,v). What is the value of the rank correlation coefficient between u and v ?

r = 1 b) r = 0.5 c) r = 0 d) r = -1

Ans key : 1) b 2) d 3) b 4) d 5) c

PARUL UNIVERSITY - FACULTY OF IT & COMPUTER SCIENCE

Department of Applied Science & Humanities

SYLLABUS FOR 3rd SEM BCA/IMCA PROGRAMME

Statistics (05191206)

Fill in the blanks :

Range of correlation coefficient is _______

The arithmetic mean is 12 and the number of observations are 20 then the sum of all the values are

For two variables u and v, if an increment in values of u ensures a decrement in values of v then correlation is said to be _____________

______________correlation between variable u and v if(-2,4),(-1,8),(0,12),(1,16),(2,20).

The central tendency median to be measured must lie in______________

Ans key : 1) -1 to 1
2) 240
3)Negative correlation
4) positive correlation
5) Second quartile

Questions : Solve the following

Ten competitors in a beauty contest are ranked by three judges in the following order. Use rank correlation coefficient to determine which of the two judges have similar approach.

1st judge	1	6	5	10	3	2	4	9	7	8
2nd judge	3	5	8	4	7	10	2	1	6	9
3rd judge	6	4	9	8	1	2	3	10	5	7

Ans : r1 = -0.212, r2 = 0.284 , r3 = 0.637
First and Third judge have similar approach .

2)From the following table calculate the coefficient of correlation by Karl Pearson's method. Arithmetic means of X and Y series are 6 and 8 respectively.

X	6	2	10	4	8
Y	9	11	?	8	7

Ans : Missing value = 5, r = -0.920

Find Mean, Median and Mode for below observation.

15,17,12,13,14,16,1,8,18,14

Ans : Mean : 12.8 , Median : 14 , Mode 14

Find mode of the following data

X	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Y(No of student)	3	5	7	10	12	15	12	6	2	8

Ans : 55

5) The scores for student are 40, 45, 49, 53, 61, 65, 71, 79, 85, 91. What is the percentile for score 71?

Ans: 60

6 Calculate Quartile-2, Percentiles-45 from the following data
85,96,76,108,85,80,100,85,70,95)

Ans : 85 , 85