

## Appendix: I

---



Department of Education [CASE]  
Faculty of Education and Psychology  
The Maharaja Sayajirao University of Baroda  
Vadodara 390 002  
Phone: 0265 2795516, 2792631

---

Date: 17<sup>th</sup> Aug 2015

To

---

---

---

---

Sub: Validation of the Tool

Respected Sir/Madam,

I am Sonia Rohilla, Doctral Scholar at the Centre of Advanced Study in Education, Faculty of Education and Psychology, The Maharaja Sayajirao University of Baroda. The title of my study is “Development of an Educational Program on Data Analysis Techniques for M.Ed. Students through Cooperative Learning”. May I request you to kindly validate my tool “Entry level check on Statistical Data Analysis Techniques”. This tool is based on four components. These four components are:

- i. Frequency distribution
- ii. Diagrammatic and graphical representation of data
- iii. Measures of central tendency
- iv. Measures of dispersion

Validation of this tool by you will facilitate my doctoral study.

Thanking you,

Yours truly,

(Sonia Rohilla)  
Research Scholar

(Dr. D. R. Goel)  
Guide

## ENTRY LEVEL CHECK ON STATISTICAL DATA ANALYSIS TECHNIQUES

Name of the student: \_\_\_\_\_

College Name: \_\_\_\_\_

**Note: Tick only one most appropriate answer out of the given options (a), (b), (c) and (d).**

1. Data can be stored in the form of
  - a) Text
  - b) Tables
  - c) graphs and diagrams
  - d) all of the above
2. Data should be classified or tabulated first?
  - a) Classification follows tabulation
  - b) Classification precedes tabulation
  - c) Both are done simultaneously
  - d) None of the above
3. In an exclusive type frequency distribution, the limits excluded are
  - a) Lower limits
  - b) Upper limits
  - c) Either of the lower or upper limits
  - d) Lower limits and upper limits
4. If the lower and upper limits of the class are 10.50 and 40.50 respectively, the mid-points of the class is
  - a) 25.50
  - b) 12.50
  - c) 15.25
  - d) 30.25
5. Class interval is measured as
  - a) The sum of the upper and lower limit
  - b) Half of the sum of the upper and lower limit
  - c) Half of the difference between the upper and lower limit
  - d) The difference between the upper and lower limit
6. A group frequency distribution with uncertain first or last classes is known as
  - a) Exclusive class distribution
  - b) Inclusive class distribution
  - c) Open end distribution
  - d) Discrete frequency distribution
7. Frequency of a variable is always
  - a) In percentage
  - b) A fraction
  - c) An integer
  - d) A whole number
8. Classification is applicable in case of
  - a) Quantitative characters
  - b) Qualitative characters
  - c) Both (a) and (b)
  - d) None of the above
9. Graphs and charts facilitate
  - a) Comparison of values
  - b) To know the trend
  - c) To know the relationship
  - d) All of the above
10. Choice of a particular chart depends on
  - a) The purpose of the study
  - b) The nature of the data
  - c) The type of the audience
  - d) All of the above
11. Year-wise production of rice , wheat, maize for last six years can be displayed by
  - a) Simple bar chart
  - b) Broken bar diagram
  - c) Subdivided column chart
  - d) Multiple bar diagram

12. Which of the following statement is not correct?
- The bars in a histogram touch each other
  - The bar in column chart touch each other
  - There are bar diagrams which are known as broken bar diagrams
  - Multiple bar diagrams also exist
13. In case of frequency distribution with classes of unequal widths, the heights of bars of a histogram are proportional to
- Class frequency
  - Class intervals
  - Frequencies in percentage
  - Frequency densities
14. Histogram is suitable for
- Time series data
  - Chronological distribution
  - Neither (a) nor (b)
  - Both of (a) and (b)
15. In a histogram with equal class intervals, heights of bar are proportional to
- Mid-values of the classes
  - Cumulative frequency
  - Frequencies of respective classes
  - Neither (a) nor (b)
16. The data relating to the number of registered allopathic and homeopathic doctors in six different states can be most appropriately represented by diagram
- Line graph
  - Histogram
  - Pie-diagram
  - Double bar diagram
17. The most appropriate diagram to represent the data relating to the monthly expenditure on different items by a family is
- Histogram
  - Pie-diagram
  - Frequency- polygon
  - Line graph
18. Mean is a measure of
- Location
  - Dispersion
  - Correlation
  - None of the above
19. Which of the following is a measure of central values?
- Median
  - Standard deviation
  - Mean deviation
  - Quartile deviation
20. If a constant value 5 is subtracted from each observation of a distribution. The mean of the changed distribution is
- Increased by 5
  - Decreased by 5
  - 5 times the original mean
  - Not affected
21. If each observation of a set is multiplied by 10, the mean of the new set of observations
- Remain the same
  - Is 10 times the original mean
  - Is one- tenth of the original mean
  - Is increased by 10
22. If each value of a series is multiplied by 10, the median of the coded value is
- Not affected
  - Is 10 times the original median
  - Is one- tenth of the original median
  - Is increased by 10

23. If the grouped data has open end classes, one cannot calculate
- Median
  - Mode
  - Mean
  - Quartiles
24. Extreme value have no effect on
- Average
  - Median
  - Geometric mean
  - Harmonic mean
25. Expenditure during first five months of a year is Rs. 96 per month and during last seven months is Rs. 120 per month. The average expenditure per month during whole year is
- Rs. 180 per month
  - Rs. 110 per month
  - Rs. 100 per month
  - Rs. 216 per month
26. The average age of 50 students in a bus is 20 years. When the age of conductor is included, the average age is increased by one year. The age of the conductor is
- 51
  - 55
  - 71
  - 50
27. If the sum of N observations is 630 and their mean is 42, then the value of N (no. of observations) is
- 21
  - 30
  - 15
  - 20
28. For a set of observations, the empirical relationship between mean, median and mode is
- Mean = median + mode
  - Mode = 3median - 2 mean
  - Mode = median + mean
  - Median = 3(mean - mode)
29. Mode is that value in a frequency distribution which possesses
- Minimum frequency
  - Maximum frequency
  - Frequency one
  - None of the above
30. Shoe size of most of the people in India is No. 8. Which measure of central value does it represent?
- Mean
  - Second quartile
  - Eighth decile
  - Mode
31. To find the median, it is necessary to arrange the data in
- Ascending order
  - Descending order
  - Either (a) or (b)
  - None of the above
32. For the distribution given below, the modal class is
- | <i>classes</i> | <i>No. of persons</i> |
|----------------|-----------------------|
| 1500-1600      | 78                    |
| 1600-1700      | 80                    |
| 1700-1800      | 90                    |
| 1800-1900      | 55                    |
| 1900-2000      | 33                    |
- 1500- 1600
  - 1600-1700
  - 1700- 1800
  - none of the above



33. Mean of a set of values is based on
- a) All values
  - b) 50 % values
  - c) First and last values
  - d) Maximum and minimum value
34. The sum of the deviations of all observations about their mean is always
- a) Zero
  - b) Minimum
  - c) Maximum
  - d) One
35. Histogram is useful to determine graphically the value of
- a) Mean
  - b) Median
  - c) mode
  - d) All of the above
36. For comparison of two different series, the best measure of dispersion is
- a) Range
  - b) Mean deviation
  - c) Standard deviation
  - d) Coefficient of variation
37. If the standard deviation (s.d.) of variable X is 20 then its variance is
- a) 400
  - b) 200
  - c) 20
  - d) 10
38. Which measure of dispersion ensures highest degree of reliability?
- a) Range
  - b) Mean deviation
  - c) Quartile deviation
  - d) Standard deviation
39. The range of the set of values 15, 12, 27, 6, 9, 18, 21 is
- a) 21
  - b) 4.5
  - c) 0.64
  - d) 3
40. If all the values in a sample are same. Then their variance is
- a) Zero
  - b) One
  - c) Not calculable
  - d) None of the above

-----\*\*\*\*\*-----

## Appendix: II

### POST ACHIEVEMENT TEST

*Name of the Student:*

*Name of the Institution:*

**Note: Tick only one most appropriate answer out of the given options (a), (b), (c) and (d).**

- 1) \_\_\_\_\_ is not a graph.
  - a) Histogram
  - b) Line diagram
  - c) Ogive curve
  - d) Frequency polygon
- 2) \_\_\_\_\_ is not a chart.
  - a) Sub divided bar diagram
  - b) Deviation bar diagram
  - c) Duo- directional bar diagram
  - d) Histogram
- 3) From the histogram we can trace the value of \_\_\_\_\_ as a measure of an average.
  - a) Mean
  - b) Median
  - c) Mode
  - d) None of the above
- 4) For the following information which one is more suitable to represent the information?

Class	VI-A	VI-B	VI-C	VI-D	VI-E	VI-F
No. of students in class	65	64	58	54	57	60
No. of passed student in class	60	54	57	50	55	52

- a) Pie chart
- b) Line diagram
- c) Histogram
- d) Multiple bar diagram
- 5) Median can be traced from \_\_\_\_\_.
  - a) Histogram
  - b) Ogive curve
  - c) Frequency curve
  - d) Frequency polygon
- 6) For a data if third quartile  $Q_3 = 67.68$  it means that \_\_\_\_\_.
  - a) 75% of cases are below the value 67.68.
  - b) 25% of cases are below the value 67.68.
  - c) 50% of cases are below the value 67.68.
  - d) None of the above
- 7) For a data if 35<sup>th</sup> percentile  $P_{35}$  is 60.33 it means that \_\_\_\_\_.
  - a) 35 % of cases are below the value 60.33.
  - b) 65% of cases are below the value 60.33.
  - c) 35% of cases are above the value 60.33.
  - d) None of the above
- 8) Year wise production of ores of different metals like iron, aluminium, copper and silver in terms of percentage for past 5 years can be displayed by:
  - a) Simple bar diagram
  - b) Sub-divided column chart
  - c) Broken bar diagram
  - d) Multiple column chart
- 9) The census data published for state wise population of India will be termed as \_\_\_\_\_.
  - a) Qualitative classification of data
  - b) Geographical classification of data
  - c) Chronological classification of data
  - d) None of the above

- 10) Distribution of students in different groups according to their percentage scored in the just previous passed class is known as \_\_\_\_\_.
- a) Quantitative classification of data                      b) Chronological classification of data  
c) Geographical classification of data                      d) Qualitative classification of data
- 11) If the middle value of a class is 100 and the difference between two consecutive middle values is 10, then the class limits are \_\_\_\_\_.
- a) 90-110    b) 95-105  
c) 80-100    d) 100-120
- 12) A systematic distribution of frequencies with respect to the values of a variate is known as \_\_\_\_\_.
- a) Tabulation    b) Classification  
c) Frequency distribution    d) Graphical presentation
- 13) \_\_\_\_\_ is not a part of table.
- a) Stub    b) Caption  
c) Heading    d) References
- 14) Given the following information on 30 people, insert the missing frequencies.

Class intervals	Frequency	Cumulative frequency
20-40	2	2
40-60	4	6
60-80	6	12
80-100	-----	-----
100-120	6	25
120-140	5	30

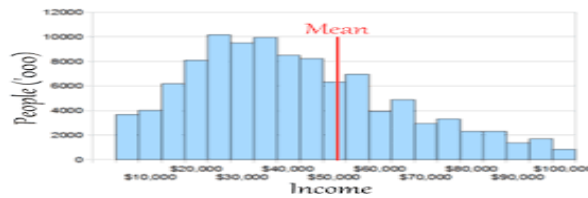
- a) 7, 19    b) 7, 18  
c) 8, 19    d) 10, 22
- 15) In a class of 50 students, science subject exam was conducted and it is found that 25 students have scored more than 73 marks and rest 25 students have scored less than 73 marks. Then 73 marks is \_\_\_\_\_ score of the class.
- a) Mean    b) Median  
c) Mode    d) Standard deviation
- 16) Every year some students dropped out from the B.Ed. Course in ABC College. The data of number of students dropped out per year is given below. Calculate mean, median and mode of students dropped out.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
No. of dropped students	4	3	3	2	3	4	3	3	2	4

- a) 3.1, 3, 3    b) 3.1, 3.5, 3  
c) 3.25, 3, 4    d) 3, 3, 3

- 17) In a class of 35 students with average score of 70, two more students added with score 74 and 63 respectively. Calculate the new average score of the class.
- 65.47
  - 70.43
  - 75.34
  - 69.91
- 18) Forty students of class IX participated in mathematics Olympiad test. The least score was 22 and the highest score was 71. If the median score of these 40 students is 56 and least score is changed to 15 and the highest score is changed to 75 then the new median score will be \_\_\_\_\_.
- 45
  - 57
  - 56
  - 70
- 19) For a set of 20 observations in a data the mode was found to be 85. If every value is increased by 5 then new mode value will be \_\_\_\_\_.
- 90
  - 85
  - 80
  - 105
- 20) The mean and standard deviation of 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7 observations is \_\_\_\_\_ & \_\_\_\_\_ respectively.
- 7 & 1
  - 7 & 0
  - 7 & 7
  - 0 & 7
- 21) The variance of observations 5, 7 and 10 is \_\_\_\_\_.
- 4.2
  - 5.0
  - 3.6
  - 7.33
- 22) NPC stands for \_\_\_\_\_
- Normal Percentile Curve
  - Negative Probability Curve
  - Normal Probability Curve
  - None of the above
- 23) NPC is \_\_\_\_\_ shaped curve.
- Symmetrical bell
  - Symmetrical ball
  - Symmetrical conical
  - None of the above
- 24) According to area property of NPC curve  $\mu - \sigma$  to  $\mu + \sigma$ , \_\_\_\_\_ % of observations fall under the curve.
- 95
  - 68
  - 97
  - 50
- 25) Lack of symmetry in the distribution of data is known as \_\_\_\_\_.
- Median
  - s.d.
  - Kurtosis
  - Skewness

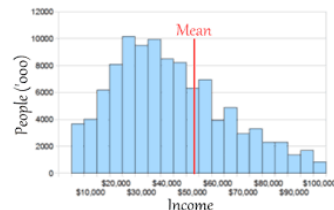
- 26) \_\_\_\_\_ skewness is present in the following histogram of frequency distribution of data.



- a) Zero  
b) Positive  
c) Negative  
d) None of the above
- 27) If coefficient of skewness of X- series is 0.14 and coefficient of skewness of Y- series is -1.5 then which series is more skewed?
- a) Series X is more skewed than series Y  
b) Series Y is more skewed than series X  
c) None of the series is skewed  
d) None of the above
- 28) Two series of data, namely A and B are plotted below. Identify which series is less skewed?



Series A



Series B

- a) Series A is less skewed than series B  
b) Series B is less skewed than series A  
c) Cannot compare them  
d) None of the above
- 29) If the coefficient of skewness is -1.9. It indicates that there is \_\_\_\_\_ skewness in the distribution of data.
- a) Positive  
b) Negative  
c) No  
d) None of the above
- 30) On 600 students of FYBA of the academic year 2015-2016, an achievement test of English grammar was conducted, the coefficient of skewness of scores of students was calculated as -1.65. It indicates that \_\_\_\_\_.
- a) Large number of students have good score in English grammar.  
b) Less number of students have good score in English grammar.  
c) Both (a) and (b).  
d) None of the above
- 31) Identify the odd one in the following.
- a) Platy kurtic  
b) meso kurtic  
c) Probabilistic  
d) leptokurtic

- 
- The graph displays two overlapping bell-shaped curves. The curve labeled 'Series U' is taller and narrower, peaking at a higher value. The curve labeled 'Series T' is shorter and wider, peaking at a lower value. Both curves are centered around the same horizontal position. Arrows point from the labels to the respective curves.

- | Terms                         | Symbols        |
|-------------------------------|----------------|
| A) Karl Pearson's correlation | I. $r_{xy.z}$  |
| B) Biserial correlation       | II. $R_{x.yz}$ |
| C) Partial correlation        | III. $r_{xy}$  |
| D) Multiple correlation       | IV. $r_{bis}$  |
| E) Regression coefficient     | V. $b_{xy}$    |

- 154

38) Match the following terms with appropriate types of scales used to calculate them.

Correlation coefficient	Types of scale
A) Pearson's Product moment	I) both scales ordinal
B) Spearman's rank- order	II) one scale naturally dichotomous (nominal), one scale interval or ratio
C) Point – biserial	III) Both the scales are interval (or Ratio)
D) Biserial	IV) Both scales are naturally dichotomous (nominal)
E) Phi	V) One scale artificially dichotomous (nominal), one scale interval (or ratio)

- a) (A) - III , (B) - I , (C) - II , (D) - V , (E) -IV  
b) (A) - III , (B) - I , (C) - V , (D) - II , (E) -IV  
c) (A) - II , (B) -III , (C) -IV , (D) -V , (E) -I  
d) (A) -IV , (B) - III , (C) -I , (D) -II , (E) -V

39) Correlation is used to study the degree of \_\_\_\_\_ between the variables.

- a) Relationship  
b) Partnership  
c) Association  
d) Prediction

40) If the correlation coefficient  $r = 0.84$  between IQ score and age of school students. It means that \_\_\_\_\_.

- a) There is high positive correlation between IQ scores and age of students.  
b) There is low positive correlation between IQ scores and age of students.  
c) There is lack of correlation between IQ scores and age of students.  
d) None of the above.

41) Match the following correlation coefficients with its appropriate range.

Correlation coefficient	Range
A) Pearson's Product moment	I) 0 to +1
B) Spearman's rank- order	II) -1 to +1
C) Point – biserial	III) -1 to +∞
D) partial	IV) -∞ to +∞
E) multiple	V) +1 to +∞

- a) (A)-II , (B) -III , (C) -IV , (D) -V , (E) -II  
b) (A)-II , (B) -II , (C) -II , (D) -II , (E) -I  
c) (A)-I , (B) -II , (C) -III , (D) -III , (E) -I  
d) (A)- I , (B) -II , (C) -II , (D) -IV , (E) -V

42) If the correlation between scores in Physics subject and Mathematics subject is 0.60 then coefficient of determination will be \_\_\_\_\_.

- a) 12 %  
b) 30%  
c) 36%  
d) 50%

- 43) Two judges gave judgment on the performance of 10 candidates in a dance competition. Rank correlation is  $r = 0.77$ , it means that \_\_\_\_\_.
- Opinion of both the judges is highly positively correlated.
  - Opinion of both the judges is moderately positively correlated.
  - Opinion of both the judges is extremely different.
  - Cannot say about their judgments.
- 44) If  $\text{Cov}(x,y) = 2.45$ ,  $v(x) = 8.5$ ,  $v(y) = 2.22$  and  $n = 20$  then the correlation coefficient  $r_{xy}$  will be \_\_\_\_\_.
- 0.56
  - 0.54
  - 0.63
  - 0.27
- 45) If the two regression coefficients  $x$  on  $y$  and  $y$  on  $x$  are 0.20 and 0.45 then the correlation coefficient will be \_\_\_\_\_.
- 0.30
  - 0.65
  - 0.90
  - 0.50
- 46) If the regression equation of  $y$  (science project scores) on  $x$  (science achievement scores) of class  $X$  is given by  $Y = 1.25X + 13.5$  then estimate the science project score of a student if he has 35 as science achievement score.
- 57.25
  - 60.25
  - 40.25
  - 45.00
- 47) The relationship between percentage scored and gender of class XII grade students scored by \_\_\_\_\_.
- Product moment
  - Rank order
  - point biserial
  - Biserial
- 48) In a simple regression analysis \_\_\_\_\_ independent and \_\_\_\_\_ dependent variables are participating.
- One , one
  - one , two
  - two , one
  - two, two

### **Basic Concepts of inferential statistics**

- 49) The type - I error is defined as
- reject null hypothesis/ null hypothesis is true
  - reject alternate hypothesis/ null hypothesis is true
  - accepting null hypothesis/ null hypothesis is true
  - reject null hypothesis/ alternate hypothesis is true
- 50) If  $\beta$  is the P[type – II error] then Power of the test is given by \_\_\_\_\_.
- $1 - \beta$
  - $1 - \alpha$
  - $1 + \beta$
  - none of the above
- 51) Standard error of the mean gives an idea about \_\_\_\_\_.
- Sampling error.
  - The difference between the sample mean and the population mean.
  - How far is the Sample mean from the population mean.
  - None of the above.
- 52) If the s.d. of a sample of 25 observations is 3.5 then the standard error of mean is \_\_\_\_\_.
- 28.5
  - 0.14
  - 87.5
  - 0.7



- 53) The educational researches are generally carried out at \_\_\_\_\_ % of level of significance.  
 a) 5% and 1%      b) 2% and 10%      c) 5% and 10%      d) 10% and 20%
- 54) If 5% level of significance is defined for some research results it means that \_\_\_\_\_.  
 a) If the experiment is repeated 100 times then at the most 5 times research results may change.  
 b) If the experiment is repeated 100 times then at least 95 times research results remains the same.  
 c) If the experiment is repeated 100 times then exactly 5 times research results will change.  
 d) Both (a) & (b) are true.
- 55) In testing of hypothesis table value is also known as \_\_\_\_\_.  
 a) Degrees of Freedom      b) Critical Value  
 c) P- Value      d) None of the above
- 56) A statistical constant describing population is called \_\_\_\_\_ and of sample is called \_\_\_\_\_.  
 a) Statistic , parameter      b) parameter , statistic  
 c) Hypothesis, test statistic      d) none of the above
- 57) Non parametric tests are also known as \_\_\_\_\_ tests.  
 a) Useless      b) Distribution free  
 c) Difficult      d) None of the above
- 58) \_\_\_\_\_ is not a parametric test.  
 a) T- test      b) z- test  
 c) F- test      d) U- test
- 59) \_\_\_\_\_ is a parametric test.  
 a) Sign test      b) Wilcoxon Matched Pairs test  
 c) Mann – Whitney U- test      d) Paired t - test
- 60) If sample size increases then S.E. will \_\_\_\_\_.  
 a) Increase      b) Decrease  
 c) Unaffected      d) None of the above
- 61) For a 3x4 contingency table in chi-square test for testing the independence of two attributes, the degrees of freedom will be \_\_\_\_\_.  
 a) 5      b) 6      c) 12      d) 1
- 62) ANOVA stands for \_\_\_\_\_.  
 a) Analysis of variable      b) Analysis of values  
 c) Analysis of Variance      d) None of the above

- 63) Match the following test statistic distribution to the appropriate degrees of freedom and level of significance to determine the critical value from the statistical table.

Test statistic distribution	Table value at
A) t	i) $(r-1)(c-1)$ d.f. and $\alpha\%$ level of significance
B) z	ii) $n_1$ & $n_2$ d.f. and $\alpha\%$ level of significance
C) $\chi^2$	iii) $(n-1)$ d.f. and $\alpha\%$ level of significance
D) F	iv) $\alpha\%$ level of significance

- a) A)- iii , B)- iv , C)-i , D)-ii  
b) A)- iii, B)- i , C)-ii , D)-iv  
c) A)- iv , B)- iii , C)-i , D)-ii  
d) A)-iv , B)- iii, C)- ii, D)-i

- 64) Draw the conclusion if

Null hypothesis: The mean length of the tables produced by the company is 120 cm.

(i.e.  $H_0: \mu = 120$ )

Alternate hypothesis: the mean length of the tables produced by the company is greater than 120 cm. (i.e.  $H_1: \mu > 120$ )

A Sample of 26 table with mean length of 120.6 cm and s.d.= 1.002 cm

Use t- table value = 2.06 at 5% level of significance and 25 d.f.

- a) Null hypothesis is rejected and mean length of tables produced from company is of greater than 120cm.  
b) Null hypothesis is not rejected and mean length of tables produced from company is of length 120cm.  
c) Neither (a) nor (b)  
d) None of the above

- 65) \_\_\_\_\_ is the only test used for both parametric as well as non-parametric testing of hypothesis.

- a)  $\chi^2$  – test    b) t –test    c) z – test    d) F- test

- 66) Eight patients were tested for hemoglobin level and a diet plan was implemented to improve the hemoglobin level in them. Readings were obtained before and after the implementation of diet plan. To study the effectiveness of diet plan which test is being used?

- a) Median test    b)  $\chi^2$  – test    c) paired t- test    d) z-test

- 67) A study was carried on VI class students to know the best teaching technique. Three parallel classes VI-A, VI-B and VI-C were selected and three different techniques T1, T2 and T3 respectively were used to teach them mathematics. After the implementation of these three techniques to different groups an achievement test was carried out. Researcher wants to test whether mean score of all the three classes are significantly same or not? To test this claim which data analysis technique should the researcher use?
- a) Mann Whitney U- test                      b) ANOVA  
c) t- test    d) sign test
- 68) In  $\chi^2$  – testing for independence of attributes. The null hypothesis is \_\_\_\_\_.  
a) The two attributes under the study are independent.  
b) The two attributes under the study are dependent.  
c) The two attributes under the study are unassociated.  
d) Both (a) and (c) are true.
- 69) In two sample testing of mean. If the null hypothesis is  $H_0: \mu_1 = \mu_2$ . It means that \_\_\_\_\_.  
a) There is no significant difference between the means of two populations.  
b) There is a significant difference between the means of two populations.  
c) Mean of population -1 is equal to the mean of population -2.  
d) None of the above.
- 70) In a one sample problem for testing the mean scores of X - class students in English subject following information is obtained.  
 $H_0: \mu = 68$     $H_1: \mu > 68$   
 $N=60$ ,  $Z_{cal} = 6.44$ ,  $Z_{tab} = 1.6449$  at 5% level of significance. Then,
- a) Null hypothesis is rejected at 5% level of significance and concluded that mean score of the class X students in English subject is significantly higher than score 68.  
b) Null hypothesis is not rejected at 5% level of significance and concluded that mean score of the class X students in English subject is not significantly different than 68.  
c) Cannot conclude from the given information.  
d) None of the above.

---

\*\*\*\*\*

---

### **Appendix: III**

#### **RATING SCALE ON COOPERATIVE LEARNING**

<i>Name of the Student:</i>
<i>Name of the Institution:</i>

Kindly register your level of agreement by putting tick (✓) in the provided space against each statement on a five point scale.

Sl.No.	Components of Cooperative learning	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
<b>1.</b>	<b>POSITIVE INTERDEPENDENCE</b>					
i.	Every member was having positive outlook to accept the task.					
ii.	Every member helped each other to complete the task.					
iii.	Every member was fully involved in the task.					
iv.	Every member respected the other ones.					
v.	Encouragement and support were provided mutually.					
vi.	All the members converged on the solution.					
<b>2.</b>	<b>EQUAL PARTICIPATION</b>					
i.	All members were involved to achieve the task.					
ii.	Every member was treated equally.					
iii.	Participation in team brought self confidence and fearlessness.					
iv.	Every member participated and presented.					
<b>3.</b>	<b>FACE TO FACE PROMOTIVE INTERACTION</b>					
i.	Members posed questions to each other.					

ii.	Members listened to each other.					
iii.	All the members got chance to express their ideas to one another.					
iv.	There was discipline during the interaction.					
v.	Members discussed in-depth to understand thoroughly.					
vi.	Members were probing deeply together.					
vii.	Members were explaining thoroughly.					
viii.	Very often interactions occurred during presentations.					
<b>4.</b>	<b>INDIVIDUAL ACCOUNTABILITY</b>					
i.	Students were always interested in learning in cooperative setup.					
ii.	Every member of the team was eager to complete the task.					
iii.	Every one accepted the assigned role in the team.					
iv.	Every one completed the accepted task.					
v.	Every one contributed ideas, thoughts and suggestions to the team.					
vi.	Members helped other team members if they faced difficulty.					
vii.	Personal assignments were completed regularly.					
viii.	Everyone got chance to represent their own team in the presentation.					
ix.	All were regular in the class.					
<b>5.</b>	<b>APPROPRIATE USE OF COLLABORATIVE SKILLS</b>					
<b>5.1</b>	<b>LEADERSHIP</b>					
i.	All the team members were engaged in the completion of task.					

ii.	The team members were treated respectfully.					
iii.	All the team members observed high moral.					
iv.	Tasks were distributed properly among the team members.					
v.	Conducive environment of learning was created.					
vi.	Time was managed properly.					
vii.	Suggestions of all the members were considered.					
viii.	Team members were properly instructed.					
ix.	It was a collective learning through participatory approach.					
<b>5.2</b>	<b>COMMUNICATION</b>					
i.	Interactions were done in a healthy learning environment.					
ii.	Every member was free to ask and respond to the questions.					
iii.	Every member got chance to express the ideas.					
iv.	Members were free to interact in different languages (Hindi, English & Gujarati).					
v.	Members paid attention to the speaker.					
<b>5.3</b>	<b>TRUST BUILDING</b>					
i.	Members were ready to work in randomly selected teams.					
ii.	All members were allowed to express their ideas.					
iii.	Ideas of all were used to solve a problem.					
iv.	There was full faith in the work done by others.					
v.	Other's explanations were relied on.					
vi.	Team work was fully observed.					
vii.	Credit of success/failure was attributed to all members of the team.					

<b>5.4</b>	<b>DECISION MAKING</b>					
i.	All the ideas were comprehended to arrive at a common solution.					
ii.	Team members were directed to carry out the distributed task.					
iii.	Results were drawn by summarizing the work of all team members.					
<b>5.5</b>	<b>RESOLVING CONFLICTS</b>					
i.	All were made emotionally & mentally ready to work in a team.					
ii.	Members were convinced logically on their arguments.					
iii.	Necessary arrangements were made to work in a team.					
iv.	Conflicts were resolved amicable.					
<b>6.</b>	<b>GROUP PROCESSING</b>					
i.	New teams were constituted in the progressive class.					
ii.	Members were selected randomly for team formation.					
iii.	Team goals and objectives were made clear to all the team members.					
iv.	Each team work was assessed periodically by the teacher.					
v.	Actions facilitating learning in this setup were promoted.					
vi.	Futile actions were dropped.					

## **Appendix: IV**

### **List of Experts for Tools Validation**

**1. Prof. D.R. Goel**

Department of Education (CASE & IASE)  
Faculty of Education and Psychology  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat

**2. Prof. Chhaya Goel**

Department of Education (CASE & IASE)  
Faculty of Education and Psychology  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat

**3. Prof. Mahesh Yagnik**

MB Patel College of Education  
Sardar Patel University  
Anand, Gujarat

**4. Prof. K. Muralidharan**

Department of Statistics  
Faculty of Education and Psychology  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat

**5. Associate Professor Sangita J. Parikh**

Department of Statistics  
Faculty of Commerce  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat

**6. Assistant Professor Arti Manish Khabia**

Faculty of Commerce  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat



- 7. Assistant Professor Rina M. Shah**  
Faculty of Commerce  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat
  
- 8. Assistant Professor Pratima Bavagosai**  
Faculty of Commerce  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat
  
- 9. Teaching Assistant Vaishali V. Mehta**  
Faculty of Commerce  
The Maharaja Sayajirao University of Baroda  
Vadodara, Gujarat

## Appendix: V

### Attendance Record of Students

Sonia Rohilla (1)

Centre for Advanced studies in Education

M.Ed. Programme 2015-16: Semester II: Attendance of Students

Subject: Methods of Educational Research subject Code: EEA 2242

Roll No.	Name of Student	21/12	22/12	26/12	29/12	2/1	5/1
1	ARAB RIZWANABANU AHMADMIYA	-	-	-	P	-	P
2	BARIYA SUNITA GOVINDBHAI	P	P	P	P	-	P
3	BHAGAT KUNJAL ASHOKKUMAR	P	P	P	P	P	P
4	BHAVANABEN MAGANBHAI PARMAR	P	P	-	-	P	P
5	DEVILABEN HIRABHAI ROHIT	P	P	P	P	P	P
6	DHARMISTA SHANKARBHAI PARMAR	P	P	P	-	P	P
7	DHIMMAR SHEFALIBEN JASHVANTBHAI	-	-	-	-	P	P
8	GANDHI DARSHANA GAJENDRAKUMAR	P	P	P	-	P	P
9	JAYASHREE PARAG JOSHI	P	P	P	P	P	P
10	JADAV RAKESHKUMAR HASMUKHLAL	P	P	-	-	-	P
11	JAIN ANUBEN DWARKAPRASAD <del>SHAH ANJU SIREN</del>	P	P	-	-	P	P
12	JAYALAXMI M	P	P	-	P	P	P
13	KAMLESHKUMAR DALSUKHBHAI MAKWANA	-	P	-	-	-	P
14	KHEVRA MANISHABEN CHETANBHAI	P	P	-	P	P	P
15	KOTIYA BHAGYASHREE KAMALCHANDRA	P	P	P	P	P	-
16	MADHU KUMARI (left) *	P	P	-	P	P	-
17	MAKWANA MINAL CHANDULAL	P	P	P	-	-	P
18	MANISHA RAVJI KAVALE	P	-	-	P	P	P
19	MEHTA JIGNASHABEN RAMANLAL	P	-	-	P	P	P
20	MERCHANT CANUT STEPHAN *	-	-	-	-	-	-
21	MESARIYA DIPIKABEN NAGINBHAI	P	P	-	P	-	P
22	PANCHAL ARPITABEN GHANSHYAMBHAI	-	P	P	-	P	P
23	PANGU SUKHVINDER KAUR HARBANSINGH	P	P	-	-	P	P
24	PATEL PAYALBEN PRAKASHBHAI	P	P	-	P	-	P
25	PRAJAPATI KALPANA KUMARI MANOHARLAL *	P	P	P	P	P	-
26	RAI ANJU RARMASHISH	-	-	-	-	-	P
27	RANA SHRUTI KANUBHAI	P	P	P	P	P	P
28	SAROJKUMARI BALBIRSINH PUNIYA	P	P	-	P	-	P
29	SHOURYA CHATURVEDI	P	P	P	P	-	-
30	TASHI YANGZOM	P	P	P	P	P	P
31	ANITA VITTAL KHADE	P	P	P	P	P	P
32	NAYEE BHAVISHA AMITKUMAR	P	P	P	P	P	P
33	NAYEE BHAVISHA AMITKUMAR	P	P	-	P	P	-
34	DAMOR DIVYABEN NARESHBHAI	P	P	-	P	P	-
35	GHANCHI HASINA ALTAFBHAI	P	P	P	P	-	P
36	THOMAS JOHN JEROME	-	-	-	-	P	P
	SIGNATURE OF TEACHERS						

21/12/15    22/12/15    26/12/15    29/12/15    2/1/16    5/1/16



Sonia Rohilla (2)

Centre for Advanced studies in Education  
M.Ed. Programme 2015-16: Semester II: Attendance of Students  
Subject: Methods of Educational Research subject Code: EEA 2242

Roll No.	Name of Student	7/1/16	9/1/16	16/1	19/1	23/1	30/1
1	ARAB RIZWANABANU AHMADMIYA	-	P	-	P	P	P
2	BARIYA SUNITA GOVINDBHAI	-	-	-	-	-	-
3	BHAGAT KUNJAL ASHOKKUMAR	P	P	P	-	P	P
4	BHAVANABEN MAGANBHAI PARMAR	P	P	P	P	-	P
5	DEVILABEN HIRABHAI ROHIT	P	P	P	P	-	P
6	DHARMISTA SHANKARBHAI PARMAR	P	P	P	P	-	P
7	DHIMMAR SHEFALIBEN JASHVANTBHAI	P	P	-	P	P	P
8	GANDHI DARSHANA GAJENDRAKUMAR	P	P	P	-	P	P
9	JAYASHREE PARAG JOSHI	P	P	P	P	P	P
10	JADAV RAKESHKUMAR HASMUKHLAL	P	P	-	P	P	P
11	JAIN ANUBEN DWARKAPRASAD	P	P	P	P	P	P
12	<del>JAIN ANUBEN DWARKAPRASAD</del> SHAH ANJIB SIREN	P	P	P	-	P	P
12	JAYALAXMI M	P	P	P	-	P	P
13	KAMLESHKUMAR DALSUKHBHAI MAKWANA	-	-	-	-	P	-
14	KHEVRA MANISHABEN CHETANBHAI	P	P	P	P	P	P
15	KOTIYA BHAGYASHREE KAMALCHANDRA	P	P	P	P	P	P
16	MADHU KUMARI (left) *	*	*	*	*	*	*
17	MAKWANA MINAL CHANDULAL	P	P	P	P	P	P
18	MANISHA RAVJI KAVALE	-	-	-	-	-	-
19	MEHTA JIGNASHABEN RAMANLAL	-	P	-	P	P	P
20	MERCHANT CANUT STEPHAN *	-	-	-	-	-	P
21	MESARIYA DIPIKABEN NAGINBHAI	-	-	P	P	-	P
22	PANCHAL ARPITABEN GHANSHYAMBHAI	-	-	-	-	-	-
23	PANGU SUKHVINDER KAUR HARBANSINGH	-	-	P	P	P	P
24	PATEL PAYALBEN PRAKASHBHAI	P	P	-	-	-	-
25	PRAJAPATI KALPANAKUMARI MANOHARLAL *	P	P	-	P	P	P
26	RAI ANJU RARMASHISH	-	-	-	P	-	P
27	RANA SHRUTI KANUBHAI	P	P	P	P	P	P
28	SAROJKUMARI BALBIRSIH PUNIYA	P	-	P	P	-	-
29	SHOURYA CHATURVEDI	P	-	P	P	P	P
30	TASHI YANGZOM	P	P	P	P	P	P
31	ANITA VITTAL KHADE	P	P	P	P	P	-
32	NAYEE BHAVISHA AMITKUMAR	P	P	P	P	-	P
33	NAYEE BHAVISHA AMITKUMAR	P	P	P	P	P	P
34	DAMOR DIVYABEN NARESHBHAI	P	P	-	P	P	P
35	GHANCHI HASINA ALTAFBHAI	P	P	P	P	P	P
36	THOMAS JOHN JEROME	P	P	-	P	P	P
	SIGNATURE OF TEACHERS	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>



(3)

## Centre for Advanced studies in Education

M.Ed. Programme 2015-16: Semester II: Attendance of Students

Subject:

subject Code:EEA

Roll No.	Name of Student	2/2/16	6/2	11/2	13/2	18/2	20/2
1	ARAB RIZWANABANU AHMADMIYA	P	P	P	P	P	P
2	BARIYA SUNITA GOVINDBHAI	—	—	—	P	P	—
3	BHAGAT KUNJAL ASHOKKUMAR	P	P	P	P	P	P
4	BHAVANABEN MAGANBHAI PARMAR	P	P	—	P	P	P
5	DEVILABEN HIRABHAI ROHIT	P	P	P	P	P	P
6	DHARMISTA SHANKARBHAI PARMAR	P	P	P	P	P	P
7	DHIMMAR SHEFALIBEN JASHVANTBHAI	P	P	P	P	P	P
8	GANDHI DARSHANA GAJENDRAKUMAR	P	P	P	P	P	P
9	JAYASHREE PARAG JOSHI	P	P	P	P	P	P
10	JADAV RAKESHKUMAR HASMUKHLAL	P	P	P	P	P	P
11	JAIN ANJUBEN DWARKAPRASAD SHAH ANJU ISIKEN	P	P	P	P	P	P
12	JAYALAXMI M	P	P	P	P	P	P
13	KAMLESHKUMAR DALSUKHBHAI MAKWANA	—	P	P	—	P	P
14	KHEVRA MANISHABEN CHETANBHAI	P	P	P	P	P	P
15	KOTIYA BHAGYASHREE KAMALCHANDRA	P	P	P	P	P	P
16	MADHU KUMARI (left)	—*	—*	—*	—*	—*	—*
17	MAKWANA MINAL CHANDULAL	P	P	P	P	P	P
18	MANISHA RAVJI KAVALE	—	P	P	P	P	P
19	MEHTA JIGNASHABEN RAMANLAL	P	P	P	P	P	P
20	MERCHANT CANUT STEPHAN (*)	—	—	—	—	—	—
21	MESARIYA DIPIKABEN NAGINBHAI	P	P	P	P	P	P
22	PANCHAL ARPITABEN GHANSHYAMBHAI	P	P	P	P	P	P
23	PANGU SUKHVINDER KAUR HARBANSINGH	P	P	P	P	P	P
24	PATEL PAYALBEN PRAKASHBHAI	—	P	—	P	P	P
25	PRAJAPATI KALPANAKUMARI MANOHARLAL (*)	P	P	P	P	P	P
26	RAI ANJU RARMASHISH	P	P	P	P	P	P
27	RANA SHRUTI KANUBHAI	P	P	P	P	P	P
28	SAROJKUMARI BALBIRSINH PUNIYA	—	P	—	P	P	P
29	SHOURYA CHATURVEDI	P	P	P	P	P	P
30	TASHI YANGZOM	P	P	P	P	P	P
31	ANITA VITTAL KHADE	—	P	P	P	P	P
32	NAYEE BHAVISHA AMITKUMAR	P	P	P	P	P	P
33	NAYEE BHAVISHA AMITKUMAR	P	P	P	P	P	P
34	DAMOR DIVYABEN NARESHBHAI	—	P	P	P	P	P
35	GHANCHI HASINA ALTAFBHAI	—	P	P	P	P	P
36	THOMAS JOHN JEROME	P	P	P	P	P	P
	SIGNATURE OF TEACHERS	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>



Centre for Advanced studies in Education  
M.Ed Programme 2015-16: Semester II: Attendance of Students

Subject: *RME*

subject Code: EEA

2242

(4)

Roll No.	Name of Student	23/2	25/2	27/2	21/3	22/3	23/3
1	ARAB RIZWANABANU AHMADMIYA	P	P	P	P	P	P
2	BARIYA SUNITA GOVINDBHAI	P	P	P	P	P	P
3	BHAGAT KUNJAL ASHOKKUMAR	P	P	P	P	P	P
4	BHAVANABEN MAGANBHAI PARMAR	P	P	P	P	P	P
5	DEVILABEN HIRABHAI ROHIT	P	P	P	P	P	P
6	DHARMISTA SHANKARBHAI PARMAR	P	P	P	P	P	P
7	DHIMMAR SHEFALIBEN JASHVANTBHAI	P	P	P	P	P	P
8	GANDHI DARSHANA GAJENDRAKUMAR	P	P	P	P	P	P
9	JAYASHREE PARAG JOSHI	P	P	P	P	P	P
10	JADAV RAKESHKUMAR HASMUKHLAL	P	P	P	P	P	P
11	<del>JAIN ANJUBEN DWARKAPRASAD</del> SHAB ANJU BIREN	P	P	P	P	P	P
12	JAYALAXMI M	P	P	P	P	P	P
13	KAMLESHKUMAR DALSUKHBHAI MAKWANA	P	P	P	P	P	P
14	KHEVRA MANISHABEN CHETANBHAI	P	P	P	P	P	P
15	KOTIYA BHAGYASHREE KAMALCHANDRA	P	P	P	P	P	P
16	MADHU KUMARI (left) *	*	*	*	*	*	*
17	MAKWANA MINAL CHANDULAL	P	P	P	P	P	P
18	MANISHA RAVJI KAVALE	P	P	P	P	P	P
19	MEHTA JIGNASHABEN RAMANLAL	P	P	P	P	P	P
20	MERCHANT CANUT STEPHAN (*)	-	-	-	-	-	-
21	MESARIYA DIPIKABEN NAGINBHAI	P	P	P	P	P	P
22	PANCHAL ARPITABEN GHANSHYAMBHAI	P	P	P	P	P	P
23	PANGU SUKHVINDER KAUR HARBANSINGH	P	P	P	P	P	P
24	PATEL PAYALBEN PRAKASHBHAI	P	P	P	P	P	P
25	PRAJAPATI KALPANAKUMARI MANOHARLAL (*)	-	-	-	-	-	-
26	RAI ANJU RARMASHISH	-	P	P	P	-	P
27	RANA SHRUTI KANUBHAI	P	P	P	P	P	P
28	SAROJKUMARI BALBIRSINH PUNIYA	-	-	P	-	P	P
29	SHOURYA CHATURVEDI	P	P	P	P	P	P
30	TASHI YANGZOM	P	P	P	P	P	P
31	ANITA VITTAL KHADE	P	P	P	P	P	P
32	NAYEE BHAVISHA AMITKUMAR	P	P	P	P	P	P
33	NAYEE BHAVISHA AMITKUMAR	P	P	P	P	P	P
34	DAMOR DIVYABEN NARESHBHAI	P	P	P	P	P	P
35	GHANCHI HASINA ALTAFBHAI	P	P	P	-	P	P
36	THOMAS JOHN JEROME	P	P	P	P	-	P
	SIGNATURE OF TEACHERS	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>	<i>[Signature]</i>

23/2

25/2

27/2

21/3

22/3

23/3

**Centre for Advanced studies in Education**  
M.Ed. Programme 2015-16: Semester II: Attendance of Students  
Subject: Res. Methodology subject Code: EEA

(5)  
2242

Roll No.	Name of Student	26/3	27/3	29/3	30/3		
1	ARAB RIZWANABANU AHMADMIYA	P	P	P	P		
2	BARIYA SUNITA GOVINDBHAI	P	P	P	P		
3	BHAGAT KUNJAL ASHOKKUMAR	P	P	P	P		
4	BHAVANABEN MAGANBHAI PARMAR	P	P	P	P		
5	DEVILABEN HIRABHAI ROHIT	P	P	P	P		
6	DHARMISTA SHANKARBHAI PARMAR	P	P	P	P		
7	DHIMMAR SHEFALIBEN JASHVANTBHAI	P	—	P	P		
8	GANDHI DARSHANA GAJENDRAKUMAR	P	P	P	P		
9	JAYASHREE PARAG JOSHI	P	P	P	P		
10	JADAV RAKESHKUMAR HASMUKHLAL	P	P	P	P		
11	JAIN ANJUBEN DWARKAPRASAD	P	P	P	P		
12	JAYALAXMI M	P	P	P	P		
13	KAMLESHKUMAR DALSUKHBHAI MAKWANA	—	P	P	P		
14	KHEVRA MANISHABEN CHETANBHAI	P	P	P	P		
15	KOTIYA BHAGYASHREE KAMALCHANDRA	P	P	P	P		
16	<del>MADHU KUMARI</del> (left) (*)						
17	MAKWANA MINAL CHANDULAL	P	P	P	P		
18	MANISHA RAVJI KAVALE	P	P	P	P		
19	MEHTA JIGNASHABEN RAMANLAL	P	P	P	P		
20	MERCHANT CANUT STEPHAN (*)	—	—	P	—		
21	MESARIYA DIPIKABEN NAGINBHAI	P	P	P	P		
22	PANCHAL ARPITABEN GHANSHYAMBHAI	P	P	P	P		
23	PANGU SUKHVINDER KAUR HARBANSINGH	P	P	P	P		
24	PATEL PAYALBEN PRAKASHBHAI	P	P	P	P		
25	PRAJAPATI KALPANAKUMARI MANOHARLAL (*)						
26	RAI ANJU RARMASHISH	P	—	P	P		
27	RANA SHRUTI KANUBHAI	P	P	P	P		
28	SAROJKUMARI BALBIRSINH PUNIYA	P	P	P	P		
29	SHOURYA CHATURVEDI	P	P	P	P		
30	TASHI YANGZOM	P	P	P	P		
31	ANITA VITTAL KHADE	—	P	P	P		
32	NAYEE BHAVISHA AMITKUMAR	P	P	P	—		
33	NAYEE BHAVISHA AMITKUMAR	P	P	P	P		
34	DAMOR DIVYABEN NARESHBHAI	P	P	P	P		
35	GHANCHI HASINA ALTAFBHAI	P	—	P	P		
36	THOMAS JOHN JEROME	P	P	P	P		
	SIGNATURE OF TEACHERS						

26/3      27/3      29/3      30/3



## **Appendix: VI**

### **List of Resources for Learning**

#### **Books in Library**

- Aggarwal, Y.P. (1998). Statistical Methods Concept, Application and Computation. New Delhi: Sterling Publishers (Pvt.) Ltd.
- Best, J.W., & Kahn, J.V. (2009). Research in Education. New Delhi: Prentice Hall of India Pvt. Ltd.
- Creswell, J.W. (2011). Educational Research: Planning, conducting, and Evaluating, Quantitative and Qualitative Research. New Delhi : PHI learning Pvt. Ltd.
- Creswell, J.W. (2011). Educational Research: Planning, conducting, and Evaluating, Quantitative and Qualitative Research. New Delhi: PHI learning Pvt. Ltd.
- Fox, D. J. (1969). The Research Process in Education. New York: Holt Rinchart and winstoninc.
- Garrett, H.E. (1966). Introduction to Statistics in Psychology and Education. New York: Longman's Green and Co.
- Gay, L.R., Mills, G. E., & Airasian, P. (2009). Educational Research. Competencies for Analysis and Applications. New Jersy: Merrill and Pearson.
- Ghose, B. N. (1969). Scientific Method and Social Research. New Delhi: Sterling publisher Pvt. Ltd.
- Graziano, M. & Raulin, M. (1980). Research Methods, A process of Inquiry. New York: Harper and Row.
- Guilford, J. P. (1978). Fundamentals of Statistics in Psychology and Education. New York: Mcgraw Hill Series.
- Gupta, I. C. (2010). Business Statistics. Himalaya Publishing House.
- Hogg, R.V., Mckean, J.W. & Craig, A.T. (2012). Introduction to Mathematical Statistics. Pearson Education India.
- Hollander, M., Wolfe, D.A. & Chicken, E. (2014). Nonparametric Statistical Methods. Wiley Publication.
- Kapoor, S. C. (2014). Fundamentals of Mathematical Statistics. Sultan Chand and Sons.
- Kapoor, V. (2019). Modern Approach to Fundamentals of Statistics for Business and Economics. Sultan Chand and Sons.

- Keeves, J. P. (Ed.) (1990). Educational Research Methodology and Measurement: An International Handbook. New York: Pargamon Press.
- Kerlinger, F.N. (1967). Foundations of Behavioural Research, Education and Psychological Inquiry. New York: Richard and Winston.
- Lovell, K. & Lawson, K.S. (1970). Understanding Research in Education. London : University of London.
- Mangal, S. K. (2002). Statistics in Psychology and Education. PHI Learning Pvt. Ltd.
- Mann, P. S. (2010). Introductory Statistics. Wiley.
- Mouly, G.T. (1963). The Science of Educational Research. New Delhi: Eurasia Publishing House.
- Nagar, A. (1985). Basic Statistics. OUP India.
- Prasant Kumar, G. A. (2018). Introduction to Statistics Including Statistics Practical. Academic Publishers.
- Sharma, J. K. (2014). Business Statistics . Vikas Publishing House.
- Singh, K. (2001). Methodology and Techniques of Social Research. New Delhi: Kanishka publishers.
- Travers, R.M. (1969). Introduction to Educational Research. London: Macmillan Publishing Co.
- Tuckman, B.W.(1972).Conducting Fundamental Research. New York: Harcourt Brace Javonovich Inc.
- Van, D.B., and Meyer, W.J. (1962).Understanding Educational Research: An introduction. New York: Mcgraw Hill Book Company.

### **Online e-books (free to download) in Pdf format**

- Leon, R.V. (2004). *Unit 14: Nonparametric Statistical Methods. Statistics 571: Statistical Methods*. web.utk.edu. <https://web.utk.edu/~leon/stat571/2004SummerPDFs/571Unit14.pdf>
- Rao, S.R. (2015). *Business research methodology by SRINIVAS R Rao*. Free-eBooks.net. <https://www.free-ebooks.net/business-textbooks/Business-Research-Methodology>
- Scanlan,C.L. (n.d.). *Introduction to Nonparametric statistics*. Academia.edu - Share research. [https://www.academia.edu/7805173/Introduction\\_to\\_Nonparametric\\_Statistics?s](https://www.academia.edu/7805173/Introduction_to_Nonparametric_Statistics?s)  
=t



- Sheskin, D.J. (2000). *Handbook of PARAMETRIC and NON-PARAMETRIC STATISTICAL PROCEDURES*. Fakultas MIPA dan Kesehatan. [https://fmipa.umri.ac.id/wp-content/uploads/2016/03/David J. Sheskin David Sheskin Handbook of Parametric and Nonparametric Statistical Methods.pdf](https://fmipa.umri.ac.id/wp-content/uploads/2016/03/David_J._Sheskin_David_Sheskin_Handbook_of_Parametric_and_Nonparametric_Statistical_Methods.pdf)
- Sprent, P. & Smeeton, N.C. (2001). *Applied Nonparametric Statistical Methods*. <https://spu.fem.uniag.sk/cvicenia/ksov/prokeino/Business%20Statistics%20and%20Econometrics/Literature/20089702653110.pdf>
- Kanji, G.K. (2006). *100 statistical tests : Kanji, Gopal K : Free download, borrow, and streaming : Internet archive*. Internet Archive. <https://archive.org/details/100statisticaltests0000kanji>
- Kothari, C.R. (2004). *Research Methodology: Methods and Techniques*. Tarbiat Modares University. <https://www.modares.ac.ir/uploads/Ag.Oth.Lib.17.pdf>
- Dowdy, S., Wearden, S. & Chilko, D. (2015, April 5). *Statistics for research - PDF free download*. epdf.pub. <https://epdf.pub/statistics-for-research.html>
- Singh, Y.K. (2015, April 21). *Fundamental of research methodology and statistics - PDF free download*. epdf.pub. <https://epdf.pub/fundamental-of-research-methodology-and-statistics.html>
- Kothari, C.R. (2015, September 15). *Research methodology: Methods and techniques - PDF free download*. epdf.pub. <https://epdf.pub/research-methodology-methods-and-techniques.html>
- Beins, B.C. & McCarthy, M.A. (2015, November 19). *Research methods and statistics - PDF free download*. epdf.pub. <https://epdf.pub/research-methods-and-statistics.html>
- Hesse, C.A., Ofosu, J.B. & Nortey, E.N. (2018, January 24). *(PDF) Introduction to nonparametric statistical methods*. ResearchGate. [https://www.researchgate.net/publication/322677728\\_INTRODUCTION\\_TO\\_NONPARAMETRIC\\_STATISTICAL\\_METHODS](https://www.researchgate.net/publication/322677728_INTRODUCTION_TO_NONPARAMETRIC_STATISTICAL_METHODS)

## Appendix- VII

### Lesson Designs

**Announcement of topic in Class:** This announcement was made three days prior to the class. For the coming class read about Graphs, their types and applications.

#### Lesson No.1: Graphs

##### Teaching Points:

- Histogram
- Frequency polygon
- Frequency curve
- Cumulative frequency Curve / Ogive curve

##### Instructional Objectives:

After completion of this class students will be able to

- i. List various types of graphs.
- ii. Distinguish among various types of graphs.
- iii. Use various types of graphs.
- iv. Interpret the drawn graph and trace useful information from them.

##### Lesson Presentation:

[Tr: Teacher and St: student]

Tr: How many ways in print form information can be presented?

St: Through text.

St: With the help of tables.

St: using graphs, charts and pictures or figures.

Tr: Yes, you all are true we can represent information by text, pictures, tables, graphs and charts.

Tr: Have you ever used graph? If yes, name them.

St: Histogram, Frequency polygon, Frequency curve and Cumulative curve or Ogive curve.

Tr: Yes, you are true. Now, when do we use histogram?

St: Histogram is used to plot the frequency of score occurrences in a continuous data set that has been divided into classes, called bins.

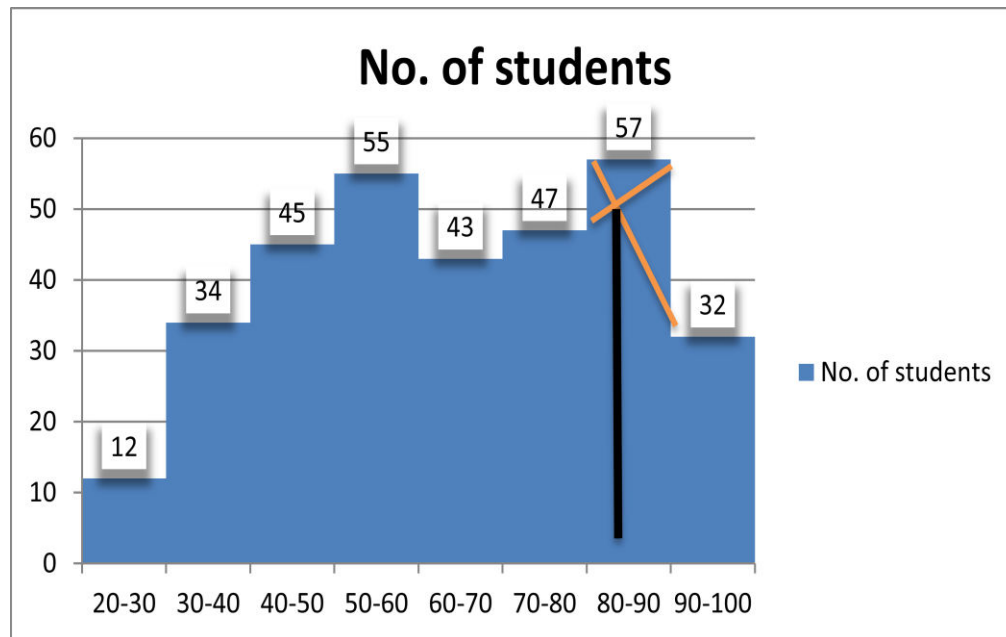
Tr: Very true. Do you know we can trace the value of mode from the histogram?

(if no one responded then only teacher will explain this and respond otherwise teacher will bring the discussion to the conclusion.

Tr: Draw a histogram for the following data and trace the value of mode from it.

Marks	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of Students	12	34	45	55	43	47	57	32

St:



Choose the column of maximum frequency and join the corner of adjutant columns as shown above. Where so ever these two lines intersect, draw a perpendicular from that point of intersection to the X- axis, read that value where it touches the X- axis. Here the value is 83. So the mode value is 83.

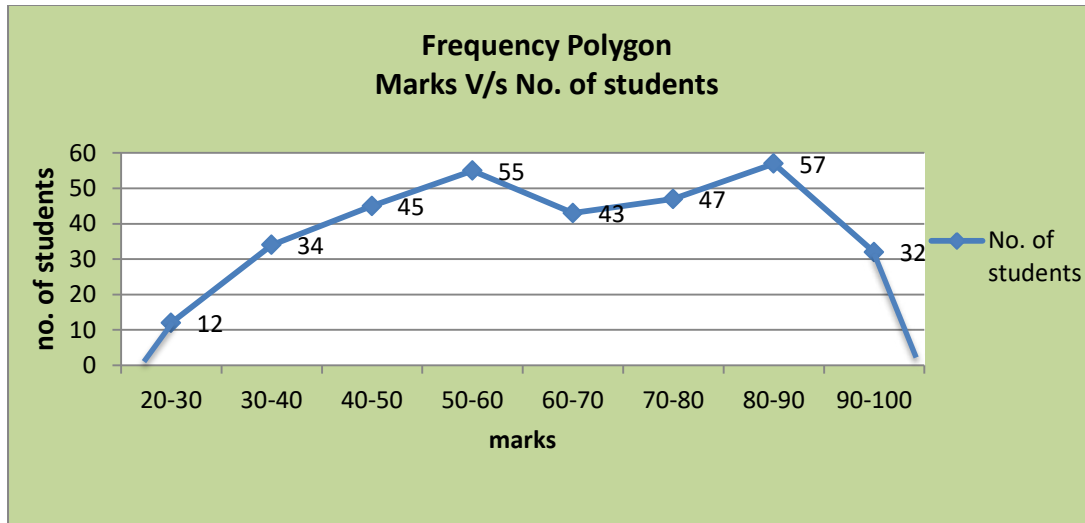
St: Madam, if there are two columns of same maximum heights then how to trace mode value from the histogram?

Tr: In that case there will be two modes. Both the columns will be used for tracing mode values.

Such data where two modes exist we call them as bimodal data.

Now how do you draw frequency polygon for the same data given above?

St: (students will draw the following graph)



Tr: You have rightly plotted the points of data at the mid - point of the classes given in the data.  
And you have also joined the end points of the curve with the x-axis.

Can you draw frequency curve for the same data?

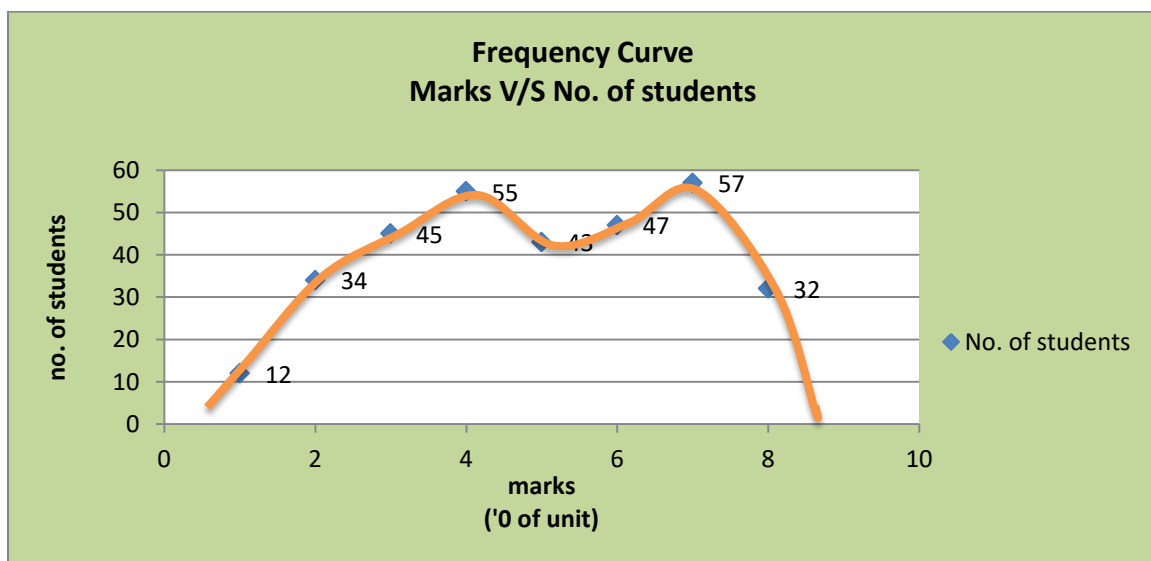
St: Yes madam!

Tr: What change will you make in drawing frequency curve?

St: In frequency curve we have to join the points with a smooth hand curve instead of joining points with line segments.

Tr: You are absolutely right. So now draw the frequency curve for the same data.

St: (students will draw this graph)



Tr: Good. Now do you know we can also trace partition values like quartiles, deciles and percentiles from a graph?

St: There may be right response. (But in case students don't know about this teacher will explain it with the help of an example.)

Tr: Cumulative frequency curve is also known as Ogive curve.

Where we can trace decile, quartile and percentile values from the graph.

Following calculation is required for that:

$$Q_i = i * (N/4)$$

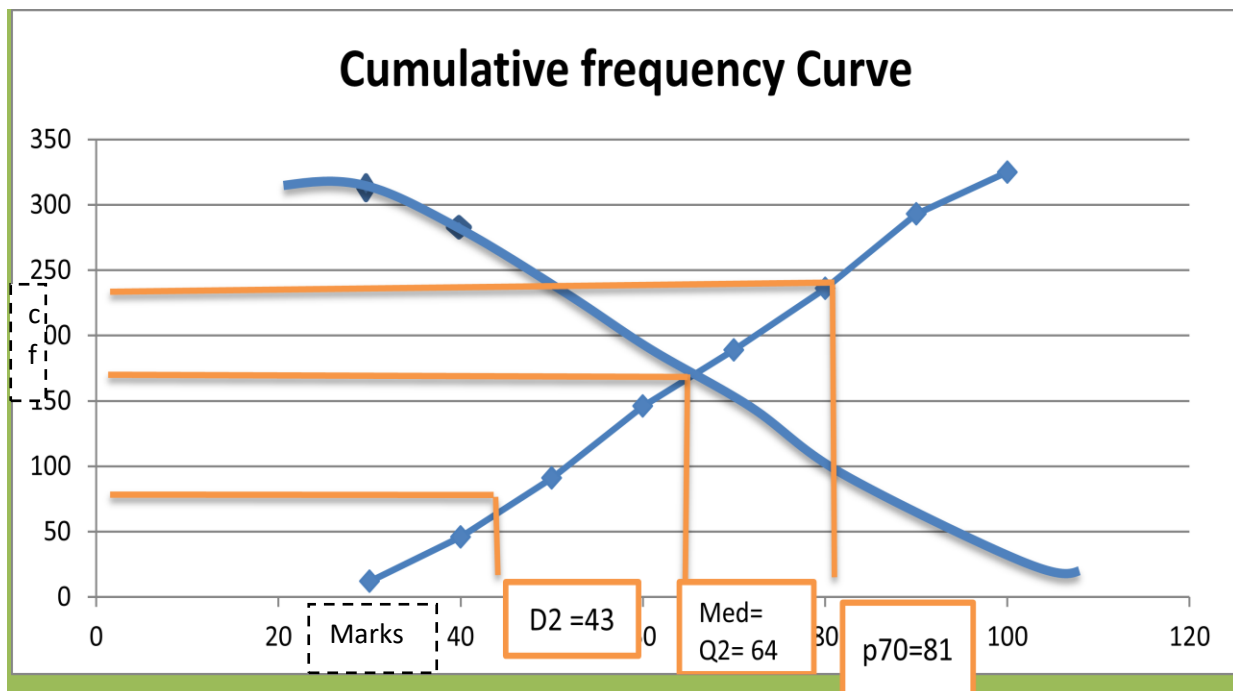
$$D_i = i * (N/10)$$

$$P_i = i * (N/100)$$

Here  $Q_2 = 2 * (325/4) = 162.5$  is traced on Y axis and meet the less than type curve and then perpendicular is drawn to meet X- axis and obtained value for  $Q_2$  is 64.

$D_2 = 2 * (325/10) = 65$  is traced on Y axis and meet to the less than type curve and then a perpendicular is drawn to meet X- axis. This will meet the X-axis at a point, say here 43.

Similarly for  $P_{70} = 70 * (325/100) = 227.5$  is traced on Y axis and meet to the less than type curve and then perpendicular is drawn to meet X- axis. So, the obtained value for  $P_{70}$  is 81.



Tr: Now interpret the values  $Q_2 = 64$ ,  $D_2 = 43$  and  $P_{70} = 81$ .

St:  $Q_2 = 64$  means that 50% of the students have scored more than 64 score and rest 50% of the students have scored less than 64 score.

St:  $D_2 = 43$  means that 20% of the students in class have scored less than 43 score or we can also state that 80% of the students have more than 43 score.

St:  $P_{70} = 81$  means that 70 % of the students have scored less than 81 score or 30 % of the students have more than 81 score.

Tr: Yes you all are right.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### Task allotment for each group:

1. Construct frequency distribution for the following data and draw Histogram. Also trace the value of Mode from the histogram.

### For Group No. 1, 3, 5

78	98	88	78	79	78	89	89	90	90
77	79	87	76	67	78	79	98	92	84
25	84	84	46	77	83	45	34	34	34
45	56	54	57	67	68	77	77	67	69
45	46	57	58	67	78	79	70	45	45
45	34	45	46	57	58	78	79	70	56
34	45	56	76	6	63	61	52	48	49
23	43	4	41	45	34	65	28	63	45

**For Group No. 2, 4, 6**

30	23	45	42	65	62	31	39	47	45
34	23	34	23	20	21	04	34	32	22
45	21	67	67	29	36	3	33	12	34
44	22	35	23	64	55	35	33	54	51
44	49	46	45	34	33	34	56	15	16
57	23	45	46	65	42	23	14	24	24
67	43	34	24	21	65	60	12	14	25
17	45	44	34	31	32	353	61	37	41

**(All Groups)**

2. Construct frequency distribution for the following data and draw Ogive curve. Also trace the following partition values from the Ogive curve and interpret them:

$Q_1, Q_2, D_4, D_7, P_{35}, P_{65}$

12	24	33	33	23	34	42	43	23	34
35	44	54	53	32	33	33	43	42	32
25	27	3	7	38	29	35	36	47	34
32	34	35	33	45	46	47	46	45	34
35	36	44	45	43	24	35	43	44	44
24	34	14	15	16	23	24	25	35	45
46	42	41	41	32	43	52	4	3	43

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about charts, their types and applications.

## Lesson No.2: Charts

### Teaching Points:

- Line Chart
- Multiple Line Chart
- Simple Bar Chart
- Multiple Bar Chart
- Pie Chart
- Band chart
- Percentage Bar Chart

### Instructional Objectives:

After completion of this class students will be able to

- i. List various types of charts.
- ii. Draw various types of charts.
- iii. Choose an appropriate type of chart to represent information.

### Lesson Presentation:

Tr: For the following information suggest suitable charts which can represent this information in a meaningful way.

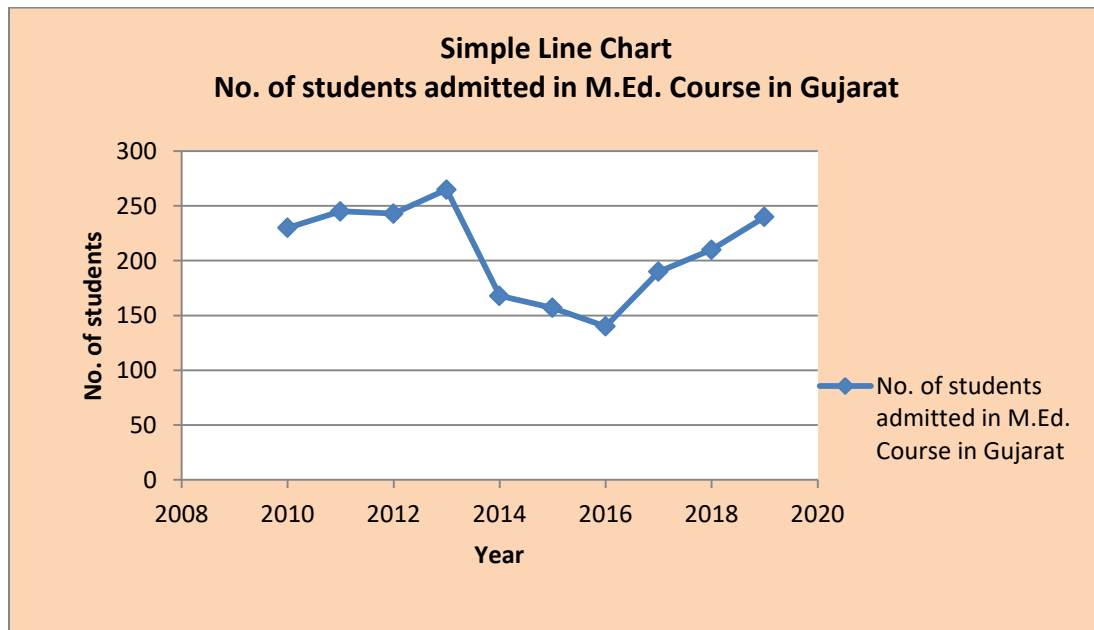
- Looking upon the given data if students will be able to identify the appropriate chart to be used then it will be fine otherwise teacher will suggest the chart and describe the cause of its selection.
- Students may face difficulty in selection of Axis in the charts. Where mutual discussion among students and if needed teachers intervention may be employed.



(a)

Year	No. of students admitted in M.Ed. Course in Gujarat
2010	230
2011	245
2012	243
2013	265
2014	168
2015	157
2016	140
2017	190
2018	210
2019	240

St:



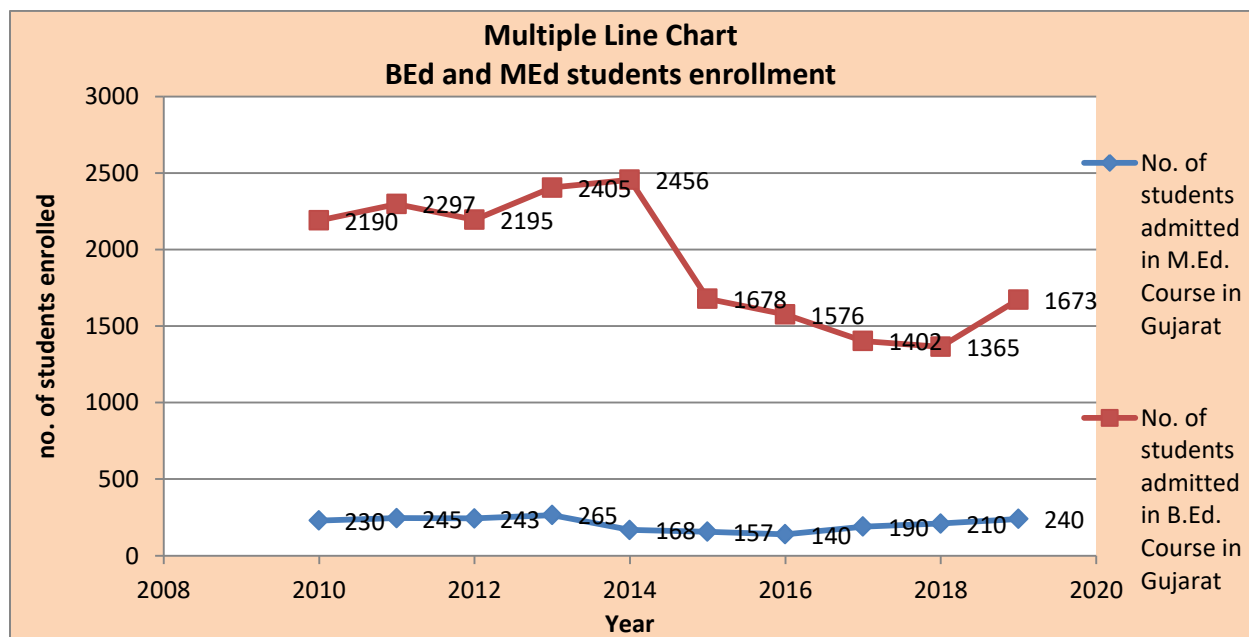
Tr: When to draw a simple line chart?

St: Line graphs are used to track changes over short and long periods of time. When smaller changes exist, line graphs are better to use than bar graphs. Line graphs can also be used to compare changes over the same period of time for more than one group.

(b) Tr:

Year	No. of students admitted in M.Ed. Course in Gujarat	No. of students admitted in B.Ed. Course in Gujarat
2010	230	2190
2011	245	2297
2012	243	2195
2013	265	2405
2014	168	2456
2015	157	1678
2016	140	1576
2017	190	1402
2018	210	1365
2019	240	1673

St:

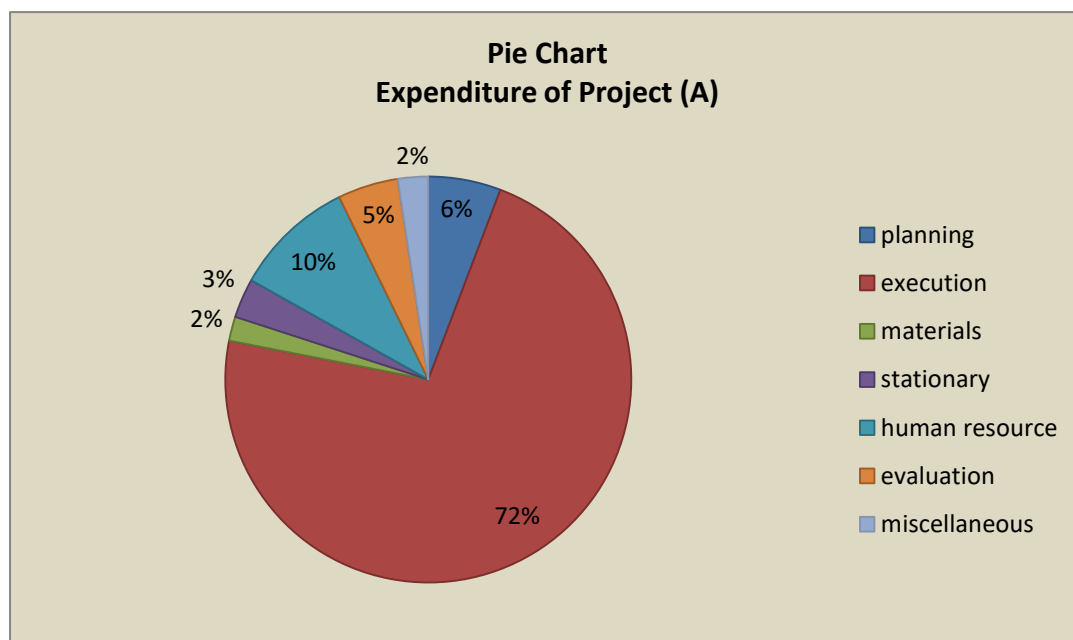


Tr: When to draw a multiple line chart?

St: A multiple line graph shows the relationship between independent and dependent values of multiple sets of data. Usually multiple line graphs are used to show trends over time. In the graph, each data value is represented by a point in the graph that are connected by a line

(c)

Items	Expenditure of Project (Rs)
Planning	120000
Execution	1500000
Materials	40000
Stationary	65000
Human Resource	200000
Evaluation	100000
Miscellaneous	50000
Total	2075000



Tr: When to draw a pie chart?

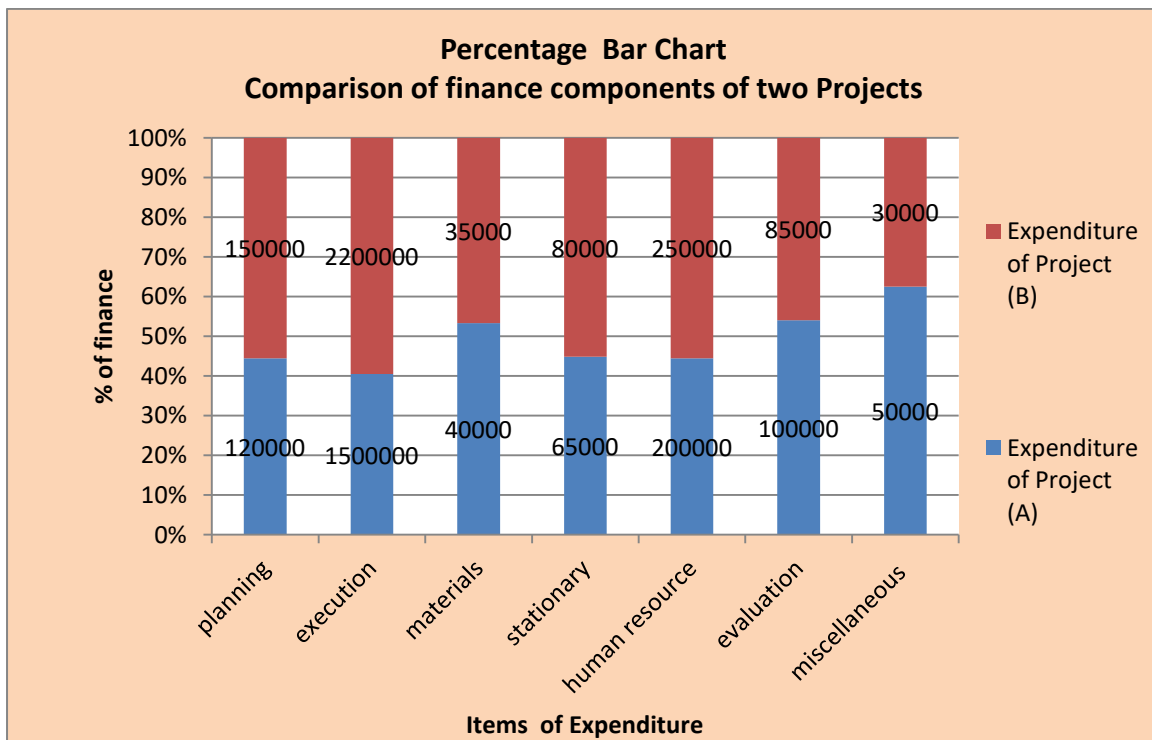
St: Pie charts are best to use when we are trying to compare parts of a whole.

Tr: Yes, you are absolutely right. When we want to compare the component of a whole in terms of proportion or percentage we use pie chart.

(d) Tr:

Items	Expenditure of Project (A)	Expenditure of Project (B)
Planning	120000	150000
Execution	1500000	2200000
Materials	40000	35000
Stationary	65000	80000
Human Resource	200000	250000
Evaluation	100000	85000
Miscellaneous	50000	30000
Total	2075000	2830000

St:



St: Madam, can we use two pie charts to represent the above information?

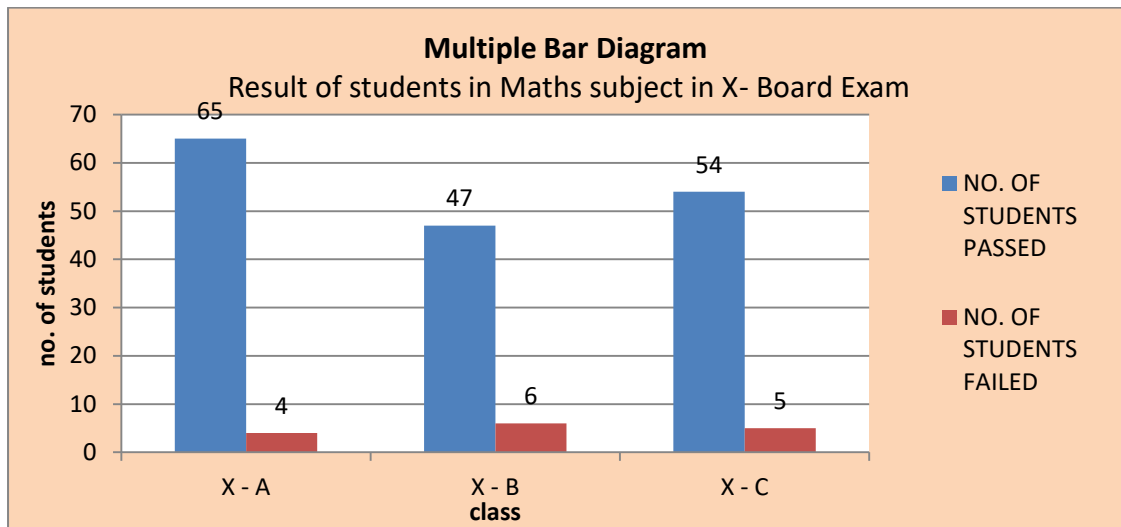
Tr: You can use two pie charts but for comparison purpose Percentage Bar chart is more advisable rather making two pie charts. If you want to represent the expenditures of two separate

projects where you don't want to compare the two project's expenditures, just for showing their within bifurcation of expenditure in each project in different components two separate pie charts can be use. Now suggest chart for the following data:

(e)

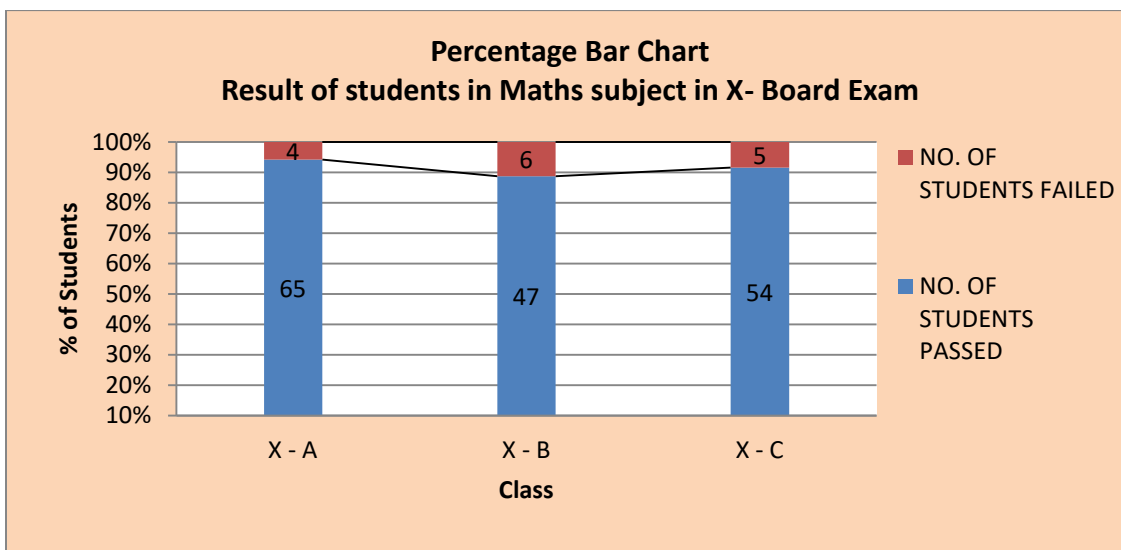
Result of Students in Maths Subject in X- Board Exams			
Class / Strength of students	X - A	X - B	X – C
No. of Students Passed	65	47	54
No. of Students Failed	4	6	5

St:



St:

Or

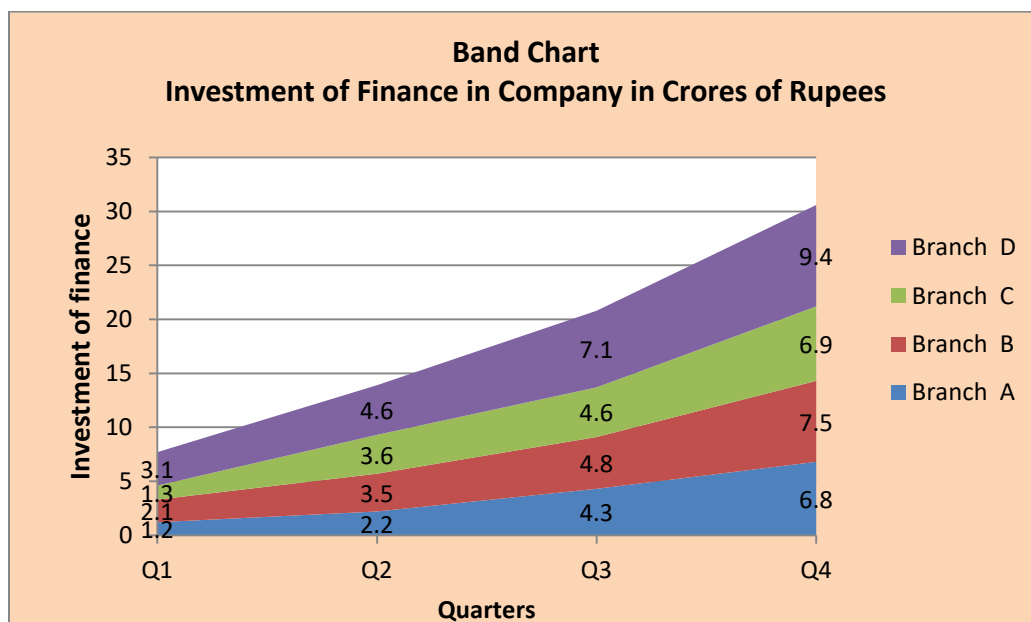


Tr: You can see the two charts shown above. Both are expressing the same information but different charts are used. In the first chart multiple bar chart is used and in the second chart percentage bar chart is used. In the multiple bar chart absolute data values are used where as in Percentage bar chart relative values are used to compare the data values. Depends upon the requirement of the reader different charts can be used to explain the context by the write. So for a same data multiple charts can be used. It depends upon the nature and demand of the writer to choose and use a chart. But the chosen chart should be meaningful enough to explain the intact information in pictorial form.

(f) Tr:

<b>Investment of Finance in Company in Crores of Rupees</b>				
Quarters/ Branches	Q1	Q2	Q3	Q4
Branch A	1.2	2.2	4.3	6.8
Branch B	2.1	3.5	4.8	7.5
Branch C	1.3	3.6	4.6	6.9
Branch D	3.1	4.6	7.1	9.4

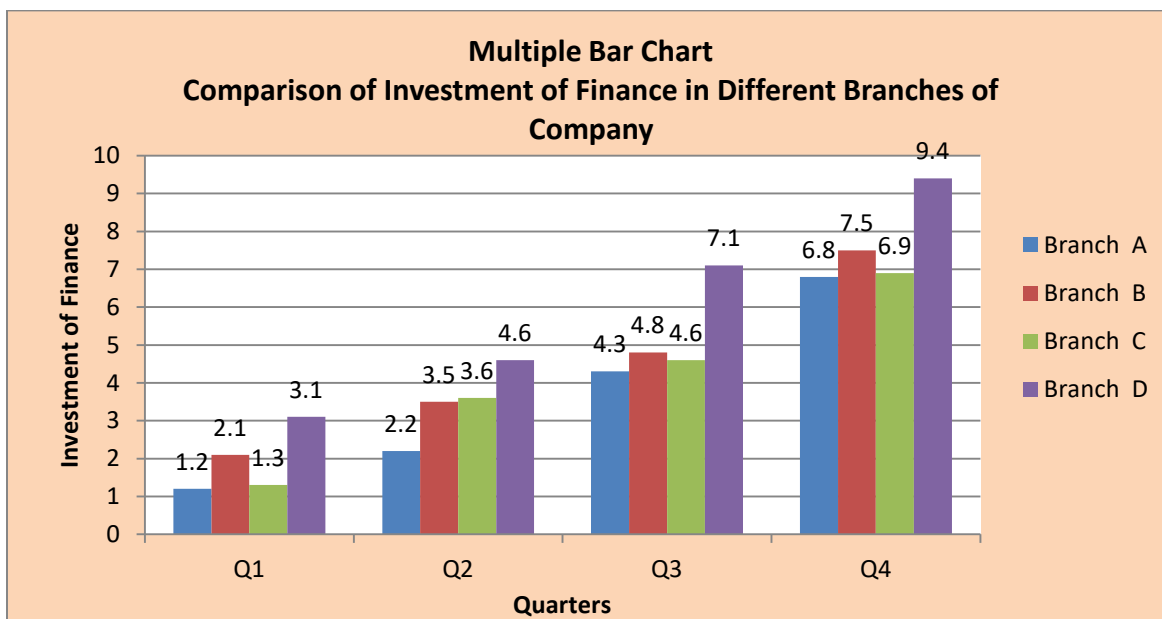
St:



St: Madam, here can we use Multiple Bar Chart?

Tr: Yes, why not? Just draw a multiple Bar chart for this data.

St:



You can use Band char as well as Multiple Bar chat. It depends upon the requirement of the writer. If you want to compare the investment of each quarter then Band chart is more suitable but if you wants to compare investment of finance in each branch per quarter than Multiple Bar chart is more efficient.

There are many more charts available to us for representing data. These are few charts which are often in use. Therefore I chose these few charts you can explore more and use different charts. For a same data multiple charts can be formed. The nature of data available and the purpose of study decide the appropriateness in selection of chart for a problem.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### Task allotment for each group:

1. Use the appropriate chart to represent the following data:

(a) **Group No: 1, 3, 5**

Roll No. of Student	1	2	3	4	5	6	7	8	9
% of marks	70	62	86	54	52	69	70	76	85

**(b) Group No. 2, 4, 6**

Roll No. of Student	1	2	3	4	5	6	7	8	9
Marks in Biology Practical Exam	40	42	36	47	37	43	47	36	36
Marks in Physics Practical Exam	43	43	42	43	39	48	43	33	35
Marks in Chemistry Practical Exam	45	47	41	44	32	49	32	39	32

**(c) Group: 1, 3, 5**

Date	10/2	11/2	12/2	13/2	14/2	15/2	16/2
Maxi. Temp. of Delhi city	17 °C	18 °C	18.5°C	19°C	21°C	22.6°C	24°C

**(d) Group: 2, 4, 6**

Items	Expenditure of National Seminar ('00000 Rs.)
TA	9
DA	1
Accommodation	1.5
Catering	3
Stationary	1.3
Publications	1.7
Decoration	0.09
Miscellaneous	0.05

**Presentation and Discussion:**

After group work one student from each group presents their task. Depending upon the presentation the teacher recapitulates and concludes the class. At the end teacher will give assignment related to the present topic to the students. Assignment work related to this topic is mentioned in the appendix- VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Measures of Central Tendency – Mean, Median and Mode.



## Lesson No.3: Measures of Central Tendency

### Teaching Points:

- Mean
- Median
- Mode

### Instructional Objectives:

After completion of this class students will be able to:

- Define Mean, Median and Mode.
- State the properties of Mean, Median and Mode.
- Identify the best suitable measure of central tendency among mean, median and mode for a given data.
- Calculate Mean, Median and Mode for Grouped and Ungrouped data.
- Interpret the obtained value of Mean, Median and Mode in context of given problem.

### Lesson Presentation:

Tr: What are the different measures we use to calculate average?

St: Mean.

Tr: Else.

St: Median, Mode.

Tr: Among these measures of average mean is most exploited measure of average or central tendency of data. Do you know why?

(Some student may answer this question but in case students would not answer then teacher will reply this.) In order to understand this first we need to understand the requisites of an ideal measure of an average.

- It should be easy to understand, easy to calculate.
- It should be well defined and most appropriate.
- It should be based upon all observations.
- It should not be affected by extreme values.
- It should not be influenced by fluctuations of sampling.
- It should be useful for the purposes of further algebraic analysis.

Now among the above mentioned requisites which one (mean, median or mode) satisfies the most?

St: Mean satisfy most of the requisites.

Tr: Yes! You are true. That is why Mean is used at most. Only it cannot be saved himself from the effect of extreme values in the data. Now define mean, median and mode?

St: Mean: It is the quantity obtained by sum of all observations divided by the total number of observations.

St: Median: It is the middle most value of the data in an arranged set of data. It divides the data array into two equal parts.

St: Mode: it is the most frequent occurring value in the data.

Tr: Yes, you all are right. For the following Raw score/ Ungrouped data how you will you calculate Mean, Median and mode?

Scores of Viva-Voce Exam:

23	24	32	34	33	33	25	26
32	36	37	38	33	21	23	24

St:

$$\text{Mean } \bar{x} = \frac{\sum x}{N}$$

$$= 474 / 16$$

$$= 29.625$$

$$\text{Median } M = \text{value of } \left( \frac{N+1}{2} \right)^{\text{th}} \text{ observation} \quad [\text{when } N \text{ is odd}]$$

$$= \text{value of } \frac{\left( \left( \frac{N}{2} \right)^{\text{th}} \text{ observation} + \left( \frac{N}{2} + 1 \right)^{\text{th}} \text{ observation} \right)}{2} \quad [\text{when } N \text{ is even}]$$

Here n= 16 and data is arranged in increasing order:

21	23	23	24	24	25	26	32	32	33	33	33	34	36	37	38
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$M = (8^{\text{th}} \text{ observation} + 9^{\text{th}} \text{ observation}) / 2$$

$$= (32 + 32) / 2 = 32$$

Mode Z = Number which is repeated maximum times in the data series.

Here Mode Z = 33 [as 33 is repeated maximum times in the data series]

Tr: Yes, you all are right. But how will u interpret these values?

St: Mean: The average score of 16 students in Viva – Voce Practical Exam is 29.625.

Or

On an average students have scored 29.625 score in Viva – Voce Practical Exam.

Tr: How will you interpret Median value?

St: Median is the middle most value of data which divides the data in two equal parts.

Tr: Can you try to interpret this Median value in the given context?

St: Half of the students have scored more than 32 score in Viva – Voce Practical Exam.

Tr: Good. Here Median is 32 which means that 50% of students in the class of 16 students have scored more than 32 and rest 50% students have scored less than the 32 score.

And how would you interpret mode value?

St: Here mode is 33 which can be interpreted as, mostly students have scored 33.

Tr: Yes, it's true. 33 is the most repeated value. Now calculate the Mean, Median and Mode for the following grouped data (discrete frequency data) and interpret the values also:

Given the following age frequency distribution of 10<sup>th</sup> standard students of a particular school.

Age (Years)	13	14	15	16	17
Number of Students	2	5	13	7	3

St: Here Variable involved is ages of 10<sup>th</sup> standard students. While the number of students' represents frequencies.

<b>Ages (Years)</b> x	<b>Number of Students</b> f	fx	Cf
13	2	26	2
14	5	70	7
15	13	195	20
16	7	112	27
17	3	51	30
<b>Total</b>	30	454	

$$\text{Mean of X} = \bar{x} = \frac{\sum fx}{\sum f}$$

$$= \frac{454}{30} = 15.13 \text{ year}$$

On an average the age of X standard students is 15.13 years.

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

[First we form a cumulative frequency distribution and median is that value which corresponds to the cumulative frequency in which  $\left( \frac{N+1}{2} \right)^{\text{th}}$  observation lies.]

$$\text{Median M} = (30 + 1)/2 = 15.5^{\text{th}} \text{ observation} = 15$$

There are 50% of students whose age is more than 15 years in this group of students.

Mode Z = Mode is calculated by inspecting the given data. We pick out that value which corresponds to the maximum frequency.

Mode Z= 15

In this group mostly students are of 15 years of age.

Tr: All the calculations and interpretation are correct. Now if you have grouped data with continuous frequency distribution then how will you find mean, median and mode values for the data. Let us find mean, median and mode values for the following continuous frequency distribution data:

The following data shows distance covered by 100 persons to perform their routine jobs.

Distance (Km)	0-10	10-20	20-30	30-40
Number of Persons	10	20	40	30

St: Here the variable involved is “distance covered”. While the “number of persons” Represent frequencies.

Distance (Km)	Number of Persons (f)	Mid Points (x)	Fx
0-10	10	5	50
10-20	20	15	300
20-30	40	25	1000
30-40	30	35	1050
Total	100		2400

$$\begin{aligned}\text{Mean } \bar{x} &= \frac{\sum fx}{\sum f} \\ &= \frac{2400}{100} = 24km\end{aligned}$$

Here mean is 24 km which interprets that on an average people travel 24 km to perform their routine jobs.

Tr: Yes, you are absolutely right. It means that 24 km is the average distance travelled by people to perform their routine jobs.

Now for the same data calculate median and mode also.

Median: For this we require to calculate cumulative frequencies of less than type. After than we calculate  $N / 2$ , with its help we determine the class whose cumulative frequency is nearly equal to  $N / 2$ . This class is known as median class. Then, the median is calculated by the following formula:

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

Where  $l$  = lower bound of median class

$C_f$  = cumulative frequency of class prior to median class.

$f$  = frequency of median class.

$h$  = class width.

Mode is calculated by inspecting the given data. We pick out that value which corresponds to the maximum frequency.

St:

Distance (Km)	Number of Persons (f)	Cf
0-10	10	10
10-20	20	30
20-30	40	70
30-40	30	100
Total	100	

$$N/2 = 100/2 = 50$$

$$M = 20 + (50 - 30) * 10/40$$

$$= 25$$

Here median is 25 which mean that 50 % people travel more than 25 km to perform their routine jobs and rest 50 % people are such which travel less than 25 km to perform their routine jobs.

**Mode:** We first find the Modal Class, the class with the highest frequency then mode we calculate from the modal class using the formula given below:

$$\text{Mode} = L + \left( \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$$

L : lower class bound of the modal class

$f_1$ : frequency of the modal class

$f_0$ : frequency of the class before the modal class in the frequency table

$f_2$ : frequency of the class after the modal class in the frequency table

h: class interval of the modal class

Distance (Km)	Number of Persons (f)
0-10	10
10-20	20 $f_0$
20-30	40 $f_1$
30-40	30 $f_2$
Total	100

$$\text{Mode} = 20 + [(40-20) * 10 / (80-20-30)] = 26.666$$

Here the mode value is 26.666 which means that most of the people travel 26.66km to perform their routine jobs.

Tr: Now you all are well acquainted with calculation and interpretation of mean, median and mode. But how will you decide which measure is best among mean, median and mode measures to describe the central value of your data.

As you know mean is very much affected by extreme values as it is based upon all the observations in the data. So if some extreme observations are found in the data which are affecting a lot to the central value of data then mean is avoided and instead of mean median or mode can be calculated. Even when end classes are open-ended mean is avoided to be use. And again median or mode is more suitable measures. It is also advisable not to use mean as a measure of central tendency for qualitative characteristics such as intelligence, honesty, beauty, or loyalty. More over mean cannot be located graphically where as median and mode can be located graphically. But mean is least affected of sampling fluctuations and It is capable of father algebraic treatments it is preferred for measures of central tendency. When quantitative variable is taken under study mean is more preferred than other measures of central tendency.

Let us know the properties of Mean:

Tr: State the Properties of Arithmetic Mean (Or Mean).

**St: 1.** The sum of the deviations of all observations from their arithmetic mean is always zero.

$$\text{i.e. } \sum (x - \bar{x}) = 0$$

**St: 2.** Arithmetic mean is depending on change of origin and scale both.

That is, if a fixed number is subtracted from each observation, their mean is diminished by this number and if each observation is divided by a fixed number, their mean is divided by this number. i.e If  $Y = a + b X$  then  $\bar{Y} = a + b \bar{X}$ .

**St: 3.** If  $\bar{x}_1$  and  $\bar{x}_2$  be arithmetic mean of two groups of observations  $N_1$  and  $N_2$  then the

combined mean of these two groups can be computed by  $\bar{x}_{12} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$

This can also be generalized in the same way for more than two groups of different observations having different arithmetic means.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.



**Task allotment for each group: (All Groups)**

1. The information regarding the no. of children per family is given in the following table. Find the Mean, Median and Mode of the following data. Interpret the values:

No. of children	0	1	2	3	4	5
No. of families	3	20	15	8	3	1

2. The frequency distribution of the marks obtained by 100 students in a test of Statistics subject is given below. Find the Mean, Median and Mode of the following data and interpret the values:

Marks obtained	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49
No. of students	8	15	20	45	12

3. A person had the following monthly bills for electricity. What are the mean and median of the collection of bills?

January \$67.92, February \$59.84, March \$52.00, April \$52.50, May \$57.99, June \$65.35, July \$81.76, August \$74.98, September \$87.82, October \$83.18, November \$65.35, December \$57.00.

4. Find the missing information in the following table:

Group	A	B	C	Combine
Number	10	8	----	24
Mean	20	----	6	15

5. Which average would be more suitable in the following cases?

- a) Average size of ready-made garments.
- b) Average intelligence of students in a class.
- c) Average production per shift in a factory.

d) The distribution has open end classes.

[Ans: 1. Mode, 2. Median, 3. Mean, 4. Median or Mode]

6. A HR team of Civil hospital began a study of the overtime hours of the enrolled nurses. Data was taken from fifteen nurses which were selected at random. Following overtime hours during Jan month were recorded: 13, 13, 12, 15, 7, 15, 5, 12, 6, 7, 12, 10, 9, 5, 9.

(Answer the following questions)

- i. Average overtime hours of all nurses is \_\_\_\_\_.
    - a) 15 hours
    - b) 10 hours
    - c) 12 hours
    - d) None
  - ii. Most of the nurses have done \_\_\_\_\_ hours of overtime.
    - a) 13
    - b) 12
    - c) 7
    - d) None
  - iii. \_\_\_\_\_ Percentage of nurses have done overtime less than or equal to 12 hours.
    - a) 73.33
    - b) 26.67
    - c) 50
    - d) None
  - iv. 50 % of nurses have done \_\_\_\_\_ hours of overtime.
    - a) 9
    - b) 10
    - c) 9.5
    - d) None
7. Median can easily be obtained through \_\_\_\_\_.
  - a) Frequency curve
  - b) Frequency polygon
  - c) Histogram
  - d) Cumulative frequency curves.

### **Presentation and Discussion:**

After group work one student from each group presents their task. Depending upon the presentation the teacher recapitulates and concludes the class. At the end teacher will give assignment related to the present topic to the students. Assignment work related to this topic is mentioned in the appendix- VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Measures of Dispersion.

## Lesson No.4: Measures of Dispersion / Variation

### Teaching Points:

- Absolute measures and relative measures of dispersion
- Range
- Quartile Deviation
- Standard deviation
- Coefficient of Variation

### Instructional Objectives:

After completion of this class students will be able to:

- i. Explain the significance of measures of dispersion.
- ii. Differentiate between absolute and relative measures of dispersion.
- iii. Define Range, Quartile Deviation and Standard deviation.
- iv. Calculate Range, Quartile Deviation and Standard deviation for grouped and ungrouped data.
- v. Calculate the coefficient of variation for a given data.
- vi. Interpret the value of coefficient of variation for a given data.

### Lesson Presentation:

Tr: Why do we need to study measures of dispersion? Are measures of central tendency alone are not enough to describe the nature of the distribution of data?

St: I think.... measures of central tendency alone are not enough to describe the nature of the distribution of data. I read that measures of dispersion determine reliability of an Average. Is it so?

Tr: Yes! You are right. Let us take an example to understand it better.

Following table shows the cricket scores of two batsmen. Compare the performance of the two:

Match No.	Batsman A	Batsman B
1	70	20
2	50	35
3	60	28

4	65	194
5	72	40
<b>Total</b>	<b>317</b>	<b>317</b>
<b>Average</b>	<b><math>317/5 = 63.4</math></b>	<b><math>317/5 = 63.4</math></b>

Looking upon the above situation both the batsmen have scored same total runs hence both the averages are same. Now can you tell who has played well?

St: Definitely A has played well.

Tr: Why?

St: Because in every match player A has scored good score where as player B has scored low score in almost all the matches except one.

Tr: Your observation is quite valid. Can we say that player A is more consistent in playing than player B.

St: Yes.

Tr: The spread of runs in first series is in between 72 to 50 whereas in series B the spread is in between 194 to 20. So the consistency of series first is more or we can say that the variations of runs in series first (batsman A) is less compared to variations of series (batsman B) second. So we can say that when averages are close enough then for comparing the series of data it is better to study the variation of the data also. Now what are different measures to study variation in data?

St: Range

St: Mean deviation

St: Quartile Deviation

St: Standard deviation

Tr: Else?

St: (may be no response)

Tr: The above mentioned measures of variation are true. But all these measures are known as absolute measures of variations. In case if you want to compare two series but of different in measurement units like comparing butter in gm and milk in liters, wheat in quintals and spices in kg etc. now can you use these measures of variation for comparison?

St: (may be or may not be any response)

Tr: For such cases we use relative measures of dispersions. These measures are free from their units in which the original data is measured. If the original data is in dollar or kilometers, we do not use these units with relative measures of dispersion. These measures are a sort of ratio and are called coefficients. Each absolute measure of dispersion can be converted into its relative measure. Thus the relative measures of dispersion are:

- i. Coefficient of Range or Coefficient of Dispersion.
- ii. Coefficient of Quartile Deviation or Quartile Coefficient of Dispersion.
- iii. Coefficient of Mean Deviation or Mean Deviation of Dispersion.
- iv. Coefficient of Standard Deviation or Standard Coefficient of Dispersion.

Coefficient of Variation (a special case of Standard Coefficient of Dispersion)

Tr: In this class we will learn about Range, Quartile Deviation, Standard deviation and Coefficient of Variation. For that let us find Range, Quartile Deviation, Standard deviation and Coefficient of Variation for the following raw data (ungrouped data):

The wheat production (in Kg) of 20 acres is given below.

1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750 and 1885.

Range = Highest value – lowest value

St: Here range =  $1960 - 1040 = 920$  kg

The wheat production of 20 acres of land varies in the range of 920 kg.

The arranged observations are given below:

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

Tr:

The quartiles are given by:

$$Q_1 = \text{value of } \left( \frac{N+1}{4} \right)^{th} \text{ observation}$$

$$Q_2 = \text{value of } 2 \left( \frac{N+1}{4} \right)^{th} \text{ observation}$$

$$Q_3 = \text{value of } 3 \left( \frac{N+1}{4} \right)^{th} \text{ observation} \quad N: \text{no. of observations of random variable } X$$

$$\text{Quartile Deviation Q.D.} = (Q_3 - Q_1) / 2$$

$$\text{St: } Q_1 = \text{value of } \left( \frac{20+1}{4} \right)^{th} \text{ observation}$$

$$= 5.25^{th} \text{ observation in the arranged data.}$$

$$= 5^{th} \text{ observation} + 0.25(6^{th} \text{ observation} - 5^{th} \text{ observation})$$

$$= 1240 + 0.25(1320 - 1240)$$

$$= 1260 \text{ kg}$$

$$Q_3 = \text{value of } 3 \left( \frac{20+1}{4} \right)^{th} \text{ observation}$$

$$= 15.75^{th} \text{ observation in the arranged data.}$$

$$= 15^{th} \text{ observation} + 0.75(16^{th} \text{ observation} - 15^{th} \text{ observation})$$

$$= 1750 + 0.75(1775 - 1750)$$

$$= 1753.75 \text{ kg}$$

Tr: How do you calculate Quartile Deviation?

St: Quartile Deviation Q. D. =  $(Q_3 - Q_1) / 2 = (1753.75 - 1260) / 2 = 246.875$

The quartile deviation is 246.875.

Tr: How do you define Standard deviation?

### Standard Deviation (s.d.)

The standard deviation is defined as **the positive square root of the mean of the square deviations taken from arithmetic mean of the data**. It is denoted by  $\sigma$  (sigma) and is defined as:

$$\text{Thus, } \sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{N}} \quad \text{or} \quad \sigma = \sqrt{\frac{1}{N} \sum X^2 - \bar{X}^2} \quad \text{for ungrouped data}$$

St: Calculation of s.d.

X	$X^2$
1120	1254400
1240	1537600
1320	1742400
1040	1081600
1080	1166400
1200	1440000
1440	2073600
1360	1849600
1680	2822400
1730	2992900

1785	3186225
1342	1800964
1960	3841600
1880	3534400
1755	3080025
1720	2958400
1600	2560000
1470	2160900
1750	3062500
1885	3553225
<b>1517.9</b>	<b>47699139</b>

$$\text{s.d.} = \sigma = \sqrt{\frac{47699139}{20} - 1517.9^2}$$

$$= 284.76$$

The s.d. of wheat production is 284.76 kg.

Tr: What is coefficient of variation?

St:

### Co-efficient of Variation (C. V.)

To compare the variations (dispersion) of two different series, relative measures of standard deviation must be calculated. This is known as co-efficient of variation or the co-efficient of s. d. Its formula is:

$$\text{C. V.} = \frac{\sigma}{x} \times 100$$

Coefficient of variation C.V. =  $(284.76/1517.9) \times 100 = 18.76 \%$

The coefficient of variation for wheat production is 18.76%

Tr: Say for another 20 acres of land the coefficient of variation for rice production is 20.6%.

Therefore, it means that wheat production is comparatively more uniform as compare to rice production. Since  $CV_{\text{wheat}} < CV_{\text{rice}}$ .



Tr: Now Calculate Range, Quartile Deviation, Standard Deviation and Coefficient of Variation for the following (continuous data) grouped data.

<b>Hb level</b>	<b>Number of students in a X Standard</b>
9.3 - 9.7	2
9.8 – 10.2	5
10.3 – 10.7	12
10.3 – 11.2	17
11.3 – 11.7	14
11.8 – 12.2	6
12.3 – 12.7	3
12.8 – 13.2	1

St:

The range is  $13.2 - 9.3 = 3.9$ .

For calculating Quartile Deviation:

<b>Maximum Load (hemoglobin level)</b>	<b>Number of students in a X Standard</b>	<b>Class Boundaries</b>	<b>Cumulative Frequencies</b>
9.3 - 9.7	2	9.25-9.75	2
9.8 – 10.2	5	9.75-10.25	7
10.3 – 10.7	12	10.25-10.75	19
10.3 – 11.2	17	10.75-11.25	36
11.3 – 11.7	14	11.25-11.75	50
11.8 – 12.2	6	11.75-12.25	56
12.3 – 12.7	3	12.25-12.75	59
12.8 – 13.2	1	12.75-13.25	60

### Quartile Deviation:

If  $L_1 - U_1, L_2 - U_2, \dots, L_n - U_n$  are  $n$  exhaustive and exclusive class of random variable  $X$  with frequency  $f_1, f_2, \dots, f_n$  respectively, then first find C.F less than type and find  $i^{\text{th}}$  quartile class i.e. class having C.F just greater than  $[i(N/4)]$ .

Then  $i^{\text{th}}$  Quartile is given by, 
$$Q_i = L + \frac{i\left(\frac{N}{4}\right) - C_f}{f} \times h \quad \text{where, } i = 1, 2, 3$$

Where,  $L$  : Lower boundary of  $i^{\text{th}}$  quartile class

$C_f$  : C.F of class above  $i^{\text{th}}$  quartile class

$f$  : Frequency of  $i^{\text{th}}$  quartile class

$h$  : Class width of  $i^{\text{th}}$  quartile class

**Quartile Deviation: Q.D. = (Q3- Q1)/2**

$$Q_1 = \text{value of } \left(\frac{N}{4}\right)^{\text{th}} \text{ item} = \text{value of } \left(\frac{60}{4}\right)^{\text{th}} \text{ item} = 15^{\text{th}} \text{ item}$$

Therefore,  $Q_1$  lies in the class 10.25 – 10.75

Now 
$$Q_1 = L + \frac{\frac{N}{4} - C_f}{f} \times h$$

Where,  $L=10.25$ ,  $h=0.5$ ,  $f = 12$ ,  $\frac{N}{4} = 15$  and  $C_f = 7$

$$Q_1 = 10.25 + \left(\frac{15-7}{12}\right) \times 0.5 = 10.25 + 0.33 = 10.58$$

$$Q_3 = \text{value of } \left(\frac{3N}{4}\right)^{\text{th}} \text{ item} = \text{value of } \left(\frac{3 \times 60}{4}\right)^{\text{th}} \text{ item} = 45^{\text{th}} \text{ item}$$

Therefore,  $Q_3$  lies in the class 11.25 - 11.75

$$\text{Now } Q_3 = L + \frac{3\left(\frac{N}{4}\right) - C_f}{f} \times h$$

Where,  $L=11.25$ ,  $h=0.5$ ,  $f = 14$ ,  $\frac{3N}{4} = 45$  and  $C_f = 36$

$$Q_3 = 11.25 + \left(\frac{45-36}{14}\right) \times 0.5 = 11.25 + 0.32 = 11.57$$

**Quartile Deviation: Q.D. =  $(11.57 - 10.58)/2 = 0.495$**

**Standard Deviation:**

$$\sigma_x = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}} \quad \text{or} \quad \sigma = \sqrt{\frac{1}{N} \sum fX^2 - \bar{X}^2} \quad \text{for grouped data}$$

Hb level	Number of students in a X Standard (f)	Mid value (X)	fX	fX <sup>2</sup>
9.3 - 9.7	2	9.5	19	180.5
9.8 – 10.2	5	10	50	500
10.3 – 10.7	12	10.5	126	1323
10.3 – 11.2	17	11	187	2057
11.3 – 11.7	14	11.5	161	1851.5
11.8 – 12.2	6	12	72	864
12.3 – 12.7	3	12.5	37.5	468.75
12.8 – 13.2	1	13	13	169
<b>Total</b>	<b>60</b>		<b>665.5</b>	<b>7413.75</b>

$$\text{s.d.} = \sigma_x = \sqrt{[7413.75 / 60 - (665.5 / 60)^2]} = 0.7331$$

$$\text{Mean } \overline{X} = 665.5/60 = 11.0917$$

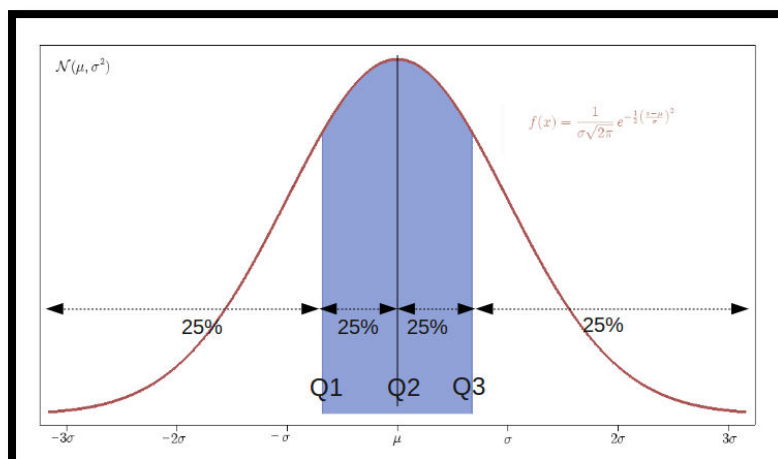
$$\text{Coefficient of Variation} = (0.7331 / 11.0917) * 100 = 6.609\%$$

Tr: Here you have calculated various values which shows the amount of variation in the data. We get range = 3.9, Q.D. = 0.495, s.d. = 0.7331, C.V.= 6.609%. How will you interpret all these values for the given problem?

The range is interpreted as the overall dispersion of values in a data set or, more literally, as the difference between the largest and the smallest value in a dataset. The range is measured in the same units as the variable of reference and, thus, has a direct interpretation as such. The range can only tell you basic details about the spread of a set of data. It gives a rough idea of how widely spread out of the most extreme observations are, but gives no information as to where any of the other data points lie.

St: Here range= 3.9 indicates that the hemoglobin level of students varies in the amount of 3.9 units.

Tr: The Quartile Deviation is a simple way to estimate the spread of a distribution about a measure of its central tendency (usually the mean). So, it gives you an idea about the range within which the central 50% of your sample data lies. When data is open ended quartile deviation is the only measure used to study the dispersion of data. For a better understanding, look at the representation given below:



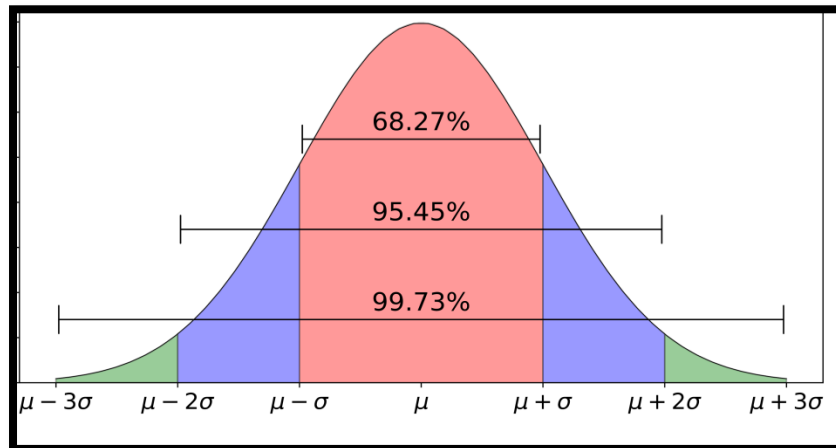
$$Q.D. = \frac{Q_3 - Q_1}{2}$$

St: Here Q.D. = 0.495, it means 25% of observations of the data lies in the range of 0.495 units.

Or

St: We can say that 50% of student's hemoglobin level lies in the range of 0.99 units.

Tr: Good. Now let us understand about Standard deviation. A small standard deviation means that the values in a statistical data set are close to the mean of the data set, on an average, and a large standard deviation means that the values in the data set are farther away from the mean, on an average. Let us take an example to make it more understandable. An IQ test score is calculated based on a norm group with an average score of 100 and a standard deviation of 15. The standard deviation is a measure of spread, in this case of IQ scores. A standard deviation of 15 means 68% of the norm group has scored between 85 (100 – 15) and 115 (100 + 15).



Specifically, if a set of data is normally distributed about its mean, then about 2/3 of the data values will lie within 1 standard deviation of the mean value, and about 95/100 of the data values will lie within 2 standard deviations of the mean value.

St: Here s.d. is 0.7331 which is not a big value which means that the observations of data are not much spread from the average value of that data.

St: It means that 68% of students have hemoglobin level between 10.3586 to 11.8248.

[i.e. (μ - σ) and (μ + σ) = (11.0917 - 0.7331) and (11.0917 + 0.7331)]

Tr: Yes both of you are alright. Now express the meaning of C.V.= 6.609%

The coefficient of variation shows the extent of variability of data in a sample in relation to the mean of the population. The coefficient of variation (CV) is the ratio of the standard deviation to the mean. When higher the coefficient of variation, greater is the level of dispersion around the mean. It is generally expressed as a percentage. The lower the value of the coefficient of variation, the more precise is the estimate. C.V. is generally used to compare two data series with different units of measurements.

St: Here the C.V. is 6.609% which is not a very big in terms of percentage. This means that the observations are not much dispersed from mean.

Tr: Variance is also used as a measure of studying data variation.

**Variance:** The term variance was used to describe the square of the standard deviation. The concept of variance is of great importance in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variations in their original series. Variance is defined as follows:

$$\text{Variance} = \frac{\sum (x - \bar{x})^2}{N}$$

**Note :** if Variance  $V(x) = \frac{\sum (x - \bar{x})^2}{N}$

Then s. d.  $(\sigma_x) = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$

Thus  $V(x) = \sigma_x^2$

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group:**

**Group No. 1, 3, 5**

1. The following data represents the average prices of gold (in dollars per fine ounce) for the years 1981 to 2000. Find out the range, quartile deviation, s.d . and coefficient of variation . Interpret the results.

460, 376, 424, 361, 318, 368, 478, 438, 383, 385, 363, 345, 361, 385, 386, 389, 332, 295, 280, 280

**Group No. 2, 4, 6**

2. The following table gives the distribution of wages in the two branches of a factory:

Monthly wages (Rs)	Number of workers	
	Branch A	Branch B
100-150	167	63
150-200	207	93
200-250	253	157
250-300	205	105
300-350	168	82

Find mean and standard deviation for the two branches for the wages separately.

(a) Which branch pays higher average wages?

(b) Which branch has greater variability in wages in relation to the average wages?

(c) What is the average monthly wage of the factory as a whole?

3. Multiple choice questions. **(All Groups)**

i. If the mean and standard deviation of series A and B are as:  $\bar{X}_A = 15.0$ ,  $\bar{X}_B = 20.0$ ,  $\sigma_A^2 = 25$ ,  $\sigma_B^2 = 16$  which of the series is more consistent?

- (a) Series A
- (b) Series B
- (c) Series A and series B are equally consistent
- (d) None of the above

ii. If Variance of distribution is 16 then standard deviation will be\_\_\_\_\_.

- (a) 4
- (b) 16
- (c) 8
- (d) 2

4. Differentiate between standard deviation and coefficient of variation. (**All Groups**)

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment related to the present topic to the students. Assignment work related to this topic is mentioned in the appendix no. VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Skewness and their types.



## Lesson No.5: Skewness and its Types

### Teaching Points:

- Concept of Skewness
- Types of skewness: positive and negative
- Measures of coefficient of skewness

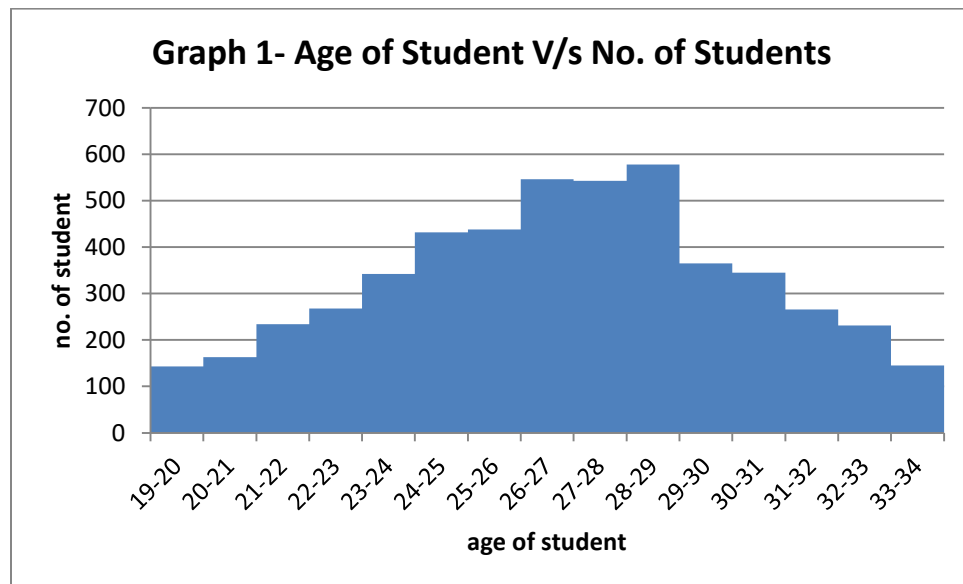
### Instructional Objectives:

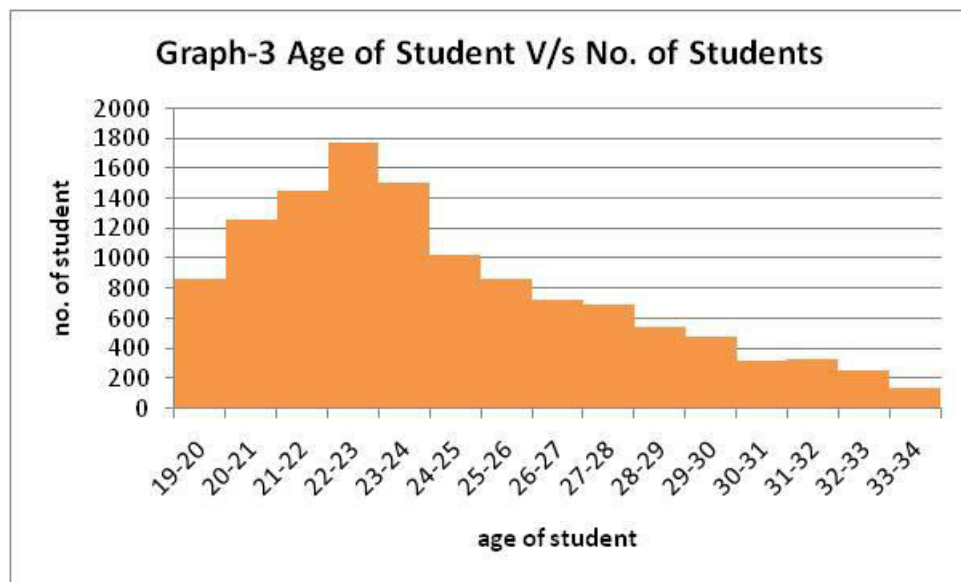
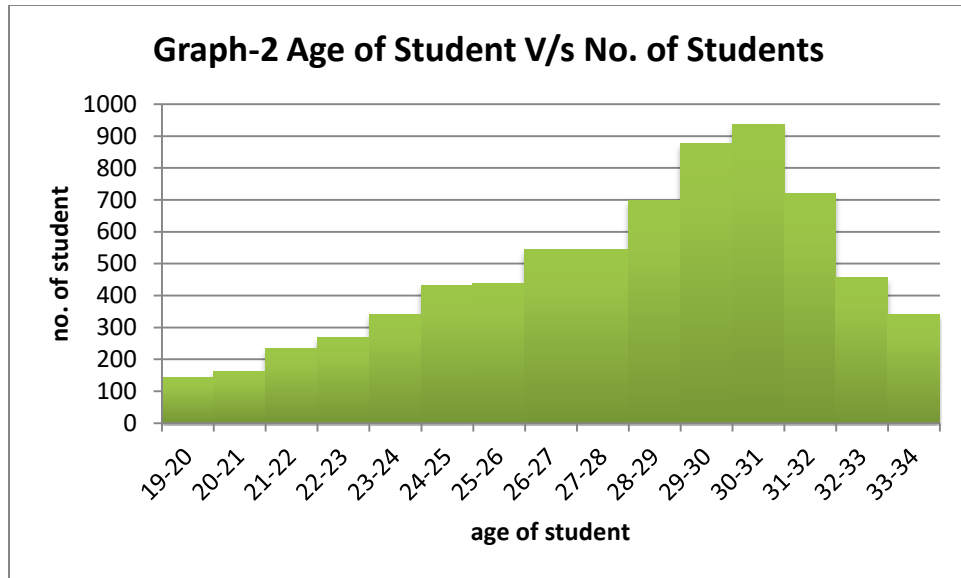
After completion of this class students will be able to:

- Define skewness.
- Explain the types of skewness.
- List the measures of coefficient of skewness.
- Calculate coefficient of skewness for a given data.
- Interpret the value of coefficient of skewness.

### Lesson Presentation:

Tr: When you are drawing Frequency Curve or histogram, the distribution of frequencies over the X- axis is sometimes symmetrically distributed and sometimes may not be symmetrically distributed. See the following graphs and identify which are symmetric and which are not symmetrically distributed.





St: Graph 2 and 3 are not symmetrically distributed.

Tr: and what about graph 1?

St: To some extent it looks like symmetrically distributed.

Tr: You are right. Can you see that in graph 2 and 3 both the tails of graph are of different lengths?

St: Yes! In graph 2 and 3 tails of curves are of different length.

Tr: Do you know the name some characteristic of frequency distribution of curve where both tails of curve are not equally elongated towards the x- axis?

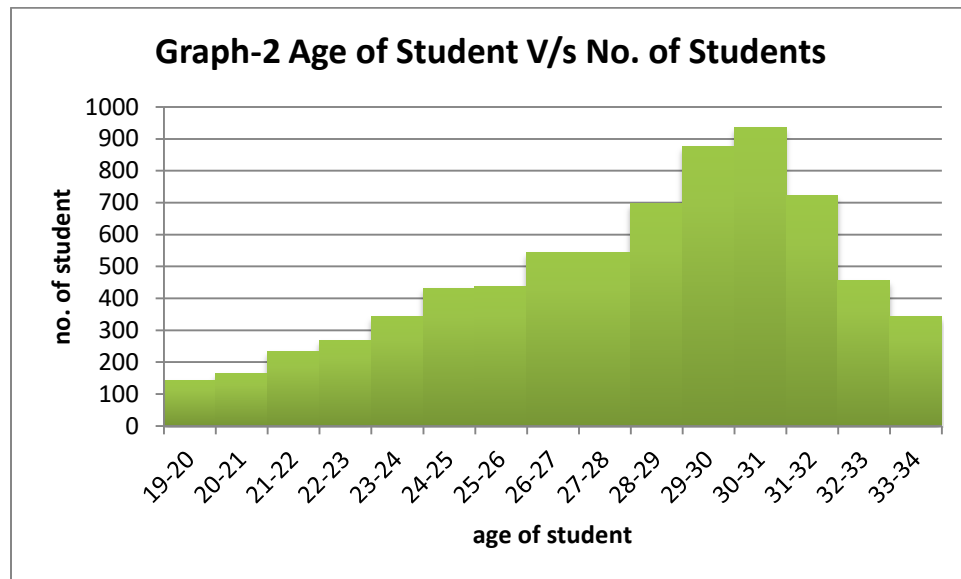
St: Yes! We call such frequency curves as skewed curves.

Tr: Very true! They are called skewed distributions. So can you define skewness.

St: Lack of Symmetry in the frequency distribution of curve is called skewness.

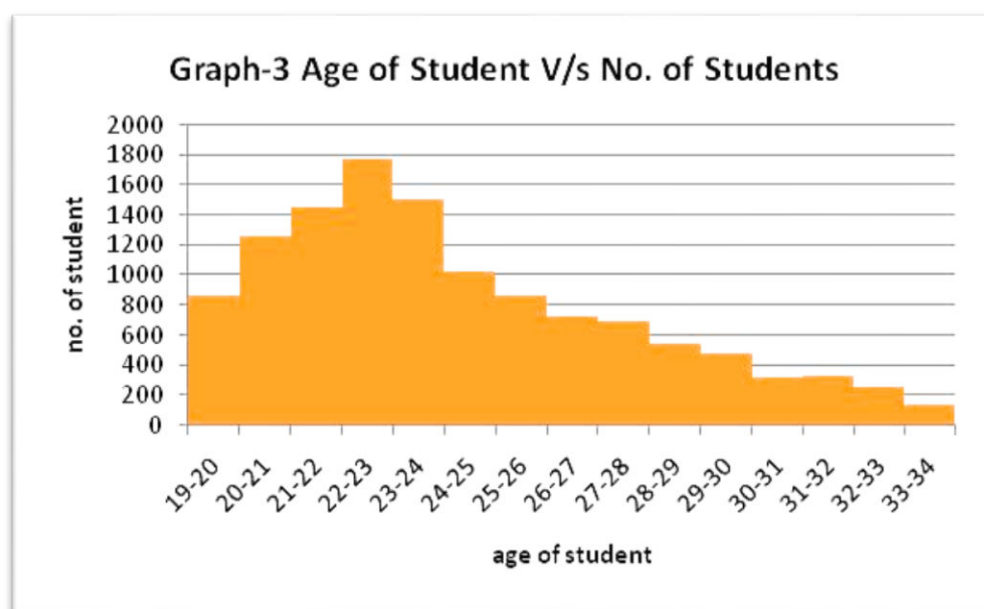
Tr: Good. Let us read some more properties of skewness.

### Negatively Skewed Distribution:



Here you can see that in graph 2 left tail is more longer than the right tail of the frequency distribution curve. Such curve is known as negatively skewed distribution. In negatively skewed distribution most of the observations attain values towards high values of X- axis (i.e. towards  $+\infty$ ). We can also say that high frequencies are towards right side of the curve.



### Positively skewed Distribution:



Here you can see that in graph 3 right tail is more longer than the left tail of the frequency distribution curve. Such curve is known as positively skewed distribution. In positively skewed distribution most of the observations attain values towards lower values of X- axis (i.e. towards  $-\infty$ ). We can also say that high frequencies are towards left side of the curve.

Tr: Now give some more properties for positively and negatively skewed distributions.

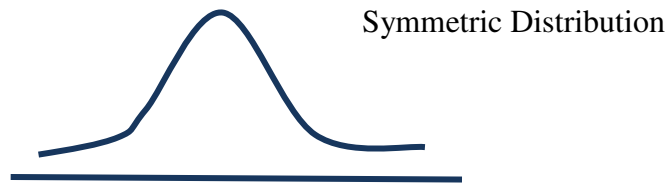
St:

Sl. No.	Positively Skewed Distribution	Negatively Skewed Distribution
1.		
2.	For a positively skewed distribution curve rises rapidly, reaches the maximum and falls slowly. In other words, if the frequency curve has longer tail to right the distribution is known as positively skewed distribution.	A negatively skewed distribution curve rises slowly reaches its maximum and falls rapidly. In other words, if the frequency curve has longer tail to left the distribution is known as negatively skewed distribution.
3.	$Mean > Median > Mode$	$Mean < Median < Mode$
4.	$Q_3 - Q_2 > Q_2 - Q_1$	$Q_3 - Q_2 < Q_2 - Q_1$
5.	Coefficient of skewness $\beta_1 > 0$	Coefficient of skewness $\beta_1 < 0$

Tr: well you have stated all the properties of positively and negatively skewness of distribution correctly. On the similar lines could you state the properties of symmetric distribution?

St: Yes.

1. Shape of the distribution:



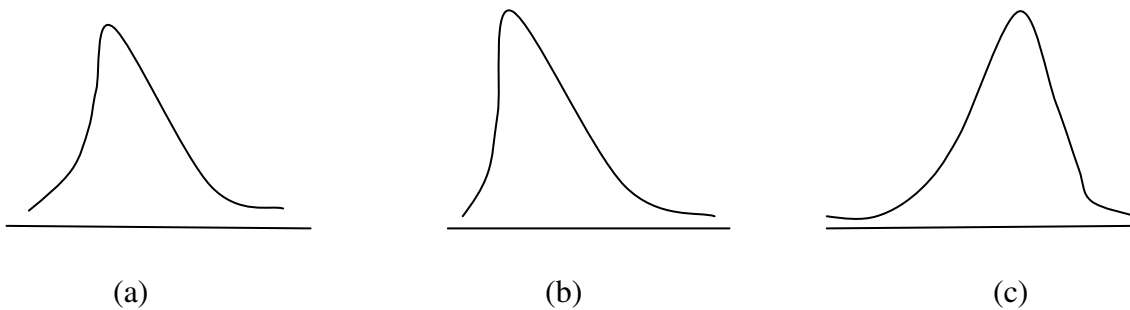
2. For a positively skewed distribution curve rises rapidly, reaches the maximum and falls slowly. In other words, if the frequency curve has longer tail to right the distribution is known as positively skewed distribution.

3.  $Mean = Median = Mode$

4.  $Q_3 - Q_2 = Q_2 - Q_1$

5. Coefficient of skewness  $\beta_1 = 0$ .

Tr: Great. Now answer the following which curve is more skewed:



St: It's difficult to say that which one is more skewed.

St: Looking upon this Figure (a) and (b) are positively skewed and Figure (c) is negatively skewed. But probably (b) is more skewed.

Tr: I think you find it difficult to say about it because the amount of skewness is very close. But yes you are true direction i.e. positive and negative can be identified by looking upon these figures. Therefore we need to have some measures to study the degree of skewness present in the data. For that we have different measure of coefficient of skewness. Do you know how coefficient of skewness can be calculated?

St: **i. Karl Pearson coefficient of Skewness- I**

$$SK = \frac{Mean - Mode}{S.D}$$

Sometimes the mode is difficult to find. So we use another formula

## ii. Karl Pearson coefficient of Skewness-II

$$SK = \frac{3(\text{Mean} - \text{Median})}{S.D}$$

Tr: If  $SK = 0$ , then there is no skewness

If  $SK > 0$ , the skewness is positive.

If  $SK < 0$ , the skewness is negative.

The sign '+' or '-' of the coefficient of skewness would indicate the direction of skewness and its numerical value would give the extent of skewness.

**Note :** Although the co-efficient of skewness is always within  $\pm 1$ , but the Karl Pearson's co-efficient lies within  $\pm 3$ .

## St: iii. Measure of Skewness given by Bowley:

$$SK = \frac{Q_1 + Q_3 - 2\text{Median}}{Q_3 - Q_1}$$

Tr: This measure is based on quartiles. For a symmetrical distribution, it is seen that  $Q_1$  and  $Q_3$  are equidistant from median.

Thus, an absolute measure of skewness  $= (Q_3 - Md) - (Md - Q_1)$

A relative measure of skewness, known as Bowley's coefficient ( $Sk$ ) is given above.

If  $SK = 0$ , then there is no skewness

If  $SK > 0$ , the skewness is positive.

If  $SK < 0$ , the skewness is negative.

**Note:** Limits for Bowley's Coefficient of Skewness is  $-1 \leq Sk \leq +1$

### St: iii. Measure of Skewness based on Moments

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

Tr: The coefficient  $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$  is the relative measure of Skewness and  $\beta_1$  is always positive

so, type of skewness can achieve based on sign of third central moment ( $\mu_3$ ).

If  $\mu_3$  is positive ( $\mu_3 > 0$ ) then, the distribution is positively skewed,

if  $\mu_3$  is negative ( $\mu_3 < 0$ ) then the distribution is negatively skewed.

**The measure of skewness is also sometimes represented by  $\gamma_1$ :**

$$\begin{aligned} \text{Gamma coefficient is } \gamma_1 &= \pm \sqrt{\beta_1} = \pm \frac{\mu_3}{\sqrt{\mu_2^3}} \\ &= \pm \frac{\mu_3}{\sigma^3} \end{aligned}$$

If  $\gamma_1 < 0$  skewness is negative

If  $\gamma_1 > 0$  Skewness is positive

If  $\gamma_1 = 0$  distribution is symmetric (no skewness)

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group:**

**Group No. 1, 3, 5.**

1. Calculate coefficient of skewness and interpret the results.

Marks	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
Frequency	8	28	35	17	12

**Group No. 2, 4, 6.**

2. Find coefficient of skewness from the following data and show which section is more skewed.

Income(Rs.)	5-58	8-61	61-64	64-67	67-70
Section A	12	17	23	18	11
Section B	20	22	25	13	4

### Multiple Choice Questions. (All Groups)

- The Coefficient of Skewness of a series A is -0.15 and that of series B is 0.062 which of the two series is less skewed?
  - Series A
  - Series B
  - No decision
  - None of the above
- Distribution having moment coefficient of skewness 1.45 and third central moment -34 is
  - Positively skewed
  - Normal
  - Negatively skewed
  - Symmetric
- The Coefficient of Skewness of a series A is 0.45 and that of series B is - 0.562 which of the two series is more skewed?
  - Series A
  - Series B
  - No decision
  - None of the above

### Presentation and Discussion:

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Kurtosis and their types.



## Lesson No.6: Kurtosis and its types

### Teaching Points:

- Concept of Kurtosis
- Types of Kurtosis: Platy kurtic, Messo kurtic and Lepto kurtic
- Coefficient of Kurtosis

### Instructional Objectives:

After completion of this class students will be able to:

- i. Define Kurtosis.
- ii. Explain the types of Kurtosis.
- iii. Calculate coefficient of Kurtosis for a given data.
- iv. Interpret the value of coefficient of Kurtosis.

### Lesson Presentation:

Tr: When you draw frequency curve or histogram you may come across with different shapes of distributions with different heights of their Peakness. Based upon the peakness of the distribution we can categories various distribution of data. Do you know the property associated with peakness of distribution of curve?

St: Yes. Kurtosis based upon peakness of the distribution of curve.

Tr: Very good. You said well.

Could you try to define kurtosis?

St: The degree of Peakness of distribution of data is known as kurtosis.

Tr: Do you know its types?

St: They are of three types.

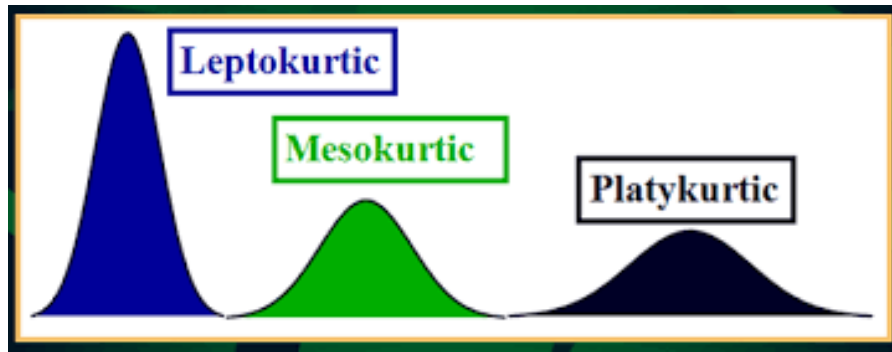
St: Lepto Kurtic.

St: Messo Kurtic.

St: Platy Kurtic.

Tr: How will you differentiate them?

St: Depends upon their shapes we can differentiate them. Following picture shows the relative peakness of the distribution of data:



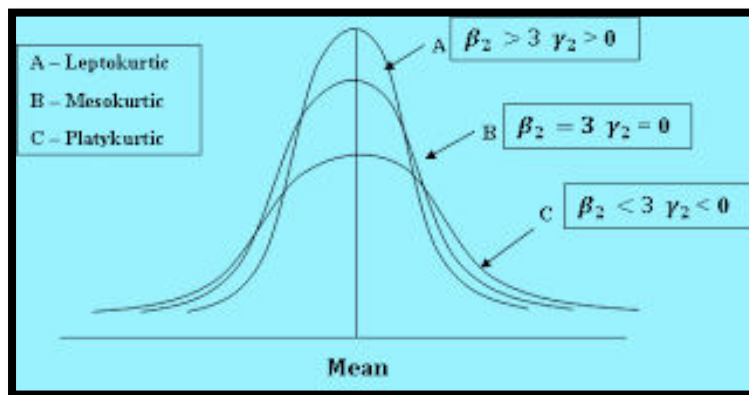
$\beta_2$  gives the measure of Kurtosis or flatness of the mode.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

If  $\beta_2 = 3$  then the curve is normal which is neither flat nor peaked i.e. Meso Kurtic.

If  $\beta_2 > 3$  then the curve is more peaked than a normal curve and is called Lepto Kurtic.

If  $\beta_2 < 3$  then curve is flatter than a normal curve and is called Platy Kurtic.



**The measure of kurtosis is also sometimes represented by  $\gamma_2$ :**

Gamma Coefficient,  $\gamma_2 = \beta_2 - 3$

$\gamma_2 = 0$  curve is Mesokurtic

$\gamma_2 > 0$  curve is Leptokurtic

$\gamma_2 < 0$  curve is Platykurtic

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group: (All Groups)**

1. If  $\beta_1 = +1$  and  $\beta_2 = 4$  and variance = 9. Comment upon nature of distribution.
2. Fill in the blanks:
  - i. If for a distribution coefficient of kurtosis  $\gamma_2 < 0$  the frequency curve is \_\_\_\_\_.
  - ii. If  $\mu_4$  is 200 and s.d. 4.5 the coefficient of Kurtosis is \_\_\_\_\_.
  - iii. If  $\beta_2 = 3$  then distribution is \_\_\_\_\_.
  - iv. For Normal distribution  $\beta_1 =$  \_\_\_\_\_ and  $\beta_2 =$  \_\_\_\_\_.
  - v. If  $\gamma_2$  is 2.3 it means that distribution is \_\_\_\_\_.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment related to the present topic to the students. Assignment work related to this topic is mentioned in the appendix no. VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Correlation, types of correlation, Scatter Diagram and Pearson's Correlation Coefficient.

## Lesson No.7: Correlation

### Karl Pearson Product Moment Correlation

#### Teaching Points:

- Concept of Correlation
- Types of correlation based on: Sign, Magnitude, No. of variables under the study, nature of variables and pattern of plotted points.
- Karl Pearson Correlation coefficient

#### Instructional Objectives:

After completion of this class students will be able to:

- i. Define correlation.
- ii. State the properties of correlation coefficient.
- iii. Draw scatter plot for depicting nature of correlation.
- iv. Explain the types of correlation based on Sign, Magnitude, No. of variables under the study, nature of variables and pattern of plotted points.
- v. Compute the value of Karl Pearson Correlation coefficient.
- vi. Interpret the value of correlation coefficient in context to the given problem.
- vii. Coefficient of determination.

#### Lesson Presentation:

Tr: Till now we described characteristics of Univariate data by means of measures of central tendency, measures of dispersion, skewness and kurtosis. But we may come across with a problem where more than one variable are coming into the picture. Such data where two related series of data are considered for some study is known as bi-variate data. If more than two variables are considered simultaneously in the study we call them as multi-variate data. Here in this class we will learn data analysis technique for bi- variate data.

Sometimes it appears that the values of the various variables, so obtained are interrelated. It is likely that such relationship may be obtained in two series relating to the heights and weights of a group of persons. It may be observed that weight increases with increase in height. So that tall people are heavier than short sized people. Similarly, if the data are collected about the prices of a commodity and quantities sold at different prices, two series would be obtained. In two such series we are again likely to find some relationship. With increases in the price of the commodity

the quantity sold is bound to decrease. We can thus conclude that there is some relationship between price and demand. Such relationship can be found in many types of series.

**The term correlation (or co-variation) indicates the relationship between two such variables in which with changes in the values of one variable, the values of the other variable also changes.**

Thus correlation is statistical tool of studying the relationship between two variables. For correlation it is essential that the two phenomena should have cause-effect relationship. If such relationship does not exist then one should not talk of correlation.

Do you know the types of Correlation?

St: Yes. Positive correlation and negative correlation

St: Strong correlation and weak correlation

St: No correlation, linear correlation, non- linear correlation.

St: Karl Pearson's correlation.

St: Rank correlation, etc.

Tr: Very nice. Now explain about them one by one.

St: Positive Correlation: While studying the relationships of any two related variables, if we find the deviation of the value of variables are in the same direction i.e. if one variable increases (or decreases), the corresponding value of the second variable also increases (or decreases), then it is called a Positive Correlation. For e.g. Height and weight of human beings, demand and supply, amount of rain fall and yield of crop have positive correlation.

St: Negative Correlation: While studying the relationships of any two related variables, if we find the deviation of the value of variables in the opposite direction i.e. if one variable increases (or decreases), the corresponding value of the second variable decreases (or increases), then it is called a Negative Correlation. For e.g. price and demand of commodity, temperatures and sales of woolen clothes have negative correlation.

St: Linear and Non-Linear Correlation: When 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 called a Linear Correlation. In such a case if the values of the variables are plotted on a graph paper, then a straight line is obtained. But when the amount of change in one variable does not tends to bear a constant ratio to the amount of change in the other variable, and then the correlation is called a Non-Linear Correlation or Curvilinear. In such situation if the values of the variables are plotted

on a graph paper, then a curve is obtained. Correlation can either be simple correlation or it can be partial correlation or it can be multiple correlation.

St: Simple Correlation: When we study the relationship between only two variables then it is called simple correlation. e.g. Let two variables be, volume of sale and price of item then correlation between them is simple.

St: Partial Correlation and Multiple Correlation: When more than two variables are involved in a study relating to correlation then it can either be multiple correlation or partial correlation. Partial correlation may be defined as the correlation between one dependent variable with one independent variable by keeping the effect of other independent variables constant. e.g Let three variables are, volume of sale, expenditure on advertisement and price of item then correlation between Volume of sale and advertisement expense, by keeping the effect of price of item constant is called partial correlation. Multiple correlation may be defined as correlation between one dependent variable with all other independent variables. e.g multiple correlation is the study of joint effect of price and advertisement expenditure on volume of sale.

Tr: We can classify Correlation based on:











- i. By direction of Change or Sign of correlation coefficient :Positive and Negative Correlation
- ii. Pattern of plotted points Linear and Non-Linear Correlation
- iii. By number of variables under study: Simple, Partial and Multiple Correlation

We can also classify the correlation based on Magnitude of correlation coefficient and nature of variable (or scale of measurement of variable).

Based on magnitude we can classify them as strong, weak and moderate correlation. The range of correlation lies between  $-1$  to  $+1$ . Therefore we can classify them on magnitude basis as follow:

Sl. No.	Value of correlation coefficient	Type of Correlation
1	1	Perfect correlation
2	0.999 to 0.6	High correlation
3	0.599 to 0.3	Moderate correlation
4	0.299 to 0.1	Low correlation
5	0 .0999 to 0.00	Lack of correlation

When value of correlation coefficient is positive or negative we associate the type of correlation as mentioned above with positive or negative prefix correspondingly for interpretation purpose. Based upon the nature of variable under the study or the scale of measurement we select the correlation type. Following table helps you to decide which correlation technique should be used in studying the correlation between the variables:

<b>Variables Y/ X</b>	<b>Nominal Variable</b>	<b>Ordinal Variable</b>	<b>Interval or Ratio Variable</b>
<b>Nominal Variable</b>	 Phi Correlation Coefficient (both are naturally dichotomous nominal variables)  Tetrachoric correlation coefficient (both are artificially dichotomous nominal variables)	-	-
<b>Ordinal Variable</b>	 Gamma Correlation Coefficient	 Spearman rank – order correlation coefficient  Kendall's tau coefficient	-
<b>Interval or Ratio Variable</b>	 Biserial correlation (X is artificially dichotomous nominal variable)  Ponit Biserial correlation (X is naturally dichotomous nominal variable)	 Spearman rank – order correlation coefficient  Kendall's tau coefficient	 Karl Pearson Product moment correlation coefficient .

Tr: Do you know Scatter diagram? How scatter plot is useful to study Correlation?

Let us learn about Scatter diagram.

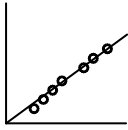
St: Some students may be nodding their heads. They may reply. If not teacher will help them.

**Scatter diagram** is simplest method to study the correlation between two variables. Take value of one variable on x-axis and another variable on y-axis and the values of each pair we plot on the graph paper and the diagram so obtained are called scatter diagram or dot diagram.

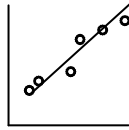
- If plotted dots lie on the straight line rising from the lower left-hand corner to the upper right hand corner then the correlation is said to be perfect positive correlation.
- If plotted dots lies on the straight line from the upper left hand corner to the lower right hand corner then correlation is said to be perfect negative correlation.
- If plotted dots fall in a narrow band showing a rising tendency from the lower left hand corner to the upper right hand corner, then correlation is high degree positive correlation. As the band becomes wider the degree of correlation becomes low and we called low degree positive correlation.
- If plotted dots fall in a narrow band showing a decreasing tendency from the upper left hand corner to the lower right hand corner, then correlation is high degree negative correlation. As the band becomes wider the degree of correlation becomes low and we called low degree negative correlation.
- If the dots are widely scattered in haphazard manner, it indicates no correlation between two study variables.



## Scatter Plots



**Positive  
Perfect  
Correlation**



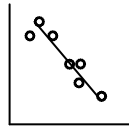
**High  
Degree  
Positive  
Correlation**



**Low degree  
Positive  
Correlation**



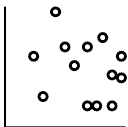
**Perfect  
Negative  
Correlation**



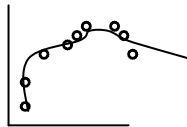
**High  
Degree  
Negative  
Correlation**



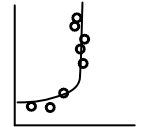
**Low Degree  
Negative  
Correlation**



**No  
Correlation  
or  
Lack of  
correlation**



**Non Linear  
Correlation**



**Non Linear  
Correlation**

Tr: It is simplest method to study the correlation between two variables. It helps to visualize the relationship between two related variables but does not enable us to measure the degree to which the variables are linearly related. To study the degree of relationship Correlation Coefficient need to be calculated. So, what is **Karl Pearson's Correlation Coefficient?**

St: Karl Pearson's Correlation Coefficient: It measures the degree of correlation between two variables. It is denoted by  $r_{xy}$  or  $r$  denoting the measure of correlation between two variables x and y. It can be written as

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

Tr: How do you interpret its value?

St: Interpretation of r:

If  $r_{xy} = +1$  means perfect positive correlation between variables x and y,

If  $r_{xy} = -1$  means perfect negative correlation between variables x and y,

If  $r_{xy}$  lies between 0 and 1 means positive correlation between variables x and y,

If  $r_{xy}$  lies between -1 and 0 means negative correlation between variables x and y.

If  $r_{xy} = 0$  means no linear correlation between variables x and y.

If the correlation coefficient is close to +1 that means you have a strong positive relationship.

If the correlation coefficient is close to -1 that means you have a strong negative relationship

Tr: You may use any of the following formula for calculating the Karl Pearson's Correlation Coefficient. You will get the same result as these are equivalent forms of the same formula.

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$

$$r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2} \sqrt{\sum y^2 - n\bar{y}^2}}$$

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \sqrt{n\sum y^2 - (\sum y)^2}}$$

Tr: Do you know the properties of Correlation coefficient r?

**St:** Properties of Correlation Coefficient are:

- (i) Karl Pearson's Correlation coefficient lies between -1 and +1, i.e.  $-1 \leq r \leq +1$
- (ii) Correlation coefficient is independent of the change of origin and scale.
- (iii) Two independent variables are uncorrelated but converse is not true.

Hence  $r_{xy} = 0$  for independent variables.

(Teacher explains each property with some illustrations)

**Tr:** What is Coefficient of Determination?

(Probably students don't know about this)

**Tr:** Coefficient of Determination is useful to measure the *strength of the relationship*. This is done by calculating the *coefficient of determination*  $R^2$ . In other words, the coefficient of determination gives the ratio of the explain variance to the total variance. The coefficient of determination is the square of the coefficient of correlation i.e  $r^2$ . Thus.

$$\text{Coefficient of determination} = r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

**Remark:** This is true for models with only one independent variable.

When  $R^2$  has a value of 0.6483. This means 64.83% of the variation in the Y is explained by your regression model. The remaining 35.17% is ***unexplained***, i.e. due to error.

In general the higher the value of  $R^2$ , the ***better*** the model fits the data.

$R^2 = 1$ : Perfect match between the line and the data points.

$R^2 = 0$ : There are no linear relationship between X and Y.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group: (All Groups)**

1. A test in Mathematics was given to 14 students who were about to bring a course in statistics. The scores (X) in their test were examined in relations to score (Y) in the final examination in Statistics. The following Statistical data were obtained:

$$\sum x = 79, \sum y = 88, \sum x^2 = 755, \sum y^2 = 640, \text{ and } \sum xy = 765.$$

2. Calculate correlation coefficient from the following results:

$$N=10, \sum (x - 15)^2 = 186, \sum (y - 16)^2 = 255, \text{ and } \sum (x - 15)(y - 16) = 78.$$

3. If coefficient of correlation between X and Y is -0.50 then find coefficient of correlation between (i)  $U = 3X + 4$  and  $V = 5Y - 10$ . (ii)  $U = 3X + 5$  and  $V = -6Y + 3$

4. From the following data, compute the coefficient of correlation and interpret it.

	X	Y
No. of pairs of observations	16	16
Arithmetic mean	23	19
Standard deviation	3.21	1.33
Sum of squares of deviations from mean	156	128
Sum of product of deviations of X and Y from their respective means	142	

5. Following is the distribution of students according to their heights and weights. Find out the correlation coefficient between height and weight and interpret it.

Height (cm)	153	159.3	153.53	153	153.5	154.3	157.5	158.5	160.4
weight (kg)	47	45	47	57	56	62	65	63	64

## 6. Multiple Choice Questions:

- Simple correlation is the degree of \_\_\_\_\_ relationship between the two variables.
  - positive
  - negative
  - linear
  - non-linear
- If two variables X (temperature of the day) and Y (cost of share) are independent then
  - $\text{Cov}(X, Y) = 1$
  - $\text{Cov}(X, Y) = 0$
  - $\text{Cov}(X, Y) > 1$
  - $\text{Cov}(X, Y) < 0$

- iii. Cov (x,y) is given by
- (a)  $\frac{1}{n} \sum (x - \bar{x})^2$  (c)  $\frac{1}{n} \sum (x - \bar{x})(y - \bar{y})$   
 (b)  $\frac{1}{n} \sum (x - \bar{x})^2 \sum (y - \bar{y})^2$  (d)  $\sum (x - \bar{x})^{1/2}$
- iv. The coefficient of correlation is denoted by:
- (a)  $r^2$  (c)  $r/2$   
 (b)  $r$  (d) none of the them
- v. Variance of variable y is given by:
- (a)  $\frac{1}{n} \sum (y - \bar{y})^2$  (c)  $\frac{1}{n} \sum (y - \bar{y})^1$   
 (b)  $\frac{1}{n} \sum (y - \bar{y})^{1/2}$  (d) none of the them
- vi. The correct expression for Karl Pearson's coefficient of correlation is:
- (a)  $\frac{\text{Cov}(x,y)}{\sqrt{v(x)}\sqrt{v(y)}}$  (c)  $\frac{\text{Cov}(x,y)}{\sqrt{\sigma(x)}\sqrt{\sigma(y)}}$   
 (b)  $\frac{\text{Cov}(x,y)}{v(x).v(y)}$  (d) none of them
- vii. If  $r_{xy} = + 0.89$  it means that there is \_\_\_\_\_ correlation between x and y variables.
- (a) positive (c) moderate positive  
 (b) high (d) high positive
- viii. If  $r_{xy} = -1$  it means that there is \_\_\_\_\_ correlation between x and y variables.
- (a) negative (c) perfect negative  
 (b) high negative (d) no
- ix. If  $\text{cov}(x,y) = 3.5$  , standard deviation of x is 10 and standard deviation of y is 20 then  $r_{xy}$  will be:
- (a) 0.001 (c) - 0.247  
 (b) 0.17 (d) 0.71
- x. If  $r_{xy} = 0.65$ ,  $\text{cov}(x,y) = 10.45$ , s.d. of x is 6 then s.d. of y is:
- (a) 2.45 (c) 4.67  
 (b) 2.67 (d) 3.45

- xi. If standard deviation of  $X$  is 10 then its variance is:
- (a)  $\sqrt[2]{10}$  (c) 100  
(b) 5 (d) 20
- xii. If correlation between  $x$  and  $y$  is 0.45 and  $u = 10x + 5$ ,  $v = -2y + 10$  then  $r_{uv}$  will be:
- (a) 0.45 (c) - 0.090  
(b) - 0.45 (d) 0.90
- xiii. If  $\text{cov}(x, y)$  is zero it means that correlation between  $x$  and  $y$  is:
- (a) Zero (c) greater than zero  
(b) less than zero (d) different from zero
- xiv. If  $x$  and  $y$  both variables are independent then the correlation between  $x$  and  $y$  will be:
- (a) +1 (c) between 0.5 to +1  
(b) -1 (d) zero
- xv. If correlation between  $x$  and  $y$  is 0.80 then coefficient of determination is:
- (a) 40% (c) 64%  
(b) 80% (d) 34%

### **Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment related to the present topic to the students. Assignment work related to this topic is mentioned in the appendix no. VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Rank Correlation and its applications.

## Lesson No.8: Rank Correlation

### Teaching Points:

- Rank correlation.
- Without Tie Case and with Tie Case of rank correlation

### Instructional Objectives:

After completion of this class students will be able to

- Identify the problems for use of rank correlation.
- Calculate the rank correlation coefficient for Without Tie Case and with Tie Case of ranks.
- Interpret the value of rank correlation coefficient.

### Lesson Presentation:

Tr: As we came to know that when variables are measured on ordinal scale we should use Spearman's Rank Correlation Technique. Give some concrete examples where Spearman's Rank Correlation can be used.

St: Correlation between the scores given by two judges in a drawing competition.

St: Correlation between the industry employee size (in different categories) and their revenue size (indifferent categories).

St: Correlation of percentage of students who have free university meals in different states and their CGPA scores more than 8.5.

St: Correlation between Price of Water bottle at shop and distance of shop to the Golden temple.

St: Good. You have given very nice examples where rank correlation is more suitable as compared to Pearson correlation. Let us take an example for calculation of rank correlation.

### Without Tie Case of Ranks:

Calculate the correlation between English Vocabulary test and English Spelling Test. Marks for the students are given below.

Exam	Marks									
Scores of English Vocabulary Test	55	74	44	72	60	65	57	88	75	59
Scores of English Spelling Test	65	75	43	61	64	57	59	79	66	63

St:

English Vocabulary Test (Marks)	English Spelling Test (Marks)	Rank (Vocabulary Test)	Rank (Spelling Test)	d	d <sup>2</sup>
55	65	9	4	5	25
74	75	3	2	1	1
44	43	10	10	0	0
72	61	4	7	3	9
60	64	6	5	1	1
65	57	5	9	4	16
57	59	8	8	0	0
88	79	1	1	0	0
75	66	2	3	1	1
59	63	7	6	1	1
				Total =	54

Where d = difference between ranks and d<sup>2</sup> = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

Tr: How will you interpret this r = 0.67 value?

The value of rank correlation is also interpreted in the similar way as Pearson's coefficient of correlation is being interpreted. The value of rank correlation coefficient is also varies in the range of -1 to +1. So now you may try to interpret rank correlation coefficient r = 0.67.



St: Interpretation of rank correlation  $r = 0.67$ . There is high positive correlation between English vocabulary and English spelling test scores of students. It means that a strong positive relationship between the ranks of individuals obtained in the English vocabulary test and English spelling test exam. That is, the higher you ranked in English vocabulary test, the higher you ranked in English spelling test also, and vice versa. Now in case ranks will tie then, how will you calculate this rank correlation?

**With Tie Case of Ranks:** Let us take an example.

Calculate the strength of the link between the price of a 200 ml of cold water pouch and distance from the shop to the Contemporary Art Museum in Delhi.

Shop No.	Distance from shop to CAM (m)	Rank distance	Price of 200 ml cold water pouch (Rs)	Rank price	Difference between ranks (d)	d <sup>2</sup>
1	57	10	1.85	2	8	64
2	174	9	1.25	3.5	5.5	30.25
3	275	8	2.10	1	7	49
4	372	7	1.10	6	1	1
5	423	6	1.10	6	0	0
6	584	5	1.25	3.5	1.5	2.25
7	712	4	0.82	9	-5	25
8	795	3	0.65	10	-7	49
9	892	2	1.10	6	-4	16
10	985	1	0.84	8	-7	49
					d <sup>2</sup> = 285.5	

$$r = 1 - \frac{6}{n(n^2-1)} * [\sum d^2 + \sum m(m^2-1)/12]$$

$$\sum m(m^2-1)/12 = \frac{1}{12} * [2*(4-1) + 3*(9-1)] = \frac{30}{12} = 2.5$$

$$r = 1 - \frac{6}{10(100-1)} * [285.5 + 2.5] = 1 - \frac{6}{990} * 288 = 1 - 1.7454$$

$$r = -0.7454$$

Tr: What do you mean by  $r = -0.75$ ?

St: It means that there is a strong negative correlation between Price of 200 ml cold water pouch and Distance from shop to Contemporary Art Museum in Delhi. It means that as the distance from shop to Contemporary Art Museum in Delhi reduces the price of cold water pouch increases and vice versa.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### Task allotment for each group:

#### Group No.: 1, 3, 5.

1. Information is collected from state universities about the percentage of students who have university scholarships and their CGPA scores. Calculate the Spearman's Rank Correlation between the two and interpret the result.

State University	% of students having university scholarships	% of students scoring above 9.0 CGPA
New Delhi	0.13	23
Jaipur	0.23	25
Bhopal	0.22	22
Mumbai	0.25	26
Vadodara	0.26	28
Ahmadabad	0.31	33
Kolkata	0.42	36

#### Group No.: 2, 4, 6.

2. Scientists wanted to know whether social dominance was associated with the number of nematode eggs of Colobus monkeys, so they converted eggs per gram of feces to ranks and used Spearman rank correlation. Calculate the rank correlation between dominance rank and eggs per gram for the following data and interpret the results.

Monkey Name	Dominance Rank	Eggs Per Gram	Eggs Per Gram (Rank)
A	1	5787	1
B	2	4265	2
C	3	2654	3
D	4	1259	4
E	5	709	8
F	6	840	6
G	7	832	7
H	8	865	5

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Partial Correlation and Multiple Correlation.

## Lesson No.9: Partial and Multiple Correlations

### Teaching Points:

- Partial correlation
- Multiple correlation

### Instructional Objectives:

After completion of this class students will be able to:

- i. Identify the problems for use of partial correlation and multiple correlations.
- ii. Calculate the partial and multiple correlation coefficients.
- iii. Interpret the value of partial and multiple correlation coefficients.

### Lesson Presentation:

Tr: Dear students, till now we considered cases with two variables under the study and but considered that we are dealing 3 or 4 or even more number of variables and we want to find correlation among them. In that case we may use partial correlation or the multiple correlation for data analysis. Could you give some examples where we can use **partial correlations**?

St: Partial correlation measures the strength of a relationship between two variables, while controlling for the effect of one or more other variables. For example, you might want to see if there is a correlation between amount of food eaten and blood pressure, while controlling for weight or amount of exercise.

St: correlation of sale value of a particular commodity is related to the expenditure on advertising when the effect of price is controlled.

St: Correlation between test scores and GPA scores after controlling for hours spent studying.

St: Correlation between body height and body weight keeping gender as constant.

St: Correlation between diet and body weigh keeping exercise variable as constant.

St: Correlation between learning strategy and achievement scores keeping IQ scores constant.

Tr: Yes very true. All examples are correct.

Give some examples of Multiple Correlation?

St: Correlation between achievement score and the joint effect of IQ scores and EQ scores.

St: Correlation between body weight and joint effect of diet and exercise.

St: Correlation between mortality rate per year and the joint effect of finance invested in health sector and education sector, etc.

Tr: Yes! All examples are correct. So we can say that Partial correlation is a study of the relationship between one dependent variable and one independent variable by keeping the effect of other independent variables constant.

Let  $X_1$ ,  $X_2$  and  $X_3$  are three variables. Then relationship between  $X_1$  and  $X_2$  by keeping effect of  $X_3$  constant is called partial correlation coefficient and denoted by  $r_{12.3}$  and defined as

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Similarly  $r_{13.2}$  = Partial correlation coefficient between  $X_1$  and  $X_3$  by keeping effect of  $X_2$  constant

On the similar lines try to write for  $r_{13.2}$  and  $r_{23.1}$

St:

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

St:  $r_{23.1}$  = Partial correlation coefficient between  $X_2$  and  $X_3$  by keeping effect of  $X_1$  constant

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Tr: State the properties of Partial correlation coefficient.

St: Properties of partial correlation coefficient are:

- i. Partial correlation coefficient lies between  $-1$  and  $+1$
- ii.  $r_{13.2} = r_{31.2}$  ,  $r_{23.1} = r_{32.1}$  and  $r_{12.3} = r_{21.3}$
- iii. Standard error (S.E.) of partial correlation coefficient is  $Z = 1/(N - 3)$

Tr: The Limitations of partial correlation coefficient are as follow:

The utility of the partial correlation analysis is great in inter-related series and in various experimental designs where inter-related phenomena are to be study. Partial correlation has following limitation.

- i. In partial correlation it is assumed that correlation between two variables is linear but in practice there may not exist linear relation between variables.
- ii. In calculation of partial correlation it is assumed that the various independent variables are independent of each other but in practice this may not be true.

Tr: As we define partial correlation, define multiple correlation.

St: Multiple correlation is the study of joint effect of all independent variables on dependent variable.

Tr: Yes, well attempted. It can also be explained as Multiple correlation coefficient is correlation coefficient between dependent variable and its estimated value which is obtained from multiple regression line. e.g. Let,  $X_1$ ,  $X_2$  and  $X_3$  are three variables  $X_1$  dependent variable and  $X_2$ ,  $X_3$ , are independent variables then Multiple correlation coefficient is same as correlation coefficient between  $X_1$  and its estimated value ( $\hat{X}_1$ ), which is obtained by multiple regression equation of  $X_1$  on  $X_2$  and  $X_3$ . It is denoted by  $R_{1.23}$  and defined as

$$R_{1.23} = r(X_1, \hat{X}_1) = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Similarly, if  $X_2$  dependent variable  $X_1$  and  $X_3$  are independent variables then multiple correlation coefficient of  $X_2$  on  $X_1$  and  $X_3$  is,

$$R_{2.13} = r(X_2, \hat{X}_2) = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

Similarly try to write for  $R_{3.12}$ .

St: if  $X_3$  dependent variable,  $X_1$  and  $X_2$  independent variables then multiple correlation coefficient of  $X_3$  on  $X_1$  and  $X_2$  is

$$R_{3.12} = r(X_3, \hat{X}_3) = \sqrt{\frac{r_{23}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

Tr: Where,  $r_{12}$  = Total correlation coefficient between  $X_1$  and  $X_2$  =  $r(X_1, X_2)$

$r_{13}$  = Total correlation coefficient between  $X_1$  and  $X_3$  =  $r(X_1, X_3)$

$r_{23}$  = Total correlation coefficient between  $X_2$  and  $X_3 = r(X_2, X_3)$

Tr: State the properties of Multiple correlation.

St: Following are the properties of Multiple Correlation Coefficient:

- i. It is non-negative coefficient. **It's value lies between 0 and 1.**
- ii.  $R_{1.23} > r_{12}$  and  $R_{1.23} > r_{13}$
- iii. If  $R_{1.23} = 0$  then  $r_{12} = 0$  and  $r_{13} = 0$ .
- iv.  $R_{1.23} = R_{1.32}$
- v. If multiple correlation coefficients are zero then the variables are not linearly related. Closer to one multiple correlation coefficient indicate better linear relationship between variables.

Tr: What are the Limitations of Multiple correlation coefficient?

St: The main limitation of multiple correlation is that it assumes linear relationship between the variables and it also assumes that there does not exist any relationship between independent variables. But in practice there may exist inter-relation between independent variable and there may not be linear relationship between them.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### **Task allotment for each group: (All Groups)**

Q1. A dataset was taken of the confidence scales of 15 employees some years ago using 4 facets of confidence (Appearance, Emotional and Problem Solving, as well as their gender and their citizenship status). For the following data calculate  $r_{13.2}$ ,  $r_{12.3}$ ,  $R_{1.23}$  and  $R_{2.13}$  and interpret the results.

Sl. No.	Appearance Confidence (X <sub>3</sub> )	Emotional Confidence (X <sub>2</sub> )	Problem Solving (X <sub>1</sub> )	Gender M=0 & F=1 (X <sub>4</sub> )	citizenship status 0: Indian 1: Non- Indian (X <sub>5</sub> )
1	31	32	54	0	0
2	33	58	66	0	0
3	34	55	64	1	0
4	32	57	62	1	0
5	34	53	65	0	1
6	36	55	65	1	1
7	37	54	63	0	0
8	34	56	67	0	0
9	33	52	64	1	0
10	36	56	65	1	0
11	37	35	36	0	0
12	35	48	57	0	0
13	48	44	48	1	1
14	49	49	59	1	0
15	36	47	55	1	0

### **Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment related to the present topic to the students. Assignment work related to this topic is mentioned in the appendix no. VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Bi- serial and Point Bi-serial Correlations.



## Lesson No.10: Bi-Serial & Point Bi-Serial Correlation

### Teaching Points:

- Bi –serial correlation
- Point bi-serial correlation

### Instructional Objectives:

After completion of this class students will be able to:

- i. Identify the problems for use of Bi-serial correlation and Point bi-serial correlation.
- ii. Calculate the Bi-serial correlation and Point bi-serial correlation.
- iii. Interpret the value of Bi-serial correlation and Point bi-serial correlation coefficients.

### Lesson Presentation:

Tr: Dear Students, till now we have studied Pearson's and Spearman's correlations. Now consider a situation in which one variable is on nominal scale and another is on interval or ratio scale of measurement. Which type of correlation will you use to calculate their degree of relationship between them?

St: We will use Bi-serial correlation.

Tr: Yes. You replied correctly. But can we find bi-serial correlation with any kind of nominal variable?

St: No. The nominal variable should be of dichotomous in nature.

Tr: What do you mean by dichotomous in nature?

St: A variable which can take only two values.

Tr: Good. Give some examples of Dichotomous variable.

St: Gender: Male and Female.

St: Resident: Indian and Non- Indian.

St: Education level: Graduate or Non- Graduate.

St: Religion: Hindus or Non- Hindus.

St: Marital Status: Married or Unmarried.

St: Result: Pass or Fail.

St: Height of person: less than 5 feet or more than 5 feet.

St: Age: less than 10 years or more than 10 years.

St: Health status: infected or uninfected.

St: Health status of child at the time of birth: Thalassemia patient or Non-Thalassemia patient.  
Species: Poisonous or non Poisonous, etc.

Tr: Here all examples are true. But you can see that some are naturally dichotomous and some are artificially. Identify the natural and artificial dichotomous variables as mentioned above.

St: Gender, Health Status, Species and Health status of child at the time of birth are naturally dichotomous.

St: Resident, educational level, religion, Marital Status, result, Height of person and Age as described above are treated as artificially dichotomous.

Tr: Well classified. You are true.

If one variable is naturally dichotomous and another is on interval / ratio scale then we should use point Biserial Correlation where as when one variable is artificially dichotomous and another is on interval / ratio scale then we should use Bi-serial correlation. Lets us learn the process of calculating Bi-serial Correlation:

Following data is related with the scores of 145 students on Music Appreciation Test. The total distribution of 145 scores has been broken down into subdivisions, the first made up of 21 students who had training in music and the second of 124 students without any formal musical training. Find whether there is correlation between test scores and previous training in music.

Scores	Training Group	Non- Training Group	Total
	F	f	f
<b>85-89</b>	5	6	11
<b>80-84</b>	2	16	18
<b>75-79</b>	6	19	25
<b>70-74</b>	6	27	33
<b>65-69</b>	1	19	20
<b>60-64</b>	0	21	21
<b>55-59</b>	1	16	17
<b>Total</b>	$N_1 = 21$	$N_2 = 124$	$N = 145$

$M_r = 71.35$ , Mean of all 145 scores.

$\sigma = 8.8$ , s.d. of all scores.

$M_p = 77.00$ , mean of trained group.

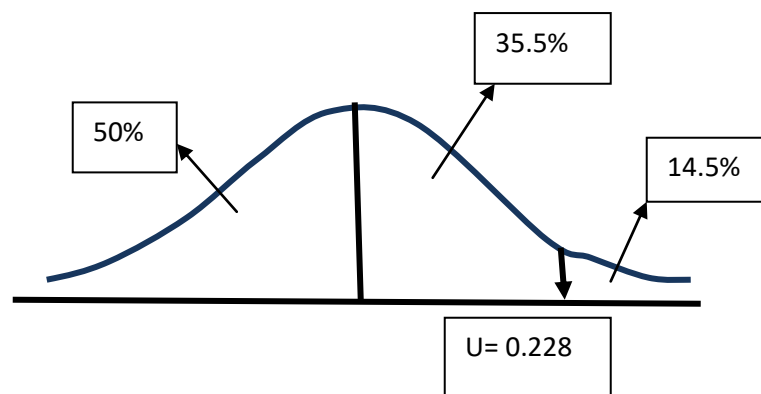
$M_q = 70.39$ , mean of un-trained group.

$p = 0.145$ , proportion in group 1 (trained)

$q = 0.855$ , proportion in group 2 (un-trained)

$u = 0.228$ , height of ordinate separating 0.145 and 0.855 in a unit normal distribution.

Area from mean $\mu$	Ordinate ( $u$ )
0.35	0.233
0.36	0.223



For 0.355  $u = ?$

$$U = (0.233 + 0.223) / 2 = 0.228$$

**Bi-serial Correlation:**

$$r_{bis} = \frac{M_p - M_q}{\sigma} - \frac{p \cdot q}{u}$$

$$r_{bis} = \frac{77.00 - 70.39}{8.8} - \frac{0.145 \cdot 0.855}{0.228} = 0.41$$

Conclusion: Here the nominal variable is artificially dichotomized or split into two categories namely trained (graduate in Music) and un-trained (non-graduate in Music). Also another variable is on interval scale i.e. Music Appreciation Test. Therefore bi-serial correlation is used to study the relationship between them. The value of  $r_{bis} = 0.41$  which means that there is a moderate association between training and test scores on Music Appreciation Test.

There may be a situation in which nominal variable is naturally dichotomized like male – female, living – dead, loyal – disloyal, delinquent – non- delinquent, psychotic – normal, colour blind – normal, etc. for such cases we use point bi-serial correlation. The formula is slightly different from bi-serial correlation.

***Point Bi-serial Correlation:***

$$r_{pbis} = \frac{M_p - M_q}{\sigma} - \sqrt{pq}$$

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group: (All Groups)**

1. In an IQ test of 155 students scores were gained and students were split in two categories as socially adjusted and socially maladjusted. Find the relationship between IQ scores and Social adjustment of students for the following data:

IQ Scores	Socially Adjusted Group	Socially Maladjusted group
	F	f
<b>100-110</b>	10	2
<b>110-120</b>	15	4
<b>120-130</b>	19	3
<b>130-140</b>	22	6
<b>140-150</b>	16	5
<b>150-160</b>	34	3
<b>160-170</b>	14	2
<b>Total</b>	<b>N<sub>1</sub> = 130</b>	<b>N<sub>2</sub> = 25</b>

2. In an Achievement test of Mathematics of 288 students scores were gained and students were split in two categories as Indian and NRI. Find the relationship between Achievement Scores in Mathematics and Residency of students for the following data:

Achievement Scores	No. of Indian Students F	No. of NRI Students F
<b>30-40</b>	12	4
<b>40-50</b>	20	7
<b>50-60</b>	37	3
<b>60-70</b>	46	8
<b>70-80</b>	57	14
<b>80-90</b>	38	6
<b>90-100</b>	27	9
<b>Total</b>	<b>N<sub>1</sub> =237</b>	<b>N<sub>2</sub> = 51</b>

#### **Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment related to the present topic to the students. Assignment work related to this topic is mentioned in the appendix no. VIII.

**Announcement of topic in Class:** This announcement was made three days prior to the class. For the coming class read Simple Regression Analysis and Concept of Multiple Regression.

## Lesson No.11: Simple Regression Analysis and Concept of Multiple Regression

### Teaching Points:

- Concept of Regression Analysis
- Simple Linear Regression
- Concept of Multiple Regression

### Instructional Objectives:

After completion of this class students will be able to:

- i. Explain the meaning of regression analysis.
- ii. Explain the simple linear regression analysis.
- iii. Predict the value of a variable using regression equation.
- iv. Explain the use of multiple regression analysis.

### Lesson Presentation:

Tr: Dear Students, as you know that Correlation is the degree of relationship between the variables. But if we want to express this relationship in some mathematical model it is known as Regression Equation. And the process by which we can get these regression equations is all about regression analysis. Correlation is used to study the relationship between variables but regression is used for basically estimation or prediction purpose. Now with this basic information about regression you can understand its importance. Let me listen from you, what do you understand by regression?

St: The general meaning of regression is stepping back or moving back.

Tr: So, what does regression analysis mean?

St: may be no response...

Tr: Here regression means stepping back towards mean or average. As the definition of regression analysis states that "Regression Analysis is the mathematical measure of the average relationship between two or more variables in terms of the original units of the data".

Tr: In regression analysis we deal with different types of variables. Do you know various types of variables used in regression analysis?

St: Dependent Variable - The Variable whose value is to be predicted.

Tr: What are the other names of dependent variables.

St: Regressed or Explained Variable.

St : Independent Variable - The variable which influences the values or is used for prediction.

Tr: What are the other names of independent variable.

St: Regressor or Predictor or Explanatory Variable.

Tr: So, in a regression analysis we must have at least one independent and one dependent variable. Therefore Simple Linear Regression is the technique for estimation of unknown value of the dependent variable from the known value of independent variable.

There may be a situation where one dependent and more than one independent variables are under the study and the value of dependent variable is estimated using the set of independent variables, this technique is known as multiple regression analysis.

What is regression Equation?

St: Regression equations: The Regression equation also known as estimating equations, are algebraic expressions of the regression lines. There are two regression equations – the regression equation of X on Y is used to describe the variations in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

For simple regression analysis:

Regression Equation of Y on X:

The regression equation of Y on X is expressed as follows:

$$Y = a + bX$$

It may be noted that in this equation:

‘y’ is a dependent variable and ‘x’ is independent variable. ‘a’ is Y-intercept and ‘b’ is the slope of the line and it represents the change in Y variable for a unit change in X variable.

The value of numerical constants ‘a’ and ‘b’ are obtained with the help of the best fit curve and this is based on the principle of least square. The principle of least square is that we minimize the sum of squares of the deviations or the errors of estimates. Thus the deviations between the given observed values of the variable and their corresponding estimated values are given by the line of best fit.

What are the two regression lines?

St: Line of Regression of Y on X is given by:

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

Where  $b_{yx} = r_{xy} \frac{\sigma_y}{\sigma_x}$  is called regression coefficient of y on x.

St: Line of Regression of X on Y is given by:

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

Where  $b_{xy} = r_{xy} \frac{\sigma_x}{\sigma_y}$  is called regression coefficient of x on y.

Tr: How will you interpret regression coefficient?

St: It is called the slope of the line.

St: It gives the rate of change of the dependent variable when independent variable changes by one unit.

St:  $b_{yx}$  measures the how much unit change in variable y when x change by one unit.

and  $b_{xy}$  measures the how much unit change in variable x when y change by one unit.

Tr: If we have data then following formulas can be used to calculate regression coefficients.

$$b_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} \quad \text{and} \quad b_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$b_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum y^2 - n\bar{y}^2} \quad \text{and} \quad b_{yx} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} \quad \text{and} \quad b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Tr: There are few important properties of regression coefficients:



(i) Correlation coefficient is geometric mean of both regression coefficients. i.e

$$r_{xy} = \pm \sqrt{b_{yx} \times b_{xy}}$$

(ii) If one regression coefficient is greater than one than other regression coefficient must be less than one. i.e.,  $b_{xy} \times b_{yx} \leq 1$ .

(iii) Sign of both regression coefficients and correlation coefficients are same.

(With illustrations these properties will be described by the teacher)

Let us consider one example:

1. Given the following information:

Year	2009	2010	2011	2012	2013	2014
Dropout of students in schools of Vadodara District per year	234	217	183	130	102	78
Annual investment of Government per student (in '000 Rs.)	11	11.5	21.3	21.5	12.9	14.6

- Develop the estimating equation that best describes the given data.
- Estimate the dropouts of students when annual investment is Rs.15500 per student per year by the government.
- How much variation in dropouts of students is explained by the variation in Annual investment of Government per student?

Tr: Here which one is independent variable and which one is dependent variable?

St: Annual investment of Government per student (in '000 Rs.) is the independent variable.

St: Dropout of students in schools of Vadodara District per year is dependent variable.

Tr: Independent variable is usually denoted by Capital X and dependent variable is denoted by capital Y.

So answer the (a) part?

St: So, in this problem we have to estimate the value of Dropout of students in schools of Vadodara District per year (y) for given value of Annual investment of Government per student (in '000 Rs.) (X).

(a) The regression equation Y on X is given by:

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

Where  $\bar{x} = \frac{\sum x}{n}$ ,  $\bar{y} = \frac{\sum y}{n}$  and  $b_{yx} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$

Year	Dropout of Students in Schools of Vadodara District Per Year (Y)	Annual Investment of Government Per Student (in '000 Rs.) (X)	X <sup>2</sup>	XY
2009	234	11	121	2574
2010	217	11.5	132.25	2495.5
2011	183	21.3	453.69	3897.9
2012	130	21.5	462.25	2795
2013	102	12.9	166.41	1315.8
2014	78	14.6	213.16	1138.8
Total	944	92.8	1548.76	14217

$$\bar{X} = \frac{92.8}{6} = 15.46666667 \quad \bar{Y} = \frac{944}{6} = 157.3333333$$

$$b_{yx} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = -3.380538254$$

Therefore,  $(y - 157.33) = -3.381(x - 15.467)$

$$\mathbf{Y = 105.047 - 3.381 X}$$

(b) For X= 15.5, Y estimate is:

$$Y^{\wedge} = 105.047 - 3.381(15.5) = 52.649$$

(c)  $r^2$  % amount of variation in dropouts of students is explained by the variation in Annual investment of Government per student.

Where  $r_{xy} = \pm \sqrt{b_{yx} b_{xy}}$

$$b_{yx} = -3.380 \text{ and } b_{xy} = \frac{n \sum xy - \sum x \sum y}{n \sum y^2 - (\sum y)^2} = -0.018987425$$

Therefore  $r^2 = (-3.380) * (-0.018987425) = 0.06419649$

$r^2 = 6.419649$  % amount of variation in dropouts of students is explained by the variation in Annual investment of Government per student.

Tr: In case of simple linear regression we study the effect of one independent variable on other dependent variable while multiple regression is a study of more than one independent variables' effect on one dependent variable i.e Let  $X_1$ ,  $X_2$  and  $X_3$  are three variables,

- (i)  $X_1$  is depends on  $X_2$  and  $X_3$ . Then for given value of  $X_2$  and  $X_3$  one can estimate  $X_1$  by assuming linear relationship. Their relation can be expressed by the equation

$$X_1 = a + bX_2 + cX_3$$

Using method of least square we can find constants a, b and c.

- (ii)  $X_2$  is depends on  $X_1$  and  $X_3$ . Then for given value of  $X_1$  and  $X_3$  one can estimate  $X_2$  by assuming linear relationship. Their relation can be expressed by the equation

$$X_2 = a + bX_1 + cX_3$$

Using least square method find constants a, b and c.

- (iii)  $X_3$  is depends on  $X_1$  and  $X_2$ . Then for given value of  $X_1$  and  $X_2$  one can estimate  $X_3$  by assuming linear relationship. Their relation can be expressed by the equation

$$X_3 = a + bX_1 + cX_2$$

Using least square method find constants a, b and c.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group: (All Groups)**

1. Given the following information:

Year	2007	2008	2009	2010	2011	2012	2013
No. of new secondary schools opened in the Gujarat state	14	19	34	19	29	13	12
Permanent Recruitment of teachers in secondary schools	130	165	256	159	213	140	110

- (i) Develop the estimating equation that best describes the given data.
- (ii) Estimate the Permanent Recruitment of teachers in secondary schools when number of new secondary schools opened in the Gujarat state is 40.
- (iii) How much variation in Permanent Recruitment of teachers in secondary schools is explained by the variation in number of new secondary schools opened in the Gujarat state?

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Z- Score and their applications.

## Lesson No.12: Z-Score

### Teaching Points:

- Concept of Z-score
- Application of Z- score

### Instructional Objectives:

After completion of this class students will be able to

- Define Z-score.
- Use Z-score for solving various problems.

### Lesson Presentation:

Tr: What is a difference between a raw score and a Z- Score?

St: A raw score X is a score gained from measurement on a variable collected from a tool when administered on some sample units. But Z score is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean. Z- Score is defined as  $Z = \frac{x-\mu}{\sigma}$ .

Tr: So, you can see that Z-score is a simple transformation of the raw score X. Now looking upon this transformation, what do you mean by Z- score is 0?

St: It means that  $X = \mu$ .

St: It means data point's score is identical to the mean score.

Tr: Yes! You are right. What do you say if Z- scores are positive?

St: If a z-score is positive, its' corresponding raw score X is above (greater than) the mean  $\mu$ .

Tr: When do the Z- score is negative?

St: A Z-score is negative, when its' corresponding raw score X is below (lesser than) the mean  $\mu$ .

Tr: So, can we say that a Z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The absolute value of the Z-score tells you how many standard deviations you are away from the mean. If a Z-score is equal to 0, it is on the mean. If a Z-Score is equal to +1, it is 1 Standard Deviation above the mean. If a z-score is equal to +2, it is 2 Standard Deviations above the mean. If a Z-score is equal to -1, it is 1 Standard Deviation below the mean. If a Z-score is equal to -2, it is 2 Standard Deviations below the mean.

95% of scores are going to be no more than 2 standard deviation units away from the mean. That means that most scores will fall between  $Z=-2$  to  $Z=+2$ . However, some scores will be greater than the absolute value of 2. You can interpret these scores to be very far from the mean.

Why Z-scores are important?

St: Z- Scores are useful to standardize the values (raw scores) of a normal distribution by transforming them into Z-scores because:

- Z- Scores allows researchers to calculate the probability of a score occurring within a standard normal distribution;
- Enables us to compare two scores that are from different samples (which may have different means and standard deviations).

Tr: How will you calculate a raw score when a Z-score is known?

St:  $X = (Z) * (SD) + \text{mean}$

Tr: Let us take some concrete example to study the applications of Z-Scores.

Identify which student has performed best and which has performed least from the following information. Also arrange the name of students on merit basis.

Sl. No.	Name of The Student	Obtained Marks	Maximum Marks	Mean	S.D.
1	Ashok	35	50	40	7
2	Rina	58	75	65	10
3	Jinal	18	25	17	4
4	Hemant	70	75	58	2
5	Urvashi	27	50	30	3
6	Rita	34	50	36	5
7	Mahek	44	50	41	3
8	Pinal	56	75	56	5
9	Payal	57	75	44	3
10	Lavina	68	75	45	5

St: Using the formula:  $z = \frac{x-\mu}{\sigma}$  values will be calculated. Following table shows the Z-scores. Here, the maximum Z- score is 6.5 (shown in yellow band) of Rita and the minimum Z- score is -1 of Urvashi. Therefore the best performer is Rita and the least performer is Urvashi.

Name of The Student	Obtained Marks (x)	Maximum Marks	Mean ( $\mu$ )	S.D.	Z - Score
Ashok	35	50	40	7	-0.714286
Rina	58	75	45	2	6.5
Jinal	18	25	17	4	0.25
Hemant	70	75	58	2	6
Urvashi	27	50	30	3	-1
Rita	34	50	36	5	-0.4
Mahek	44	50	41	3	1
Pinal	56	75	56	5	0
Payal	57	75	44	3	4.3333333
Lavina	68	75	45	5	4.6

The merit of students is given below:

Name of The Student	Merit No.	Z – Score
Rina	1	6.5
Hemant	2	6
Lavina	3	4.6
Payal	4	4.3333333
Mahek	5	1
Jinal	6	0.25
Pinal	7	0
Rita	8	-0.4
Ashok	9	-0.714286
Urvashi	10	-1

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group:**

**Q1.** Select the 5 best students from the following Information:

Name of The Student	Obtained Marks (x)	Maximum Marks	Mean ( $\mu$ )	S.D.
A	30	40	25.4	2.3
B	33	40	26.8	3.3
C	24	40	26.2	3.5
D	32	40	27	4.2
E	23	30	22	5.3
F	26	50	29	4.23
G	35	50	31.8	3.7
H	36	50	28.44	4.6
I	45	50	43	4.83
J	35	40	34	4.1
K	35	50	32	2.9
L	36	30	26	3.0
M	37	30	37	4.12
N	22	30	20.8	4.5
P	19	40	15	3.47
Q	24	30	17	3.33

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about probability sampling methods.



## **Lesson No.13: Sampling Methods**

### **(Probability Sampling Techniques)**

#### **Teaching Points:**

- **Sampling**
- **Types of Sampling: Probability Sampling and Non-probability Sampling**
- **Simple Random Sampling: Simple Random Sampling with replacement (SRSWR) and Simple Random Sampling without replacement (SRSWOR)**
- **Cluster Sampling**
- **Systematic Sampling**
- **Stratified Sampling: Neyman Allocation Stratified Random Sampling and Proportional Allocation Stratified Random Sampling**
- **Multi Phase Sampling:**
- **Multi Stage Sampling**

#### **Instructional Objectives:**

After completion of this class students will be able to:

- i. Explain the meaning of sampling.
- ii. Describe the significance of sampling.
- iii. Distinguish between probability and non-probability sampling techniques.
- iv. Explain simple random sampling, cluster sampling, systematic sampling, stratified sampling, multi phase sampling and multi stage sampling technique with illustrations.

#### **Lesson Presentation:**

Tr: In Sampling theory, what is population?

St: In studying various characteristics related to items or individuals or objects belong to a particular group. This group of individuals under study is known as Population.

Tr: What is a sample? And how sample is different from sampling?

St: For every inquiry complete enumeration of the entire population is quite difficult and therefore a sample needs to be selected in a scientific manner. All the information contained in the selected sub-part (sample) is examined critically with respect to question under study. Finally, on the basis of available results predictions are drawn about whole population characteristics is known as Sampling. The most important aim of the sampling studies is to

obtain maximum information about the phenomena under study with minimum use of money, time and energy.

Tr: Therefore we can say that Sampling is a process and sample is the output of sampling. Population is the aggregate of statistical data forming a subject of investigation. Sample is a finite subset of population, selected from it with a view of estimation the characteristic of population is known as sample.

Keep in mind that a sample is not studied for its own sake. The main objective of its study is to draw inferences and estimating about population under consideration.

Hence, Sampling is defined as the process of learning about the population on the basis of sample draw from it and the process of sampling is consists of:

- Selecting the sample,
- Examining it and collecting information required and
- Drawing conclusions or inferring about the population.

Do you know various types of populations?

St: Finite Population, Infinite Population.

St: Existent Population, Hypothetical Population.

Tr: Explain each term with some example.

St: **Finite Population:** The finite population contains only finite number of elements in it. Eg. A study under which V standard students of CBSE of the academic year 2013. Say there are 5,02,304 students. Then population Size  $N = 502304$ . For this study suppose 3000 students were taken then sample size  $n = 3000$ .

St: **Infinite Population:** The population having an infinite number of elements or with the number of objects as large as appear practically infinite. Eg: A study under which People consuming tobacco in the Vadodara Distric in 2010.

St: **Existent Population:** A population consisting of concrete objects is termed as existent population.

St: **Hypothetical Population:** A population consisting of imaginary objects is termed as hypothetical population.

Tr: what is parameter and statistic?

St: **Parameter ( $\theta$ ):** The statistical constants of population are known as parameters.

E.g. population mean =  $\mu$ , Population variance =  $\sigma^2$

**St: Statistic (t):** Statistic is any function of sample observations.

E.g. Sample mean =  $\bar{x}$ , Sample variance =  $s^2$

Tr: What is the relationship between Parameter and Statistic?

St: may not be able to answer.

Tr: Statistic is used to estimate the value of parameter. Generally statistic is denoted by  $t$  and parameter is denoted by  $\theta$ . What is an unbiased estimator?

A statistic  $t=f(x_1, x_2, \dots, x_n)$  is said to be an unbiased estimator of population parameter  $\theta$  if  $E(t) = \theta$ . i.e.  $E(\text{statistic}) = \text{Parameter}$

eg. . Average of sample statistics is same as population parameter.

Here,  $t$  is said to be an unbiased estimate of the parameter  $\theta$ .

Basically, the technique of sampling estimates different characteristics of the population from which it is selected. Therefore there may exist some errors in predicted results about the population. The errors characterized during such estimation are known as sampling error. This error is inherent and unavoidable in any sampling scheme.

When error can be introduced in sampling?

St: The sampling error can be involved at the time of collection, processing and analysis of the sample data.

Tr: What are different types of errors in sampling surveys?

St: In a sample survey errors may be classified as follows:

Sampling Error: The errors which arise due to only a sample being used to estimate the population parameter is termed as sampling error or sampling fluctuation. Whatever the degree of cautiousness in selecting a sample, there will always be a difference between the population parameter and its corresponding estimate. This error is inherent and unavoidable in any and every sampling scheme. A sample with a smallest sampling error is considered a good representative for the population. Increasing sample size can reduce sampling error. Generally, sampling errors are due to the following reasons:

- (i) Improper selection of sample
- (ii) Substitution
- (iii) Faulty demarcation of statistical units
- (iv) Errors due to variability of population and wrong method of estimation.

St: Non-sampling Error: The sample estimates may be subject to other errors. Following are some of the main sources that may result in Non-Sampling errors.

- (i) Failure in measuring some of the units in the selected sample.
- (ii) Observational errors due to defective measurement technique.
- (iii) Errors introduced in editing, coding and tabulating the results.

In practice, the population census may result in non-sampling errors, but it is free from sampling errors. The magnitude of non-sampling error is likely to increase with increase in sampling size.

Data can be collected in two ways, first is through census method or complete enumeration and second is through Sample method.

Do you know what is Census Method or Complete Enumeration?

St: The census method or Complete Enumeration.

This method deals with complete inspection of all the units/elements/objects of population. Here all the units of population are examined and information is collected. Since all the population units are to be examined, we get most precise information. But on the other hand this method consumes lots of time, labor (manpower), money. This method is administratively inconvenient in certain cases. More difficulties in terms of time, labor, funds etc., are added when whole population is scattered over large geographical area. Non-sampling errors are those, which are attributed due to faulty planning, incorrect execution, processing and analysis of data. Since in census method all the units of population are to be examined, magnitude of non-sampling error is high as compared to partial enumeration. Census method is recommended when

- Complete information from all the units of population is required,
- N, the population size is small,
- Accuracy is of ultimate requirement, which otherwise may cause loss of life or serious causality of an object/unit/personnel under consideration.

Tr: What is Sample Method? State the conditions when sample method is more advisable.

St: The sample method has large number of advantages over the previous one. Those are in terms of Speed, Economy, Administrative convenience, Greater scope in infinite as well as hypothetical population as well as in destructive experiments, Reliability, Adaptability etc. This method may also attribute non-sampling errors but the magnitude is very less as compared to earlier one. A properly designed and executed sample survey yields in fairly reliable and good

results often better than that of population survey. This method is also not free from Sampling and Non-sampling errors. But the magnitude of these errors is relatively less as compared to census method. Partial enumeration is preferred when,

- Population consists of infinite number of units or it is hypothetical population.
- In the process of inspection, the population unit itself is destroyed,
- Entire population is scattered over wide geographical area etc.

Tr: What do you mean by a sample size?

St: The size of a sample is the number of sampling units which are selected from a population by a random method.

Tr: The problem that arises is how to decide what should actually be the number of sampling units to be selected from a population. The sample size depends on a number of considerations which are as follows:

- i. The purpose for which the sample is drawn.
- ii. The types of population from which the sample is to be drawn.
  - If the sampling units constituting the population are highly variable, then a large sample is requiring.
  - If the population has less variable units then a small sample is good enough.
  - For perfectly homogeneous population, a single unit is sufficient to get the correct result for the whole population. e.g the blood of a person is perfectly homogeneous and hence a drop of a blood is taken for investigation.
- iii. Availability of technical people or equipment needed.
- iv. Resources allotted for the study in terms of time and money.

What are the characteristics of a good sample?

St: Characteristics of a good sample are:

- i. It should be a good representative of population.
- ii. It should be free from bias. (The selection should be done by scientific method.)
- iii. The sample should be of appropriate size.
- iv. Selection of sample units should be done independently from one another.
- v. If heterogeneous population is given, stratification must be carried out.
- vi. Units of sample should be selected during the same time period.

Tr: What are the merits and demerits of sampling?

St: Merits of sampling:

- i. Reduce cost
- ii. Greater speed
- iii. More accuracy
- iv. More scientific
- v. Provides greater scope of study
- vi. In large population or in some destructive experiment it is the only feasible method.

St: Demerits of sampling:

- 1) The results may be inaccurate and misleading if sampling is carried out and executed in non-scientific manner.
- 2) Requires experienced and qualified personals.
- 3) Sometimes due to complicated technique, it may even consume more time, labour and fund in comparison to complete enumeration.

Tr: What are different types of Sampling techniques?

St: There are two types: Probability Sampling, Non-Probability Sampling.

Tr: Describe each of its types.

St: Non-probability sampling: It is also known as Purposive or Subjective of Judgment sampling. In this method some criterion of selection is first laid down and sample is selected according to that purpose. Here the probability of selecting some units of population for the sample is very high whereas the probability for the rest of units, not satisfying predetermined criteria completely, is very less.

Probability or random sampling: Here all the units from the population are selected at random. Random selection does not mean haphazard selection. By random selection we mean that, the probability of inclusion of each item of the population, in the sample is equal.

Tr: State the names of Probability Sampling and Non-Probability Sampling methods?

St: Following table shows the methods of sampling falling under probability sampling and non-probability sampling.

Probability Sampling (Random)	Non-probability Sampling
<ul style="list-style-type: none"> <li>• Simple random sampling</li> <li>• Stratified random sampling</li> <li>• Systematic sampling</li> <li>• Cluster sampling</li> <li>• Multi-Stage sampling</li> <li>• Multi phase sampling</li> </ul>	<ul style="list-style-type: none"> <li>• Convenience sampling</li> <li>• Purposive sampling</li> <li>• Quota sampling</li> <li>• Judgment sampling</li> <li>• Snowball sampling</li> </ul>

Tr: Yes, your bifurcation is true. Now explain Simple Random Sampling?

St: Simple Random (Unrestricted) Sampling:

Simple random sampling or unrestricted random sampling will refer to that technique of random sampling in which each unit of the population has an equal and independent chance of being included in the sample. Whenever population is homogeneous, sample can be selected by SRS. The method is completely free from bias of investigator and it is completely dependent upon an element of chance. The sample may be selected by any of the following techniques.

- Simple Random Sampling Without Replacement (SRSWOR): This is the method of selecting  $n$  units out of  $N$  units such that each of  ${}^N C_n$  samples has an equal and independent chance of being selected. Here unit once selected is examined is not replaced back in the population before next selection is performed.

$$\text{Number of simple random samples in case of without replacement} = {}^N C_n = \frac{N!}{n!(N-n)!}$$

- Simple Random Sampling With Replacement (SRSWR): This is the method of selecting  $n$  units out of  $N$  units such that each of  $N^n$  samples has an equal and independent chance of being selected. Here unit once selected is examined and replaced back in the population before next selection is performed.

$$\text{Number of simple random samples in case of with replacement} = N^n$$

Tr: What are different methods of selecting simple random sample from the population? Or how will you administer simple random sampling technique?

St: A simple random sample can be drawn by any of the following methods.

(a) Using Lottery method: This is the simplest method of selecting SRS. This consists in indentifying each population unit by a distinctive number, which is recorded on a chits/slip of paper/card. Therefore, there are as many slips as total number of population units. All such slips are homogenous and identical in shape, colour and size because of criterion of randomness. Thus bunch of  $N$  slips represent the miniature population for the purpose of sampling. All the slips are kept in a bag/container. After through shuffling slips are selected one by one up to require number of times. The sampling unit corresponding to a number on a selected slip will constitute a random sample. Practically this method is difficult to use when size of population ( $N$ ) is very large.

(b) Rotatory disk

(c) Using Random Number tables:

Tr: In practice if population size small the lottery method or rotatry disk methods are frequently used but when population size is very big Random table are advisable. Let me tell you the way these random tables can be used.

A random number table is an arrangement of 0 to 9, in such a way that each number appears with approximately same frequency and independent of each other. Some random number tables in common use are listed below.

- i. Tippet's random number tables,
- ii. Fisher and Yates tables
- iii. Kendall and Smith tables
- iv. Tables of RAND Corporation
- v. A million random number digits.

Most popular random number tables are given by Fisher and Yates. Here steps for selecting random sample of size  $n$  from population of  $N$  units are:

Step-1 Each population unit is number from 0 to  $(N-1)$ .

Step-2 If  $N$  is a two digit figure, the units can be numbered as 00, 01, 02, ..., 98. In case  $N$  is three digit figure, the units can be numbered as 000, 001, 002, 003,..., 998 and so on. In this way the list of serially numbered population units is known as the sampling frame.



Step-3 Now the sample of size  $n$  can be selected. Let  $N$  be a  $d$ -digit number. Make use of  $d$ -digit random number table . Read numbers one by one from the random number table .

Step-4 If this observation is say  $K$ , is less than or equal to  $(N-1)$ , then select  $K^{\text{th}}$  unit ,

If  $K = N$  select the unit at serial number 00.

If  $K > N$  , then divide  $K$  by  $N$  and get the remainder ' $R$ '. Thus  $R^{\text{th}}$  unit is selected.

Step-5 Continue this process till  $n$  sampling units are selected.

e.g.  $N=18$  and  $n=5$  use 2 digit random numbers. Numbered all 18 units as 00,01, 02,...,17

Five random numbers from tables are say, 65, 43, 62, 54, 06.

On dividing 65(  $K > N$ ) the remainder is 11 so 11 th unit is selected. Similarly 43( $K > N$ ) the remainder is 7 so 7th unit is selected, 62( $K > N$ ) remainder is 8 so 8th unit is selected, 54( $K > N$ ) the remainder is 0 so 00th unit is selected, 06( $K < N-1$ ) so 6th unit is included in the sample.

Tr: What is Stratified Random Sampling?

St: According to the procedure of Stratified Random Sampling, whole population is first divided into  $k$  homogeneous, mutually exclusive and collectively exhaustive sub classes called STRATA or STRATUM. The procedure of division of heterogeneous population into relatively homogeneous sub classes is called STRATIFICATION. Now, from each stratum of the population of size  $N_1, N_2, N_3, \dots, N_k$  such that  $N_1 + N_2 + N_3 + \dots + N_k = N$  (population size). Select samples of size  $n_1, n_2, n_3, \dots, n_k$  such that  $n_1 + n_2 + n_3 + \dots + n_k = n$  (sample size) either by proportional allocation or by disproportional allocation. These samples of sizes  $n_1, n_2, n_3, \dots, n_k$  from each of  $k$  stratum respectively (of population) are called STRATIFIED SAMPLES. The process of selecting stratified random sample is called STRATIFIED RANDOM SAMPLING.

The sample taken in such a way represents the characteristic of the whole population and hence reliable results can be obtained by employing such a sampling technique. In most of the surveys the number of samples taken from each stratum, i.e. the sample size of each stratum is proportional to the size of stratum. But this is not necessary in case of all surveys.

Let us understand it with the help of an example. For determining the standard of statistics subject of the students, we can divide the number of students into different strata, vis. F.Y Students, S.Y Students, T.Y Students. From each class, i.e. from each stratum, we select randomly the number of students. When the three samples of each stratum are combined they form a single sample, which we call a stratified random sampling.

Tr: What are the advantages of stratified random sampling?

St: Advantages of Stratified Random Sampling:

- i. Since each stratum is internally homogeneous, we get reliable information about the stratum even when the sample size is small.
- ii. Since the population is divided into different strata, accurate information can be obtained from all the parts of the population.
- iii. When different standard of accuracy is required for different strata, this method is convenient.
- iv. In the case where the cost of survey is fixed this method helps in reduction of error.

Tr: What are the disadvantages of stratified random sampling?

St: Disadvantages of Stratified Random Sampling:

- i. It is not always possible to divide the population into homogeneous strata.
- ii. This method may lead to faulty results if the stratification carried out is improper.
- iii. The calculation involved in this method, in order to estimate the characteristics of the population is more, and so the method becomes difficult to use.
- iv. Simple random samples are to be taken from strata. In case of no-availability of the number of efficient persons the desired standard of accuracy cannot be attained.

Tr: How will you select sample from population using Stratified Random Sampling technique?

St: This is the Structure of Stratified Random Sampling Data:

St: Strata #	Population		Sample	
	Units	Size	Units	Size
1	$Y_{11}, Y_{12}, Y_{13}, \dots, Y_{1N_1}$	$N_1$	$y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}$	$n_1$
2	$Y_{21}, Y_{22}, Y_{23}, \dots, Y_{2N_2}$	$N_2$	$y_{21}, y_{22}, y_{23}, \dots, y_{2n_2}$	$n_2$
...	...	...	...	...
K	$Y_{k1}, Y_{k2}, Y_{k3}, \dots, Y_{kN_k}$	$N_k$	$y_{k1}, y_{k2}, y_{k3}, \dots, y_{kn_k}$	$n_k$
	Total	N	Total	N

There are two ways sample can be drawn by Stratified Random Sampling, (i) Proportional Allocation, and (ii) Disproportional allocation.

Allocation in Stratified Random Sampling : To obtain efficient results, the allocation of sample size  $n_i$  i.e. number of units selected in  $i$ th stratum ( $i = 1, 2, \dots, k$ ), is such that the total sample size denoted as  $n$ , can be  $n_1 + n_2 + n_3 + \dots + n_k = n$ . This can be done in the following ways.

(a) Proportional Allocation:

According to this allocation the units in the sample are selected from each population stratum in the same proportion, as they exist in the population. The allocation of the sample size is termed as proportional if the sample fraction i.e. the ratio of the sample size to the population size of  $i$ th stratum remains same in all the strata. Mathematically,

$$n_i \propto N_i \Rightarrow n_i = c N_i \text{ where } c \text{ is a constant of proportionality.}$$

We know that,  $n_1 + n_2 + n_3 + \dots + n_k = n$

$$\sum n_i = \sum c N_i$$

$$\Rightarrow n = c N_1 + c N_2 + c N_3 + \dots + c N_k$$

$$\Rightarrow n = c N$$

$$\Rightarrow c = n/N$$

Substitute value of  $c$  in  $n_i = c N_i$

$$\Rightarrow n_i = (n/N) N_i$$

Thus,

$$n_1 = \frac{n N_1}{N}, \quad n_2 = \frac{n N_2}{N}, \quad n_3 = \frac{n N_3}{N}, \quad \dots, \quad n_k = \frac{n N_k}{N}$$

(b) Disproportional allocation (Neyman Allocation):

In this case an equal number of units from each population stratum is selected regardless of how the stratum is represented in the population or proportion to variance of each stratum or fixing cost and minimizing variance or by other rule. In short, a stratified sample in which the number of units selected from each stratum is independent of its size is called disproportional allocation.

Tr: What is Systematic sampling?

St: Systematic sampling:

Systematic sampling is used when the information from cards or registers which are in serial order is collected or in case when a sample of trees from forest of houses in a city is needed. In this scheme first unit is selected randomly and the rest such at equal interval.

Tr: How will you select sample by Systematic sampling?

St: There are two ways of selection of sample by systematic sampling, Linear systematic sampling and another is Circular Systematic sampling.

- i. Linear Systematic Sampling: Suppose a population consists of  $N$  units and from this a systematic sample of  $n$  units is to be selected. Also  $N = kn$  i.e.  $N$  is multiple of  $n$  then Systematic sampling is known as Linear. Here select a random number between 1 to  $k$ , let it be  $j$  then  $j^{\text{th}}$  and every subsequent  $j+k, j+2k, \dots, j+(n-1)k^{\text{th}}$  positional units are selected. e.g.  $N=15$  and  $n=3$  the first unit is selected using random number table and every  $k = (N/n) = 5^{\text{th}}$  unit is selected. Let random number is 03 so  $3^{\text{rd}}$  unit and  $8^{\text{th}}(3+5)$  and  $13^{\text{th}}(3+(2*5))$  unit is selected in the sample
- ii. Circular Systematic sampling: This is applied when  $N \neq kn$ . Here  $k = \text{nearest integer of } N/n$ .

Eg. Select a Random Number between 1 to  $N$ . Let this number be  $m$ . Now select every  $(m+jK)$  th unit when  $m+jk < N$  and select every  $(m+jk-N)$  th unit when  $m+jk \geq N$  putting  $j = 1, 2, 3, \dots$  till  $n$  units are selected. e.g.  $N = 13$ ,  $n=4$  then  $k = 13/4 = \text{nearest integer of } 3.25 = 3$ . Now select random number between 1 to 13 let it be 09 then  $12^{\text{th}}$ ,  $2^{\text{nd}}$ ,  $5^{\text{th}}$  and  $8^{\text{th}}$  unit is selected.

Tr: Give some illustration of Systematic sampling.

St: Suppose a dissertation topic is “A Study into the Impact Leadership Style on Employee Motivation in ABC Company” and you have chosen semi-structured in-depth interview as primary data collection method. ABC Company has 200 operational level employees who could be potentially interviewed. Suppose our identified sample size as 24 subjects, i.e. you will interview 24 employees.

You will have to do the following:

**1. Label each employee with a unique number.**

**2. Calculate the sampling fraction.**

Sampling fraction = Actual Sample Size/Total Population =  $24/200 = 3/25$ .

This sampling fraction can be narrowed down to  $1/8$ . Accordingly, every  $8^{\text{th}}$  member of the sampling frame needs to be selected to participate in the study.

**3. Choose the first sample randomly.** Suppose you randomly selected the sample #47 as the starting point for selecting samples. Accordingly, your sample group will comprise of ABC Company employees under the following numbers: #47; #55; #63; #71; #79; #87; #95; #103; #111; #119; #127; #135; #143; #151; #159; #167; #175; #183; #191; #199; #7; #15; #23; #31.

Tr: What are advantages of systematic sampling?

St: Advantages of Systematic sampling are:

- i. The method of selection is very simple and is not very expensive.
- ii. The sample is evenly distributed over whole population and hence all contiguous parts of the population are represented in the sample.
- iii. It has an advantage over other sampling plans because it has managerial control of field work.

Tr: What are disadvantages of Systematic sampling?

St: Disadvantages of Systematic sampling are:

- i. If the variation in units is periodic i.e. the units at regular intervals are correlated, then the sample becomes highly biased. For example, if the houses are in blocks and if the first randomly selected house is corner house then all other houses selected in the sample will be corner ones and this definitely gives biased sample.
- ii. No single reliable formula for estimating standard error of sample mean is available.

Tr: What is Cluster Sampling?

St: In cluster sampling, groups of elementary units are formed generally on location, class or area basis. These groups are called Cluster. These clusters are referred as mini populations and have all the features of population. Then a simple random sample of few clusters is selected and all units in the selected clusters are studied. Here within the cluster units are homogenous and between the cluster units are heterogeneous.

Eg: An organization is looking to survey the performance of smart phones across India. They can divide the entire country's population into cities (clusters) and further select cities with the highest population and also filter those using mobile devices. This multiple stage sampling is known as cluster sampling. The selection of clusters can be done by using simple random sampling or by systematic random sampling.

Tr: How Stratified and Cluster samplings are different?

St: Difference between Stratified and Cluster sampling is this, in stratified random sampling, all the strata of the population is sampled while in cluster sampling, the researcher only randomly selects a number of clusters from the collection of clusters of the entire population. Therefore, only a number of clusters are sampled, all the other clusters are left unrepresented.

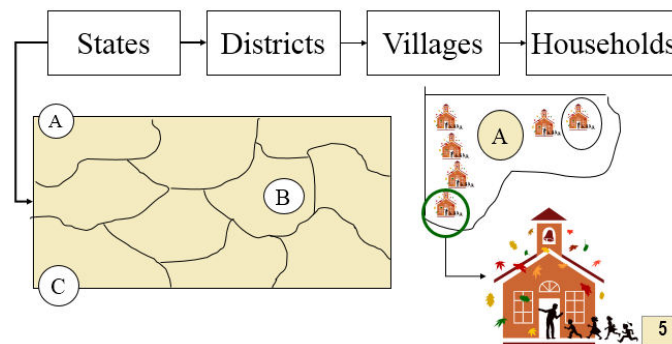
Tr: What is Multi-stage Sampling?

**St:** When a large scale surveys on district, state or national level are to be conducted, it is one of the most suitable sampling plan. For example, if the government or some other agency wants to collect information about the tax payer household in a city, it is better to select a sample of various mohallas (wards) or localities before selecting a sample of households from these selected localities. In this way, the localities are first stage unit and households are second stage unit. Such a sampling procedure is known as two-stage sampling.

Again if a survey is conducted to estimate the crop production of a district, it is preferable to select villages as first stage units, sample of farms in selected villages as second stage units, and select a sample of plots from each selected farms as third stage units. In this case selection procedure is known as three-stage sampling. Selection procedure can be extended to any number of stages. Hence the sampling in various stages in general is known as multi-stage sampling.

Another example: to evaluate online spending patterns of households in the US through online questionnaires. You can form your sample group comprising 120 households in the following manner:

- i. Choose 6 states in the USA using simple random sampling (or any other probability sampling).
  - ii. Choose 4 districts within each state using systematic sampling method (or any other probability sampling).
  - iii. Choose 5 households from each district using simple random or systematic sampling methods.
- This will result in 120 households to be included in your sample group.



**Tr:** What is Multiphase sampling?

**St:** Multiphase sampling is essentially, applying various tests for the characteristic on a population in series. So at each level, you would apply a characteristic which would decide who would be a part of the next subset. This way, your subsequent subsets will be smaller in number than the previous subsets.

Multiphase sampling is an extension of two-phase sampling, also known as double sampling. Multiphase sampling must be distinguished from multistage sampling since, in multiphase sampling, the different phases of observation relate to sample units of the same type, while in multistage sampling, the sample units are of different types at different stages.

St: Multiphase sampling is defined as, Method that collects basic information from a large sample of units and then, for a subsample of these units, collects more detailed information. The most common form of multi-phase sampling is two – phase sampling (or double sampling), but three or more phases are also possible.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group:**

**Group No.: 1 and 2**

Q1. State two illustrations when Simple Random Sampling can be used effectively.

Q2. State two illustrations when systematic sampling can be used effectively.

**Group No. 3 and 4**

Q3. State two illustrations when Stratified Random Sampling can be used effectively.

Q4. State two illustrations when Cluster Sampling can be used effectively.

**Group No. 5 and 6**

Q5. State two illustrations when Multistage Sampling can be used effectively.

Q6. State two illustrations when Multiphase Random Sampling can be used effectively.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Non- Probability sampling methods.

## Lesson No.14: Sampling Methods

### (Non-Probability Sampling Techniques)

#### Teaching Points:

- **Purposive Sampling**
- **Judgmental Sampling**
- **Convenient Sampling**
- **Quota Sampling**
- **Snow Ball Sampling**

#### Instructional Objectives:

After completion of this class students will be able to

- Explain Purposive Sampling with suitable illustrations.
- Explain Judgmental Sampling with suitable illustrations.
- Explain Convenient Sampling with suitable illustrations.
- Explain Quota Sampling with suitable illustrations.
- Explain Snow Ball Sampling with suitable illustrations.

#### Lesson Presentation:

Tr: Dear Students, The non-probability sampling techniques may also use explicitly in cases where it is not feasible to use probability based methods. The main defect of this sampling is that, in these techniques, the extent of bias in selecting a sample is not known. The choice of sampling units thus there is room for bias. This makes it difficult to say anything about the representativeness or accuracy of the sample. However, this method is not scientific one, but still it is widely used by research in solving their problems. However, if the investigator is experienced, skilled and this sampling is carefully applied, then may yield valuable results.

Let me ask from you, what are different types of Non-Probability Sampling Techniques.

St: Purposive Sampling

St: Judgmental Sampling

St: Convenient Sampling

St: Quota Sampling

St: Snow Ball Sampling

Tr: Explain about purposive sampling?



St: Purposive Sampling: In purposive sampling researcher chooses a sample based on their knowledge about the population and the study itself. The study participants are chosen based on the study's purpose. Participants are selected according to the needs of the study (hence the alternate name, *deliberate* sampling), applicants who do not meet the profile are rejected. For example, you may be conducting a study on why high school students choose community college over university. You might canvas high school students and your first question would be "Are you planning to attend college?" People who answer "No," would be excluded from the study.

Tr: Do you know various types of Purposive Sampling techniques?

St: Critical Case Sampling: collecting cases which are likely to give the most information about the phenomenon you are studying.

St: Expert Sampling: Sampling to include only those with expertise in a certain area.

St: Extreme Case Sampling: this technique focuses on participants with unique or special characteristics.

St: Homogeneous Sampling: collecting a very specific set of participants. For example, age 20 to 24 years of college educated male students.

St: Maximum Variation Sampling: collecting a wide range of participants with different viewpoints to study a certain phenomenon. So that researcher can uncover with some common themes.

St: Typical Case Sampling: allows the researcher to develop a profile about what is normal or average for a particular phenomenon.

Tr: Explain about Judgment Sampling?

St: Judgment Sampling: In the judgmental sampling method, researchers select the samples based purely on the researcher's knowledge and credibility. In other words, researchers choose only those people who they deem fit to participate in the research study. It is expected that these samples would be better as the experts are supported to know the population. However, as the use of randomness is not there and moreover there is no way to find the accuracy of the samples, hence the method has its limitations and is used mainly for situations requiring extremely small size of samples, i.e. use of rare events, members having extreme positions, etc.

Tr: Judgmental or purposive sampling is not a scientific method of sampling and the downside to this sampling technique is that the preconceived notions of a researcher can influence the results. Thus, this research technique involves a high amount of ambiguity.

Tr: Explain about Convenient Sampling?

St: Convenient Sampling: Convenience sampling is a non-probability sampling technique where samples are selected from the population only because they are conveniently available to the researcher. Researchers choose these samples just because they are easy to recruit, and the researcher did not consider selecting a sample that represents the entire population. Ideally, in research, it is good to test a sample that represents the population. But, in some research, the population is too large to examine and consider the entire population. It is one of the reasons why researchers rely on convenience sampling, which is the most common non-probability sampling method, because of its speed, cost-effectiveness, and ease of availability of the sample.

St: The selection of sample is left to the researcher who is to select the sample. The researcher normally interviews persons in groups at some retail outlet, supermarket or may stand at prominent point and interview the persons who happen to there. This type of sampling is also called 'accidental sampling' as the respondents in the sample are included merely because their presence on the spot. The data collection and sample cost is minimum in this sampling. However method suffers greatly from the quality. This type of sampling is more suitable in exploratory research where the focus is on getting new ideas/insights into a given problem.

Eg: An example of convenience sampling would be using student volunteers known to the researcher. Researchers can send the survey to students belonging to a particular school, college, or university, and they would act as a sample in this situation.

For example, you could divide a population by the state they live in, income or education level, or sex. The population is Quota Sampling?

St: Quota sampling means to take a much tailored sample that's in proportion to some characteristic or trait of a population.

Quota Sampling: this sampling involves the fixation of certain quotas, which are to be fulfilled by the interviewers. For example, you could divide a population by the state they live in, income or education level, or sex. The population is divided into segments on the basis of certain characteristics. Here the segments are termed as cells. A quota of unit is selected from each cell. The units are selected without randomization from each cell. But units are selected from each cell in the predetermined proportion.

Tr: State the advantages and disadvantages of Quota sampling.

St: Advantages:

- i. Quota sampling does not require prior knowledge about the cell to which population unit belongs. Therefore this sampling has a distinct advantage over the stratified sampling where every population units must be placed in the appropriate stratum before the sample selection.
- ii. It is simple to administer. Sampling can be done very quickly.
- iii. The necessity of the researcher going to various geographical locations is avoided and thus cost is reduced.

St: Disadvantages:

- i. It may not possible to get a representative sample within the quota as the selection depends entirely on the mood and convenience of the interviewer.
- ii. Since too much liberty is being allowed to the interviewer, the quality of work suffers if they are not competent.

Tr: What is Snow Ball Sampling?

St: Snowball Sampling: Snowball sampling helps researchers find a sample when they are difficult to locate. Researchers use this technique when the sample size is small and not easily available. In this method, the initial group of respondents is selected randomly. Subsequent respondents are being selected based on the opinion or referrals provided by the initial respondents. Further referrals will lead to more referrals, thus leading to snowball sampling. The referrals will have demographic and psychographic characteristics that are relatively similar to the person referring them. Researcher asks them for assistance to seek similar subjects to form a considerably good size sample.

Tr: Give some illustration of Snow ball sampling.

St: A Research study on a particular illness in patients or a rare disease. Now researcher can seek help from subjects to refer to other subjects suffering from the same ailment to form a subjective sample to carry out the study.

St: Researches on Prostitutes.

St: Researches on Gays.

St: Researches on lesbians.

St: Researches on homosexuals, etc.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group:**

**Group No.: 1, 3, 5**

Q1. State two illustrations when Purposive Sampling can be used effectively.

Q2. State two illustrations when Judgmental sampling can be used effectively.

Q3. State two illustrations when Convenient Sampling can be used effectively.

**Group No.: 2, 4, 6**

Q1. State two illustrations when Quota sampling can be used effectively.

Q2. State two illustrations when Snow Ball Sampling can be used effectively.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Meaning of inference in statistical Analysis and Basic terms used in inferential statistics.

## Lesson No.15: Introduction to Inferential Statistics -I

### Teaching Points:

- Meaning of inference
- Basic terms used in inferential statistics

### Instructional Objectives:

After completion of this class students will be able to:

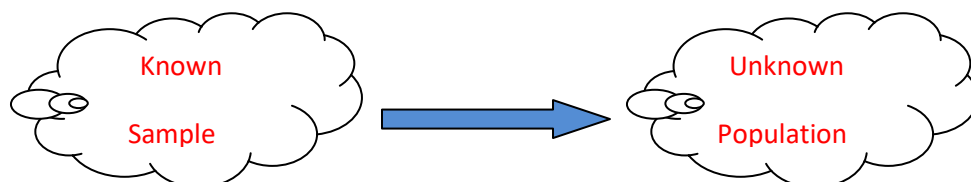
- Explain the meaning of inference in Statistical data analysis.
- Explain the following terms:
  - Parameter and statistic
  - Hypothesis and its types
  - Level of Significance
  - Degrees of Freedom
  - Sampling Distribution
  - Standard Error
  - Sampling Error.

### Lesson Presentation:

Tr: Dear Students, today we will study about inferential statistics. Do you know the meaning of inference?

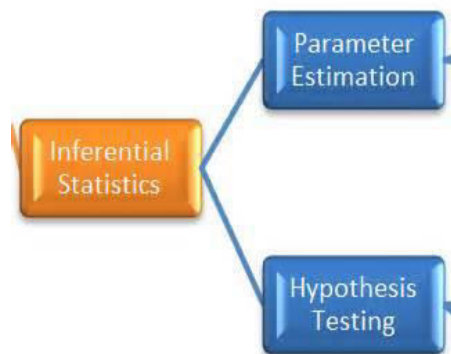
St: The English dictionary meaning of inference is conclusion.

Tr: Yes, that's true. But in inferential statistics its meaning is quite specific. The population may be real or imaginary, it may be finite, countable or unaccountably infinite. It may be homogenous or heterogeneous. When the population is homogeneous it is not necessary to test sample for its significances. It gives complete information about the population. When population is heterogeneous then the one particular sample will not give complete information about the population. So, here problem of inference is arises. In statistics, Inference refers to the process of selecting and using sample information to draw conclusion about a population parameter. It is the process of moving to unknown population from known sample.



According to nature of problem and techniques statistical inference is divided in main two parts.

- (1) Estimation: Here population parameter is estimated on the basis of sample information.
- (2) Hypothesis Testing: Here some hypothetical statement is made about population parameter and it is tested at certain significance level whether the made hypothesis is right or wrong on the basis of sample information.



### Estimation

Statistical technique of estimating unknown population parameters like mean, variance, correlation and regression coefficient etc. from the corresponding sample is referred as *Estimation*. e.g. a manufacturer may be interested in estimating the average life of his product, proportion of defective item in his lot, average demand of his product etc. The Two ways of estimation of parameters are suggested.

- i. Point Estimation: A single value of a statistic (function of sample observation) that is point estimate is called an *estimator* and the value of statistic is the *estimate*.

Let,  $\mu$  = population mean

.  $\sigma$  = population s.d

$x_1, x_2, \dots, x_n$  be the sample of size  $n$  then

$\bar{x}$  = sample mean

$s$  = sample s.d.

Then sample mean  $\bar{x}$  is an estimator of  $\mu$  and  $s$  is estimator of  $\sigma$ .

Different samples will have generally different estimates. But a good estimator must be closer to the true value of the parameter as far as possible.

- ii. Interval Estimation: Unlike point estimates, in interval estimation it is desired to find out an interval which is expected to include the unknown parameter with a specified probability. The interval estimation method consists of the determination of two values of parameter say  $t_1$  and  $t_2$  such that

$$P(t_1 < \theta < t_2) = 1 - \alpha$$

Where  $\alpha$  = is the level of significance

$\theta$  = Unknown parameter

The limits  $t_1$  and  $t_2$ , so determined are known as confidence limits.

$(1 - \alpha)$  100% confidence limits (C.L) for  $\theta$  is

$$t \pm S.E.(t) t_{\alpha}$$

Where,  $t$  is the sample statistic.

Tr: What is hypothesis Testing?

St: Theory of testing of hypothesis employs statistical technique to arrive at decision in certain situations where there is element of uncertainty on the basis of a sample whose size is fixed in advance.

Tr: What do you mean by Hypothesis?

St: The term hypothesis is nothing but some assumption made about a parameter. In other words a hypothesis in statistics is simply a quantitative statement about a population. It is denoted by 'H'.

Tr: What are different types of hypothesis?

St: Simple hypothesis: and composite hypothesis.

St: Null hypothesis and alternative hypothesis?

St: Directional hypothesis and non – directional hypothesis.

Tr: Elaborate each term.

St: If the hypothesis completely specifies the population then it is called a simple hypothesis.

e.g.

- $H : \mu = 30, \sigma = 25$  in case of normal distribution
- $H: X \sim B (n = 5, P = 0.5)$
- $H: X \sim P (\lambda = 1.5)$

St: If the hypothesis does not completely specifies the population then it is called composite hypothesis e.g .

- $H: \mu = \mu_0, \sigma$  is unknown
- $H: n$  is unknown,  $P = p_0$
- $H: \mu > \mu_0, \sigma = \sigma_0$
- $H: \mu < \mu_0, \sigma = \sigma_0$
- $H: \mu = \mu_0, \sigma > \sigma_0$

St: A statistical hypothesis which is stated for the purpose of possible acceptance is call Null Hypothesis. In other words null hypothesis is the hypothesis of no difference between true value and assume value of parameter. It is denoted by  $H_0$ .

Alternative hypothesis:

St: Any hypothesis which is complementary to the null hypothesis is called as alternative hypothesis and usually denoted by  $H_1$ .

e.g. if  $H_0 : \mu = 163 \text{ cm} = \mu_0, \sigma$  is known then  $H_1$  could be

- $H_1 : \mu \neq 163$  (Two tailed alternative)
- $H_1 : \mu > 163$  (one tailed right sided alternative)
- $H_1 : \mu < 163$  (one tailed left sided alternative )

Tr: What is critical region?

St: Critical region or Rejection region:



Suppose several samples of the same size from a given population are taken. For each of these samples some statistic  $t$  is computed. Let  $t_1, t_2, \dots, t_k$  be the values of the statistic for these samples. Each of these values may be used to test some null hypothesis  $H_0$ . Some values of  $t$  lead to the rejection of  $H_0$ , while others may lead to acceptance of  $H_0$ . Thus statistic values  $t_1, t_2, \dots, t_k$  may be divided into mutually disjoint groups. One leads to rejection of null hypothesis and other leading to acceptance of null hypothesis. The values of statistic that lead to rejection of null hypothesis is called rejection region or Critical region (  $C$  ), while those lead to acceptance of null hypothesis is called acceptance region (  $A$  ). Thus, if the value of statistic  $t \in C$ ,  $H_0$  is rejected and if value of statistic  $t \in A$ ,  $H_0$  is accepted. (where  $C$  and  $A$  are complementary sets of each other,  $C \cap A = \emptyset$ ,  $C \cup A = S$  sample space).

Tr: In brief, A region (corresponding to a statistic  $t$  ) in the sample space  $S$  in which null hypothesis is rejected is term as *critical region*. The region under the normal curve which is not covered by the rejection region is known as *Acceptance region*. The value of the test statistic computed to test the null hypothesis is known as the *Critical value*. The critical value separates the rejection region from acceptance region.

Tr: What do you understand for Errors in Testing?

St: The decision to accept or reject null hypothesis  $H_0$  is taken on the basis of information supplied by sample data. There some error is involved. There are four possibilities with hypothesis testing.

- (i) Hypothesis is true but we reject it
- (ii) Hypothesis is true and we accept it
- (iii) Hypothesis is false but we accept it
- (iv) Hypothesis is false & we reject it

In case of (ii) and (iv) we are not committing any error.

Tr: You can better understand from this Decision table from sample:

	<b><math>H_0</math> is Accepted</b>	<b><math>H_0</math> is Rejected</b>
<b><math>H_0</math> is True</b>	No Error	Type I Error
<b><math>H_0</math> is False</b>	Type II Error	No Error

Type I error: When we reject the correct Null hypothesis

Type II error: When we accept the wrong Null hypothesis

$$P(\text{Type I error}) = P(\text{Reject } H_0 \text{ when } H_0 \text{ is true}) = P(\text{Reject } H_0 / H_0 \text{ is true})$$

$$= P(t \in C / H_0)$$

$$= \alpha$$

$$P(\text{Type II error}) = P(\text{Accept } H_0 \text{ when } H_0 \text{ is false}) = P(\text{Accept } H_0 / H_0 \text{ is false})$$

$$= P(t \in A / H_1)$$

$$= p(\text{Accept } H_0 / H_1 \text{ is true})$$

$$= \beta$$

Thus

$$\alpha = P(\text{Rejecting good lot})$$

$$= \text{Producer's Risk}$$

$$\beta = P(\text{Accepting bad lot})$$

$$= \text{consumer's Risk}$$

Note: (1) Practically it is not possible to minimize both these errors simultaneously.

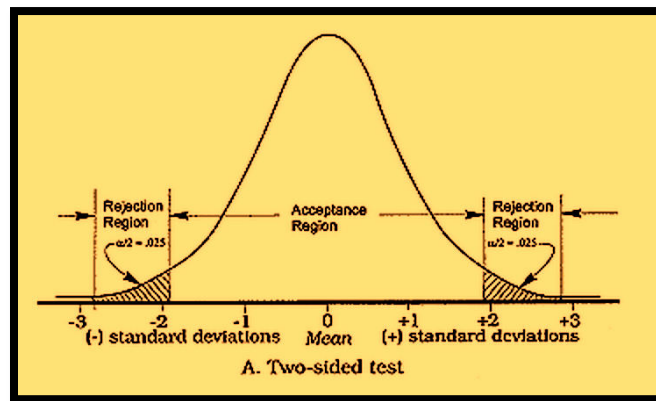
(2) In most of decision making problems in business and social science it is more risky to accept a wrong hypothesis than to reject a correct one i.e. consequences of Type II errors are more likely to be more serious than the Type I error. Therefore it is more advisable to minimize more serious error after fixing up the less serious error. Thus fix  $\alpha$  i.e.  $P(\text{Type I error})$  and minimize  $P(\text{Type II error})$ .

Tr: Depending upon rejection region, whether on one side or two side there are two types of tests: two tailed test and one tailed test.

What do you understand with Two tailed test or Two sided Test?

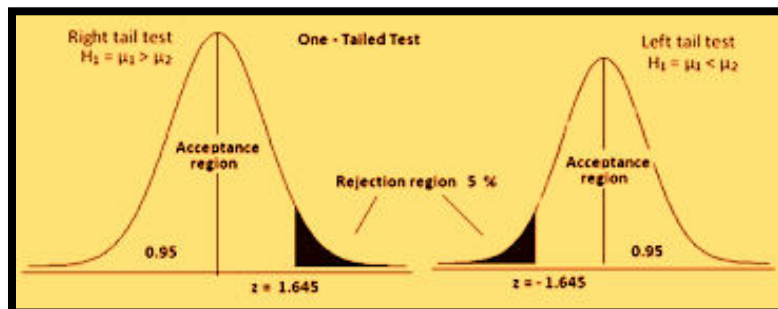
St: When the test of hypothesis is made on the basis of rejection region represented by both sides of the sampling distribution curve then it is called two tailed test. In other words, a test of statistical hypothesis where the alternative hypothesis  $H_1$  is two sided such as:

Null Hypothesis  $H_0 : \mu = \mu_0$  against Alternate Hypothesis  $H_1 : \mu \neq \mu_0$  ( $\mu > \mu_0$  and  $\mu < \mu_0$ ) is called two tailed test.



Tr: What do you understand with One tailed Test or One sided Test?

St: A test of statistical hypothesis, where the alternative hypothesis is one sided is called one tailed or one sided test. There are two types of one tailed test.



- (i) Right tailed test: Here the rejection or critical region lies entirely on the right tailed of the sampling distribution curve.

e.g. Null Hypothesis  $H_0 : \mu = \mu_0$  against Alternative Hypothesis  $H_1 : \mu > \mu_0$ .

- (ii) Left tailed test: Here the rejection or critical region lies entirely on the left tailed of the sampling distribution curve.

e.g. Null Hypothesis  $H_0 : \mu = \mu_0$  against Alternative Hypothesis  $H_1 : \mu < \mu_0$

Tr: What is Level of Significance?

St: The level of significance, usually denoted by  $\alpha$  (alpha), is specified before the samples are drawn, so that the results so obtained should not influence the choice of the decision-maker. It is

specified in terms of the maximum probability of making type I error. Popular levels of significance are 1%, 5%, 10 %, depending upon accuracy desired. 5% level of significance means in 5 samples out of 100, we are likely to reject  $H_0$  whenever  $H_0$  is true i.e. we are 95% confident that our decision to reject  $H_0$  is correct.

Tr: You have already studied about Population and Sample. Just recapitulate their meanings.

St: The target group under investigation, about which we want to get information, is *population*.

St: Population is the totality of persons, objects, items etc. corresponding to certain characteristics. The part of the population pertaining to which data is available is called a *sample*. The drop of blood examined in the laboratory is a sample from the population of all blood in the body.

Tr: Similarly define Parameter and Statistics also?

St: An exact but generally unknown measure (or value), which describe the entire population or process characteristics is called *parameter*. For example, quantities such as mean  $\mu$ , variance  $\sigma^2$ , standard deviation ( $\sigma$ ), median, mode and proportion  $P$  computed from population data set are called *parameter*.

St: A measure (or value) found from analyzing sample data is called a *statistic*. For example, quantities such as sample mean ( $\bar{x}$ ), sample variance ( $s^2$ ), sample standard deviation ( $s$ ), sample median, sample mode and sample proportion ( $p$ ) computed from sample data set are called *statistic*.

Tr: The value of every statistic varies randomly from one sample to another whereas the value of a parameter is considered as constant. The value for statistic calculated from any sample depends on the particular random sample drawn from a population. Inferential statistical methods attempt to estimate population parameters using sample statistic.

Population data  $X_1, X_2, \dots, X_N$

Population size =  $N$

Examples of Parameter :

$$\mu = \text{population mean} = \frac{\sum X}{N}$$

$$\sigma = \text{population standard deviation} = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

$$P = \text{Population proportion} = \frac{X}{N}$$

$$\text{where, } P = \frac{\text{Number of population observation having particular characteristic}}{\text{Total number of population observation}}$$

Sample data  $x_1, x_2, \dots, x_n$

Sample size =  $n$

Examples of Statistic :

$$\bar{x} = \text{sample mean} = \frac{\sum x}{n}$$

$$s = \text{sample standard deviation} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

$$p = \text{Sample proportion} = \frac{\text{Number of sample observation having particular characteristic}}{\text{Total number of sample observation}} = \frac{x}{n}$$

Tr: What is Sampling distribution?

St: If a population is very large and the description of its characteristics is not possible by the census method, then to arrive at the statistical inference, sample of a given size are drawn repeatedly from the population and a particular '*statistic*' is computed for each sample. The computed value of a particular statistic will differ from sample to sample. Thus, theoretically it is possible to construct a frequency table showing the values assumed by the statistic and the frequency of their occurrence. This distribution of values of a statistic is called a *sampling distribution*. Since the values of statistic are the result of several simple random samples,

therefore statistic is a random variable. Such sampling distributions are: Z-distribution, Student's t-distribution, chi-square  $\chi^2$ -distribution, F-distribution.

For example; let  $X_1, X_2, \dots, X_N$  are population having probability distribution as Normal with population mean  $\mu$  and population standard deviation  $\sigma$ , then statistic sample mean ( $\bar{x}$ ) has probability distribution as Normal with mean  $\mu$  and standard deviation  $\frac{\sigma^2}{n}$ . Here Normal probability distribution with mean  $\mu$  and standard deviation  $\frac{\sigma^2}{n}$  is called sampling distribution of statistic  $\bar{x}$ . Similarly sampling distribution of sample variance when population mean ( $\mu$ ), is Chi-square probability distribution with  $n$  degree of freedom.

In notation  $X \sim N(\mu, \sigma^2)$  then sampling distribution of statistic  $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$  when population mean ( $\mu$ ) known.

Tr: What is a difference between Standard Error (S.E) of statistic and standard deviation (S.D.)?

St: (probably students won't answer this therefore teacher will explain this with some illustration.)

Tr: Standard deviation of sampling distribution of sample statistic measures sampling error and is also known as *standard error of statistic*. While, scatteredness of the population observations from population mean is measured by standard deviation of population.

e.g. (i) Standard Error of statistic sample mean ( $\bar{x}$ ) is  $\frac{\sigma}{\sqrt{n}}$  or  $\frac{s}{\sqrt{n}}$

(ii) Standard error of statistic sample proportion ( $p$ ) is  $S.E(p) = \sqrt{\frac{pq}{n}}$

Thus, the population standard deviation describes the variation among values of the members of the population, whereas the standard deviation of sampling distribution (standard error) measures the variability among values of the sample statistic (such as mean values, proportion values) due to sampling error.

So, we can conclude that the difference between S.D. and S.E is that the former concern original values and the latter concerns the statistic computed from sample of original values.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group: (All Groups)**

**In the multiple choice questions choose the most appropriate option from the given choices:**

- 1) If you drew all possible samples from some population, calculated the mean for each of the samples, and constructed a line graph (showing the shape of the distribution) based on all of those means, what would you have?
  - a. A population distribution
  - b. A sample distribution
  - c. A sampling distribution
  - d. A parameter distribution
  
- 2) What does it mean when you calculate a 95% confidence interval?
  - a. The process you used will capture the true parameter 95% of the time in the long run.
  - b. You can be “95% confident” that your interval will include the population parameter.
  - c. You can be “5% confident” that your interval will not include the population parameter
  - d. All of the above statements are true
  
- 3) What would happen (other things equal) to a confidence interval if you calculated a 99 percent confidence interval rather than a 95 percent confidence interval?
  - a. It will be narrower
  - b. It will not change
  - c. The sample size will increase
  - d. It will become wider
  
- 4) Which of the following statements sounds like a null hypothesis?
  - a. The coin is not fair
  - b. There is a correlation in the population
  - c. There is no difference between male and female incomes in the population
  - d. The defendant is guilty
  
- 5) What is the standard deviation of a sampling distribution called?
  - a. Sampling error
  - b. Sample error
  - c. Standard error
  - d. Simple error

- 6) A \_\_\_\_\_ is a subset of a \_\_\_\_\_.
- a. Sample, population
  - b. Population, sample
  - c. Statistic, parameter
  - d. Parameter, statistic
- 7) A \_\_\_\_\_ is a numerical characteristic of a sample and a \_\_\_\_\_ is a numerical characteristic of a population.
- a. Sample, population
  - b. Population, sample
  - c. Statistic, parameter
  - d. Parameter, statistic
- 8) A sampling distribution might be based on which of the following?
- a. Sample means
  - b. Sample correlations
  - c. Sample proportions
  - d. All of the above
- 9) As a general rule, researchers tend to use \_\_\_\_\_ percent confidence intervals.
- a. 99%
  - b. 95%
  - c. 50%
  - d. none of the above
- 10) Which of the following is the researcher usually interested in supporting when he or she is engaging in hypothesis testing?
- a. The alternative hypothesis
  - b. The null hypothesis
  - c. Both the alternative and null hypothesis
  - d. Neither the alternative or null hypothesis
- 11) When  $p < .05$  is reported in a journal article that you read for an observed relationship, it means that the author has rejected the null hypothesis (assuming that the author is using a significance or alpha level of .05).
- a. True
  - b. False
- 12) When  $p > .05$  is reported in a journal article that you read for an observed relationship, it means that the author has rejected the null hypothesis (assuming that the author is using a significance or alpha level of .05).
- a. True
  - b. False



13) \_\_\_\_\_ are the values that mark the boundaries of the confidence interval.

- a. Confidence intervals
- b. Confidence limits
- c. Levels of confidence
- d. Margin of error

14) \_\_\_\_\_ results if you fail to reject the null hypothesis when the null hypothesis is actually false.

- a. Type I error
- b. Type II error
- c. Type III error
- d. Type IV error

15) A good way to get a small standard error is to use a \_\_\_\_\_.

- a. Repeated sampling
- b. Small sample
- c. Large sample
- d. Large population

16) The car will probably cost about 16,000 dollars; this number sounds more like a(n):

- a. Point estimate
- b. Interval estimate

17) A \_\_\_\_\_ is a range of numbers inferred from the sample that has a certain probability of including the population parameter over the long run.

- a. Hypothesis
- b. Lower limit
- c. Confidence interval
- d. Probability limit

18) The use of the laws of probability to make inferences and draw statistical conclusions about populations based on sample data is referred to as \_\_\_\_\_.

- a. Descriptive statistics
- b. Inferential statistics
- c. Sample statistics
- d. Population statistics

19) The cutoff the researcher uses to decide whether to reject the null hypothesis is called the:

- a. Significance level
- b. Alpha level
- c. Probability value
- d. Both a and b are correct

20) As sample size goes up, what tends to happen to 95% confidence intervals?

- a. They become more precise
- b. They become narrower
- c. They become wider
- d. Both a and b

21) What is the key question in the field of statistical estimation?

- a. Based on my random sample, what is my estimate of the population parameter?
- b. Based on my random sample, what is my estimate of normal distribution?
- c. Is the value of my sample statistic unlikely enough for me to reject the null hypothesis?
- d. There is no key question in statistical estimation

22) Assuming innocence until “proven” guilty, a Type I error occurs when an innocent person is found guilty.

- a. True
- b. False

23) This is the difference between a sample statistic and the corresponding population parameter.

- a. Standard error
- b. Sampling error
- c. Difference error
- d. None of the above

24) ‘Children can learn a second language faster before the age of 7’. Is this statement:

- a. A non-scientific statement
- b. A one-tailed hypothesis
- c. A two-tailed hypothesis
- d. A null hypothesis

25) The p-value used in statistical significance testing should be used to assess how strong a relationship is. For example, if relationship A has a  $p=.04$  and relationship B has a  $p=.03$  then you can conclude that relationship B is stronger than relationship A.

- a. True
- b. False

26) What is the relationship between sample size and the standard error of the mean?

- a. The standard error decreases as the sample size decreases.
- b. The standard error is unaffected by the sample size.
- c. The standard error increases as the sample size increases.
- d. The standard error decreases as the sample size increases.

27) If my experimental hypothesis were 'Eating cheese before bed affects the number of nightmares you have', what would the null hypothesis be?

- a. Eating cheese before bed gives you more nightmares.
- b. Eating cheese before bed gives you fewer nightmares.
- c. Eating cheese is linearly related to the number of nightmares you have.
- d. The number of nightmares you have is not affected by eating cheese before bed.

28) If my null hypothesis is 'Dutch people do not differ from English people in height', what is my alternative hypothesis?

- a. All of the statements are plausible alternative hypotheses.
- b. Dutch people are taller than English people.
- c. English people are taller than Dutch people.
- d. Dutch people differ in height from English people.

### **Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read the difference between parametric and non parametric tests and Steps of doing hypothesis testing.

## Lesson No.16: Introduction to Inferential Statistics -II

### Teaching Points:

- Difference between parametric and non parametric tests.
- Steps of doing hypothesis testing.

### Instructional Objectives:

- i. State the major steps of doing hypothesis testing.
- ii. Differentiate between parametric and non parametric tests.

### Lesson Presentation:

Tr: Dear students, You must have gone through with many tests used for hypothesis testing. But you can definitely observe that there are certain steps which every hypothesis testing is following. So, today we will study the steps of doing hypothesis test. Let me ask from you, what should be those steps?

St: Following are the steps in testing of hypothesis

- I. Formulating null hypothesis ( $H_0$ ) and Alternate hypothesis ( $H_1$ ).** Alternative hypothesis ( $H_1$ ) will help in deciding whether the test is one sided or two sided.
- II. Fixing Level of significance:** Choose appropriate level of significance ( $\alpha$ ) depending on the reliability of the estimates and permissible risk. This is to be decided before sample is drawn. The commonly used levels of significances in practice are 5% and 1%. If 5% level of significance is used, it means that the probability of making type one error is 0.05. In other words, there is 95 % confident that a correct decision has been made.
- III. Computation of test statistics.** The test statistic is a statistic based on appropriate probability distribution. It is used to test whether the null hypothesis is accepted or rejected.

Eg: Most commonly used test statistic is Z-statistic, t-statistic etc.

The Z-statistic is defined as 
$$Z = \frac{t - E(t)}{S.E.(t)}$$

- IV. Establish Critical region.** The region in which null hypothesis is rejected is called critical region. The value of the sample statistic that separates the region of rejection

and acceptance is called critical value. Generally critical value is obtained from table of sampling distribution of test statistic for given parameter and level of significance.

- V. **Defining Rejection Criteria of Null Hypothesis:** Formulate the Decision rule to accept or reject the null hypothesis. Rejection Criteria depends upon whether the test is one sided or two sided.
- VI. **Drawing inference:** Compare the calculated value of the test statistic with the critical value. The decision rule for the null hypothesis is Accept  $H_0$  if the test statistic value falls within the area of acceptance and reject otherwise.

Tr: These steps of hypothesis testing are applicable for both the types of tests i.e. Parametric and non parametric tests. Do you know, what is a parametric and non parametric test?

To make the generalization about the population from the sample, statistical tests are used. These hypothetical testing related to differences are classified as parametric and nonparametric tests. On one hand, the parametric test is one which has information about population parameter while on the other hand, nonparametric tests, is one where the researcher has no idea regarding the population parameter.

The parametric test is the hypothesis test which provides generalizations for making statements about the mean of the parent population. A t-test based on Student's t-statistic, which is often used in this regard. The t-statistic rests on the underlying assumption that there is the normal distribution of variable and the mean is known or assumed to be known. The population variance is calculated for the sample. It is assumed that the variables of interest, in the population are measured on an interval scale.

The nonparametric test is defined as the hypothesis test which is not based on underlying assumptions, i.e. it does not require population's distribution to be denoted by specific parameters. The test is mainly based on differences in medians. Hence, it is alternately known as the distribution-free test. The test assumes that the variables are measured on a nominal or ordinal level. It is used when the independent variables are non-metric.

State the differences you know.

St: The Key Differences between Parametric and Nonparametric Tests are following:

- I. A statistical test, in which specific assumptions are made about the population parameter is known as the parametric test. A statistical test used in the case of non-metric independent variables is called nonparametric test.
- II. In the parametric test, the test statistic is based on distribution. On the other hand, the test statistic is arbitrary in the case of the nonparametric test.
- III. In the parametric test, it is assumed that the measurement of variables of interest is done on interval or ratio level. As opposed to the nonparametric test, wherein the variable of interest are measured on nominal or ordinal scale.
- IV. In general, the measure of central tendency in the parametric test is mean, while in the case of the nonparametric test is median.
- V. In the parametric test, there is complete information about the population. Conversely, in the nonparametric test, there is no information about the population.
- VI. The applicability of parametric test is for variables only, whereas nonparametric test applies to both variables and attributes.
- VII. For measuring the degree of association between two quantitative variables, Pearson's coefficient of correlation is used in the parametric test, while spearman's rank correlation is used in the nonparametric test.

Tr: **Parametric statistics** is a branch of statistics which assumes that sample data come from a population that can be adequately modeled by a probability distribution that has a fixed set of parameters. Conversely a **non-parametric model** differs precisely in that the parameter set is not fixed and can increase, or even decrease, if new relevant information is collected.

Most of the statistical tests we perform are based on a set of assumptions. When these assumptions are violated the results of the analysis can be misleading or completely erroneous. State those typical assumptions used for parametric tests.

St: typical assumptions used for parametric tests are following:

- **Normality:** Data have a normal distribution (or at least is symmetric)
- **Homogeneity of variances:** Data from multiple groups have the same variance
- **Linearity:** Data have a linear relationship
- **Independence:** Data are independent

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group: (All Groups)**

Q1. Differentiate Between Parametric and non parametric Test in tabular form.

Q2. State the equivalent non-parametric test for the given parametric test mention below:

Independent t - test (difference between the means of two independent groups), Paired t- test, One way ANOVA, Repeated Measures ANOVA, Pearsons' correlation coefficient, Simple linear regression, Assessing the relationship between two categorical variables.

Answer:

Parametric tests	Non-Parametric tests
Independent t - test (difference between the means of two independent groups)	Mann Whitney Test
Paired t- test	Wilcoxon Signed Rank test
One way ANOVA	Kruskal Wallis Test
Repeated Measures ANOVA	Friedman Test
Pearsons' correlation coefficient	Spearman's rank correlation
Simple linear regression	Not applicable
For Assessing the relationship between two categorical variables – no applicable in parametric test.	Chi- square test

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about Z- distribution and their applications and hypothesis testing for mean (one sample problem and two sample problem).

## PARAMETRIC TESTS

### Lesson No.17: Z-test and its applications

#### Teaching Points:

- Z- distribution
- Applications of Z- distribution.
- Testing for mean (one sample problem)
- Testing for mean (two sample problem)

#### Instructional Objectives:

After completion of this class students will be able to

- Describe Z- distribution.
- Use Z- test for hypothesis testing of population mean for one sample problem.
- Use Z- test for hypothesis testing of population mean for two samples problem.
- Infer the results after using Z- test for hypothesis testing of population mean for one sample and two sample problems.

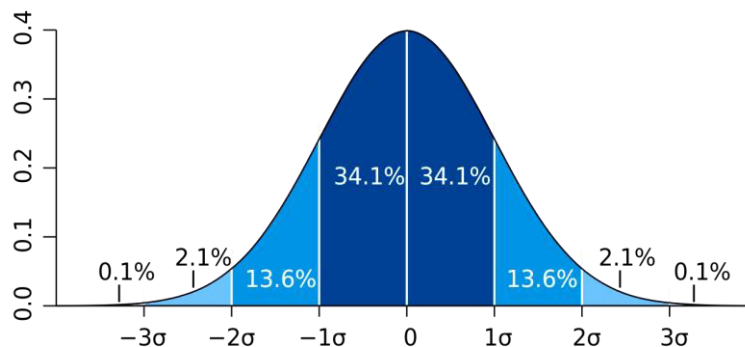
#### Lesson Presentation:

Tr: Dear students Z- distribution is one of the most exploited distribution in the entire Sampling distributions family.

What is a Z- distribution?

St: Z- distribution is also known as standard normal distribution (SND).

SND is a special case of normal distribution with mean  $\mu = 0$  and s.d.  $\sigma = 1$ . Following graph shows the curve of SND (or Z- distribution):



Tr: Who proposed normal distribution and what is the other name of this distribution?



St: The **normal distribution** is also called **Gaussian distribution** because it was discovered by Carl Friedrich Gauss.

Tr: What is a Z- test?

St: A **Z-test** is any statistical test for which the distribution of the test Statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples.

Tr: What are the applications or uses of Z- test?

St: z-test is based on the normal probability distribution and is used for judging the significance of several statistical measures, particularly the mean. The relevant test statistic\*,  $z$ , is worked out and compared with its probable value (to be read from table showing area under normal curve) at a specified level of significance for judging the significance of the measure concerned. This is a most frequently used test in research studies. This test is used even when binomial distribution or  $t$ -distribution is applicable on the presumption that such a distribution tends to approximate normal distribution as ' $n$ ' becomes larger.

- Z-Test is generally used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when population variance is known.
- Z-Test is also used for judging the significance of difference between means of two independent samples in case of large samples, or when population variance is known.
- Z-Test is also used for comparing the sample proportion to a theoretical value of population proportion or for judging the difference in proportions of two independent samples when  $n$  happens to be large.
- Z- Test may be used for judging the significance of median, mode, coefficient of correlation and several other measures.

Tr: Now we will learn hypothesis testing for mean (for one sample problem and two sample problem).

➤ **Testing mean of single Normal Population:**

$\mu$ : population mean	$\sigma$ : population s.d.	
$\bar{x}$ : sample mean	$s$ : sample s.d.	$n$ : sample size

→  $\sigma$  is Known

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \mu = \mu_0$	(1) $H_1 : \mu > \mu_0$ (2) $H_1 : \mu < \mu_0$ (3) $H_1 : \mu \neq \mu_0$	$Z_{cal} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	(1) $Z_{cal} > Z_{\alpha}$ (2) $Z_{cal} < -Z_{\alpha}$ (3) $ Z_{cal}  > Z_{\alpha/2}$

→  $\sigma$  is unknown, large sample ( $n \geq 30$ )

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \mu = \mu_0$	(1) $H_1 : \mu > \mu_0$ (2) $H_1 : \mu < \mu_0$ (3) $H_1 : \mu \neq \mu_0$	$Z_{cal} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	(1) $Z_{cal} > Z_{\alpha}$ (2) $Z_{cal} < -Z_{\alpha}$ (3) $ Z_{cal}  > Z_{\alpha/2}$

➤ Testing Means of two Independent normal populations (Two population test)

→  $\sigma_1$  and  $\sigma_2$  Known

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \mu_1 = \mu_2$	(1) $H_1 : \mu_1 > \mu_2$ (2) $H_1 : \mu_1 < \mu_2$ (3) $H_1 : \mu_1 \neq \mu_2$	$Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	(1) $Z_{cal} > Z_{\alpha}$ (2) $Z_{cal} < -Z_{\alpha}$ (3) $ Z_{cal}  > Z_{\alpha/2}$

→  $\sigma_1$  and  $\sigma_2$  unknown, large sample ( $n_1 \geq 30$  or  $n_2 \geq 30$ )

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \mu_1 = \mu_2$	(1) $H_1 : \mu_1 > \mu_2$ (2) $H_1 : \mu_1 < \mu_2$ (3) $H_1 : \mu_1 \neq \mu_2$	$Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	(1) $Z_{cal} > Z_{\alpha}$ (2) $Z_{cal} < -Z_{\alpha}$ (3) $ Z_{cal}  > Z_{\alpha/2}$

Tr: In the above tables' null hypothesis, alternate hypothesis, test statistic and rejection criteria are shown. Let us take some examples to learn hypothesis testing:

Tr: Example (1) From Bajaj insurance company an agent has claimed that the average age of the policy holders who insure through him is less than the average age for all the agents, which is 30.5 years. A random sample of 100 policy holders who had insured through him has mean 28.8 years and standard deviation 6.35 years. Test his claim at 5% level of significance. (Given table value 1.645).

Null hypothesis:  $H_0: \mu = 30.5$  years (claim is not true) against

Alternative Hypothesis:  $H_0: \mu < 30.5$  years (claim is true)

Sample mean:  $\bar{x} = 28.8$  Years

Sample standard deviation:  $s = 6.35$  years

Sample observations:  $n = 100$  (large sample, population s.d. ( $\sigma$ ) unknown)

Test statistic 
$$Z_{cal} = \frac{\bar{x} - \mu_o}{s/\sqrt{n}} = \frac{28.8 - 30.5}{6.35/\sqrt{100}} = -2.677$$

Rejection Criteria: Reject  $H_0$  if  $Z_{cal} < -Z_{tab}$ ,

Here  $Z_{tab} = Z_{0.05} = 1.645$

i.e. Reject  $H_0$  if  $Z_{cal} < -1.645$

Conclusion: Here  $Z_{cal} = -2.677 < -1.645$ .

So we reject null hypothesis at 5% level of significance and conclude that claim of agent is true. That is, the average age of the policy holders who get insurance through him is less than the average age for all the agents, which is 30.5 years.

Tr: Example (2): The average breaking strength of the cables supplied by a manufacturer is 1800 with s.d. 100. Through some new modification technique in manufacturing process, test whether the average breaking strength of the cables increased or not. In order to test this, a sample of 50 cables is examined. It is found that the mean breaking strength is 1850. Use  $\alpha = 0.01$ . (Given table value 2.33)

St: Solution: Null hypothesis:  $H_0: \mu = 1800$  against

Alternative Hypothesis:  $H_0: \mu > 1800$

Sample mean:  $\bar{x} = 1850$

Population standard deviation:  $\sigma = 100$

Sample observations:  $n = 50$  (population s.d. ( $\sigma$ ) known)

$$\text{Test statistic } Z_{cal} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.5355$$

Rejection Criteria: Reject  $H_0$  if  $Z_{cal} > Z_{tab}$

Here  $Z_{tab} = Z_{0.01} = 2.33$

i.e Reject  $H_0$  if  $Z_{cal} > 2.33$

Conclusion: Here  $Z_{cal} = 3.5355 > 2.33$

So, we reject null hypothesis at 5% level of significance and conclude that by a new technique in manufacturing process, the mean breaking strength of the cable increased.

Tr: Example (3): The average production of wheat from a sample of 100 fields in 200 lbs. per acre with a s.d. of 10 lbs. Another sample of 150 fields gives the average of 220 lbs. with a s.d. of 12 lbs. Can the two samples be considered to have been taken from the same population whose s.d. is 11 lbs? Use 5% level of significance.

Solution: Taking the null hypothesis that the means of two populations do not differ, we can write  $H_0 : \mu_1 = \mu_2$

$$H_1: \mu_1 \neq \mu_2$$

$$n_1 = 100; n_2 = 150;$$

$$\bar{x}_1 = 200 \text{ lbs.}; \quad \bar{x}_2 = 220 \text{ lbs.}$$

$$\sigma_{s1} = 10 \text{ lbs.} \quad \sigma_{s2} = 12 \text{ lbs.} \quad \text{and } \sigma_p = 11 \text{ lbs.}$$

$$\text{Test Statistic: } Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_p^2 \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

$$Z_{cal} = \frac{200 - 220}{\sqrt{11^2 \left[ \frac{1}{100} + \frac{1}{150} \right]}} = -14.08$$

$$|Z_{\text{cal}}| = 14.08$$

$$Z_{\text{tab}} = Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$$

Rejection Criteria: Reject  $H_0$  if  $|Z_{\text{cal}}| > Z_{\alpha/2}$

Conclusion: Since if  $|Z_{\text{cal}}| > Z_{\alpha/2}$  at 5% level of significance we reject  $H_0$  and conclude that the two samples have been taken from the different populations.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### Task allotment for each group: (All Groups)

1. A random sample of hand gloves worn by 200 combat soldiers in a hill region showed an average life of 2.2 years with a standard deviation of 0.04. Under the standard conditions, the hand gloves are known to have an average life 2.53 years. Is there reason to assert at a level of significance of 0.05 that use in the hills causes the average life of such hand gloves to decrease?

2. A simple random sampling survey in respect of weekly wages earned by balloon sailors' workers in two cities gives the following statistical information:

City	Weekly Earning	s.d. of sample data of weekly earning	Sample Size
Mumbai	2560	259	400
Ahmadabad	2678	342	370

Test the hypothesis at 5 per cent level that there is no difference between weekly earnings of workers in the two cities.

### Presentation and Discussion:

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read hypothesis testing for proportion (one sample and two sample problems).

## Lesson No.18: Z-test Application

### Teaching Points:

- Testing for proportion (one sample problem)
- Testing for proportion (two sample problem)

### Instructional Objectives:

After completion of this class students will be able to:

- i. Use z-test for hypothesis testing of population proportion for one sample problem.
- ii. Use z-test for hypothesis testing of population proportion for two sample problem.
- iii. Infer the results after using Z- test for hypothesis testing of population proportion for one sample and two sample problems.

### Lesson Presentation:

Tr: Dear students, today we will learn about hypothesis testing of population proportion for one sample problem. and two sample problems. Following tables will help you to decide about null hypothesis, alternative hypothesis, test statistic and rejection criteria for a given problem.

#### ➤ Testing for proportion (one population test)

P: population proportion

X: no. of observations in favour of certain characteristics in population

N: population size

$p$ : sample proportion

$x$ : no. of observation in favour of certain characteristics in sample

$n$ : sample size

$P: X / N$

$P: x / n$

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : P = P_0$	(1) $H_1 : P > P_0$ (2) $H_1 : P < P_0$ (3) $H_1 : P \neq P_0$	$Z_{cal} = \frac{p - P_0}{\sqrt{P_0 Q_0 / n}}$ where $Q_0 = 1 - P_0$	(1) $Z_{cal} > Z_{\alpha}$ (2) $Z_{cal} < - Z_{\alpha}$ (3) $ Z_{cal}  > Z_{\alpha/2}$

Tr: Solve the following example:

A famous producer of an advertising company claims that 40% of the people saw an advertisement put out on the television by the company, remembered the brand name of the product even 24 hours after they had seen the show. In a sample survey conducted 24 hours after the show found that 152 out of 400 persons remembered the brand name of the product advertised. Test if the claim of the company at 1 % level of significance. (Table value 2.57)

St: Solution:

Null hypothesis:  $H_0: P = 0.4$  (claim) against

Alternative Hypothesis:  $H_0: P \neq 0.4$

$$n = 400, \quad x = 152, \quad p = \frac{x}{n} = \frac{152}{400} = 0.38$$

$$\text{Test Statistic: } Z_{cal} = \frac{p - P_0}{\sqrt{P_0 Q_0 / n}} = \frac{0.38 - 0.4}{\sqrt{(0.4 \times 0.6) / 400}} = -0.8165$$

*where  $Q_0 = 1 - P_0$*

$$\text{Rejection Criteria: Reject } H_0 \text{ if } |Z_{cal}| > Z_{\alpha/2}$$

$$Z_{\alpha/2} = Z_{0.01/2} = Z_{0.005} = 2.57$$

$$\text{i.e. Reject } H_0 \text{ if } |Z_{cal}| > 2.57$$

Conclusion: Here  $|Z_{cal}| = 0.8165 < 2.57$ . So we do not reject null hypothesis at 1 % level of significance and conclude that claim of company is true. That is, 40% of the people who saw an advertisement put out on the television by the company, remembered the name of the product even after 24 hours they had seen the show.

➤ **Testing for differences in proportions of two samples (two population test)**

$P_1$ : proportion of success in population one

$P_2$ : proportion of success in population two

$X_1$ : no. of observations in favour of certain characteristics in 1<sup>st</sup> population

$X_2$ : no. of observations in favour of certain characteristics in 2<sup>nd</sup> population

$N_1$ : 1<sup>st</sup> population size

$N_2$ : 2<sup>nd</sup> population size

$p_1$ : proportion of success in sample one

$p_2$ : proportion of success in sample two

$x_1$ : no. of observations in favour of certain characteristics in sample taken from 1<sup>st</sup> population

$x_2$ : no. of observations in favour of certain characteristics in sample taken from 2<sup>nd</sup> population

$n_1$ : sample size from 1<sup>st</sup> population

$n_2$ : sample size from 2<sup>nd</sup> population

$P_1$ :  $X_1 / N_1$        $P_2$ :  $X_2 / N_2$

$P_1$ :  $x_1 / n_1$        $P_2$ :  $x_2 / n_2$

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : P_1 = P_2$	(1) $H_1 : P_1 > P_2$ (2) $H_1 : P_1 < P_2$ (3) $H_1 : P_1 \neq P_2$	$Z_{cal} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left\{ \frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2} \right\}}}$ <p>Is used in case of heterogeneous populations. But when populations are similar with respect to a given attribute, we work out the best estimate of the population proportion as under:</p> $p_0 = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ <p><math>q_0 = 1 - p_0</math> in which case we calculate test statistic as:</p> $Z_{cal} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0 q_0 \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}}$	(1) $Z_{cal} > Z_{\alpha}$ (2) $Z_{cal} < - Z_{\alpha}$ (3) $ Z_{cal}  > Z_{\alpha/2}$

Tr: Example: A clinical industry is testing two drugs newly developed to reduce blood pressure levels. The drugs are administered to two different sets of animals. In group one, 350 of 600



animals tested respond to drug A and in group two, 260 of 500 animals tested respond to drug B. The clinical industry wants to test whether there is a difference between the efficacies of the said two drugs at 5 % level of significance.

St: Null hypothesis:  $H_0: P_1 = P_2$

Alternate hypothesis:  $H_1: P_1 \neq P_2$

$$p_1 = 350/600 = 0.583 \quad q_1 = 1 - p_1 = 1 - 0.583 = 0.417$$

$$p_2 = 260/500 = 0.520 \quad q_2 = 1 - p_2 = 1 - 0.520 = 0.480$$

$$n_1 = 600 \quad n_2 = 500$$

Test Statistic:

$$Z_{cal} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left\{ \frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2} \right\}^{1/2}}}$$

$$Z_{cal} = \frac{0.583 - 0.520}{\sqrt{\frac{0.583 \times 0.417}{600} + \frac{0.520 \times 0.480}{500}}} = 2.093$$

$$Z_{table} : Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$$

Rejection Criteria: Reject  $H_0$  if  $|Z_{cal}| > Z_{\alpha/2}$ .

Conclusion: Since  $Z_{cal} = 2.093 > 1.96 = Z_{\alpha/2}$  at 5% level of significance we reject  $H_0$  and conclude that there is a significant difference between the efficacies of the said two drugs A and B.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### Task allotment for each group: (All Groups)

1. A certain process produces 6% defective articles (bottles). A supplier of new raw material claims that the use of his material would reduce the proportion of defectives. A random sample of 450 units using this new material was taken out of which 27 were defective units. Can the supplier's claim be acceptable at 1% level of significance?

2. A railway ticket Master said that 30% of the passengers go in first class, but management recognizes the possibility that this percentage could be more or less. A random sample of 1550 passengers includes 430 passengers holding first class tickets. Can the claim of railway ticket Master be rejected at 5% level of significance?

3. In a City Mall on a certain day 156 out of a random sample of 1500 men were found to be smoking. After awareness program on effects of smoking another sample was drawn after three days at the same City Mall. Sample was of 1900 men and it was found that 120 were smokers. Was the observed decrease in the proportion of smokers significant? Test at 5% level of significance.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read hypothesis testing for t- distribution and its Applications, Testing for mean for one sample problem and for two sample problem, Confidence Interval.

## Lesson No.19: t- test and its applications

### Teaching Points:

- **t- Distribution.**
- **Applications of t- distribution.**
- **Testing for mean (one sample problem)**
- **Testing for mean (two sample problem)**
- **Confidence Interval**

### Instructional Objectives:

After completion of this class students will be able to:

- i. Use t-test for hypothesis testing of population mean for one sample problem.
- ii. Use t-test for hypothesis testing of population mean for two sample problem.
- iii. Infer the results after using t- test for hypothesis testing of population mean for one sample and two sample problems.

### Lesson Presentation:

Tr: What did you know about t- distribution?

St: The t- distribution, also known as the Student's t-distribution.

St: t-distribution is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails.

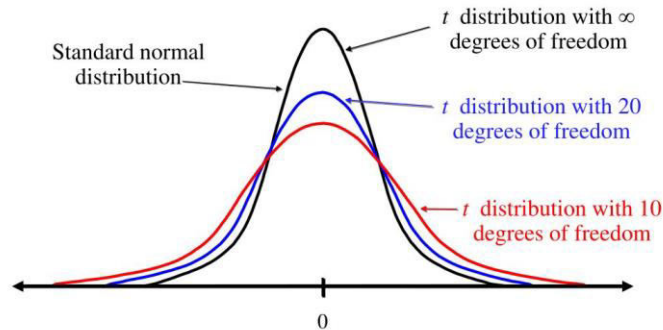
St: t- distributions have a greater chance for extreme values than normal distributions, hence the fatter tails. Tail heaviness is determined by a parameter of the t- distribution called degrees of freedom, with smaller values giving heavier tails and with higher values making the t-distribution resemble a standard normal distribution with a mean of 0 and a standard deviation of 1.

Tr: You are right. Then, what is difference between a t - distribution and a normal distribution?

St: Normal distributions are used when the population distribution is assumed to be normal. The t distribution is similar to the normal distribution, just with fatter tails. Both assume a normally distributed population. t - Distribution has higher kurtosis than normal distributions. The probability of getting values very far from the mean is larger with a t distribution than a normal distribution. Following graph will help you to understand when n tends to infinity or very large, t- distribution converges to standard normal distribution.

### $t$ Distribution

The  $t$ -distribution is used when  $n$  is **small** and  $\sigma$  is **unknown**.



Tr: What is  $t$ -test? State their applications.

St:  $t$ -test is based on  $t$ -distribution and is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of small sample(s) when population variance is not known (in which case we use variance of the sample as an estimate of the population variance).

St: In case two samples are related, we use paired  $t$ -test (or what is known as difference test) for judging the significance of the mean of difference between the two related samples.

St: It can also be used for judging the significance of the coefficients of simple and partial correlations. The relevant test statistic,  $t$ , is calculated from the sample data and then compared with its probable value based on  $t$ -distribution (to be read from the table that gives probable values of  $t$  for different levels of significance for different degrees of freedom) at a specified level of significance for concerning degrees of freedom for accepting or rejecting the null hypothesis.

Tr: It may be noted that  $t$ -test applies only in case of small sample(s) when population variance is unknown.

Following table will help you to learn  $t$ -test for Testing of hypothesis for mean for one sample and two sample problems.

➤ **Testing for mean for one sample:**

→ **population s.d.( $\sigma$ ) unknown and small sample( $n < 30$ )**

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \mu = \mu_0$	(1) $H_1 : \mu > \mu_0$ (2) $H_1 : \mu < \mu_0$ (3) $H_1 : \mu \neq \mu_0$	$t_{cal} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$	(1) $t_{cal} > t_{n-1, \alpha}$ (2) $t_{cal} < -t_{n-1, \alpha}$ (3) $ t_{cal}  > t_{n-1, \alpha/2}$

Tr: Let us solve some examples.

Example: Mohan Samosa Stall near the railway station at Vadodara has been having average sales of 500 per day. Because of the development of City Mall nearby, it expects to increase its sales. During the first 12 days after the start of the City Mall, the daily sales were as under: 550, 570, 490, 615, 505, 580, 570, 460, 600, 580, 530, 526. On the basis of this sample information, can one conclude that Mohan Samosa's sales have increased? Use 5% level of significance.

St: Solution: Null hypothesis  $H_0: \mu = 500$

Alternative hypothesis  $H_1: \mu > 500$  (as we want to conclude that sales have increased)

Test Statistics: 
$$t_{cal} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Sl. No.	Xi	$(Xi - \bar{x})$	$(Xi - \bar{x})^2$
1	550	2	4
2	570	22	484
3	490	-58	3364
4	615	67	4489
5	505	-43	1849
6	580	32	1024
7	570	22	484
8	460	-88	7744
9	600	52	2704
10	580	32	1024
11	530	-18	324
12	526	-22	484
<b>Total</b>	<b>6576</b>		<b>23978</b>

$$n = 12 \quad \bar{x} = 548 \quad \sigma = \sqrt{\frac{23978}{12-1}} = 46.68 \quad \text{Degree of freedom} = n - 1 = 12 - 1 = 11$$

$$t_{\text{cal}} = \frac{548 - 500}{\frac{46.68}{\sqrt{12}}} = 48 / 13.49 = 3.558$$

$$t_{\text{tab}} = t_{n-1, \alpha} = t_{11, 0.05} = 1.796$$

Rejection Criteria: Reject  $H_0$  if  $t_{\text{cal}} > t_{n-1, \alpha}$ .

Conclusion: Since  $t_{\text{cal}} = 3.558 > t_{n-1, \alpha} = 1.76$  we reject  $H_0$ . at 5% level of significance we conclude that Mohan Samosa's sales has significantly increased or claim of the shopkeeper is true.

Tr: Example: A random sample of 36 New Delhi civil service personnel, the average age and the sample Standard deviation was found to be 40 years and 4.5 years respectively. Construct a 95% Confidence interval for the mean age of civil servants in New Delhi.

Solution: Here  $n = 36$ ,  $\bar{X} = 40$  years,  $\sigma_s = 4.5$  years

$Z_{0.05} = 1.96$  (as per the normal curve area table).

Thus, 95% confidence interval for the mean age of population is:

$$\bar{X} \pm Z_{\alpha} \frac{\sigma_s}{\sqrt{n}}$$

$$\text{Therefore, } 40 \pm 1.96 \frac{4.5}{\sqrt{36}} = 40 \pm 1.47 = (38.53, 41.47)$$

At 95 % C.I. the mean age of civil servants in New Delhi is lies in (38.53, 41.47).

### ➤ Testing of means of two independent normal population means

→ Population s.d. unknown but equal i.e.  $\sigma_1 = \sigma_2$  but unknown, small samples ( $n_1 < 30$  &  $n_2 < 30$ )

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \mu_1 = \mu_2$	(1) $H_1 : \mu_1 > \mu_2$ (2) $H_1 : \mu_1 < \mu_2$ (3) $H_1 : \mu_1 \neq \mu_2$	$t_{\text{cal}} = \frac{\bar{x}_1 - \bar{x}_2}{s^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ <p>where</p> $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	(1) $t_{\text{cal}} > t_{v, \alpha}$ (2) $t_{\text{cal}} < -t_{v, \alpha}$ (3) $ t_{\text{cal}}  > t_{v, \alpha/2}$  where, $v = n_1 + n_2 - 2$

Tr: Example: A Sample of sales in similar shops in two towns are taken for a new product with the following results:

Town	Mean Sales	Variance	Size of sample
A	57	5.3	5
B	61	4.8	7

Is there any evidence of difference in sales in the two towns? Use 5% level of significance for testing this difference between the means of two samples. Construct 95% C.I. for the difference between the means of sales of two towns.

St: Null hypothesis  $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Sample from town A as sample one	$\bar{x}_1 = 57$	$\sigma_{s_1}^2 = 5.3$	$n_1 = 5$
Sample from town B As sample two	$\bar{x}_2 = 61$	$\sigma_{s_2}^2 = 4.8$	$n_2 = 7$

Since in the given question variances of the population are not known and the size of samples is small, we shall use t-test for difference in means, assuming the populations to be normal.

Test statistic t:

$$t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{s * \sqrt{n}} = \frac{57-61}{\sqrt{5*0.3428}} = -4 / 1.309 = -3.05576$$

$$|t_{cal}| = 3.05576$$

where

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(5-1)*5.3 + (7-1)*4.8}{5+7-2} = 5$$

$$s = 2.236$$

$$n = (1/n_1) + (1/n_2) = (1/5) + (1/7) = 0.342857$$

$$t_{\text{tab}} = t_{v, \alpha/2} = t_{10, 0.05/2} = 2.228$$

where,

$$\text{d.f.} = v = n_1 + n_2 - 2 = 5 + 7 - 2 = 10$$

Rejection Criteria: Reject  $H_0$  if  $|t_{\text{cal}}| > t_{v, \alpha/2}$

Conclusion: Since  $|t_{\text{cal}}| = 3.05576 > 2.228 = t_{\text{tab}}$  we reject  $H_0$  at 5% level of significance and conclude that there is significant difference in the sales of the product in two towns.

The 95% C.I. for the difference between the means of sales of two towns is given by:

When  $n_1 > 30$  and  $n_2 > 30$ :

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

When  $n_1 < 30$  and  $n_2 < 30$ :

$$(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2, \alpha} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$\begin{aligned} \text{Therefore, } (61 - 57) \pm t_{n_1+n_2-2, \alpha\%} * 2.236 * 0.342857 \\ = 4 \pm 2.228139 * 0.7666 \\ = 4 \pm 1.7081 \\ = (2.291, 5.708) \end{aligned}$$

(2.291, 5.708) is the 95% C.I. for the difference between the means of sales of two towns.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.



**Task allotment for each group: (All Groups)**

1. The specimen of Silver wires drawn from a large lot have the following breaking strength (in kg weight): 522, 542, 550, 548, 562, 558, 530, 542, 556, 534 Test whether the mean breaking strength of the lot may be taken to be 540 kg weight. Test at 2% level of significance.
2. The management of big bazar market wanted to investigate whether the male customers spend more money on average than the female customers. A sample of 220 male customers who shopped at this supermarket showed that they spend an average of Rs. 530 with s.d. of Rs.96. Another sample of 300 female customers, who shopped at the same big bazaar, showed that they spend an average of Rs. 266 with s.d. of Rs.80. Assume that the amounts spent at this big bazar by all male and female customers are normally distributed with equal but unknown population standard deviation. Construct 95% C.I. for the difference between the mean amount spent by all male and female.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read hypothesis testing for paired observations and testing for correlation (one sample test)

## Lesson No.20: t- test Applications

### Teaching Points:

- Paired t – test
- Testing for Correlation (one sample problem)

### Instructional Objectives:

After completion of this class students will be able to

- Use t-test for hypothesis testing for differences of paired observations in population.
- Use t-test for hypothesis testing of population correlation coefficient for one sample problem.
- Infer the results after using paired t- test.
- Infer the results of hypothesis testing of population correlation coefficient for one sample problem.

### Lesson Presentation:

Tr: Dear Students, when did we use paired t – test?

St: Paired t-test is a way to test for comparing two related samples, involving small values of n that does not require the variances of the two populations to be equal, but the assumption that the two populations are normal must continue to apply.

St: For a paired t-test, it is necessary that the observations in the two samples be collected in the form of what is called matched pairs i.e., “each observation in the one sample must be paired with an observation in the other sample in such a manner that these observations are somehow ‘matched’ or related, in an attempt to eliminate extraneous factors which are not of interest in test.”

Tr: Such a test is generally considered appropriate in a before-and-after-treatment study. For instance, we may test a group of certain students before and after training in order to know whether the training is effective, in which situation we may use paired t-test. Following table will help you to know about the null and alternate hypothesis, test statistic and rejection criteria.

→ Testing of means of two dependent normal population (Paired t-test)

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \mu_D = 0$  Where, $\mu_D = \mu_1 - \mu_2$	(1) $H_1 : \mu_D > 0$ (2) $H_1 : \mu_D < 0$ (3) $H_1 : \mu_D \neq 0$	$t_{cal} = \frac{\bar{d}}{s_d / \sqrt{n}}$ where, $d = x_1 - x_2$ $\bar{d} = \sum d / n$ $s_d = \sqrt{\frac{\sum d^2}{n} - \bar{d}^2}$	(1) $t_{cal} > t_{n-1, \alpha}$ (2) $t_{cal} < -t_{n-1, \alpha}$ (3) $ t_{cal}  > t_{n-1, \alpha/2}$

Tr: Let us take an example.

Example: Memory capacity of 9 students was tested before and after training. Test at 5 percent level of significance whether the training was effective from the following scores:

Student	1	2	3	4	5	6	7	8	9
Before	10	15	9	3	7	12	16	17	4
After	12	17	8	5	6	11	18	20	3

St: Solution: Null hypothesis:  $H_0: \mu_D = 0$       Where,  $\mu_D = \mu_1 - \mu_2$

$H_1: H_1 : \mu_D < 0$

$$t_{cal} = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{-0.77777}{1.617802 / \sqrt{9}} = -1.44229$$

where,

Test Statistic:

$$d = x_1 - x_2$$

$$\bar{d} = \sum d / n = -0.77778$$

$$s_d = \sqrt{\frac{\sum d^2}{n} - \bar{d}^2} = 1.617802$$

$$t_{cal} = -1.44229$$

$$t_{tab} = t_{n-1, \alpha} = t_{8, 0.05} = 1.860$$

$$-t_{n-1, \alpha} = -1.860$$

Student No.	Before (X1)	After (X2)	d= X1-X2	d <sup>2</sup>	d- $\bar{d}$	(d- $\bar{d}$ ) <sup>2</sup>
1	10	12	-2	4	-1.22222	1.493822
2	15	17	-2	4	-1.22222	1.493822
3	9	8	1	1	1.77778	3.160502
4	3	5	-2	4	-1.22222	1.493822
5	7	6	1	1	1.77778	3.160502
6	12	11	1	1	1.77778	3.160502
7	16	18	-2	4	-1.22222	1.493822
8	17	20	-3	9	-2.22222	4.938262
9	4	3	1	1	1.77778	3.160502
<b>Total</b>			<b>-7</b>	<b>49</b>		<b>23.55556</b>

Rejection Criteria: Reject  $H_0$  if  $t_{cal} < -t_{n-1, \alpha}$

Conclusion: Since  $t_{cal} = -1.44229$  is not greater than  $-t_{n-1, \alpha} = -1.860$  we do not reject  $H_0$  at 5% level of significance and conclude that there is no significant difference between before and after scores of students. Hence training is ineffective. In other words, we should conclude that the training was not effective.

### ➤ Testing for Correlation Coefficient (r):

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \rho = 0$  Where, $\rho$ is population correlation coefficient	(1) $H_1 : \rho > 0$ (2) $H_1 : \rho < 0$ (3) $H_1 : \rho \neq 0$	$t_{cal} = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$  Where r: simple correlation coefficient d.f. = n-2	(1) $t_{cal} > t_{n-2, \alpha}$ (2) $t_{cal} < -t_{n-2, \alpha}$ (3) $ t_{cal}  > t_{n-2, \alpha/2}$

Tr: Example: The correlation coefficient between achievement scores of Mathematics and Science subjects for sample of 27 students of IX standard students is 0.76. Using 5% level of

significance test whether the linear correlation coefficient between achievement scores of Mathematics and Science subject scores is positive. Assume that the population of both variables is normally distributed.

St: Null hypothesis  $H_0 : \rho = 0$  (Correlation is not significant)

Alternate hypothesis  $H_1 : \rho > 0$  (Correlation is significant)

Test Statistic:

$$t_{cal} = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t_{cal} = \frac{0.76*5}{\sqrt{1-0.5776}} = 5.7099$$

$$t_{tab} = t_{n-2, \alpha} = t_{25, 0.05} = 2.059539$$

Rejection Criteria: Reject  $H_0$  if  $t_{cal} > t_{n-2, \alpha}$

Conclusion: Since  $t_{cal} = 5.7099 > t_{tab} = 2.059539$  we reject null hypothesis and conclude that there is significant positive correlation between achievement scores of Mathematics and Science of IX std. students.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### Task allotment for each group: (All Group)

1. The sales data of an item in ten shops before and after a special promotional campaign are:

Shops	A	B	C	D	E	F	G	H	I	J
Before the promotional Campaign	50	25	30	46	57	45	33	44	32	43
After the campaign	57	28	37	57	67	48	39	46	39	58

Can the campaign be judged to be a success? Use paired t-test at 5% level of significance.

2. The correlation coefficient between achievement scores of Economics and Statistics for a sample of 25 students is 0.69. Using 1% level of significance test whether the linear correlation coefficient between achievement scores of Economics and Statistics is positive. Assume that the population of both variables is normally distributed.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read F- distribution and its application, ANOVA.

## Lesson No.21: F- test and its application: ANOVA

### Teaching Points:

- F- distribution
- Applications of F- distribution.
- ANOVA

### Instructional Objectives:

After completion of this class students will be able to:

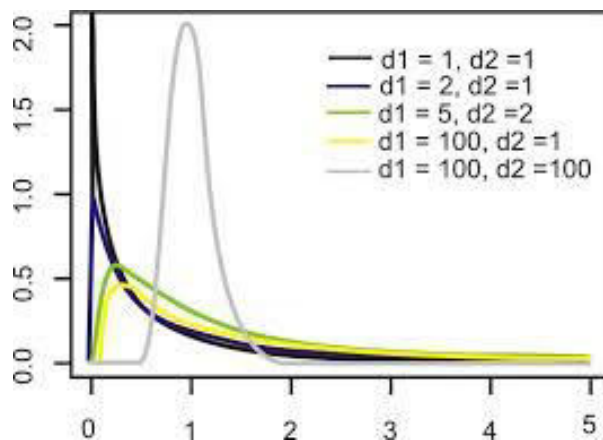
- Describe F- distribution.
- Explain the concept of ANOVA.
- Draw ANOVA table for one way and two way analysis.
- Use f-test for hypothesis testing of comparison of means from various samples.
- Infer the results of hypothesis testing of comparison of means from various samples.

### Lesson Presentation:

Tr: What do you know about F- distribution?

St: **F distribution** is a probability density function that is used especially in analysis of variance and is a function of the ratio of two independent random variables each of which has a chi-square distribution and is divided by its number of degrees of freedom.

Tr: The graphical representation of F-distribution is given below:



For F- distribution there are two degrees of freedom  $d1$  and  $d2$ .

Tr: What is F-test?

St: F-test is based on F-distribution and is used to compare the variance of the two-independent samples.

St: This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means at one and the same time.

St: It is also used for judging the significance of multiple correlation coefficients.

Test statistic,  $F$ , is calculated and compared with its probable value (to be seen in the  $F$ -ratio tables for different degrees of freedom for greater and smaller variances at specified level of significance) for accepting or rejecting the null hypothesis.

Tr: ANOVA technique is used when multiple sample cases are involved. As stated earlier, the significance of the difference between the means of two samples can be judged through either  $z$ -test or the  $t$ -test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. The ANOVA technique enables us to perform this simultaneous test. The ANOVA technique is important in the context of all those situations where we want to compare more than two populations such as in comparing the yield of crop from several varieties of seeds, the gasoline mileage of four automobiles, the smoking habits of five groups of university students and so on. In such circumstances one generally does not want to consider all possible combinations of two populations at a time for that would require a great number of tests before we would be able to arrive at a decision. This would also consume lot of time and money, and even then certain relationships may be left unidentified (particularly the interaction effects). Therefore, one quite often utilizes the ANOVA technique and through it investigates the differences among the means of all the populations simultaneously.

Tr: Who has developed this technique?

St: Professor R.A. Fisher was the first man to use the term 'Variance' and in fact it was he who developed a very elaborate theory concerning ANOVA, explaining its usefulness in practical field.

Tr: What does ANOVA explains?

St: ANOVA explains the total amount of variation in a data set broken down into two types that is amount which can be attributed to chance and that amount which can be attributed to specified causes. There may be variation between samples and also within sample items. ANOVA consists in splitting the variance for analytical purposes. Hence, it is a method of analyzing the variance



to which a response is subject into its various components corresponding to various sources of variation.

Tr: What are the assumptions of ANOVA to use?

St: The use of an ANOVA assumes that:

- All the populations are normally distributed (follow a bell shaped curve)
- All the population variances are equal,
- And all the samples were taken independently of each other and are randomly collected

from their population.

Tr: How do you frame null and alternate hypothesis for ANOVA?

St:  $H_0$ : All the population means are equal.

$H_1$ : At least one of the population means is not equal.

Tr: Give some examples where ANOVA can be used?

St: You have a group of individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.

St: You have a group of individuals randomly split into smaller groups based on an attribute they possess. For example, you might be studying leg strength of people according to weight. You could split participants into weight categories (obese, overweight and normal) and measure their leg strength on a weight machine.

Tr: Let us learn the ANOVA application of F- distribution.

Example: A study is designed to test whether there is a difference in mean daily calcium intake in adults with normal bone density, adults with osteopenia (a low bone density which may lead to osteoporosis) and adults with osteoporosis. Adults 60 years of age with normal bone density, osteopenia and osteoporosis are selected at random from hospital records and invited to participate in the study. Each participant's daily calcium intake is measured based on reported food intake and supplements. Is there a statistically significant difference in mean calcium intake in patients with normal bone density as compared to patients with Osteopenia and Osteoporosis? The data are shown below:

Normal Bone Density	Osteopenia	Osteoporosis
1200	1000	890
1000	1100	650
980	700	1100
900	800	900
750	500	400
800	700	350

Solution:

$H_0: \mu_1 = \mu_2 = \mu_3$  (There is no significant difference in mean calcium intake in patients with normal bone density as compared to patients with Osteopenia and Osteoporosis)

$H_1$ : All Means are not equal (There is significant difference in mean calcium intake in patients with normal bone density as compared to patients with Osteopenia and Osteoporosis) OR (There must be at least one such pair of means for that  $\mu_i \neq \mu_j$  where  $i \neq j$  and  $i=1,2,3$  &  $j=1,2,3$ )

Test Statistic:

$$F_{cal} = \frac{MSB}{MSE}$$

To organize our computations we will complete the ANOVA table. In order to compute the sums of squares we must first compute the sample means for each group and the overall mean.

Normal Bone Density	Osteopenia	Osteoporosis
$n_1=6$	$n_2=6$	$n_3=6$
$\bar{X}_1 = 938.3$	$\bar{X}_2 = 800.0$	$\bar{X}_3 = 715.0$

If we pool all  $N=18$  observations, the overall mean is 817.8.

We can now compute:

$$SSB = \sum n_j (\bar{X}_j - \bar{X})^2$$

Substituting:

$$SSB = 6(938.3333 - 817.7778)^2 + 6(800.0 - 817.7778)^2 + 6(715.0 - 817.7778)^2$$

$$SSB = 87,201.77 + 63,379.66 + 1,896.301 = 152,477.7$$

$$SSE = \sum \sum (X - \bar{X}_j)^2$$

SSE requires computing the squared differences between each observation and its group mean.

We will compute SSE in parts. For the participants with normal bone density:

Normal Bone Density	(X - 938.3)	(X - 938.3333) <sup>2</sup>
<b>1200</b>	261.6667	68,486.9
<b>1000</b>	61.6667	3,806.9
<b>980</b>	41.6667	1,738.9
<b>900</b>	-38.3333	1,466.9
<b>750</b>	-188.333	35,456.9
<b>800</b>	-138.333	19,126.9
<b>Total</b>	0	<b>130,083.3</b>

$$\sum (X - \bar{X}_1)^2 = 130,083.3$$

For participants with Osteopenia:

Osteopenia	(X - 800.0)	(X - 800.0) <sup>2</sup>
<b>1000</b>	200	40,000
<b>1100</b>	300	90,000
<b>700</b>	-100	10,000
<b>800</b>	0	0
<b>500</b>	-300	90,000
<b>700</b>	-100	10,000
<b>Total</b>	0	<b>240,000</b>

$$\Sigma(X - \bar{X}_2)^2 = 240,000.0$$

For participants with osteoporosis:

Osteoporosis	(X - 715.0)	(X - 715.0) <sup>2</sup>
<b>890</b>	175	30,625
<b>650</b>	-65	4,225
<b>1100</b>	385	148,225
<b>900</b>	185	34,225
<b>400</b>	-315	99,225
<b>350</b>	-365	133,225
<b>Total</b>	0	<b>449,750</b>

$$\Sigma(X - \bar{X}_3)^2 = 449,750.0$$

$$SSE = \Sigma \Sigma(X - \bar{X}_j)^2 = 130,083.3 + 240,000.0 + 449,750.0 = 819,833.3$$

**ANOVA table:**

Source of Variation	Sums of Squares (SS)	Degrees of freedom (df)	Mean Squares (MS)	F
<b>Between Treatments</b>	152,477.7	2	76,238.6	1.395
<b>Error or Residual</b>	819,833.3	15	54,655.5	
<b>Total</b>	972,311.0	17		

In order to determine the critical value of F we need degrees of freedom:

$$df1=k-1 \text{ and } df2=N-k$$

In this example,  $df1=k-1=3-1=2$  and  $df2=N-k=18-3=15$ .

$F_{3, 15, 0.05} = 3.68$  (The critical value of F at 3,15 d.f. and at 5% level of significance is 3.68.)

Rejection Criteria: Reject  $H_0$  if  $F_{cal} > F_{tab}$ .

Conclusion: Since  $F_{cal} = 1.395 < F_{tab} = 3.68$ , we do not reject  $H_0$ . Hence conclude that there is no significant difference in mean calcium intake in patients with normal bone density as compared to osteopenia and osteoporosis.

**Another Method:**

For calculating sum of squares for Total sum of squares and within sum of squares:

	<b>Normal Bone Density</b>	<b>Osteopenia</b>	<b>Osteoporosis</b>
	1200	1000	890
	1000	1100	650
	980	700	1100
	900	800	900
	750	500	400
	800	700	350
<b>sum</b>	<b>5630</b>	<b>4800</b>	<b>4290</b>
<b>mean</b>	<b>938.33</b>	<b>800</b>	<b>715</b>

Grand sum                      14720  
General mean                817.7778

<b>Normal Bone Density X<sub>1</sub></b>	<b>Osteopenia X<sub>2</sub></b>	<b>Osteoporosis X<sub>3</sub></b>	<b>X<sub>1</sub><sup>2</sup></b>	<b>X<sub>2</sub><sup>2</sup></b>	<b>X<sub>3</sub><sup>2</sup></b>
1200	1000	890	1440000	1000000	792100
1000	1100	650	1000000	1210000	422500
980	700	1100	960400	490000	1210000
900	800	900	810000	640000	810000
750	500	400	562500	250000	160000
800	700	350	640000	490000	122500
<b>Total</b>			<b>5412900</b>	<b>4080000</b>	<b>3517100</b>

$$CF = (\text{GRAND SUM})^2 / N = (\Sigma X)^2 / N = 12037688.89$$

$$SS_T = \Sigma X^2 - CF = 972311.1111$$

$$SS_B = (\Sigma X_1)^2 / n_1 + (\Sigma X_2)^2 / n_2 + \dots + (\Sigma X_n)^2 / n_n - CF = 152477.8$$

$$SS_W = TSS - BSS = 819833.3333$$

**ANOVA TABLE:**

Source of Variation	SS	DF	MSS = SS/DF	F <sub>CAL</sub>
Among the means of condition	152477.8	2	76238.89	1.394897
Within condition	819833.3	15	54655.56	
Total	972311.1	17	57194.77	

In both the methods you will get same Fcal value.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group: (All Groups)**

Example: An experiment shows following results in which 48 subjects were assigned at random to 8 groups of 6 subjects each. Groups are tested under 8 different experimental conditions, designed respectively A, B, C, D, E, F, G and H. Do the mean scores achieved under the 8 experimental conditions differ significantly?

A	B	C	D	E	F	G	H
64	73	77	78	63	75	78	55
72	61	83	91	65	93	46	66
68	90	97	97	44	78	41	49
77	80	69	82	77	71	50	64
56	97	79	85	65	63	69	70
95	67	87	77	76	76	82	68

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read about ANCOVA.

## Lesson No.22: F- test Application: ANCOVA

### Teaching Points:

- Concept of ANCOVA
- Use ANCOVA

### Instructional Objectives:

After completion of this class students will be able to:

- i. Explain ANCOVA.
- ii. Define Covariate.
- iii. Draw ANCOVA table.
- iv. Use f-test for hypothesis testing of comparison of means from various samples in the presence of covariate.
- v. Infer the results of hypothesis testing of comparison of means from various samples in the presence of covariate.

### Lesson Presentation:

Tr: Dear Students, What is ANCOVA?

St: ANCOVA stands for Analysis of Covariance.

St: ANCOVA evaluates whether the means of a dependent variable (DV) are equal across levels of a categorical independent variable (IV) often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates (CV) or nuisance variables. Mathematically, ANCOVA decomposes the variance in the Dependent Variable into variance explained by the Covariate(s), variance explained by the categorical IV, and residual variance.

St: Analysis of covariance is used to test the main and interaction effects of categorical variables on a continuous dependent variable, controlling for the effects of selected other continuous variables, which co-vary with the dependent. The control variables are called the "covariates."

St: Covariates are characteristics (excluding the actual treatment) of the participants in an experiment. A covariate can be an independent variable (i.e. of direct interest) or it can be an unwanted, confounding variable. Adding a covariate to a model can increase the accuracy of your results.



Tr: In many experimental situations, especially in the fields of psychology, medicine and education, we wish to compare groups that are initially unlike, either in the variable under the study or some presumably related variable. Generally we equate the two groups in the beginning either by person to person matching method or by matching groups initially for mean and s.d. in one or more related variables. But neither of these methods is entirely satisfactory and neither is always easy to apply. Equivalent groups often necessitate a sharp reduction in size of N when matching of scores is difficult to accomplish. Moreover, in matched groups it is often difficult to get the correlation between the matching variable and the experimental variable in the population from which our samples were drawn. Therefore ANCOVA can be used to compare two non equivalent groups as ANCOVA makes statistically equivalent groups for comparison.

Tr: What are the assumptions in ANCOVA?

St: The ANCOVA technique requires one to assume that there is some sort of relationship between the dependent variable and the uncontrolled variable.

St: We also assume that this form of relationship is the same in the various treatment groups.

Other assumptions are:

- (i) Various treatment groups are selected at random from the population.
- (ii) The groups are homogeneous in variability.
- (iii) The regression is linear and is same from group to group.

Tr: Let us learn about ANCOVA with an example.

Example: An experiment was carried on 15 students of V standard class. Students have been given one trial (X) of a test. Five are then assigned at random to each of three groups, A, B and C. After one month, say Group A is praised Lavishly, Group B scolded severely and the test repeated (Y). At the same time, a second trial (Y) is also given to Group C, the control group, without any comment treated as control group.

Group A (PRAISED)		Group B (SCOLDED)		Group C (CONTROL)	
X <sub>1</sub>	Y <sub>1</sub>	X <sub>2</sub>	Y <sub>2</sub>	X <sub>3</sub>	Y <sub>3</sub>
15	30	25	28	5	10
10	20	10	12	10	15
20	25	15	20	20	20
5	15	15	10	5	10
10	20	10	10	10	10

Solution:

	Group A (PRAISED)			Group B (SCOLDED)			Group C (CONTROL)		
SL.NO.	X <sub>1</sub>	Y <sub>1</sub>	X <sub>1</sub> Y <sub>1</sub>	X <sub>2</sub>	Y <sub>2</sub>	X <sub>2</sub> Y <sub>2</sub>	X <sub>3</sub>	Y <sub>3</sub>	X <sub>3</sub> Y <sub>3</sub>
1	15	30	450	25	28	700	5	10	50
2	10	20	200	10	12	120	10	15	150
3	20	25	500	15	20	300	20	20	400
4	5	15	75	15	10	150	5	10	50
5	10	20	200	10	10	100	10	10	100
<b>Sum</b>	<b>60</b>	<b>110</b>	<b>1425</b>	<b>75</b>	<b>80</b>	<b>1370</b>	<b>50</b>	<b>65</b>	<b>750</b>
<b>Mean</b>	<b>12</b>	<b>22</b>		<b>15</b>	<b>16</b>		<b>10</b>	<b>13</b>	

For all three groups:

$$\Sigma X = 60 + 75 + 50 = 185 \quad \Sigma Y = 110 + 80 + 65 = 255 \quad \Sigma XY = 1425 + 1370 + 750 = 3545$$

$$\Sigma X^2 = 2775 \quad \Sigma Y^2 = 5003$$

STEP1: Correction Terms

$$C_x = (\Sigma X)^2 / N = 185^2 / 15 = 2282$$

$$C_y = (\Sigma Y)^2 / N = 255^2 / 15 = 4335$$

$$C_{xy} = \Sigma X * \Sigma Y / N = 185 * 255 / 15 = 3145$$

STEP 2: Total SS

$$\text{For } x = \Sigma X^2 - C_x = 2775 - 2282 = 493$$

$$y = \Sigma Y^2 - C_y = 5003 - 4335 = 668$$

$$xy = \Sigma XY - C_{xy} = 3545 - 3145 = 400$$

STEP 3: Among Group Means SS

$$\text{For } x = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] / n_1 - C_x = [60^2 + 75^2 + 50^2] / 5 - 2282 = 63$$

$$y = [\Sigma Y_1^2 + \Sigma Y_2^2 + \Sigma Y_3^2] / n_2 - C_y = [110^2 + 80^2 + 65^2] / 5 - 4335 = 210$$

$$xy = [\Sigma X_1 \Sigma Y_1 / n_1 + \Sigma X_2 \Sigma Y_2 / n_2 + \Sigma X_3 \Sigma Y_3 / n_3] - C_{xy} = [(60 * 110) + (75 * 80) + (50 * 65)] / 5 - 3145 = 25$$

#### STEP 4: Within Group SS

For x= Total SSx – Among Group Means SSx = 493- 63 = 430

For y= Total SSy – Among Group Means SSy = 668- 210 = 458

For xy= Total SSxy – Among Group Means SSxy = 400- 25 = 375

#### STEP 5: Analysis of Variance of X and Y scores taken separately

Source of Variation	Df	SSx	SSy	MSx(Vx)	MSy(Vy)
Among Means	2	63	210	31.5	105
Within Groups	12	430	458	35.8	38.2
Total	14	493	668		

$$F_x = 31.5/35.8 = 0.88$$

from F Table at 2,12 df

$$F_y = 105/38.2 = 2.75$$

F at 0.05 level = 3.88

Since  $F_{cal} < F_{tab}$  at 5% level of significance it means that neither F is significant. But  $F_x = 0.88$  shows that the experimenter was quite successful in getting random samples in groups A, B, C.

#### STEP 6: Computation of Adjusted SS for Y i.e. SSy.x

$$\text{Total SS} = 668 - [400^2 / 493] = 343$$

$$\text{Within SS} = 458 - [375^2 / 430] = 131$$

$$\text{Among M's SS} = 343 - 131 = 212$$

#### Analysis of Covariance

Source of Variation	Df	SSx	SSy	Sxy	SSy.x	MSy.x (Vy.x)	SDy.x
Among Means	2	63	210	25	212	106	
Within Groups	11	430	458	375	131	12	3.46
Total	13	493	668	400	343		

$$F_{y.x} = 106/12 = 8.83$$

From F table at 2, 11 df

F at 0.05 level = 3.98

Since  $F_{y.x} > F_{tab}$  at 5% level of Significance we reject the null hypothesis and conclude that the average scores of all the three groups differ significantly. It means that the three strategies namely praising, scolding and no action gives significant different effects on the scores of the students.

#### STEP 7: Correlation and regression

$$r_{total} = \frac{400}{\sqrt{493 \times 668}} = 0.70$$

$$b_{total} = 400/493 = 0.81$$

$$r_{among\ means} = \frac{25}{\sqrt{63 \times 210}} = 0.22$$

$$b_{among\ means} = 25/63 = 0.40$$

$$r_{within} = \frac{375}{\sqrt{430 \times 458}} = 0.84$$

$$b_{within} = 375/430 = 0.87$$

#### STEP 8: Calculation of Adjusted Y Means

Groups	N	M <sub>x</sub>	M <sub>y</sub>	M <sub>y.x</sub> (adjusted)
A	5	12	22	22.3
B	5	15	16	13.7
C	5	10	13	15
<b>General Means</b>		<b>12.3</b>	<b>17</b>	<b>17</b>

$$M_{y.x} = M_y - b(M_x - GM_x)$$

For group:

$$A: M_y - bx = 22 - 0.87(12-12.3) = 22.3$$

$$B: M_y - bx = 16 - 0.87(15-12.3) = 13.7$$

$$C: M_y - bx = 13 - 0.87(10-12.3) = 15.0$$

#### STEP 9: Significance of differences among adjusted Means

$$SD_{y.x} = \sqrt{12} = 3.46$$

$$SE_{M_{y.x}} = 3.46 / \sqrt{5} = 1.55$$

$$SE_D \text{ between any two adjusted means} = SD_{y.x} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}} = 3.46 \sqrt{\frac{1}{5} + \frac{1}{5}} = 3.46 * 0.63 = 2.18$$

For  $df = 11$ ,  $t_{0.05} = 2.20$

For A and B:  $t_{cal} = D/SE_D = (22-16)/2.18 = 2.75229 > 2.20$  therefore rejecting null hypothesis. It means that difference is significant.

For A and C:  $t_{cal} = D/SE_D = (22-13)/2.18 = 4.12 > 2.20$  therefore rejecting null hypothesis. It means that difference is significant.

For B and C:  $t_{cal} = D/SE_D = (16-13)/2.18 = 1.3761 < 2.20$  therefore do not reject null hypothesis. It means that difference is not significant.

From the above discussions it is clear that scores of Group A is different from scores of Groups B and C. Moreover it is also being found that scores of B and C are of no significant difference. Hence we can conclude that when initial differences are allowed for, praise makes for significant changes in final score, but that scolding has no greater effect than mere repetition of the test. Neither of these last two factors (scolding and control) makes for significant changes in the test scores.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

#### **Task allotment for each group: (All Groups)**

1. The following are paired observations for three experimental groups I, II and III respectively:

Group I		Group II		Group III	
X	Y	X	Y	X	Y
17	12	25	18	20	25
16	15	24	22	35	26
19	17	25	25	32	20
25	19	29	28	38	34
22	20	31	39	34	20

Consider Y as covariate variable and draw ANCOVA table. Test the significance of differences between the adjusted means on X by using the appropriate F-ratio.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read Chi – Square Distribution, Applications of chi- square distribution and Testing for Variance (one sample problem).

## Lesson No.23: Chi-square test and its applications

### Teaching Points:

- Chi – Square Distribution
- Applications of chi- square distribution.
- Testing for Variance (one sample problem)

### Instructional Objectives:

After completion of this class students will be able to

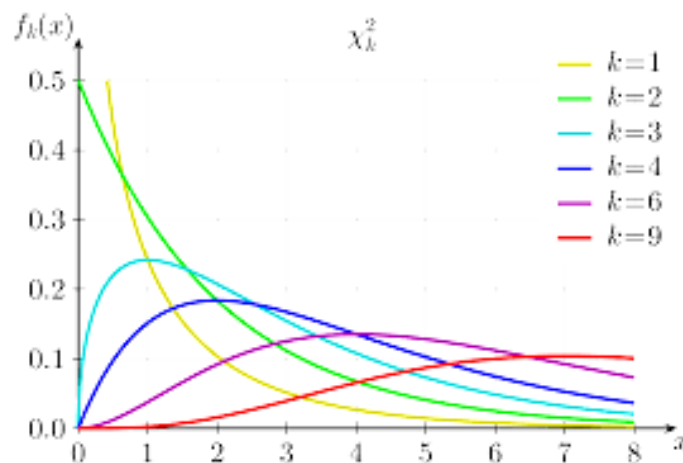
- Explain chi-square distribution
- Use chi –distribution for testing the population variance when population mean is known for one sample problem.
- Use chi –distribution for testing the population variance when population mean is unknown for one sample problem.

### Lesson Presentation:

Tr: What is Chi-square distribution?

St: The Chi Square distribution is the distribution of the sum of squared standard normal deviates. The degree of freedom of the distribution is equal to the number of standard normal deviates being summed.

Tr: You can see the graph of chi-square distribution with different degrees of freedoms.



Tr: What are the applications of Chi-square distribution?

St: The chi-square distribution is used in the common chi-square tests for goodness of fit of an observed distribution to a theoretical one.

St: For testing relationships between two categorical variables.

St: Testing for Variance for one sample problems.

St: In confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.

St: Many other statistical tests also use this distribution, such as Friedman's analysis of variance by ranks.

Tr: What is  $\chi^2$ -test?

Tr:  $\chi^2$ -test is based on chi-square distribution and as a parametric test is used for comparing a sample variance to a theoretical population variance.

### ➤ Testing for Variance

→Testing variance of normal population when  $\mu$  is known:

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \sigma = \sigma_0$	(1) $H_1 : \sigma > \sigma_0$ (2) $H_1 : \sigma < \sigma_0$ (3) $H_1 : \sigma \neq \sigma_0$	$\chi^2 = \frac{\sum (x - \mu)^2}{\sigma_0^2}$	(1) $\chi^2_{\text{cal}} > \chi^2_{n, \alpha}$ (2) $\chi^2_{\text{cal}} < \chi^2_{n, 1-\alpha}$ (3) $\chi^2_{\text{cal}} > \chi^2_{n, \alpha/2}$ or $\chi^2_{\text{cal}} < \chi^2_{n, 1-\alpha/2}$

→Testing variance of normal population when  $\sigma$  is unknown:

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0 : \sigma = \sigma_0$	(1) $H_1 : \sigma > \sigma_0$ (2) $H_1 : \sigma < \sigma_0$ (3) $H_1 : \sigma \neq \sigma_0$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	(1) $\chi^2_{\text{cal}} > \chi^2_{n-1, \alpha}$ (2) $\chi^2_{\text{cal}} < \chi^2_{n-1, 1-\alpha}$ (3) $\chi^2_{\text{cal}} > \chi^2_{n-1, \alpha/2}$ or $\chi^2_{\text{cal}} < \chi^2_{n-1, 1-\alpha/2}$



Tr: Example: Following are the prices of the Canon brand of CameraLens found at eight stores in London. \$755, 815, 789, 799, 732, 835, 799, 769. Test at 5% significance level whether the population variance is different from 500 square dollars with population mean \$800.

St: Solution:

Null hypothesis:  $H_0: \sigma^2 = 500$  square dollars against

Alternative Hypothesis:  $H_0: \sigma^2 \neq 500$  square dollars

Sample size  $n = 8$ ,

Population mean  $\mu = 800$  dollars

Population mean, known so use chi-square test with  $n$  degree of freedom.

<b>X</b>	<b>(X-800)<sup>2</sup></b>
755	2025
815	225
789	121
799	1
732	4624
835	1225
799	1
769	961
<b>Sum</b>	<b>9183</b>

Test Statistic:

$$\chi_{cal}^2 = \frac{\sum (x - \mu)^2}{\sigma_0^2} = \frac{\sum (x - 800)^2}{500} = \frac{9183}{500} = 18.66$$

Rejection Criterion:  $\chi_{cal}^2 > \chi_{n, \alpha/2}^2$  or  $\chi_{cal}^2 < \chi_{n, 1-\alpha/2}^2$

$$\chi_{n, \alpha/2}^2 = \chi_{8, 0.05/2}^2 = \chi_{8, 0.025}^2 = 17.535$$

$$\chi_{n, 1 - \alpha/2}^2 = \chi_{8, 1 - 0.025}^2 = \chi_{8, 0.975}^2 = 2.18$$

i.e. Reject  $H_0$  if  $\chi_{cal}^2 > 17.535$  or  $\chi_{cal}^2 < 2.18$

Conclusion : Here  $\chi_{cal}^2 = 18.66 > 17.535$  so we reject null hypothesis at 5% level of significance, 8 degree of freedom and conclude that population variance is different from 500 square dollars.

Tr: Let us go for one more example.

Example: A random sample of 25 customers taken from certain bank gave the variance of the waiting times equal to 4.3 square minutes. Test at the 1% significance level whether the variance of the waiting times for all customers at this bank is greater than 4.0 square minutes. Assume that the waiting times for all customers are normally distributed.

St: Solution: Null hypothesis:  $H_0: \sigma^2 = 4$  square minutes against

Alternative Hypothesis:  $H_1: \sigma^2 > 4$  square minutes

Sample size  $n = 25$ , sample variance  $s^2 = 4.3$  square minutes

Population mean  $\mu$  unknown, so use chi-square test with  $n-1$  degree of freedom.

Test Statistic:

$$\chi_{cal}^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{24 \times 4.3}{4} = 25.8$$

Rejection criterion:  $\chi_{cal}^2 > \chi_{n-1, \alpha}^2$

$$\chi_{n-1, \alpha}^2 = \chi_{24, 0.01}^2 = 42.98$$

i.e. Reject  $H_0$  if  $\chi_{cal}^2 > 42.98$

Conclusion : Here  $\chi_{cal}^2 = 25.8 < 42.98$  so we do not reject null hypothesis at 1% level of significance , 24 degree of freedom and conclude that the variance in waiting time for all customers is less than 4 square minutes.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### **Task allotment for each group: (All Groups)**

1. A random sample of 27 customers taken from CCD coffee house gave the variance of the waiting times equal to 5.5 square minutes. Assume that the waiting times for all customers are normally distributed. Test at the 2% significance level whether the variance of the waiting times for all customers at this coffee house is lesser than 4.5 square minutes.

2. Following are the prices of the clay toys of same brand found at ten stores in Gujarat Handlooms of Ahmadabad City. Rs 145, 150, 155, 146, 158, 160, 170, 150, 160, 165. Test at 5% significance level whether the population variance is different from 22.45 square Rupees with population mean Rs.155.

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read hypothesis testing for non parametric chi-square test: Testing for Independence of Two Attributes, Testing whether observations are normally distributed or not, Testing whether observations are equally distributed or not.

## NON-PARAMETRIC TESTS

### Lesson No.24: Chi –Square Test application

#### Teaching Points:

- Testing for Independence of Two Attributes
- Testing whether observations are normally distributed or not
- Testing whether observations are equally distributed or not

#### Instructional Objectives:

After completion of this class students will be able to:

- i. Use chi- square test for testing of Independence of Two Attributes.
- ii. Use chi-square test for testing the hypothesis Whether Observations Are Normally distributed or not.
- iii. Use chi-square test for testing the hypothesis whether observations are equally distributed or not.
- iv. Infer the results of testing of Independence of Two Attributes.
- v. Infer the results gain after using chi-square test for testing the hypothesis Whether Observations Are Normally distributed or not.
- vi. Infer the results gain after using chi-square test for testing the hypothesis whether observations are equally distributed or not.

#### Lesson Presentation:

Tr: Dear students, today we will study some more applications of Chi-square test. Earlier you have learnt that chi-square test is also been used for testing of Variance for one sample problems which is a parametric test. But same chi square test is also used for non parametric tests. This (chi – square) is the only distribution which used for both parametric and non parametric tests. Let us learn the application of chi-square test for testing for independence of two attributes.

Following table will help you to understand about the framing of null and alternative hypothesis, test statistic and rejection region for this test.

➤ **Non- parametric Chi square test**

→ **Testing for Independence of Two Attributes**

Null Hypothesis	Alternative Hypothesis	Test statistic	Rejection criterion Reject $H_0$ if
$H_0$ : Both the attributes are independent. Or There is no association between the two attributes.	$H_1$ : Both the attributes are not independent. Or There is association between the two attributes.	$\chi^2_{cal} = \sum \frac{(O-E)^2}{E}$ $E = \frac{MR * MC}{n}$ <div> <math>M_R</math> = represents the row marginal for that cell,  <math>M_C</math> = represents the column marginal for that cell, and  <math>n</math> = represents the total sample size. </div>	$\chi^2_{cal} > \chi^2_{(r-1) * (c-1), \alpha\%}$

Example: A company honour wants to inspect the results of injected vaccine to their employees against the pneumonia disease. The results are given below:

Health Outcome	Non-Vaccinated	Vaccinated
<b>Sick with pneumococcal pneumonia</b>	23	5
<b>Sick with non-pneumococcal pneumonia</b>	8	10
<b>No pneumonia</b>	61	77

Was there a difference in incidence of pneumonia between the two groups? Test at 5% level of significance.

Solution: Null Hypothesis:  $H_0$ : There is no significant difference in occurrence of pneumococcal pneumonia between the vaccinated and nonn-vaccinated groups.

Alternate Hypothesis:  $H_1$ : There is a significant difference in occurrence of pneumococcal pneumonia between the vaccinated and non-vaccinated groups.

Test Statistic:

$$\chi^2_{cal} = \sum \frac{(O - E)^2}{E}$$

O: Observed Frequency

E: Expected frequency

#### Calculation of Marginals:

Health Outcome	Non- vaccinated Col 1	Vaccinated Col 2	Row Marginals (Row sum)
Sick with pneumococcal pneumonia	23	5	<b>28</b>
Sick with non-pneumococcal pneumonia	8	10	<b>18</b>
Stayed healthy	61	77	<b>138</b>
Column marginals (Sum of the column)	<b>92</b>	<b>92</b>	<b>N = 184</b>

For calculation of expected frequencies:

$$E = \frac{MR * MC}{n}$$

$M_R$  represents the row marginal for that cell,

$M_C$  represents the column marginal for that cell, and

$n$  = represents the total sample size.

O	E	(O-E)	(O-E) <sup>2</sup> /E
<b>23</b>	13.92	9.08	5.922874
<b>5</b>	12.57	-7.57	4.558862
<b>8</b>	8.95	-0.95	0.100838
<b>10</b>	9.05	0.95	0.099724
<b>61</b>	69.12	-8.12	0.953912
<b>77</b>	69.88	7.12	0.725449
		Total =	<b>12.36166</b>

Cell expected values and (cell Chi-square values):

Health outcome	Non-vaccinated	Vaccinated
<b>Sick with pneumococcal pneumonia</b>	13.92 (5.92)	12.57 (4.56)
<b>Sick with non-pneumococcal pneumonia</b>	8.95 (0.10)	9.05 (0.10)
<b>Stayed healthy</b>	69.12 (0.95)	69.88 (0.73)

$$\chi^2_{\text{cal}} = 12.36166$$

$$\chi^2_{\text{tab}} = \chi^2_{(r-1) * (c-1), \alpha\%} = 5.991465$$

Where, d.f. =  $(r-1) * (c-1) = (3-1) \times (2-1) = 2 * 1 = 2$  and  $\alpha = 5\%$

Rejection Criteria: Reject  $H_0$  if  $\chi^2_{\text{cal}} > \chi^2_{(r-1) * (c-1), \alpha\%}$ .

Conclusion: Since  $\chi^2_{\text{cal}} = 12.36166 > \chi^2_{\text{tab}} = \chi^2_{(r-1) * (c-1), \alpha\%} = 5.991465$  we reject our Null hypothesis at 5% level of Significance and conclude that There is a significant difference in occurrence of pneumococcal pneumonia between the vaccinated and non-vaccinated groups.

Tr: Example: The items of an opinionnaire scale are answered as: strongly disagree, disagree, can't say, agree and strongly agree. The distribution of answers to an item .marked by 300 subjects is shown in the following table. Do these answers diverge significantly from the distribution to be expected if there are no preferences in the groups.

strongly disagree	disagree	can't say	agree	strongly agree
34	54	67	88	57

St:  $H_0$ : There is no preference among the people for different groups.

$H_1$ : There is significant preference among the people for different groups.

Level of liking	strongly disagree	disagree	can't say	agree	strongly agree	<b>Total</b>
Observed frequency(O)	34	54	67	88	57	300
Expected frequency (E)	60	60	60	60	60	300

Expected frequency (E) = sum of all observed frequencies / no. of categories

$E = 300/5 = 60$  (for all cells)

Test Statistic:

$$\chi^2_{cal} = \sum \frac{(O - E)^2}{E}$$

Level of liking	Observed frequency(O)	Expected frequency (E)	(O-E)	(O-E)^2	(O-E)^2/ E
strongly disagree	34	60	-26	676	11.2666667
Disagree	54	60	-6	36	0.6
can't say	67	60	7	49	0.81666667
Agree	88	60	28	784	13.0666667
strongly agree	57	60	-3	9	0.15
<b>Total</b>	300	300			25.9

$$\chi^2_{cal} = 25.9$$

$$\chi^2_{tab} = \chi^2_{(r-1) * (c-1), \alpha\%} = \chi^2_{4, 0.05} = 9.487729$$

Conclusion: Since  $\chi^2_{cal} = 25.9 > \chi^2_{tab} = 9.487729$  we reject the null hypothesis and conclude at 5% level of significance that there is significant preference among the people for different groups.

Tr: Example: Two hundred thirteen delivery boys have been classified into three groups namely, very good, satisfactory and poor – by a consensus of sales managers. Does this distribution of ratings differ significantly from that to be expected if delivery ability is normally distributed in our population of delivery boys.

Very Good	Satisfactory	Poor
45	103	65

St: Ho: The distribution of ratings does not differ significantly from that to be expected normally distributed in our population of delivery boys.



H<sub>1</sub>: The distribution of ratings differs significantly from that to be expected normally distributed in our population of delivery boys.

Test Statistics:

$$\chi^2_{cal} = \sum \frac{(O - E)^2}{E}$$

Level of rating	Observed frequency(O)	Expected frequency (E)	(O-E)	(O-E) <sup>2</sup>	(O-E) <sup>2</sup> / E
<b>Very good</b>	45	34.08	10.92	119.2464	3.499014
<b>Satisfactory</b>	103	144.84	-41.84	1750.586	12.08634
<b>Poor</b>	65	34.08	30.92	956.0464	28.053
<b>total</b>	<b>213</b>	<b>213</b>			<b>43.63836</b>

$$\chi^2_{cal} = 43.638$$

$$\chi^2_{tab} = \chi^2_{(r-1) * (c-1), \alpha\%} = \chi^2_{2, 0.05} = 5.991465$$

Conclusion: Since  $\chi^2_{cal} = 43.638 > \chi^2_{tab} = t_{(r-1) * (c-1), \alpha\%} = 5.99$  we reject the null hypothesis and conclude at 5% level of significance that, the distribution of ratings differs significantly from that to be expected normally distributed in our population of delivery boys.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

#### Task allotment for each group: (All Groups)

1. The table given below shows the data obtained during outbreak of Measles disease in Kanpur district of UP state:

	Attacked	Not attacked	Total
Vaccinated	130	26783	26913
Not Vaccinated	3425	76584	80009
Total	3555	103367	106922

Test the effectiveness of vaccination in preventing the attack from Measles. Test your result with the help of chi –square test at 10% level of significance.

2. The items of a rating scale are answered as: strongly disagree, disagree, moderate, agree and strongly agree. The distribution of answers to an item .marked by 210 subjects is shown in the following table.

(i) Do these answers diverge significantly from the distribution to be expected if there are no preferences in the groups.

strongly disagree	Disagree	can't say	Agree	strongly agree
28	34	53	52	43

(ii) Do these answers diverge significantly from the distribution to be expected if responses are normally distributed in our population of respondents.

strongly disagree	Disagree	can't say	Agree	strongly agree
28	34	53	52	43

### **Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read Non parametric test: Median Test and Sign Test.

## Lesson No.25: Median Test and Sign Test

### Teaching Points:

- Median test
- Sign test

### Instructional Objectives:

After completion of this class students will be able to

- i. Use Median test.
- ii. Infer the result gain after the use of Median test.
- iii. Use Sign test.
- iv. Infer the result gain after the use of Sign test.

### Lesson Presentation:

Tr: Dear Students, Mood's median test is a special case of Pearson's chi-squared test. It is a nonparametric test that tests the null hypothesis that the medians of the populations from which two or more samples are drawn are identical.

What do you know about Median test?

St: The test has low power (efficiency) for moderate to large sample sizes. The Wilcoxon–Mann–Whitney U two-sample test or its generalization for more samples, the Kruskal–Wallis test, can often be considered instead.

ST: The relevant aspect of the median test is that it only considers the position of each observation relative to the overall median, whereas the Wilcoxon–Mann–Whitney test takes the ranks of each observation into account. Thus the other mentioned tests are usually more powerful than the median test. Moreover, the median test can only be used for quantitative data.

St: Although the alternative Kruskal - Wallis test does not assume normal distributions, it does assume that the variance is approximately equal across samples. Hence, in situations where that assumption does not hold, the median test is an appropriate test.

St: Sign test is a statistical method to test for consistent differences between pair of observations, such as the weight of subjects before and after treatment.

St: If  $X$  and  $Y$  are quantitative variables, the **sign test** can be used to test the hypothesis that the difference between the  $X$  and  $Y$  has zero median, assuming continuous distributions of the two random variables  $X$  and  $Y$ , in the situation when we can draw paired samples from  $X$  and  $Y$ .

St: The sign test can also test if the median of a collection of numbers is significantly greater than or less than a specified value. For example, given a list of student grades in a class, the sign test can determine if the median grade is significantly different from, say, 75 out of 100.

Tr: let us take some example of Median test.

Example: A psychiatrist wish to study the effects of a tranquilizing drug upon leg tremor. 14 psychiatric patients are given the drug and 18 other patients matched for age and sex are given placebo. Following data was observed. Test whether the drug increases the steadiness -as shown by lower scores in the experimental group? Steadiness tester is used to measure the Tremor.

Sl. No.	Experimental group scores	Sl. No.	Control group scores
1	53	1	48
2	39	2	65
3	63	3	66
4	36	4	38
5	47	5	36
6	58	6	45
7	44	7	59
8	38	8	53
9	59	9	58
10	36	10	42
11	42	11	70
12	43	12	71
13	46	13	65
14	46	14	46
		15	55
		16	61
		17	62
		18	53

Solution: In the following table, a '+' sign indicates a score above the common median and a '-' sign indicate a score below the common median.

Common Median = 49 (calculated using both the groups of readings)

Sl. No.	Experimental group scores	Sign	Sl. No.	Control group scores	Sign
1	53	+	1	48	-
2	39	-	2	65	+
3	63	+	3	66	+
4	36	-	4	38	-
5	47	-	5	36	-
6	58	+	6	45	-
7	44	-	7	59	+
8	38	-	8	53	+
9	59	+	9	58	+
10	36	-	10	42	-
11	42	-	11	70	+
12	43	-	12	71	+
13	46	-	13	65	+
14	46	-	14	46	-
			15	55	+
			16	61	+
			17	62	+
			18	53	+

Ho: There is no significant difference between median of experimental group and the median of control group.

H1: The median of experimental group is lower than the median of control group. (as indicating more steadiness and thus less tremor.)

From the above gained + and – signs following 2 X 2 contingency table will be made as follow:

	Below Median	Above Median	Total
Experimental	10	4	14
Control	6	12	18
Total	16	16	32 = N

$$\chi^2_{\text{cal}} = \frac{32 (1120 - 241 - 32/2)^2}{16 \cdot 16 \cdot 18 \cdot 14} = 3.17$$

$\chi^2_{\text{tab}}$  value at 1 df and 0.05, 0.08 and 0.10 level of significance:

$$\chi^2_{1,0.05} = 3.84145$$

$$\chi^2_{1,0.08} = 3.064902$$

$$\chi^2_{1,0.10} = 2.705543$$

Since  $\chi^2_{\text{cal}} = 3.17 > 2.7055 = \chi^2_{\text{tab}}$  at 10% level of significance we reject the null hypothesis but the same null hypothesis is not rejected at 5% level of significance. Therefore results may vary if the level of significance changes. Here we can conclude that at 10% level of significance the median of experimental group is lower than the median of control group. This indicates that more the steadiness or fewer tremors found in experimental group or we can say that the drug produces some reduction in tremor.

Let us take an example of Sign Test for One Sample Problem.

Example: Ten women are asked to judge two brands of perfume which has a more pleasant odor. Eight of the women select Perfume A and two of the women select Perfume B. Is there a significant difference with respect to preference for the perfumes?

Solution:  $H_0$ : There is no significant difference with respect to preference for perfume A and B.

(i.e.  $H_0$ :  $P = 0.5$ )

$H_1$ : There is significant difference with respect to preference for perfume A and B.

(i.e.  $H_0$ :  $P \neq 0.5$ )

$X \sim B(n=10, p=0.5)$

$$\begin{aligned} P(X \geq 8) &= P(X=8) + P(X=9) + P(X=10) \\ &= {}^{10}C_8 (0.5^8) (0.5^2) + {}^{10}C_9 (0.5^9) (0.5) + {}^{10}C_{10} (0.5^{10}) (0.5^0) \\ &= 45 * (0.5^{10}) + 10 * (0.5^{10}) + 1 * (0.5^{10}) \\ &= 0.5^{10} [45 + 10 + 1] \end{aligned}$$

$$P(X \geq 8) = 0.5^{10} * 56 = 0.0546875$$

Rejection Criteria: Reject  $H_0$  if P- value is  $< \alpha/2$  % level of significance.

Let  $\alpha = 5\% = 0.05$ , therefore  $\alpha/2 \% = 0.025$

Conclusion: Since  $P(X \geq 8) = 0.0546875 > 0.025$  at 5% level of significance we do not reject  $H_0$  and conclude that there is no significant difference with respect to preference for perfume A and B.

If the  $H_1$  is directional hypothesis then we must compare P-value with  $\alpha\%$  level of significance. Now let us take an example of Sign Test for Paired Observations.

Example: There are two different tasks say S and C, where S refers to spelling of 25 words and C refers to spelling of 25 words of equal difficulty presented as an integral part of a sentence. A researcher wanted to study which condition is favourable to higher scores. Following information was collected from 10 randomly selected IX grade students under C and S. Here scores are being recorded in pairs. In III column the sign of the difference (C-S) as plus or minus. Under the null hypothesis  $\frac{1}{2}$  of the differences should be + and  $\frac{1}{2}$  should be -. Test the hypothesis that C is better than S.

Solution:  $H_0$ : C and S both tasks are equally efficient.

$H_1$ : C is significantly better (superior) than S task (condition).

C	S	(C-S)	Signs
15	12	+	+ 7
18	15	+	- 2
9	10	-	0 1
15	16	-	Total 10
18	18	0	
12	10	+	
15	12	+	
16	13	+	
14	12	+	
22	19	+	
C: words in context			
S: words spelled as separates			

From the 10 differences

7 are + (as C is higher than S)

2 are - (as S is higher than C)

1 is 0 (as C = S)

In the binominal expansion of  $(p + q)^9$ , p is the probability of + and q is the probability of -.

$$(p + q)^9 = p^9 + 9p^8q + 36p^7q^2 + 84p^6q^3 + 126p^5q^4 + 126p^4q^5 + 84p^3q^6 + 36p^2q^7 + 9pq^8 + q^9$$

Here, there are total  $2^9 = 512$  combinations are possible for + or - for 9 observations.

Adding the first three terms ( $p^9 + 9p^8q + 36p^7q^2$ ), we have total of 46 combinations which contain 7 or more + signs. Some 46 times in 512 trials 7 or more plus signs out of 9 will occur when the mean number of + signs under the null hypothesis is 4.5..the probability of 7 or more + signs, therefore, is  $46/512$  or 0.09 and is clearly not significant as level of significance  $0.05 < 0.09$ . This is a one -tailed test, since our hypothesis states that C is better than S.

If the hypothesis at the outset had been that C and S differ without specifying which is superior, we would had a 2 -tailed test for which  $P = 0.18$ .

When no. of paired observations is as large as 20, the normal curve may be used as an approximation to the binomial expansion or the  $\chi^2$  test applied.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

### **Task allotment for each group: (All Groups)**

1. An insulator used in a machine is to be checked for the accuracy of its design setting of 110°F. Twelve insulators were tested to determine their actual settings, resulting in the following data:

102.7, 103.4, 106.5, 112.5, 116.3, 118.0, 113.7, 110.8, 111.3, 112.0, 115.8, 117.4

Test whether the designed insulator matches with the expected level of settings?

2. Suppose playing four rounds of golf at the City Club 11 professionals totaled 280, 282, 290, 273, 283, 283, 275, 284, 282, 279, and 281. Use the sign test at 5% level of significance to test the null hypothesis that professional golfers average  $\mu_{H0} = 284$  for four rounds against the alternative hypothesis  $\mu_{H0} < 284$ .



3. At Lothal excavation site two archaeologists dug up of artifacts in last 30 days. Following information is recorded:

X	1	0	2	3	1	0	2	2	3	0	1	1	4	1	2	1	3	5	2	1	3	2	4	1	3	2	0	2	4	2
Y	0	0	1	0	2	0	0	1	1	2	0	1	2	1	1	0	2	2	6	0	2	3	0	2	1	0	1	0	1	0

Use the sign test at 1% level of significance to test the null hypothesis that the two archaeologists, X and Y, are equally good at finding artifacts against the alternative hypothesis that X is better than Y.

### **Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

**Announcement of topic in Class:** This announcement was made two days prior to the class. For the coming class read non parametric Mann Whitney U-test.

## Lesson No.26: Mann Whitney U-test

- **Mann Whitney U-test**

### **Instructional Objectives:**

After completion of this class students will be able to

- i. Use Mann Whitney U-test.
- ii. Infer the result gain after the use of Mann Whitney U-test.

### **Lesson Presentation:**

Tr: Dear students, the Mann Whitney U test is also known as Mann Whitney Wilcoxon (MWW), Wilcoxon rank sum test, or Wilcoxon Mann Whitney test. When Mann Whitney  $U$  test can be used?

St: Mann Whitney U test is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from a second population.

St: This test can be used to investigate whether two independent samples were selected from populations having the same distribution.

St: The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.

Tr: Do you know the parametric test which is parallel test of Mann Whitney  $U$  test?

St:  $t$  – test for two independent sample problems.

Tr: What if the two sample are not independent?

St: A similar nonparametric test is used on dependent samples i.e. the Wilcoxon signed-rank test.

Tr: Mann Whitney  $U$  test is more powerful and useful test than the Median Test. It is most useful alternative to the parametric  $t$  test when the parametric assumptions cannot be met and when the measurements are expressed in ordinal scale values.

Tr: Let us take one example to understand Mann Whitney U test.

Example: A researcher wanted to evaluate the effectiveness of discussion method and team teaching method of learning. For that researcher has created two groups say A and B. Each group is made of 15 students. The scores of the students are given below:

Sl. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Group A	45	46	44	34	24	34	54	54	65	57	66	57	67	65	45
Group B	56	55	54	53	56	36	47	48	58	53	47	28	47	57	60

St: Ho: There is no significant difference between discussion and team teaching methods of Learning.

H<sub>1</sub>: There is significant difference between discussion and team teaching methods of Learning.

Group A	Rank of Group A	Group B	Rank of Group B
24	1	28	2
34	3.5	36	5
34	3.5	47	11
44	6	47	11
45	7.5	47	11
45	7.5	48	13
46	9	53	14.5
54	17	53	14.5
54	17	54	17
57	23	55	19
57	23	56	20.5
65	27.5	56	20.5
65	27.5	57	23
66	29	58	25
67	30	60	26
<b>NI = 15</b>	<b>232</b>	<b>N2= 15</b>	<b>233</b>

$$\text{Test statistic: } z = \frac{U - \frac{N_1 N_2}{2}}{\sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}} = \frac{113 - \frac{15 \cdot 15}{2}}{\sqrt{\frac{15 \cdot 15 (15 + 15 + 1)}{12}}} = \frac{0.5}{24.1091} = 0.0207$$

Where  $N_1$ : number in first group =15

$N_2$ : number in second group =15

$\Sigma R_1$ : Sum of ranks in first group =232

$\Sigma R_2$ : Sum of ranks in second group =233

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - \Sigma R_1 = 15 \cdot 15 + \frac{15 \cdot 16}{2} - 232 = 113$$

$$U_2 = N_1 N_2 + \frac{N_2(N_2+1)}{2} - \Sigma R_2 = 15 \cdot 15 + \frac{15 \cdot 16}{2} - 233 = 112$$

The two U's are related by the formula:

$$U_1 = N_1 N_2 - U_2$$

$$113 = 15 \cdot 15 - 112$$

$$113 = 113$$

Here, df are  $N_1=15$ ,  $N_2=15$  both are less than 8 distribution converges to z distribution.

$$Z_{0.05} = 1.96 \text{ (two tailed test)}$$

Conclusion: Since  $Z_{cal} = 0.0207 < 1.96 = Z_{0.05}$  we don't reject null hypothesis at 5% level of significance and conclude that there is no significant difference between discussion and team teaching methods of Learning or in other words both the methods of are equally effective.

Tr: Example: The values in one sample are 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. In another sample they are 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. Test at the 10% level the hypothesis that they come from populations with the same mean. Apply U-test.

Solution:  $H_0$ : The two samples came from the populations with same means.

$H_1$ : The two samples came from the populations with different means.

From the table given below we find that the sum of the ranks assigned to sample one items or  $R_1 = 2 + 3 + 8 + 9 + 11.5 + 13 + 14 + 17 + 21.5 + 21.5 + 23 + 24 = 167.5$  and similarly we find that the sum of ranks assigned to sample two items or  $R_2 = 1 + 4 + 5 + 6.5 + 6.5 + 10 + 11.5 + 15 + 16 + 18 + 19.5 + 19.5 = 132.5$  and we have  $n_1 = 12$  and  $n_2 = 12$ .

$$\text{Hence, test statistic } U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 = 12 \times 12 + \frac{12(13)}{2} - 167.5$$

$$= 144 + 78 - 167.5 = 54.5$$

Size of sample in ascending order	Rank	Name of selected sample [A for sample one and B for sample two]
32	1	B
38	2	A
39	3	A
40	4	B
41	5	B
44	6.5	B
44	6.5	B
46	8	A
48	9	A
52	10	B
53	11.5	B
53	11.5	A
57	13	A
60	14	A
61	15	B
67	16	B
69	17	A
70	18	B
72	19.5	B
72	19.5	B
73	21.5	A
73	21.5	A
74	23	A
78	24	A

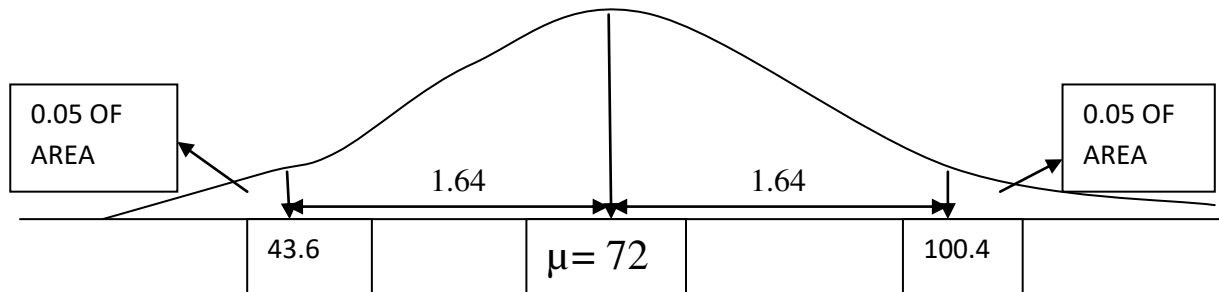
Since in the given problem  $n_1$  and  $n_2$  both are greater than 8, so the sampling distribution of  $U$  approximates closely with normal curve. Keeping this in view, we work out the mean and

standard deviation taking the null hypothesis that the two samples come from identical populations as under:

$$\mu_U = \frac{N_1 * N_2}{2} = \frac{12 * 12}{4} = 72$$

$$\sigma_U = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}} = \sqrt{\frac{12 * 12 (25)}{12}} = 17.32$$

As the alternative hypothesis is that the means of the two populations are not equal, a two-tailed test is appropriate. Accordingly the limits of acceptance region, keeping in view 10% level of significance as given, can be worked out as under:



As the z value for 0.45 of the area under the normal curve is 1.64, we have the following limits of acceptance region: Upper limit =  $\mu_U + 1.64 \sigma_U = 72 + 1.64 (17.32) = 100.40$

$$\text{Lower limit} = \mu_U - 1.64 \sigma_U = 72 - 1.64 (17.32) = 43.60$$

Conclusion: As the observed value of U is 54.5 which is in the acceptance region, we accept the null hypothesis and conclude that the two samples come from identical populations (or that the two populations have the same mean) at 10% level of significance.

Now students will be grouped in 5 or 6 members per group and group task will be allotted for each group.

**Task allotment for each group:**

1. Two samples with values 91, 96, 34 and 45 in I case and the other with values 58, 33, 8, 29, and 36 are given. Test applying Wilcoxon test whether the two samples come from populations with the same mean at 10% level against the alternative hypothesis that these samples come from populations with different means.

2. An Educationists wished to evaluate the effectiveness of two teaching methods. For that educationists has made two groups say A and B. Each group is made of 17 students. The scores given by the educationist on the basis of their responses are given below:

Sl. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Group A	23	43	33	23	43	24	38	36	36	35	38	34	45	44	34	31	41
Group B	18	43	23	27	44	43	49	47	43	42	41	43	42	38	37	42	44

**Presentation and Discussion:**

After group work one student from each group presents their task. Depends upon the presentation teacher recapitulates and concludes the class. At the end teacher will give assignment to the students. Assignment work related to this topic is mentioned in the appendix-VIII.

## Appendix- VIII

### List of Assignment Work

#### ❖ Descriptive Statistics

#### Lesson No.1: Graphs

1. Construct frequency distribution for the following data and draw Histogram. Also trace the value of Mode from the histogram.

78	98	88	78	79	78	89	89	90	90
77	79	87	76	67	78	79	98	92	84
25	84	84	46	77	83	45	34	34	34
45	56	54	57	67	68	77	77	67	69
45	46	57	58	67	78	79	70	45	45
45	34	45	46	57	58	78	79	70	56
34	45	56	76	6	63	61	52	48	49
23	43	4	41	45	34	65	28	63	45

2. Construct frequency distribution for the following data and draw Ogive curve. Also trace the following partition values from the Ogive curve and interpret them:

$Q_1, Q_2, Q_3, D_4, D_7, P_{35}, P_{70}, P_{65}, P_{80}$

12	24	33	33	23	34	42	43	23	34
35	44	54	53	32	33	33	43	42	32
25	27	3	7	38	29	35	36	47	34
32	34	35	33	45	46	47	46	45	34
35	36	44	45	43	24	35	43	44	44
24	34	14	15	16	23	24	25	35	45
46	42	41	41	32	43	52	4	3	43
21	22	13	24	35	45	45	46	47	39



## Lesson No.2: Charts

1. Use appropriate chart to represent the following information:

(a)

Roll No. of Student	1	2	3	4	5	6	7	8	9
Marks in History	23	33	24	34	42	33	26	36	42

(b)

Roll No. of Student	1	2	3	4	5	6	7	8	9
Marks in History	23	33	24	34	42	33	26	36	42
Marks in English	35	43	48	43	30	36	28	37	40
Marks in Geography	37	42	35	37	44	22	29	32	36

(c)

Days (20/05/2015 to 26/05/2015)	Sun	Mon	Tue	Wed	Thru	Fri	Sat
Temperature of Vadodara city	40 °C	42 °C	45°C	46°C	47°C	46°C	47°C

(d)

Items	Transp ortation	Catering	Booking of Tickets	Parking of Buses	Snacks	Miscell aneous	Left over
Expenditure of Educational Trip	20,000	12,000	3,000	1,000	5,500	3000	1,500

### Lesson No.3: Measures of Central Tendency

1. Define mean, median and mode and explain their merits and demerits.
2. Calculate mean, median and mode for the following data and interpret the obtained values.

(a) Raw scores of 80 students in Mathematics subject are given below:

12	24	33	33	23	34	42	43	23	34
35	44	54	53	32	33	33	43	42	32
25	27	3	7	38	29	35	36	47	34
32	34	35	33	45	46	47	46	45	34
35	36	44	45	43	24	35	43	44	44
24	34	14	15	16	23	24	25	35	45
46	42	41	41	32	43	52	4	30	43
21	22	13	24	35	45	45	46	47	39

(b) Raw scores of 80 students in Science subject are given below:

78	98	88	78	79	78	89	89	90	90
77	79	87	76	67	78	79	98	92	84
25	84	84	46	77	83	45	34	34	34
45	56	54	57	67	68	77	77	67	69
45	46	57	58	67	78	79	70	45	45
45	34	45	46	57	58	78	79	70	56
34	45	56	76	61	63	61	52	48	49
23	43	4	41	45	34	65	28	63	45

3. In a tournament series XI class students scored goals in the following manner. Calculate mean, median and mode for the following data and interpret the results:

No. of Goals	2	3	4	5	6	7	8
No. of Matches	2	1	1	4	3	3	1

4. Identify which team has played well in the following given data:

No. of Goals	Gokul School No. of Matches	Hari Prasad School No. of Matches
3	4	5
4	5	4
5	6	6
6	7	5
7	6	4
8	7	3

5. Calculate Mean, Median and Mode of the following data and interpret the results:

Marks of Assignment work of Social Studies Subject	No. of Students
10-12	1
12-14	2
14-16	5
16-18	7
18-20	12
20-22	14
22-24	9
Total	50

#### Lesson No.4: Measures of Dispersion / Variation

1. Describe the meaning of dispersion. Also explain the significance of it.
2. Explain the absolute and relative measures of dispersion.
3. Calculate the standard deviation and quartile deviation for the following data. Interpret the results.

(a) Scores on an objective test.

3	5	6	7	4
5	6	7	8	5
6	6	5	7	9
3	5	5	6	8

(b) Identify which team has played more consistently:

No. of Goals	Gokul School No. of Matches	Hari Prasad School No. of Matches
3	4	5
4	5	4
5	6	6
6	7	5
7	6	4
8	7	3

(c) Identify which Class has scored more uniformly in board exams:

Percentage of Marks in XII Board Exam	No. of Students of Class XII –A	No. of Students of Class XII –B
40-50	4	5
50-60	6	5
60-70	10	16
70-80	13	14
80-90	11	9
90-100	14	16
total	58	65

## Lesson No.5: Skewness and its Types

### ➤ Skewness

1. Explain the properties of positive and negative skewness with diagram.
2. List various measures for calculating coefficient of skewness.
3. Calculate coefficient of skewness  $\beta_1$  for the following information and interpret the results:

(a) Scores of a class test.

34	21	35	46	54	55	43	32	45	76
45	23	44	35	65	37	39	60	57	23

(b) Scores of a project work.

Scores of a Project Work	No. of students
20-25	05
25-30	12
30-35	17
35-40	08
40-45	07
45-50	06
Total	55

## Lesson No.6: Kurtosis and its types

1. Define kurtosis and explain its types
2. Draw a figure to show the types of kurtosis and explain their properties.
3. Calculate the coefficient of kurtosis  $\beta_2$  for the following data:

(a) Scores of a class test.

4	6	4	18	7	6	17	18	9	10
21	23	43	22	12	13	14	3	12	14

(b) Scores of a project work.

Scores of a Project Work	No. of students
20-25	05
25-30	12
30-35	17
35-40	08
40-45	07
45-50	06
Total	55

### Lesson No.7: Correlation

**Correlation (Simple, Rank Correlation, Partial, Multiple, Bi-Serial, Point Bi-Serial)**

➤ **Simple Correlation / Karl Pearson Correlation / Product Moment Correlation**

1. From the following data, find out the correlation coefficient between heights of fathers and sons.

Heights of fathers(inches)	65	66	67	67	68	69	70	72
Heights of sons(inches)	67	68	65	68	72	72	69	71

2. The following data refers to research expense and no. of units dropped in last six months.

Research Expense (in '000 Rs.)	14	21	26	22	15	19
No. of units dropped	31	37	50	45	33	39

Calculate the correlation coefficient and comment on the result. Also draw a scatter diagram and interpret it.

3. Compute Karl Pearson's coefficient of correlation in the following series relating to cost of living and wages.

Wages (Rs.)	100	101	102	100	99	98	97	98	96	95
Cost of living	98	99	99	97	95	92	95	94	90	91

4. A prognostic test in Mathematics was given to 10 students who were about to bring a course in statistics. The scores (X) in their test were examined in relations to score (Y) in the final examination in Statistics. The following results were obtained:

$$\sum x = 71, \sum y = 70, \sum x^2 = 555, \sum y^2 = 526, \text{ and } \sum xy = 527.$$

Find the coefficient of correlation between x and y.

### ➤ Lesson No.8: Rank Correlation

1. Two critics were allotted 10 movies to rank them from 1 to 10. Where 1 indicate best movie and 10 indicate worst movie and so on. Following table summarizes the data for the study. Is there a significant association between the two Critics of judgment?

<b>Movies No.</b>	<b>Critic 1 (<math>R_X</math>)</b>	<b>Critic 2 (<math>R_Y</math>)</b>
1	9	6.5
2	10	8
3	8	9
4	1	4
5	7	6.5
6	2	1
7	5.5	5
8	3	3
9	4	2
10	5.5	10

2. In a debate competition two judges were assigned to access the performance of 10 students.

Calculate the degree of similarity in their judgments. Results are given below:

<b>Candidate No.</b>	1	2	3	4	5	6	7	8	9	10
<b>Scores of Judge A</b>	24	26	36	45	47	34	36	34	26	45
<b>Scores of Judge B</b>	22	34	34	38	44	36	29	25	24	48

## Lesson No.9: Partial and Multiple Correlations

1. The following data are on the average weekly profit (in thousand rupees) of six Restaurants, their seating capacities and the average daily traffic (in thousands of cars) which passes their location.

Seating capacity ( $x_2$ )	116	189	147	179	236	234
Traffic count ( $x_3$ )	24	11	17	15	14	17
Weekly net profit ( $x_1$ )	30	26	21	25	32	22

- a) Find partial correlation between weekly net profit and traffic count keeping seating capacity constant.
- b) What is the multiple effects of traffic count and seating capacity on weekly net profit?

2. Following data gives information on rice yield ( $x_1$ ) (per hectare in quintals), rainfall ( $x_2$ ) (inches) and use of fertilizer ( $x_3$ ) (kg. per hectare).

$X_1$	32	31	36	34	32	34	38
$X_2$	11	21	31	41	51	61	71
$X_3$	42	43	55	67	72	71	82

- a. What is the correlation between rainfall and wheat yield when fertilizer is constant?
- b. What is the correlation between fertilizer used and wheat yielded keeping rainfall constant?
- c. What will be the joint effect of rainfall and fertilizer used on wheat yielded?



## Lesson No.10: Bi-Serial & Point Bi-Serial Correlation

1. State the conditions in which Bi- serial correlation can be used.
2. Give some illustrations in which bi- serial correlation can be used.
3. Determine the degree of association between salary and education level of the following data:

Sl. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Salary in (thousands of Rupees)	23	33	94	53	45	26	55	62	63	46	39	46	55	77	63	54
Education level	G	PG	PG	G	G	G	G	PG	G	PG	PG	G	G	PG	G	PG

### ➤ Point Bi-Serial

1. State the conditions in which Point Bi- serial correlation can be used.
2. Give some illustrations in which Point bi- serial correlation can be used.
3. Determine the degree of association between salary and gender for the following collected data:

Sl. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Salary in (thousands of Rupees)	45	34	45	53	32	34	31	24	35	46	56	43	67	78	84	43
Gender	M	F	F	F	M	M	F	M	F	F	M	F	M	M	F	M

## Lesson No.11: Simple Regression Analysis and Concept of Multiple Regression

### ➤ Simple Regression

1. Describe the uses of Simple regression analysis in Education.
2. For the given data:

Year	2001	2002	2003	2004	2005	2006	2007
Research expense (in '000 Rs.)	14	15	18	19	20	22	23
Annual Profit ( in '000 Rs.)	80	89	91	93	98	103	124

- a) Develop the estimating equation that best describes the given data.
  - b) Estimate the annual profit when research expense made will 25000.
  - c) How much variation in the annual profits is explained by the variation in the research expenditure?
3. From the following data of the age of wife and age of husband, form the regression line which can best describe the relationship of age of husband and age of wife. Calculate the husband's age when wife's age is 36.

Husband's age	35	22	28	29	24	26	34	32	34	35
Wife's age	28	19	25	24	23	22	30	26	28	32

4. Given the following results for the height (x) and weight (y) in appropriate units of 500 students.

Mean of  $X = 64.3$ , mean of  $y = 152$ ,  $\sigma_x = 2.35$ ,  $\sigma_y = 20.6$ , and  $r = 0.62$ .

Obtain the equations of two regression lines. Estimate height of a student whose weight 210 units and also estimate weight of a student whose height is 69 units.

### ➤ Concept of Multiple Regression

1. Explain the utility of multiple regression.
2. Write the Multiple Linear Regression Model and explain its components.

### Lesson No.12: Z-Score

1. Define Z-score.
2. Describe the utility of z- score with some illustrations.
3. Identify which student has scored well from the following information:

Sl. No.	Name of the student	Obtained marks	Maximum marks	mean	s.d.
1	Ashok	35	50	40	7
2	Rina	58	100	65	10
3	Jinal	18	25	17	4

### Lesson No.14: Sampling Methods

**Sampling Methods (probability sampling- simple random sampling, cluster sampling, systematic sampling, stratified sampling, multi phase sampling, multi stage sampling; non-probability sampling- purposive sampling, judgmental sampling, convenient sampling, quota sampling, snow ball sampling)**

1. Describe probability and non probability sampling techniques with some suitable illustrations.

### Lesson No.15: Introduction to Inferential Statistics -I

1. Define the following terms:
  - a) Parameter and statistic
  - b) Hypothesis and its types
  - c) Level of Significance
  - d) Degrees of Freedom
  - e) Sampling Distribution
  - f) Standard Error
  - g) Sampling Error

### Lesson No.16: Introduction to Inferential Statistics -II

1. Differentiate parametric and non parametric tests.
2. Write the major steps of doing hypothesis testing.

## PARAMETRIC TESTS

### Lesson No.17: Z-test and its applications

#### ➤ Z-test

- **testing for mean (one sample problem)**

1. One LIC insurance agent has claimed that the average age of the policy holders who insure through him is less than the average age for all the agents, which is 34.2 years. A random sample of 650 policy holders who had insured through him has mean 34 years and standard deviation 5.8 years. Test his claim at 5% level of significance. (Given table value 1.645)

2. The mean breaking strength of the RR cables supplied by a manufacturer is 2245 with s.d. 106. By a new technique in manufacturing process, test whether the mean breaking strength of the cables increased or not. In order to test this, a sample of 500 cables is examined. It is found that the mean breaking strength is 2278. Use  $\alpha = 0.01$ . (Given table value 2.33)

3. A random sample of Canopy shoes worn by 200 combat soldiers in a desert region showed an average life of 2.2 years with a standard deviation of 0.07. Under the standard conditions, the Canopy shoes are known to have an average life 2.6 years. Is there reason to assert at a level of significance of 0.05 that use in the desert causes the mean life of such Canopy shoes to decrease?

- **testing for mean (two sample problem)**

4. The average production of rice of a sample of 150 fields is 220 lbs. per acre with a standard deviation of 12 lbs. Another sample of 160 fields gives the average of 260 lbs. with a standard deviation of 15 lbs. Can the two samples be considered to have been taken from the same population whose standard deviation is 14 lbs? Use 5 per cent level of significance.

5. A simple random sampling survey in respect of monthly earnings of skilled workers in two cities gives the following statistical information:

City	Mean of monthly earning	s.d. of monthly earning	Sample size
Ahmedabad	865	70	290
Surat	845	80	280

Test the hypothesis at 5 per cent level that there is no difference between monthly earnings of workers in the two cities.

## Lesson No.18: Z-test Application

### ➤ Z-test

- Testing for proportion (one sample problem)

6. Scarletline advertising company claims that 70% of the people saw an advertisement put out on the hoardings by the company, remembered the name of the product 48 hours after they had seen on the roads. In a sample survey conducted 48 hours after they pass through the road, 1578 out of 2100 persons remembered the name of the product advertised. Test the claim of the company at 1% level of significance. (Table value 2.57)

7. A company manufacturing a *light diet khakhara* for breakfast and claims that 60% of all housewives prefer that type to any other. A random sample of 3000 housewives contains 2165 that do prefer *light diet khakhara* for breakfast. At 5 per cent level of significance, test the claim of the company.

- Testing for proportion (two sample problem)

8. A Clinical trial agency is testing two new drugs, recently developed to reduce blood pressure level among the human beings. The drugs are administered to two different sets of animals. In group one, 756 of 1200 animals tested respond to drug one and in group two, 580 of 1100 animals tested respond to drug two. The research agency wants to test whether there is a difference between the efficacies of the said two drugs at 1% level of significance.

9. In a University 400 out of a random sample of 1500 students were found to be smokers. After a heavy imposition of tax on cigarette, another random sample of 1200 students of the same University was inspected and found that 230 were smokers. Was the observed decrease in the proportion of smokers significant? Test at 5 % level of significance.

## Lesson No.19: t- test and its applications

### ➤ t- test

- **testing for mean (one sample problem)**

1. According to cricket team coach, the mean height of female college cricket players is 67.5 inches. A random sample of 26 such players produces a mean height of 68.25 inches with s.d. of 2.1 inches. Assuming that the height of all female cricket players is normally distributed, test at 5% level significance whether their mean height is different from 67.5 inches. Also obtain 99% C.I for mean height of all female cricket players.

- **testing for mean (two sample problem)**

2. A XYZ Marketing Research Company wanted to investigate whether the male customers spend less money on average than the female customers. A sample of 26 male customers who shopped at this mall showed that they spend an average of Rs.280 with standard deviation of Rs.70. Another sample of 22 female customers who shopped at same mall showed that they spend an average of Rs.320 with standard deviation of Rs.130. Assume that the amounts spend at this mall by all male and female customers are normally distributed with equal but unknown population standard deviation. Construct 99% C.I. for the difference between the mean amount spent by all male and female customers at this mall. Using 1% level of significance, can you conclude that the mean amount spent by all male customers at this mall is less than that by female customers?

3. A University Professor wanted to know if TYBSc-students at her College tend to have more free time than the TYBCA students. She took a random sample of 25 TYBSc-students and 23 TYBCA students. Each student was asked to record the amount of free time he or she had in a specified week. The mean for the TYBSc-students was found to be 29 hours of free time per week with standard deviation of 5.5 hours. For the TYBCA students the mean was 22 hours of free time per week with a standard deviation of 6.5 hours. Assume that the two populations are normally distributed with unequal but unknown population standard deviation. Make 99% C.I for the difference between the corresponding population means. Test at 1% significance level whether the two population means are differ.

## Lesson No.20: t- test Applications

- **Paired t – test**

1. A government agency IITE claims that the crash course it offers significantly increases the typing speed of people. The following table gives the scores of ten people before and after they attended this course.

Before	81	75	89	91	65	70	90	64	78	77
After	97	72	93	110	78	69	115	72	90	87

Make 95% C.I for the mean difference of the population paired differences, where a paired difference is equal to the score before attending the course minus the score after attending the course. Using the 5% level of significance, can you conclude that the attending this course increases the typing speed of people?

2. A research team of gasoline additives claims that the use of this additive increases gasoline mileage. A random sample of 10 trucks was selected and these trucks were driven for one month without the gasoline additive and then for one month with the gasoline additive. The following table gives the miles per gallon for these trucks without and with the gasoline additives.

Without gasoline additive	23.6	26.3	19.9	24.7	25.4	29.5	27.5	30.2	31.5	29.5
With gasoline additive	26.3	31.7	22.2	27.3	25.3	30.9	29.6	34.7	35.2	32.5

Construct a 99% C.I for the mean difference of two paired population. Using the 1% level of significance, can you conclude that the use of the gasoline additive increase the gasoline mileage?

- **Testing for Correlation (one sample problem)**

3. The correlation coefficient between income and food expenditure for sample of 10 households from a low income group is 0.92. Using 1% level of significance tests whether the linear correlation coefficient between incomes and food expenditure is positive. Assume that the population of both variables is normally distributed.

7. The correlation between ages of cars and its prices for eight cars of a specific model are  $-0.65$ . Test at 5% level of significance whether population correlation coefficient is negative.

## Lesson No.21: F- test and its application: ANOVA

### ➤ F-test ANOVA (one way)

1. Test at 10% level of significance. Is there significant difference among the variety of Basmati Rice produced per acre of land? Use the following information.

	Per acre production of Basmati Rice		
	Variety of Rice		
Plot of land	A	B	C
1	16	15	17
2	17	16	18
3	13	17	19
4	15	18	16
5	16	14	18

2. Test at 5% level of significance. Is there any significant difference among the teaching strategies A, B & C. Marks are given below and gained after the implementation of same test on three groups of students. Use the following information.

	Teaching strategies		
Roll No. of students	A	B	C
1	23	24	34
2	24	35	34
3	32	36	38
4	34	43	47
5	43	42	41
6	36	49	42
7	37	43	45
8	38	41	44
9	32	24	47
10	34	35	42



## Lesson No.22: F- test Application: ANCOVA

1. What is ANCOVA? Write the assumptions need to check prior use of ANCOVA.

2. The following are paired observations for three experimental groups:

GROUP I		GROUP II		GROUP III	
X	Y	X	Y	X	Y
17	2	25	8	32	15
16	5	24	12	33	16
19	7	25	15	35	20
15	9	29	18	36	24
12	10	21	19	41	30

Y is the covariate variable. Test the significance of differences between the adjusted means on X by using the appropriate F-ratio. Also calculate the adjusted means on X.

## Lesson No.23: Chi-square test and its applications

### • Testing for Variance (One Sample Problem)

1. The following are the prices of the same brand of digital camera found at eight stores in New York.

\$876, 798, 764, 785, 843, 872, 769, 870

Test at 5% significance level whether the population variance is different from 485 square dollars with population mean \$845.

2. A random sample of 28 customers taken from SBI bank gave the variance of the waiting times equal to 3.8 square minutes. Assume that the waiting times for all customers are normally distributed. Test at the 1% significance level whether the variance of the waiting times for all customers at this bank is greater than 4.0 square minutes. (Table value 42.98)

3. The body weight of 10 students is given below:

Sl. No.	1	2	3	4	5	6	7	8	9	10
Weight (kg.)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weights of all students from which the above sample of 10 students was drawn is equal to 15 kg? Test this at 1% level of significance.

## NON-PARAMETRIC TESTS

### Lesson No.24: Chi –Square Test application

#### ➤ Chi-square test

##### • Testing for Independence of Two Attributes

1. The table given below shows the data obtained during outbreak of viral infection:

	Attacked	Not attacked	Total
Vaccinated	50	450	500
Not Vaccinated	160	1340	1500
Total	210	1790	2000

Test the effectiveness of vaccination in preventing the attack from viral infection. Test your results with the help of chi square test at 1% level of significance.

2. The following information is obtained concerning an investigation of 190 ordinary shops of small size:

	Shops		
	In towns	In villages	Total
Run By Men	45	76	121
Run By Women	15	54	69
Total	60	130	190

Can it be inferred that shops run by women are relatively more in villages than in towns? Use Chi square test at 5% level of significance.

3. The following values of Chi-square from different investigations carried to examine the effectiveness of a recently invented medicine for checking dengue are obtained:

Investigation	Chi square values	d.f.
1	2.5	1
2	3.2	1
3	4.1	1
4	3.7	1
5	4.5	1

What conclusion would you draw about the effectiveness of the new medicine on the basis of the five investigations taken together?

- **Testing Whether Observations Are Normally distributed or not**

4. Seventy salesman has been classified into 3 groups – Excellent, satisfactory and poor- by a consensus of sales managers. Does this distribution of ratings differ significantly from that to be expected if selling ability is normally distribution in our population of salesmen?

(a)

	Excellent	Satisfactory	Poor
Observed frequency	20	40	10

(b)

.	Excellent	Satisfactory	Poor
Observed frequency	13	27	30

- **Testing whether observations are equally distribution or not**

5. The items in an attitude scale are answered by underlining one of the following phrases: strongly approved, approved, indifferent, disapproved, strongly disapproved. The distribution of answers to an item marked by 100 subjects is shown below. Do these answers diverge significantly from the distribution to be expected if there are no preferences in the groups?

(a)

	strongly approved	approved	Indifferent	Disapproved	strongly disapproved
Observed frequency	23	18	24	17	18

(b)

	strongly approved	approved	indifferent	Disapproved	strongly disapproved
Observed frequency	25	11	18	22	24

### Lesson No.25: Median Test and Sign Test

#### ➤ Median test

1. A clinical psychologist wants to investigate the effects of a tranquilizing drug upon hand tremor. 10 psychiatric patients are given the drug and 14 other patients matched for age and sex are given placebo. Tremor is measured by a steadiness tester. For the given below scores, test whether the drug increases the steadiness -as shown by lower scores in the experimental group?

Sl. No.	Experimental group scores	Sl. No.	Control group scores
1	37	1	39
2	45	2	50
3	43	3	56
4	42	4	47
5	48	5	49
6	65	6	65

7	42	7	46
8	53	8	59
9	54	9	68
10	74	10	80
		11	65
		12	54
		13	34
		14	36

➤ **Sign test (one sample test)**

1. In a clinical study 12 patients were observed for appearance transit times for occluded right coronary arteries. Data is given below:

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Transit time (in sec)	1.40	3.20	5.46	2.15	2.55	3.45	2.65	3.41	2.47	2.87	3.10	3.20

Can we conclude, at the 5% level of significance, that the median appearance transit time in the population from which the data were drawn, is different from 3.25 seconds?

2. The following data are recorded from 15 drug abusers whose age is 16 or older. All these drug abusers were arrested by police and their IQs' were recorded. Is there any evidence that the median IQ of drug abusers in the population is greater than 110? Use 5% level of significance.

94	103	94	95	145	128	137	121
149	114	117	129	107	115	135	

## Lesson No.26: Mann Whitney U-test

### ➤ Mann Whitney U-test

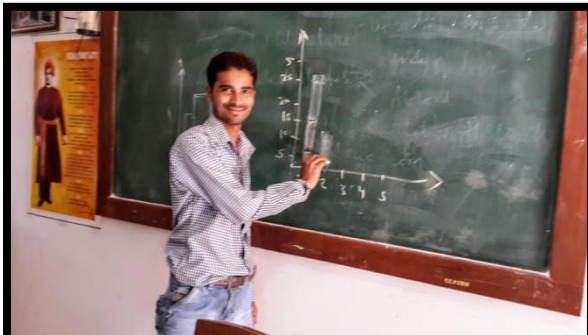
1. A teacher wishes to evaluate the effects of two methods of teaching which is applied to two different groups say Group A and Group B respectively for 20 randomly assigned students, drawn from the same population. Scores obtained by each student are given below:

Sl. No.	Group A	Group B
1	50	49
2	60	90
3	89	88
4	94	76
5	82	92
6	75	81
7	63	55
8	52	64
9	97	84
10	95	51
11	83	47
12	80	70
13	77	66
14	80	69
15	88	87
16	78	74
17	85	71
18	79	61
19	72	55
20	68	73

\*\*\*\*\*

## Appendix: IX

### Photographs taken during the Data Collection Process











## Appendix- X

### Permission Letter



Department of Education [CASE]  
Faculty of Education and Psychology  
The Maharaja Sayajirao University of Baroda  
Vadodara 390 002  
Phone: 0265 2795516, 2792631

Date: 16<sup>th</sup> Dec 2015

To  
The Head  
CASE, Department of Education  
Faculty of Education & Psychology  
The M.S. University of Baroda  
Vadodara

Subject: Permission for data collection for my Doctoral Research

Respected Sir,

I (Sonia Rohilla) am pursuing my Doctoral Study titled "Development of an Educational Program on Data Analysis Techniques for M.Ed. Students through Cooperative Learning" under the guidance of Dr. D. R. Goel. In this regard, kindly allow me to conduct classes of M.Ed. Semester-II course –Educational Research Methodology –II at CASE. The data collected will be used for research purpose only.

Thanking You.

Yours Truly,

Sonia Rohilla  
Research Scholar  
Department of Education  
Faculty of Education & Psychology  
The M.S. University of Baroda  
Vadodara

Through:

Dr. D. R. Goel  
Guide

## Appendix- XI

### Course Work Certificate



**THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA**

### CERTIFICATE

*[As per O.Ph.D. 2 under UGC (Minimum Standards and Procedure for Awards of M.Phil. Ph.D. Degree) Regulation, 2009 for 15 Credits to be earned by Ph.D. Scholars]*

This is to certify that **Rohilla Sonia Narenderkumar**, Research Scholar, registered under UGC (Minimum Standards and Procedure for Awards of M.Phil./Ph.D. Degree) Regulation, 2009, vide Registration Certificate Number **64** dated **17/07/2012**, for pursuing Ph.D. on has undertaken and completed the course work with the Grade A.

### STATEMENT OF CREDITS EARNED

Name of Research Scholar: **Rohilla Sonia Narenderkumar**

Faculty/Institution: Faculty of Education and Psychology

Department: Department of Education

Paper Number	Course Title	Course Credits	Grade Earned
<b>Core Courses – 09 Credits [Offered at University Level]</b>			
I.	Introduction to Research & Research Writing	3	A
II.	Introduction to Basic Computer Functions & Applications for Research Purposes	3	A
III.	Quantitative Research Techniques & Data Analysis	3	A
<b>Departmental Courses – 06 Credits [Offered at Departmental Level]</b>			
IV.	Review of Related Literature	3	A
V.	Conceptual and Theoretical Framework	3	A
<b>Overall Grade</b>			<b>A</b>

UC : 43 (Phase 4)

DC : 80 (090919)

FOEDU/64

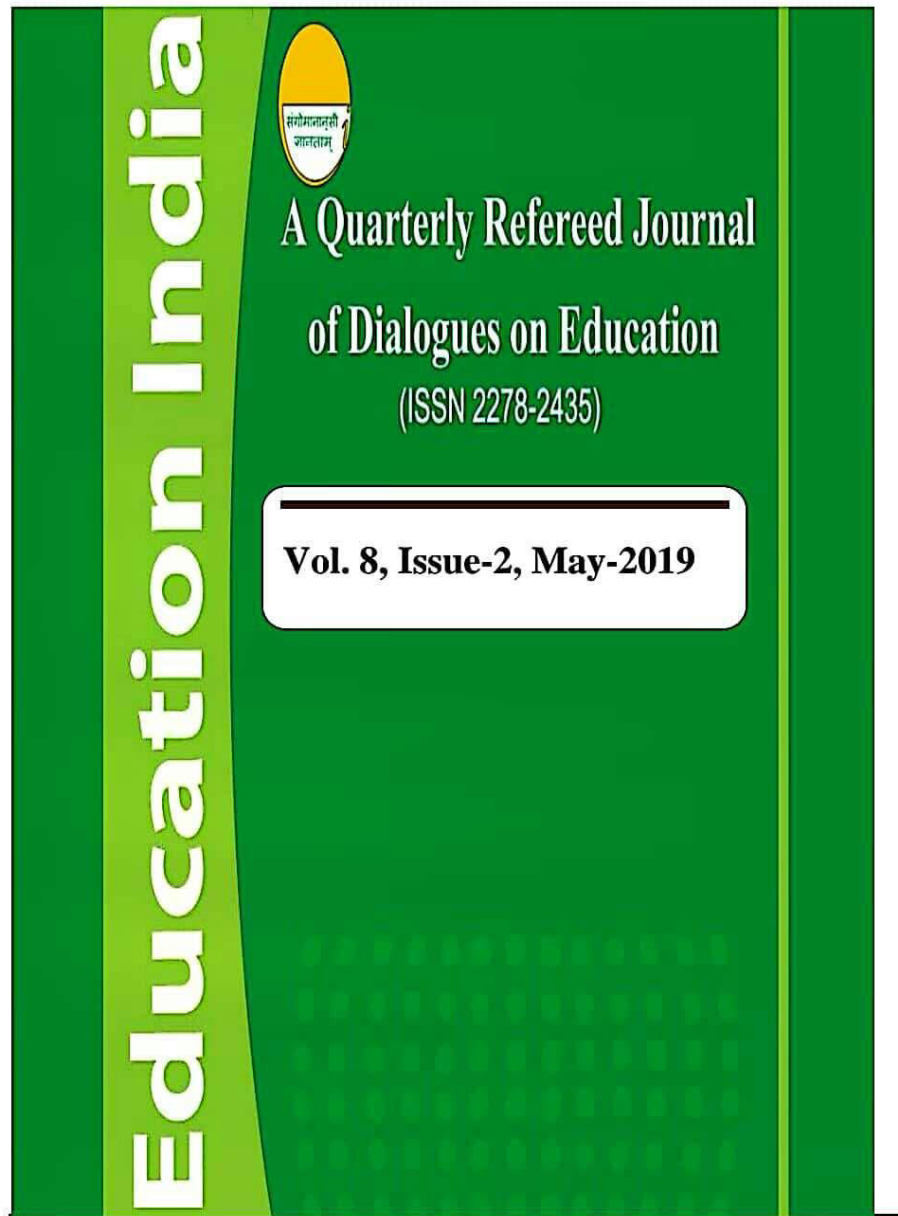
Date of Issue: 12/09/2019

Place: Vadodara

  
Registrar (I/c.)

## Appendix- XII

### Paper Published



Education India Journal: A Quarterly Refereed Journal of Dialogues on Education, ISSN 2278-2435, Vol. 8, Issue-2, May-2019. Page 1



## Index

Sl. No.	Paper Title	Author	Page No.
1.	Teaching Statistics Through Cooperative Learning: An Evaluative Study	Sonia Rohilla	03-08
2.	Conceptualisation of Pedagogical Content Knowledge (PCK) of science from Shulman's notion to Refined Consensus Model (RCM): A journey	Mr. Sudhendra Roy Dr. Shyamsundar Bairagya	09-53
3.	Environmental Awareness among Senior Secondary School Students: A Comparative Study	Arup Bhowmik Dr. Anju Verma	54-67
4.	A Study on the Emerging Trends of Teacher Education Programme in Assam	Rabbul Hussain Dr. Anita Singha	68-78
5.	A Study of the Study Habits of Secondary Level Students	Dr. Lalima Ms. Swati Vishwkarma	79-90
6.	Women's Participation In Workforce In India	<b>Neelam Kumari</b>	91-114



**A Quarterly Refereed Journal  
of Dialogues on Education  
(ISSN 2278-2435)**

**Paper-1**

**Teaching Statistics Through Cooperative  
Learning: An Evaluative Study**

Sonia Rohilla

Education India Journal: A Quarterly Refereed Journal of Dialogues on Education, ISSN 2278-2435, Vol. 8, Issue-2, May-2019.

Page 3

## Teaching Statistics Through Cooperative Learning: An Evaluative Study

Sonia Rohilla<sup>1</sup>

### Abstract

*Personalized Teacher Education is highly desirable. How a teacher educator could be impersonally personal? It is highly desirable to provide differentiated differential inputs. But how to? It is highly demanding on the part of a Teacher educator in a face to face Teacher Education to attend to each and every pupil teacher or prospective teacher educator, personally. Teaching employing cooperative learning is a resolve. The present study attempts to teach statistics at the M.Ed. level employing cooperative learning. The study demonstrates very well that how teaching statistics employing cooperative learning enhances achievement in statistics at M.Ed. level. Not only it enhances achievement in statistics, but, develops many a affect attributes, such as, fellow feeling, team work, sharing the group energy scientifically, team mind, networking, sharing the strength and weakness, adjustment, resilience, interrelation, interdependence, harmonious existence. It presumes the basic premise that the whole approach is very often greater than the individual approach because of interrelation and interdependence amongst various individuals. Unity breeds cooperation and greater returns on investment.*

\*\*\*\*\*

Education is an essential need of every individual. The basic foundation of any education system depends upon its values. Education without humane values is meaningless. There are various values which develop among the students at the time of learning. It is the pivot responsibility of teachers to develop and enhance such humane values among the students while teaching. There are various humane values, like, respecting one another, valuing perception of others, discussing academically, healthy learning environment, putting arguments logically, disagreeing scientifically, leading the team, giving opportunity to others, helping and cooperating others and

---

<sup>1</sup> Research Scholar, Department of Education, Faculty of Education and Psychology, The Maharaja Sayajirao University of Baroda, Vadodara- 390002, Gujarat\_ India

adjusting with others. All such values are very important for a student and for any individual for establishing a healthy society.

According to report to UNESCO of the International Commission on Education for the Twenty-first Century (1996), "Education must be organized around four fundamental types of learning, which, throughout a person's life will be in a way the pillars of knowledge: learning to know, that is acquiring the instruments of understanding; learning to do, so as to be able to act creatively on one's environment; learning to live together, so as to participate and co-operate with other people in all human activities; and learning to be, an essential progression which proceeds from the previous three". In order to develop and enhance such values among the M.Ed. students an initiative was taken to employ a Cooperative Learning strategy while teaching data analysis techniques to the M.Ed. students. Here cooperative learning in singular sense means that growing by giving and taking help of others. But in plural sense the meaning of cooperative learning means as described by Johnson & Johnson (1995), cooperative learning is an instruction that involves students working in teams to accomplish a common goal, under conditions that include the following six essential elements:

- The first element is Positive Interdependence. Positive interdependence means that a gain for one student is associated with gains for the other students. The discipline of using cooperative groups begins with structuring positive interdependence. It is positive interdependence that requires group members to work together to accomplish something beyond individual success.
- The second element is Equal participation. Equal participation refers to the fact that no student should be allowed to dominate a group, either socially or academically.
- The third element is Individual Accountability. Individual accountability exists when the performance of each individual member is assessed, the results are given back to the individual and the group to compare against a standard of performance, and the member is held responsible by group mates for contributing his or her fair share to the group's success.
- The fourth element is Simultaneous Interaction. In cooperative group, group members meet face to face to work together to complete assignments and promote each other's



success. Although some of the group work may be parceled out and done individually, but most of the work must be done interactively with group members providing one another with feedback, challenging reasoning and conclusions, and perhaps most importantly teaching and encouraging one another in the group.

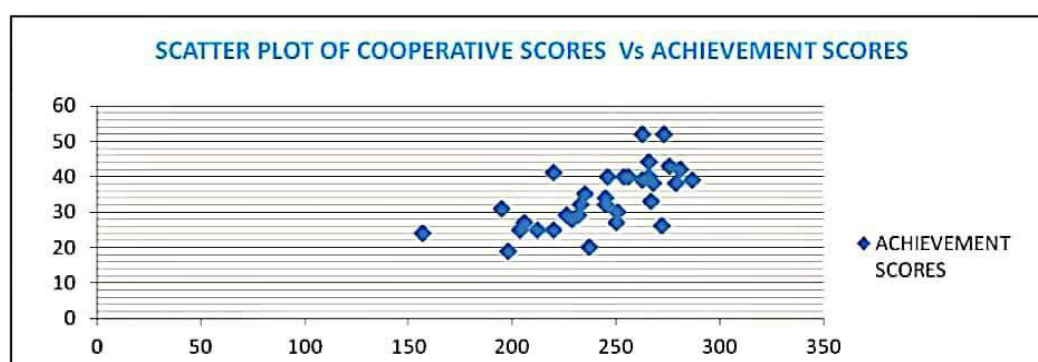
- The fifth element is interpersonal and Small Group Skills. Cooperative learning is inherently more complex than competitive or individualistic learning because students have to engage in task work and teamwork simultaneously to coordinate efforts that will achieve mutual goals. Here students are encouraged and helped to develop and practice trust-building, leadership, decision-making, communication, and conflict management skills.
- The sixth element is structuring group processing. Group processing may be defined as reflecting on a group session to (a) describe what member actions were helpful and unhelpful and (b) make decisions about what actions to continue or change. Team members set group goals, periodically assess what they are doing well as a team, and identify changes they will make to function more effectively in the future.

It is essential to make our students ready for developing cooperative spirit in learning and in carrying out our day to day life tasks. After the completion of formal education of a student the values which were developed during the formal education direct the entire life of a student significantly. In M.Ed. program students are very often heterogeneous in the sense that they differ by age, they differ by subject streams, namely, arts, commerce, science and languages, they differ by mathematical aptitude, they differ by IQ and they differ by individual choices and so on. So in such a heterogeneous group it is desirable to study the impact of cooperative learning on values and the academic achievement scores of the M.Ed. students. Therefore here in this paper an attempt has been made to focus upon the various values emerged among the students while learning in cooperative groups and also explain the relationship between the cooperative scores and the achievement scores. In order to develop and enhance cooperative environment among the students while learning cooperative learning strategy was employed which comprised of six major components, namely, positive interdependence, equal participation, individual accountability, simultaneous interaction, interpersonal and Small Group Skills and structuring group processing. The author has conducted 36 classes spanned over four

---

Education India Journal: A Quarterly Refereed Journal of Dialogues on Education, ISSN 2278-2435, Vol. 8, Issue-2, May-2019. Page 6

months with M.Ed. students of the academic year of 2014-2015 of The Maharaja Sayajirao University of Baroda, Vadodara. Depending upon these components of cooperative learning a rating scale with 61 statements was constructed and administered on all the 33 M.Ed. students and cooperative learning scores for each student were calculated. Also an achievement test with 70 items was administered on the same group of students and respective scores were calculated. From the received cooperative scores and the achievement scores relationship was studied. The relationship of cooperative scores and the achievement scores is shown below:



The correlation coefficient obtained between cooperative scores and the achievement scores is 0.65304. It is a high positive value. It means that there is a high positive correlation between cooperative scores and the achievement scores of the students. Hence we can conclude that if cooperative score increases achievement score will also increase. From the obtained correlation coefficient value, the coefficient of determination is 0.4274 i.e. 42.74% of change in the achievement scores of the students is explained by the cooperative scores of the students. This conclusion indicates that cooperative learning is an effective strategy in raising the achievement scores of the students. From the interview of the students of M.Ed. following reflections were made by them regarding the cooperative learning:

- Cooperative learning helps in the reduction of fear in learning of data analysis techniques.
- Cooperative learning raises our enthusiasm in attending classes.
- Cooperative learning gives opportunity to the talented minority as well as to the non-participating majority at the same time.

- Cooperative learning creates healthy environment of learning and sharing knowledge among the peers.
- Cooperative learning builds caring and sharing nature of peers.
- Cooperative learning enables us in understanding the technical statistical language while reading the data analysis techniques from various sources.
- Cooperative learning built team spirit and team learning.

### **Concluding Remarks**

Teaching using cooperative learning has been found to be an effective teaching- learning approach for, both, the development of affect attributes and learning of data analysis techniques at M.Ed. level.

### **References:**

- Johnson, D.W & Johnson, R.T. (1995). *Learning Together and Alone: Cooperative, Competitive and Individualistic Learning*. USA: Allyn and Bacon.
- UNESCO (1996). *Learning to Live Together*. Retrieved on 23/10/11 from <http://www.unesco.org/delors/index.html>.

\*\*\*\*\*