Department of Open & Distance Learning

Punjabi University, Patiala

Class : M.A. I (Education) **Semester : 2**
Paper : III (Methodology of Educational Research-II) **Unit : I**
Medium : English

Lesson No.

Department website : www.pbidde.org

M.A. (EDUCATION) PART-I **PAPER–III**
EDUCATIONAL STATISTICS

LESSON NO. 2.1 **AUTHOR : DR. Y.P. AGGARWAL**

Significance of Mean difference

Structure of the Lesson :

2.1.1 Objectives

The students will be able to :

(1) understand parametric and non-parametric statistics
(2) differentiate between null hypothesis and alternative hypothesis
(3) solve the problem related to difference between means
(4) distinguish Type-I and Type-II errors.

2.1.2 Parametric and Non-parametric Statistics

Parametric statistical methods are based on assumptions about the population from which the sample has been drawn. Particularly the assumptions like form of the probability distribution. Under the assumptions of normaity of population parametric tests are most powerful, i.e. with the use of parametric test, probability of researcher's correctly rejecting the null hypothesis is quite high.

If the assumptions do not hold good or the data do not meet the requirement of parametric statistical methods, then non-parametric methods can be used in alternate to the parametric methods. Non parametric methods required very mild assumptions like continuity and symmetry of the distribution.

In experimental and other research work, the determination of real and observed difference is of such magnitude that it cannot be attributed to chance factors of sampling variations, and it is often our major interest. For example, we may observe that a group of subjects tested under one set of experimental conditions has a higher mean than a compatible group tested under a different set of experimental conditions has a higher mean than a compatible group tested under a different set of experimental conditions. Is the observed difference between the means is one that might occur frequently by chance or is it as a result of sampling variations? If not then we might infer that differences is a product of the experimental conditions. For this purpose we need a statistical tested of significance of difference between the means. The critical ratio or the ‘t’ test is the only generally used in such circumstances.

2.1.3 The Null Hypothesis, (Ho) and Alternative Hypothesis (H_1)

The null hypothesis is a proposition of zero difference. Fisher has emphasized that every experiment may be said to exist only in order to give a chance of disproving the null hypothesis. Thus a hypothesis which is setup with the possibility of its being rejected at some defined probability value is called a null hypothesis. The term “null” referring to our interest in the possible rejection of the hypothesis. In statistical terms a null hypothesis may be stated as.

Ho : $\mu_1 = \mu_2$ where μ_1 and μ_2 population means.

It states that there is no significant difference between the means of the two populations. In fact it leads to the rejections or non-retention of the Ho. Then the alternative hypothesis (H_1) stated as below stands accepted.

In which the symbol ≠ means “not equal to”, and μ_1 and μ_2 are two populations means. Thus the alternative hypothesis is $H_1 : \mu_1 \neq \mu_{2.}$ It means that two populations differ significantly.

The Process

The process of testing for significance of difference between the two means includes the following steps :

(i) Set up H_0 and the H_1, according to the requirements of the problem.

(ii) Decide about the level of significance for the test. Customarily .05 and .01 level are selected.

(iii) Decide whether one tailed or two tailed test of significance was needed.

(iv) Decide whether the data warranted a test of significance for the independent or the correlated means.

(v) Decide whether large sample or the small sample was involved.

(vi) Use one of the following formulas appropriated to (iv) and (v) above, for the calculation of SE (Standard Error) of Mean Difference SE_D.

1. Independent of Uncorrelated Mean

Discription of Symbols

Large Sample

$$SE_D = \sigma(M_1 - M_2)$$

? = S.E. difference between Means

$$= \sqrt{(\sigma M_1)^2 + (\sigma M_2)^2} = \sqrt{\frac{\sigma\frac{2}{1}}{N_1} + \frac{\sigma\frac{2}{1}}{N_2}}$$

σM_1 = S.E. of the mean of the first sample

σM_2 = S.E. of the mean of the second sample

σ_1 & σ_2 are SDs of the two groups

N_1 & N_2 = number of cases in the groups,

Small Sample

$$SE_D = SD\sqrt{\frac{N_1 + N_2}{N_1N_2}}$$

SD = Pooled Standard Deviation of the groups.

M_1 & M_2 = Mean of the two groups.

$$SD = \sqrt{\frac{\sum(X_1 - M_1)^2 + \sum(X_2 - M_2)^2}{N_1 + N_2 - 2}}$$

Where $SD = \sqrt{\frac{\left(\Sigma M_1 - M_2\right)^2 + \left(\Sigma X_2 - M_2\right)^2}{\left(N_1 - 1\right) + \left(N_2 - 1\right)}}$ X_1 & X_2 = Individual raw scores in the two groups.

2. Correlated Means

Large Sample

r_{12} = Correlation Coefficient between scores of Group I and II, other symbols.

$$SE_D = \sqrt{\frac{SE^2M_1 + SE^2M_2 - 2r_{12}SE_{M1}SE_{M2}}{N_1 + N_2 - 2}} = \sqrt{\sigma M\frac{2}{1} + \sigma M\frac{2}{2} - 2r_{12}.\sigma_{M1}.\sigma_{M2}}$$

$SE_M = \frac{SD}{\sqrt{N-1}}$ as above

SD= The standard deviation of the, different Scores. N = No. of persons in the group.

(vii) Calculate the value of the critical ratio, or t by using formula.

$\frac{M_1 - M_2}{SE_D}$ In which M_1 & M_2 are the means to be compared the SE_D is the SE of the Mean Difference calculated under step (vi) above.

(viii) Calculate Degree of Freedom (df) as below

(a) For uncorrelated or Independent samples, df = $(N_1 + N_2)$ - 2.

(b) For Correlated Samples, df = N-1.

(ix) Look up the table of value with df:(as decided in step vii above) and the level significance (Step ii).

(x) Decision Rule : Compare with the calculated value of 't'

(a) If the calculated value of t is equal or more than the table value of it, rejected H_0 and accept H_1.

(b) If the calculated value of t is less than the table value of it accept H_0.

(xi) Interpret the results as below

(a) H_0 rejected : There is a significant difference between the two means.

(b) H_0 accepted : There is no significant difference between the two means.

Whatever the difference, it has arisen due to sampling fluctuation and chance factors only.

The procedural steps in the use of the t test of significance for differences between two means will now be explained with help of some numerical examples. For the purpose of convenience, the examples have been arranged in two sections. The first section is concerned with the comparison of independent uncorrelated means. The second, section is on comparison of correlated means.

2.1.4 Standard Error (SE) or Difference between two Independent Means Large Sample

When two distinct groups of subject are involved, the groups may be termed as independent. These groups are drawn at random from totally different and unrelated population. No attempt is made to equate the groups by using pair comparison or any other.

Example :

Thirty boys and forty girls selected randomly from the eighth class of a big school were given a standard test of Arthmetic ability. The means and SD's are reported below :

	N	M	SD
Boys	30	20.5	4.0
Girls	40	16.2	5.2

Is the mean difference in the Arithmetic ability significant ?

Hypothesis :

H_0 : $\mu_1 = \mu_2$

H_1 : $\mu_1 \neq \mu_2$

Decision Rules :

Given : .05 significance level & df = $(N_1 + N_2) -2 = 68$ and table value of t = 2.0

If t calculated < 2.0, accept Ho.

If t calculated > 2.0, reject Ho.

Computation

$$SE_D = \sqrt{SE^2 M_1 + SE^2 M_2}$$

Substituting the numerical values.

$$SE_D = \sqrt{\frac{(4.0)^2}{30} + \frac{(5.2)^2}{40}} = 1.06$$

$$t = \frac{M_1 - M_2}{SE_D}$$

$$= \frac{20.5 - 16.2}{1.06} = \frac{4.3}{1.06} = 4.06$$

x_1 y_2

Interpretation : Rejected H_o because 4.05 > table value of 2.0.

Note : If the 't' value table is not handy, note that if value of t is less than 1.96. it is not significant. If it is 1.96 or more, upto 2.57 it is significant at .05 level of confidence. If it is 2.58 or more it is usually significant at .01 level of confidence.

2.1.5 The SE of Difference between Means In small Independent Samples

Example

An attitude test was administered to 10 boys in English class and 5 boys in Hindi class. Their scores are given below. Is the mean difference between the groups significant?

Hypothesis is :

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Decision Rules :

Given Significant Levels = 0.5 & df = (N_1+N_2) -2 = 13 and table value of t = 2.16

If calculated value of t > 2.16 reject H_0

Computation

English Course			Hindi Course		
X	x	x^2	Y	Y	Y^2
6	-4	16	4	1	1
7	-3	9	3	0	0
8	-2	4	2	-1	1
10	0	0	1	-2	4
15	+5	25	5	2	4
16	+6	36	$\Sigma Y=15$		$\Sigma Y^2=10$
9	-1	1			
10	0	0			
9	-1	1			
$\Sigma X_1 = 100$		$\Sigma X^2=92$			

$$M_1 = \frac{90}{9} = 10 \qquad M_2 = \frac{15}{5} = 3$$

$$\text{Formula : SD} = \sqrt{\frac{\Sigma X^2 1 + \Sigma X^2 2}{(N_1 - 1) + (N_2 - 1)}} = \sqrt{\frac{92 + 10}{8 + 4}} = 2.92$$

$$SE_D = SD\sqrt{\frac{N_1 + N_2}{N_1 N_2}} = 2.92 \times \sqrt{\frac{9 + 5}{9 \times 5}} = 2.92 \times \sqrt{0.3} = 1.63$$

$$t = \frac{M_1 - M_2}{SE_D} t = \frac{M_1 - M_2}{SE_D} = \frac{(10 - 3)}{1.63} = 4.29$$

Interpretation

Reject Ho because calculated value (4.29) > table value (2.16). It shows that the two groups differed significantly on their mean attitude scores.

2.1.6 Standard Error of the Difference between Two correlated Means :

Correlation between means is introduced in the following situations :

(a) The Single Group Situations : When a single group is tested twice on the same test an equivalent form of the test is used on the second occasion.

(b) The Equivalent Group Situations : When equivalent group are formed by using “matching by pairs”

If the some group of students take Arithmetic ability test twice instead of two different group taking it we have the same individual’s score on the first testing pair of with his Score in the second testing. If in a comparison of males and females, the two Groups are standardized better by taking a brother or a sister from each family or if the boys and girls are paired with respect to age. I.Q or social status, and if such factors of common family are I.Q or social status have any relations to Arithmetic ability. They will automatically introduce correlation between the two samples.

A Correlation coefficient is computed and introduced in the relevant formula. A numerical example using the single group situations is given below :

Example : A group of 35 randomly selected students was tested before and after an experiment treatment. The data so obtained are given below :

	Pretest	Post test	
Mean	15.5	21.6	r = 70
SD	5.2	4.8	N = 35

Find out if the groups differed significantly on the two testing.

Hypothesis

$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$

Decision Rules ; Given significance level = 0.1 & df = N-l = 34. and table value of t=2.72.

If calculated t < 2.72, accept

If calculated t $\geq$ 2.72 reject Ho

Computation : Formula $SE_D = \sqrt{\sigma M\frac{2}{1} + \sigma M\frac{2}{2} - 2r\sigma_{M1}\sigma_{M2}}$

$$\sigma M_1 = \frac{\sigma_1}{\sqrt{N}} = \frac{5.2}{\sqrt{35}} = .88 \left(\sigma_2 \text{ is the SD of pre - test}\right)$$

$$\sigma M_2 = \frac{\sigma_2}{\sqrt{N}} = \frac{4.8}{\sqrt{35}} = .81 \left(\sigma_2 \text{ is the SD of pre - test}\right)$$

$$SE_D = \sqrt{(.88)^2 + (.81)^2 - 2x(0.70)x(.88)x(.81)} = 1.15$$

$$t = \frac{M_1 - M_2}{SE_D} = \frac{6.1}{1.15} = 5.3$$

Interpretation :

Reject Ho, as calculated value of t (5.3) $\geq$ 2.72

There is significant difference between the mean scores of the group on pre-test and post test.

Difference Methods : (Small Sample)

When groups are small, the procedure called the Difference Method is often to be preferred. It is quicker and easier to apply than the long method of calculating SE for each mean and the SE of the difference. It is to be preferred if the value of the correction coefficient between the two sets of scores is not required for any other purpose.

Examples :

Ten subjects were given three successive trails test. The score for the first and the last trials is shown below. Is the mean gain from the first lo the third trial significant ?

Hypothesis Ho : $\mu_1 = \mu_2$ $H_1 = \mu_1 \neq \mu_2$

Decisions Rules :

Given Significance level = .01 & df = N-l = 9 and table value of t = 3.25

If calculated t < 3.25, accept Ho

If calculated $\geq$ 3.25, reject Ho

Trial-I	Trial-II	Difference (D)	X	X2
(TI)	(TII)	(TII -TI)	(D-3)	
12	16	4	1	1
14	18	4	1	1
10	17	7	4	16
8	10	2	-1	1
16	18	2	-1	1
17	25	8	5	25
18	18	0	-3	9
20	21	1	-2	4
16	17	+1	-2	4
19	20	+1	-2	4
Σ=150	Σ=180	ΣD=30		ΣX_2=66

$$\text{Mean D} = \frac{\Sigma D}{N} = \frac{30}{10} = 3.0$$

$$SD = \sqrt{\frac{\Sigma x^2}{N-1}} = \sqrt{\frac{66}{9}} = 2.7$$

$$SE = \frac{SD}{\sqrt{N}} = \frac{2.71}{\sqrt{10}} = 0.86$$

$$t = \frac{MD - 0}{SE_D} = \frac{3-0}{0.86} = 3.49$$

Interpretation :

Reject Ho because calculated value (3.49) > tabulated value (3.25)

Hence there is a significance difference between the means on two trials.

In the above- example, non-directional hypothesis was put forward. The H_1 did not mention any direction of the difference. The difference could be in favour of the first trial also. However, if our hypothesis had been that there would be gain of successive trial, a one-tailed test would have been used. In that case critical value of t from the table would have been read as follow :

(i) For 9 df the .01 level is read from the .02 col. (p/2= .01) ; t value = 2.82

(ii) For 9 df the 0.5 level is read from the .01 col. (p/2 = 0.5); t value = 1.83

The calculated value of t in this example is much larger than the value of t required for significance at both the level with directional hypothesis. Hence the mean gain from the first to the third trial on the test is significantly larger.

Some Conceptual Issues :

The testing of the significance difference between the two means required the understanding of some concepts. The meaning and examples about the Null Hypothesis have already been given. Here an attempt is made to clarify some of the others issues :

2.1.7 Sampling Distribution and the SE

If a large number of samples are taken from the same population and same test administered to them under identical conditions the average scores of means of these samples can be calculated. If the means so obtained are arranged in the form of frequency distribution and also plotted on a graph as a frequency polygon, we obtain Distribution means, which is called the sampling distribution mean. The difference between a distribution of scores and a sampling distribution lies in the fact that the former is based on an arrangement of score while the latter on that of means of any other statistics. It has been found that even if the distribution tends to reach a normal space, the μ is the population mean of this distribution.

The Standard Error (SE) is the standard deviation of the sampling distribution and is to be interpreted in the same manner. The sampling distribution is estimated form the sample standard deviation which is the only value available to us.

2.1.8 Type I and type II Errors

Research required testing of hypothesis. In this process two wrong inference can be drawn. These are called Type I and Type II errors.

Type I error is committed when we reject a null hypothesis by making a difference significant, although no true difference exists.

Type II errors is committed when we accept a null hypothesis by making the difference not significant, when a true difference actually exists. These can be shown as below ;

	Reject H_0	Accepts H_0
H_0 is Ture	Error-Type I	No Error
H_0 is False	No Error	Error-Type II

2.1.9 Simple Regression and Prediction

The term regression was first used by Galton. He found that children of tall parents tended to be less tall and children of short parents less short than their parents. This shows that heights of the children tend to move back towards the mean height of the general population. The tendency towads maintaining the mean height is called principle of regression by Galton. The line explaining the relationship of height in parent and their children, was named regression line. Suppose in a group of 50 students, when their height and weight are known, we can estimate with regression line a students' weight if we know his height.

The line of regression is a straight line which gives the best fit in the least square to the given frequency distribution, if the straight line is so chosen that the sum of Y on X gives the best estimates of Y for any given value of X.

Calculation of Regression Equations

Regression equations can be calculated in two forms :-

1. Deviation form : The equation gives the relationship of the deviation from mean height to deviations from mean weight. It can be written as :-

$$Y = r\frac{\sigma_y}{\sigma_x} \times X$$

(regression equation of Y on X, deviations taken from the means of Y and X)

The factor is called Regression Coefficient. It gives the relationship between Y and X in deviation form.

For example :	Mx = 136.3 Ibs.	σy = 2.62
	My = 66.5 inches	σx = 15.55
		r = 0.60

$$Y = 0.60\frac{2.62}{15.55} \times X = 0.10\ X$$

The equation of the second regression line is :-

$$X = r\frac{\sigma y}{\sigma x} \times Y$$

(regression equation of X on Y, deviations taken from the means of X and Y). In the above example the equation can be :-

$$X = 0.60 \times \frac{15.55}{2.62} Y = 3.56Y$$

It gives the relationship X and Y in deviation form, when the two lines are expressed in internal units, regression equations do not give the relationship between the X and Y score deviations. These special forms of regression equation should not be used except when plotting the equations on a correlation chart. Whenever the most probable deviation in the one variable corresponding to a known deviation in the other is wanted, regression equations in which *as* are expressed in score units should be employed.

2. The Regression Equation ia Score Form : In the above regression equations the value of x and y substituted are deviations from the means X and Y. But is convenient to be able to estimate an individual's actual score in Y, from the scores in X without first converting the X score into the deviation from Mx.

This can be done by using the score form of the regression equation. This conversion is as follows : Denote mean of Y by M and Y score simply by Y. We may write the deviation of any individual from the mean is Y-M or Y = Y-M . In the same way, X = X-M_x when x is the deviation of any X score from the mean X. If we substitute Y-M_y for y and X-M_x for x, the two regression equations become as

$$\overline{Y} = r\frac{\sigma y}{\sigma x}\left(X - M_X\right) + My$$

$$\overline{X} = r\frac{\sigma y}{\sigma x}\left(Y - M_Y\right) + Mx$$

and

$\overline{Y}$ = estimated score or value in the series when score in X series is given

$\overline{X}$ = actual raw score on series.

σσy - SD of y series

σx = SD of x series

Mx = Mean of X series

My = Mean of Y series

$\overline{X}$ = estimated score of value in the X series when score in y series is given.

$\overline{Y}$ actual raw score on the y series.

y = SD of y series

x = SD of x series

Mx = Mean of X series x

My = Mean of Y series

These two equations are said to be in score form since X and Y represent actual scores and not deviation from the means of the two distributions.

Example 1.

The coefficient of correlation between History and Geography achievement scores is .72 and other statistics are as follows.

History (X)	Geography (Y)
M_x = 50	My = 60
σx = 10	σy = 12
	r=0.72

(a) Student A	X = 60	Y = ?
(b) Student B	X = ?	y = 70

(c) Develop regressionY equations.

(C) The regression equations for Y is

$$Y = .72\frac{12}{10}\left(x-50\right)+60$$

$$Y = .86X + 17$$

The regression equation for X is

$$X = .72\frac{10}{12}\left(Y-60\right)+50$$

$$X = .60Y + 14$$

(A) Y is to predicted when X = 60

Y = .86X + 17

= .86(6) + 17

= 68.60

(B) X is to be predicted when

Y=70

X =.60 Y + 14

= .60(70) + 14

= 56

Example 2.

Calculate score of the student in English when his score in Arithmetic is 15, and score of the student in Arithmetic when his score in English is 20.

Arithmetic X- 13, 16,Yl1, 16, 17, 12, 19, 22, 20, 16

English Y- 14, 19, 10, 16, 15, 18, 26, 29, 23, 10

Mean of Arithmetic scores or M =16

Mean of Epglish scores of My =18

SD of X series = 3.03

SD of Y series = 6.06

r = 0.75

(i) $$\overline{Y} = r\frac{\sigma y}{\sigma x}\left(X - M_X\right) + My$$

$$= 0.75 \times \frac{6.06}{3.03}\left(15 - 16\right) + 18$$

= 16.5

(ii) $$\overline{X} = r\frac{\sigma x}{\sigma y}\left(Y - M_X\right) + Mx$$

$$= 0.75 \times \frac{3.03}{6.06}\left(20 - 18\right) + 16$$

= 16.75

2. Data for two sets are as follows.

X	Y
Mx = 150	My = 140
σx = 12	σy = 16

r = 0.36

(i) find X when Y is 160

(ii) find Y when X is 120

(i) $\overline{X} \quad = r\dfrac{\sigma x}{\sigma y}\left(X - M_y\right) + Mx$

$= 0.36 \times \dfrac{12}{16}\left(160 - 140\right) + 150$

$= 155.4$

(ii) $\overline{Y} \quad = r\dfrac{\sigma y}{\sigma x}\left(X - M_x\right) + MY$

$= 0.36 \times \dfrac{16}{12}\left(120 - 150\right) + 140 = 125.6$

Such prediction and regression equations are useful in educational and vocational guidance or selection of workers in offices and factories. Advice on such basis measure better than subjective judgement.

2.1.10 Suggested Questions

1. Two groups of students selected from the different colleges were administrated an attitude scale, and the following data were collected Do the two groups differ significantly in their attitudes ?

	N	M	SD
College I	40	30.5	6.0
College II	50	25.5	5.0

Answer (t = 4.322. Sig.)

2. A group of 10 students was given trails of physical efficiency. Their scores onI and II trials are given below. Test whether there was significant gain from the first to the second trials.

Students	Trial-I	Trial-II
1	15	20
2	16	22
3	17	22
4	20	25
5	25	35
6	30	30
7	17	21
8	18	23
9	10	17
10	12	20

Answer : (t=6.60 Sig.)

2.1.11 Suggested Readings

Garret., .E. : Statistics in Psychology and Education, Bombay Louis J. Ferrer and Simons, 1973.

Guilford., J.P. : Fundamental Statistics in Psychology and Education, New York, Me Graw Hill, 1995.

LESSON NO. 2.2 AUTHOR : DR. S.K. BAWA

ANALYSIS OF VARIANCE

2.2.1 Objectives

This lesson will help you to :

(i) know about ANOVA
(ii) get information about the assumptions of ANOVA
(iii) know how to interpret it
(iv) learn the steps involved in this technique
(v) know the merits and demerits of ANOVA
(vi) get the opportunity to learn and practise the technique with the help of unsolved problems
(vii) explore more information regarding ANOVA through references.

2.2.2 Introduction

Analysis of Variance, often abbreviated as ANOVA, is a method which enables is to test for the significance of the differences among more than two sample means. It helps to make inferences about whether our samples are drawn from populations having the same mean. This method was devised by R. A. Fisher in 1923. It is also known as F-test; F indicates Fisher. This test examines the significance of difference between OF-ameng variance as well as within variance is used to test the significance of the difference between the means of a number of different populations. To test the effects of 'n' treatments or different methods of teaching or the effect of x variable on 'n' number of variables, F test can be applied. For Example, to know the impact of three psychological variables (a, b and c) on the professional growth of teachers, we must determine whether the three samples represented by the sample means $\overline{x}_1 = \overline{x}_2 = \overline{x}_3$ could have been drawn from populations having the same mean, μ.

The null hypothesis to test would be :

H_0: $\mu_1=\mu_2=\mu_3$

the alternative hypothesis would be :

H_1 : μ_1, μ_2 and μ_3 are not all equal i.e. at least one is different.

2.2.3 Interpretation of the Results :

If we conclude from our test that the sample means do not differ significantly it can be inferred that psychological variables a, b and c do not influence the professional growth of teachers. On the other hand, if we find differences among the sample means that are too large to attribute to chance sampling error, it can be inferred that professional growth of teachers depends upon the psychological variables a, b and c. Thus, the psychological variables should be kept in mind to increase professional growth accordingly.

2.2.4 Assumptions of ANOVA :

The method of analysis of variance has a number of assumptions. The failure of the data to satisfy these assumptions may lead to draw invalid inferences.

1. Each of the sample is to be drawn from a normal population and that each of these population has the same variance, σ^2. If however, the sample sizes are large enough, we do not need the assumption of normality.

2. The variance in the populations from which the samples are selected are equal i.e. $\sigma_1^2 = \sigma_2^2 = \sigma_3^3 = \sigma_4^2 .... = \sigma_k^2$.

   This is known as homogenity of variance.

3. The effects of variance factors on the total variance are additive. The total variance is equal to among variance nnd within variance.

The one advantage of the analysis of variance is that reasonable departure from the assumptions of normality and homogeneity may occur without seriously affecting, the validity of the inferences drawn from the obtained data. The analysis of variance is used in the analysis of data obtained from experiments which involve more than one basis of classification.

2.2.5 Steps in Analysis of Variance :

1. Determine one estimate of the population variance from the variance among the sample means.

2. Determine a second estimate of the population variance from the variance with in the samples.

3. Compare these two estimates. If they are approximately equal in value, accept the null hypothesis.

2.2.6 Example :

If there are six experimental conditions, and we wish to study the effects of these conditions on performance. The question is : Do the mean scores of these six conditions differ significantly ?

	A1	A2	A3	A4	A5	A6	
	4	9	2	8	8	4	
	5	10	4	8	3	3	
	1	9	4	6	3	7	
	2	8	2	6	2	6	
Sums	12	36	12	28	16	20	Grand Sum = 124
Mean's	3	9	3	7	4	5	General Mean = 5.16

Calculation of SUMS of SQUARES

Step-1 : Correction term $(C) = \frac{(124)^2}{24} = 640.66$

Step-2 : Total sum of squares

$= ((4)^2 + (5)^2 + (1)^2 + (2)^2 + (9)^2 + (10)^2 \ldots\ldots\ldots (7)^2 + (6)^2) - C$

$= 808 - 640.66 = 167.34$

Step-3 : Sum of Squares among Means of A_1, A_2, A_3, A_4, A_5 and A_6.

$$= \frac{(12)^2 + (36)^2 + (12)^2 + (28)^2 + (16)^2 + (20)^2}{4} - C$$

$= 756 - 640.66 = 115.34$

Step-4 : Sum of Square within conditions groups A_1, A_2, A_3, A_4 and A_6

= Total SS - Among Means SS

$= 167.34 - 115.34 = 52$

Analysis of Variance Table

Variations	df	Sums of Square	Mean Square	F
Among means of Conditions	5	115.34	23.06	$\frac{23.06}{2.88}$ = 7.98
With in Conditions	18	52	2.88	

Step-1 : Correction term is calculated by adding all the scores of all the experiments, squaring them and then divide it by N. Thus correction term is = $\frac{(\sum X)^2}{N} = 640.66$

Step-2 : To find out total sum of squares, each score is squared and summed up. The value obtained is substracted from correction term, e.g. in the above example 808-640.66 = 167.34.

Step-3 : To find the sum of square among means, first square the sum of each column (each experiment), add these sums and divide the total by N i.e. number of scores in each group, subtracting the correction term from it will give the sum of square between or among means, e.g. in the above example 756-640.66 = 115.34

Step-4 : The sum of squares with in conditions can be calculated by subtracting sum of squares among means from the total sum of squares. SS_t-SS_m = SS_w In the above example, 167.34 - 115.34 = 52.

F ratio is equal to Means of sum of squares between groups divided by the -leans of sum of squares with in groups.

$$F = \frac{M_{sb}}{M_{sw}} = \frac{23.06}{2.77} = 7.98$$

The obtained value of F=7.98 with df: (5, 18) is higher than the table value, thus, the null hypothesis is rejected at both levels of significance. It may be said that the difference among the groups or treatments is highly significant.

The significance of F value indicates that there is high significant difference either in one or two or three pairs of groups. The specification is not possible for drawing the inferences about the effectiveness of treatments. Thus t-test should be followed by ANOVA. F-test is followed when 't' is not significant both are complementary to each other because :

(i) t is followed when F value is significant for the specification of inferences.

(ii) F test is followed when t value is not significant because with in groups variance is not evaluated by t-test. It evaluates only the difference between variances.

(iii) There is fixed relationship between t and F.

F value is the square of 't' and t value is the square root of 'F'.

$$F = t^2 \text{ or } t = \sqrt{F}$$

2.2.7 Advantages of ANOVA :

The analysis of variance has the following advantages in research :

(i) It is an improved technique over t test or z test. It evaluates both types of variance between and with in.

(ii) It is used for ascertaining the difference among several groups or treatments at a time. It is an economical device.

(iii) It involves more than one independent variable in studying the main effects and interaction effects.

(iv) The experimental designs like simple random design and levels x treatment designs are based on one way analysis of variance technique.

(v) If 't' is not significant, F test must be followed to analyse the difference between two means.

2.2.8 Limitations of ANOVA :

(i) This method has certain assumptions regarding normality and hamogeneity of the distribution of data. The departure of the data from these assumptions may effect adversely on the inferences.

(ii) The F value provides global findings of difference among groups but it can not specify the interence. Therefore, t test is followed for specifying the statistical inferences when F value is formpd significant.

(iii) It is time consuming technique and requires knowledge and skills of arithmetical operations.

2.2.9 Examples for Practising ANOVA :

1. A simple random design of experiments is used for testing the difference among four treatments and five subjects are assigned to each treatment. Is the difference among the treatments as given below significant ?

A	B	C	D
14	19	12	17
15	20	16	17
11	19	16	14
10	16	15	12
12	16	12	17

Ans. 6.38

2. Apply complete analysis of variance technique to the following data and specify the inferences.

A	B	C	D	E	F	G	H
24	33	37	38	23	35	38	15
32	21	43	51	25	53	06	26
28	50	57	57	04	38	01	09

37	40	29	42	37	31	10	24
16	57	39	45	25	23	29	30
55	27	47	37	36	36	42	28

Ans. 3.557

3. Three training methods were compared and the following information was obtained.

Method-1	45	40	50	39	53	44
Method-2	59	43	47	51	39	49
Method-3	41	37	43	40	52	37

Is there any significant difference in these methods ?

Ans - 1.664

4. Following data pertains to four samples. Can we conclude that it had come from populations having the same value ?

Sample	1	2	3	4
	16	29	14	21
	21	18	15	28
	24	20	21	20
	28	19	19	22
	29	30	28	18
		21	17	

Ans - 0.384

2.2.10 TWO-WAYS ANALYSIS OF VARIANCE

The one way analysis of variance technique is used to analyse the effect of one independent variable or one type of treatment. It is used to study the main effect of one variable only. The simple random design makes use of one way analysis. It has the focus to analyse the sampling error. Other types of experimental errors are not taken into consideration. But there are many situations in which more than one independent variables are studied simultaneously. In these cases, two ways or three ways classification is used.

In two ways classification two independent variables are taken simultaneously. It has two main effects and one interaction effect of joint effect of two variables on the dependent variable. In two ways classification, three F-values are calculated - Two F-values for two main effects and One F-value for interaction effect.

	Example :	Methods
	A_1	A_2
	12	14
High	13	14
	14	13
intelligence	15	15

		14		15
		15		15
		13		13
		12		13
		14		13
		15		14
			Method	
		A_1		A_2
		2		4
High		3		4
Intelligence	4		3	
		5		5
		4		5
		5		5
		3		3
		2		3
		4		3
		5		4
		$\Sigma A_1B_1=37$		$\Sigma A_1A_2=39$
$\Sigma B_1=37+39=76$				
		4		1
		6		0
		6		2
Low		6		2
Intelligence	5		1	
(B_2)		3		3
		2		3
		5		0
		5		2
		5		2
		$\Sigma B_2A_1=47$		$\Sigma B_2A_2=16$
$\Sigma B_2=47+16=63$				

$\Sigma A_1=84$ $\Sigma A_2=55$ N=139

The obtained scores are modfied by subtracting a constant 10 from each score. The variance will not change by subtracting a constant 10 from each score, but it will help in making the computational process easy. There are four groups in four cells in 2×2 design. The scores of each cell are calculated separately by squaring the scores, sum of each cell and grand total is 139.

Then following steps are followed :

1. Correction Term = (c)

$$C = \frac{\Sigma x^2}{N}$$

$$= \frac{139 \times 139}{40} = 483$$

2. Total Sum of Squares (SS_T)

$SS_T = x_1{}^2 + x_2{}^2 + x_3{}^2....x_{40}{}^2$ - C

= $2^2 + 3^2 +$......... $2^2 + 2^2$ -483

=545-483=107

3. Sum of Squares of Treatments ‘A’ = SS_A

$F_{A_2} =$

$$SS_A = \frac{(\Sigma A_1)^2 + (\Sigma A_2)^2}{2N} - C$$

$$= \frac{(84)^2 + (55)^2}{20} - 483 \quad F_B =$$

$= 504 - 483 = 21$

4. Sum of Squares of Intellig e ‘B’ : SS_B

$F_{AB} =$

$$SS_B = \frac{(\Sigma B_1)^2 + (\Sigma B_2)^2}{2N} - C$$

$$= \frac{(76)^2 + (63)^2}{20} - 483$$

5. Sum of Squares (Cells) : SS_{cell}

$$SS_{Cell} = \frac{(\Sigma A_1 B_1)^{2+} + (\Sigma B_1 A_2)^2 + (\Sigma B_2 A_1)^2 + (\Sigma B_2 A_2)^2}{N}$$

$$= \frac{37^2 + 39^2 + 47^2 + 16^2}{10} - 483$$

= 535 - 50 - 483 = 52.50

6. Sum of Square A × B (SS_{AB})

$SS_{Cell} = SS_A + SS_B + SS_{AB}$

$SS_{AB} = SS_{Cell} – (SS_A + SS_B)$

=52.50– (21+4.50)

=27

7. Sum of Squares within Subjects (SS_W)

$SS_W = SS_T - SS_{Cell}$

= 107 - 52.50

= 54.50

ANOVA (2×2) Table

Sources	df	Sum of Squares	Mean Squares
Methods-A	1	21.00	21.00
Levels-B	1	4.50	4.50
Interaction			
A×B	1	27.00	27.00
Within Subjects	36	54.50	1.40

Main Effect A = $F_A \frac{M_{SA}}{M_{SW}} = \frac{21.00}{1.40} = 14.40$

(Method)

Main Effect B = $F_B \frac{M_{SB}}{M_{SW}} = \frac{4.50}{1.40} = 3.20$

(Intelligence)

Interaction AB = $F_{AB} \frac{M_{SAB}}{M_{SW}} = \frac{27.50}{1.40} = 18.44$

The F_A value 14.40 is higher than the table value even at .01 level of significance. The null hypothesis is rejected. It may be stated that the difference between methods is highly significant.

The F_B value 3.20 is not significant at any level. The null hypothesis is not rejected.

The F_{AB} value 18.44 is highly significant becasue F value is greater than the table value even at .01 level of significance. It may be interpreted that the joint effect of method of teaching and intelligence on criterion variable is significant.

Uses of Two-Way ANOVA

(1) It is used to analysis three effects simultaneously - main effect, simple effect and interaction effect.

(2) The obtained data of factorial design are analysed by this technique. The experimental designs are based on this technique of analysis or treatment of data.

(3) It is used in complex problems and experimental studies.

(4) More than two factors effects are analysed by this technique of ANOVA.

2.2.11 Exercise :

1. Explain the technique ANOVA. Enumerate the basic assumptions of analysis of variance.

2. Describe the steps of ANOVA. Justify that F test and t test are complimentary to each other. Give its advantages and demerits.

3. Apply ANOVA to example-3.

2.2.12 References

* Guilford, J. P. (1978) Fundamentals Statistics in Psychology and Education Tokyo : Mcgraw-Hill KOGA KUSHA Ltd.

* Ferguson, G. A. (1981) Statistical Analysis in Psychology and Education, Auckland : McGraw-Hill International Book Co.

* Garrett, H. E. (1981) Statistics in Psychology and Education, Bombay: Vakils, Feffer and Simons Ltd.

M.A. (EDUCATION) PART-I **PAPER–III**

LESSON NO. 2.3 **AUTHOR : DR. LOKESH KOUL**

Chi-Square

Structure of the Lesson :

2.3.1 Objectives

The students will help you to :

(1) know about Chi-Square

(2) learn the steps involved in this technique

(3) get the opportunity to learn and practice the technique with the help of unsolved problems

2.3.2 Introduction - Chi-Square

TESTING EXPERIMENTAL HYPOTHESIS

Some forms of the experimental hypothesis assert that the results found in an experiment do not differ significantly from result to be expected on a probability basis or stipulated in terms of some theory.

The Chi-square test represents a useful method of comparing experimentally obtained results with those to be expected theoretically on some hypothesis. The equation for chi-square (x^2) is stated as follows :

$$x^2 = \Sigma\left\{\frac{(f_o - f_e)^2}{f_e}\right\}$$

In which

f_o = Frequency of occurrence of observed or experimentally determined facts

f_e = expected frequency of occurrence on some hypothesis.

The difference between observed and expected frequencies are squared and divided by the expected, the smaller is the chi-square, closer is the agreement between observed data and the type of hypothesis being tested. Contrawise, the larger the chi-square, the greater the probability of a real divergence of experχmentally observed from expected results. The value of chi-square is evaluated with the help of given degree of freedom (df).

2.3.3 Use of Chi-Square test

1. The chi-square test is used when the data are in the form of frequencies, or when data can be reduced to frequencies. This includes proportions/percentages and probabilities.

2. Since the essential feature of chi-square is its additive property, a hypothesis involving more than one set of data can be tested for significance.

3. Since chi-square is a distribution free statistics it is used to test hypothesis of equal probability as well as hypothesis of normality. Ah important problem of statistical inference is to test the hypothesis that the given data have been obtained by random sampling from a specified population with definite values for its parameters. The data usually obtained can be arranged in the form of a frequency distribution wherein we give the observed frequencies for the various "classes" or "cells". The corresponding theoretical (expected) frequencies are obtained from the knowledge of the population and our problem is to test the compatibility of observed frequencies with expected frequencies. The theoretical (expected) frequencies are small enough to be regarded as end to fluctuations of random sampling if they indicate that the data could not have possibly come from population giving rise to the theoretical frequencies.

4. It is used to study the relationship between the two variables in which data are obtained in various categories on the basis of some attribute.

Testing the Divergence of observed results from those expected on the Hypothesis of equal Probability (Null Hypothesis)

In some research situation, our hypothesis may assert that the frequencies of events which we have observed really follow the hypothesis of equal probability.

For the sake of illustration, suppose a group of 120 college students were asked to express their judgement. These judgements were classified into categories as under.

Categories

	I	II	III	IV	V	VI	Total
Judgements	30	25	18	10	22	15	120

On the hypothesis of equal probability (null hypothesis), 320 students may be expected to express their judgements in each of the six possible categories. To test this hypothesis we may arrange the frequencies, observed as well as expected, in the two f_0 and f_e.

Judgements	I	II	III	IV	V	VI	Total
f_0	30	25	18	10	22	15	120
f_e	20	20	20	20	20	20	120
f_0-f_e	10	5	-2	-10	2	-5	
$(f_0\text{-}f_e)^2$	100	25	4	100	4	25	
$\frac{(f_0 - fe)^2}{f_e}$	5	1.25	0.2	5	0.2	1.25	

$$(\text{Chi} - \text{square})\ X^2 = \Sigma\left\{\frac{(f_0 - fe)^2}{f_e}\right\}$$

=5+1.25+0.2+5+0.2+1.25
= 12.90

2.3.4 Computation of Chi Square

Step 1. Find the difference between fo and fe for each category and enter these in the row (f_0-f_e).

Step 2. Find the square of (f_0-f_e) and enter these in the row $(f_0\text{-}f_e)^2$

Step 3. Divide each $(f_0\text{-}f_e)^2$ by the corresponding fe and enter these in the row $\frac{(f_0 - fe)^2}{f_e}$

Step 4. Find the sum $\Sigma\left\{\frac{(f_0 - fe)^2}{f_e}\right\}$ to obtain 12.90 as the value of chi-square x^2.

Step 5.

Calculate degree of freedom (df) using the formula :
df=(r-l)(c-l)
In which
r = number of rows in which data are tabulated.
c = number of columns in which data are tabulated.
Here r =2 and c = 6
dg = (2-1) (6-1)
=5

Step 6.

Entering the table of chi-square; we find in row df = 5, x^2 of 11.07 in column headed 0.5. This value is less than the obtained x2 value of 12.90.

Interpretation

Since the obtained value 8x^2, 12.90 is greater than the table value of x^2, 11.07, for 5 df at, .05 level, the hypothesis of equal probability is rejected and we may conclude that judgements expressed by 120 students in different categories differ significantly from the expected one.

Testing the Divergence of observed results from those expected on the Hypothesis of a Normal Distribution.

The Chi-square test can also be used to test the disergence of observed results from the expected ones on the hypothesis that observed results are normally distributed instead of being equally probable.

The following example illustrates how this hypothesis may be tested by chi-square.

Suppose attitude scale was adminisered to a group of 200 students and the items of the scale were to be answered by underlying one of the following five categories: Strongly Agree, Agree, Indifferent, Disagree and Strongly disagree. The distribution of answer to an item is shown in the following table.

Strongly Agree	Agree	Indifferent	Disagree	Strongly Disagree	Total
60	22	45	28	45	200

It was hypothesized that the. distribution of responses differ significantly from that to be expected if the attitude is normally distributed in our population of students.

Computation of x^2

Step 1.

Since the expected frequencies (f_e) are not computed on the basis of equal probability the value of f_e for each category with the help of normal table. In this example, the base line of the normal distribution curve is to be divided into five equal segements. The total length of the base line is 6σ and therefore the length of

each segment becomes $\frac{6\sigma}{5} = 1.2\sigma$

Using normal table, we find that 6.90 or 7 cases out of 200 fall, in the first category, of [“strong agree” and equal number in fifth category “strongly disagree”. 47.68 or 48 out of 200 cases fall in the second category of] “disagree” and consequently the remaining cases i.e. 90.28 or 90 out of 200 fall in the third category “Indifferent”.

After computing the value of expected frequencies for each category, we may arrange the frequencies observed and expected in the two rows, fo and fe

Strongly Agree	Agree	Indifference		Disagree	Strongly Disagree	Total
f_0	60	22	45	28	45	200
f_e	7	48	90	48	7	200
(f_0-f_e)	53	-26	-45	-20	38	
$(f_0-f_e)^2$	2809	676	2025	400	1444	
$\frac{(f_0 - fe)^2}{f_e}$	401.29	14.08	22.5	8.33	206.29	

X^2 = 401.29 + 14.08 + 22.5 + 8.33 + 206.29 =652.49

Step 2. Find the difference between f_0 and f_e for each category and enter these in the row (f_0-f_e)

Step 3. Find the square of $(f_0-f_e)^2$ and enter these in the row $(f_0-f_e)^2$

Step 4. Divide each (f, - f-)2 by the corresponding f,. and enter these in the row

$$\frac{(f_0 - fe)^2}{f_e}$$

Step 5. Find the sum $\Sigma\left\{\frac{(f_0 - fe)^2}{f_e}\right\}$ to obtain 652.49 as the value of chi-square (x2).

Step 6. Calculate degree of freedom (df) using of formula
df=(r- l)(c-l)
=(2-1) (5-1) =4

Step 7. Entering the table of chi-square; we find for df = 4 value of 13.277 in the column headed by . 01. This value is less than the obtained x^2 value of 652.49.

Interpretation

Since the obtained value of x^2, 652.49 is greater than the table value of 13.277 for 4 df at, .01 level, the hypothesis that the response of the students of the item of attitude scale are normally distributed is rejected.

Computation of Chi-Square When Table Entries are Small

The value of Chi-square is less stable and is subject to error when it is either computed from a table with any observed frequency (in the table) as less than 5 or when the table is 2 x 2 i.e. when there is only 1 df. Hence, to overcome the error correction for continuity known as Yates' correction is made.

The computation of x^2, using "Yates' correction is explained with the help of the following example.

Suppose a judge gave 9 right and 3 wrong judgements in 12 trials. We can use chi-square to test the hypothesis of equal probability. To test the hypothesis, we may arrange the observed and expected frequencies in two rows to and fe.

	Judgements		
	Right	Wrong	
f_0	9	3	12
f_e	6	6	12
$(f_0$-$f_e)$	3	-3	
Correction (-0.5)			
	2.5	-2.5	
$(f_0$-$f_e)^2$	6.25	6.25	
$\frac{(f_0 - fe)^2}{f_e}$	1.04	1.04	

x^2 = 1.04 + 1.04
= 2.08

Computation of x^2

Step 1. Find the difference between f_0 and f_e for each category and enter these in the row $(f_0$-$f_e)$.

Step 2. Apply 'Yates' correction, which consists in substracting 0.5 from each $(f_0$-$f_e)$ difference, and enter these in the two : correction (-0.5).

Step 3. Find the square of $(f_0$-$f_e)$ after applying the 'Yates' correction and enter these in the row $(f_0$-$f_e)^2$.

Step 4. Divide $(f_0$-$f_e)^2$ by the corresponding f_e and enter these in the row

$$\frac{(f_0 - fe)^2}{f_e}$$

Step 5. Find the sum $\Sigma\left\{\frac{(f_0 - fe)^2}{f_e}\right\}$ to obtain 2.08 as the value of x^2.

Step 6. Calculate the degree of freedom with the help of the formula
df=(r- l)(e- 1)
=(2- 1) (2- 1)
=1

Step 7. Entering the table of chi-square; we find in row df = 1, x^2 of 3.841 in the column headed .05. This value is greater than the obtained x^2 value of 2.08.

Interpretation

Since the obtained value x^2, 2,08, is less than the table value of x^2, 3.841, for 1 df at .05 level, the hypothesis of equal probability is accepted and we may conclude that the divergence of observed frequencies from the expected ones is not significant.

Note : There is, however, a shorter formula for x^2 for 2x2 table with expected frequencies. The formula is :

$$x^2 = \frac{2(f_0 - fe)^2}{f_e}$$

Substituting the value of $(f_0 - f_e)^2$ of earlier example in the above formula, we get :

$$x^2 = \frac{2(2.5)^2}{6}$$

= 2.08, which is same as we got by using the other method.

Computation of Chi-Square when Table Entries are in Percentages.

In some research situations, we obtain data in percentages. For example, the response to an item is a questionnaire may be obtained in percentages of two categories, “Yes” and “No”. This chi-square test may be used to test hypothesis of equal probability.

To illustrate, consider the following examples in which we obtain data in percentage responses to an item “Should India manufacture Atom Bomb?” From 10 politicians. The responses were obtained in “Yes and “No”

	Yes	No	Total
f_0	30%	70%	100%
f_e	50%	50%	100%

$(f_0\text{-}f_e)$	20%	20%
Correction (-0.5)	15%	15%
$(f_0\text{-}f_e)^2$	225%	225%

$$x^2\% = \frac{2(225)}{50} = 9$$

$$x^2 = \frac{9x10}{100} = 0.90$$

Computation of x^2

Step 1. Find the difference between f_0 and f_e enter these in the row $(f_0\text{-}f_e)$. **Step 2.** Apply Yates correction, which consists in subtracting 5% from each $(f_0\text{-}f_e)$ difference and enter these in the row : correction (-5%).

Step 3. Find the square of $(f_0\text{-}f_e)$ after applying Yates correction and enter these in the row $(f_0\text{-}f_e)^2$.

Step 4. Apply the formula $x^2 = \frac{(f_0 - fe)^2}{f_e}$ to obtain x^2 % = 9

Step 5. In order to bring 2x2 to its proper in term of original number, we must multiply the percent $\frac{x^2 \text{by } N(N = \text{size of the sample})}{100}$ in order to adjust it to the actual frequencies in the given sample. Using this formula, the value of chi-square in the proper value is 0.90.

Step 6. Calculate the degree of freedom df using the formula :

=(r-l)(c-l)

=(2-1) (2-1)

= 1

Step 7. Entering the table of chi square; we find in row df=l.x^2 of 3.841 in the column headed .05. This is greater than the obtained x^2 value of 0.90.

Interpretation

Since the obtained value of x^2, 0.90 is less than the table value of x^2, 3.851 for 1 df at, 0.5 level, the hypothesis of equal probability is accepted and we may conclude that the divergence of observed frequencies from the expected one is not significant.

Chi-Square test of independence in contingency tables.

We have seen that the chi-square test may be employed to test the agreement between observed results and those expected on some hypothesis. A further use of this test can be made when we want to investigate the relationship between traits or attributes which can be classified into one or more categories in a contingency table.

To illustrate the use of chi-square test, consider the following data of 1000 students who have been categorized into five groups. A, B, C, D and E on the basis of age and their preference for colours blue, green, violet, yellow and red.

	Blue	Green	Violet	Yellow	Red	Total
A	(59.6)	(48.6)	(15.6)	(40.4)	(35.8)	200
	75	47	32	30	16	
B	(59.6)	(48.6)	(15.6)	(40.4)	(35.8)	200
	42	41	10	40	67	
C	(59.6)	(48.6)	(15.6)	(40.4)	(35.8)	200
	107	36	12	22	23	
D	(59.6)	(48.6)	(15.6)	(40.4)	(35.8)	200
	44	76	16	44	20	
E	(59.6)	(48.6)	(15.6)	(40.4)	(35.8)	200
	30	43	8	66	53	
Total	298	243	78	202	179	1000

Across the first row we find percents of 200 students falling in group A : 75 have given their preference for green colour. A 47 for Violet, 32 for Yellow and 30 for Yellow and 16 for red. Reading down the column, we find that of 298 students giving their preference for blue colour, 75 belong to group A, 42 to group B, 107 to group C, 44 group D and 30 to group E. The other columns and rows are interpreted in the same way.

The hypothesis to be tested is the Null hypothesis, namely, that colour preference are essentially unrelated or independent In order to compute X- we must calculate an independence value of expected frequency for each cell in the contingency table. Independence values represented by the figures within the different cells: they give the number of students whom we should expected to fall in a particular age group, are owing their preference for a particular colour in the absence of any real association.

The calculation of independence values, (fe) and X^2 are shown as under:

1. Calculation of independence values (f)

$$\frac{298 \times 200}{1000} = 59.6 \qquad \frac{243 \times 200}{1000} = 48.6$$

$$\frac{78 \times 200}{1000} = 15.6 \qquad \frac{203 \times 200}{1000} = 40.4$$

$$\chi^2 \qquad \frac{179 \times 200}{1000} = 35.8$$

2. Computation of x^2 value
Using the formula

$$x^2 = \Sigma \left\{ \frac{\left(f_0 - fe\right)^2}{f_e} \right\}$$

$$= \frac{(75-59.6)^2}{59.6} + \frac{(47-48.6)^2}{48.6} + \frac{(32-15.6)^2}{15.6}$$

$$+ \frac{(30-40.4)^2}{40.4} + \frac{(16-35.8)^2}{35.8} + \frac{(42-59.6)^2}{59.6} + \frac{(41-48.6)^2}{48.6}$$

$$+ \frac{(10-15.6)^2}{15.6} + \frac{(40-40.4)^2}{40.4} + \frac{(67-35.8)^2}{35.8} + \frac{(107-59.6)^2}{59.6}$$

$$+ \frac{(36-48.6)^2}{48.6} + \frac{(12-15.6)^2}{15.6} + \frac{(22-40.4)^2}{40.4} + \frac{(23-35.8)^2}{35.8}$$

$$\chi^2$$

$$+ \frac{(44-59.6)^2}{59.6} + \frac{(76-48.6)^2}{48.6} + \frac{(16-15.6)^2}{15.6} + \frac{(44-40.4)^2}{40.4}$$

$$+ \frac{(20-35.8)^2}{35.8} + \frac{(30-59.6)^2}{59.6} + \frac{(43-48.6)^2}{48.6} + \frac{(8-15.6)^2}{15.6}$$

$$+ \frac{(66-40.4)^2}{40.4} + \frac{(53-35.8)^2}{35.8}$$

3. x^2 = 190.40
3. df = (r-1) (c-1)
= (5-1) (5-1)
= 16

The x^2 values for 16 df as given in the chi-square table, 26.296 and 32.000 respectively for .05 level of significance and the obtained value, 190.40

of x^2 is higher than these values. This indicates that their is a significant relationship between the age apd the colour preference and the hypothesis that age and colour preference are essentially independent is rejected at .01 level of significance.

2x2 Fold Contingency tables

The 2 x 2 table, which is comprised of 4 cells with (2 - 1) (2 - 1) df, there is a simple formula that eliminates the need to compute the expected frequencies (independence values) for each cell.

$$X^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

In which A, B, C and D are the frequencies in the first, second, third and fourth cells respectively and the vertical lines (AD - BC) means that the difference is to be taken as positive.

To illustrate the use of chi-square test let us determine whether item 20 of a test differentiates between two groups of boys and girls. The responses to the item are given in the following 2x2 tables.

Passed item 20		Failed item 20	Total
	(A)	(B)	(A+B)
Boys	30	20	50
	(C)	(D)	(C+D)
Girls	25	15	40
Total	(A+C)	(B+D)	
	55	35	90

Using formula

$$X^2 = \frac{90[(30)(15) - (20)(25)]^2}{(30 + 20)(25 + 15)(30 + 25)(20 + 15)}$$

$$= \frac{90(450 - 500)^2}{(50)(40)(55)(35)} = 0.058$$

Since the obtained value, 0.058 of X^2 does not exceed the critical values 3.81 of X^2 accept the null hypothesis at the .05 level of significance, we may conclude that item 20 of the test does not differentiate between the groups of boys and girls.

When entries in 2 x 2 tables are less than 'Yates' correction should be applied to the formula. The correct formula reads :

$$X^2 = \frac{N\left[(AD - BC) - N/2\right]^2}{(A+B)(C+D)(A+C)(B+D)}$$

2.3.5 Suggested Questions

1. From the following table, determine whether test item 25 differentiates between two groups high and low general ability.

Number of two groups differing in general ability who pass item 25 in a test.

Passed	Failed	Total
High Ability	Low Ability	
.41	29	70
34	36	70
75	65	140

2. The table below shows the number ofextraverts and introverts whose each of the three possible answers to an item on a personality questionnaire. Does these items differentiate between the two groups ? Test the independence hypothesis.

Yes	No	?	Total
Extroverts	Introverts		
24	76	20	120
37	76	17	130
61	152	37	250

3. The table below shows the number of boys and the number of girls who choose each of the following answers to an item opinionnaire. Do these data indicate a significant sex difference in opinion towards this item ? Testthe Null hypothesis.

	Strongly Approve	Approve	Indifference	Disapprove	Strongly Disapprove
Boys	35	40	20	35	20
Girls	20	25	15	25	25

2.3.6 Suggested Readings

Ferguson. G.A. (1981) : Statistics Analysis is Psychology and Education, Aukland; McGraw Hill International.

Garret., H.E. (1962) : Statistics in Psychology and Education, Bombay; Allied Pacific Pvt. Ltd.

Gilford., J.P. (1965) : Fundamental Statistics in Psychology and Education, New York, Me Graw Hill Book Co.

Koul, Lokesh. (1984) : Methodology of Educational Research, New Delhi; Vani Educational Books : A divison of Vikas Publishing House Pvt. Ltd.

Mc.Nair, Qwin (1982) : ‘Psychological Statistics’ New York : John Wiley and Sons.

M.A. (EDUCATION) PART-I **PAPER–III**

EDUCATIONAL STATISTICS

LESSON NO. 2.4 **AUTHOR : DR. LOKESH KOUL**

LINEAR CORRELATION

Structure of the Lesson

2.4.1 Objectives :

After reading this lesson you will be able to :

- Define Correlation and Coefficient of Correlation.
- Define Linear Correlation.
- Calculation Product Moment Correlation from grouped and ungrouped data.
- Define Rank Correlation.
- Interpret the results obtained.

2.4.2 Introduction :

In previous lessons, various types of analysis we have discussed i.e. how measures of central tendency and measures of dispersion are calculated in such cases for the purposes of comparison and analysis. With the help of these measures such data can easily understood. The data in which we secure measures of one variable for each individual is called a univariate distribution. If we have pairs of measures on two variables of each individual the joint presentation of the two sets of scores is called a bivariate distribution.

If we come accross a number of situations involving the study of two or more variables, then these variable are called multivariate distribution. In this lesson we will deal with bivariate distributions.

2.4.3 Linear Correlation :

The data in which we secure measure of two variables for each individual are called a bivariate data. For example, we get bivariate data if we have measures of both. l.Q. and academic achievement for a group of school children. The essential feature of the bivariate data is that, one measure can be measured for each member of the group. When we study bivariate data we may like to know the degree of relationship between variables of such data. This degree of relationship is known as correlation. It can be represented quantitatively by the co-efficient of the correlation. We observe that students who have high l.Q. tend to receive high score on an achievement test in Mathematics, whereas those with low l.Q. tend to score low. When this type of relationship is observed, the variable l.Q. and achievement in Mathematics are said to be positively and linearly correlated. In general when individuals above average in one variable tend to be well above average in the other and those below average in the one tend to be correspondingly below average in the other, the two variables show a positive correlation. This type of relationship can be described as perfect positive correlation, high positive correlation or low positive correlation.

Perfect positive linear correlation exists when there is one-to-one correspondence between the measures of scores of two variables. For example, if 20 children have exactly the same rank or performance on two tests and the child who ranks first in one test also attains the first rank on the other test; the child who ranks second in the ranking of scores on one variable and second in

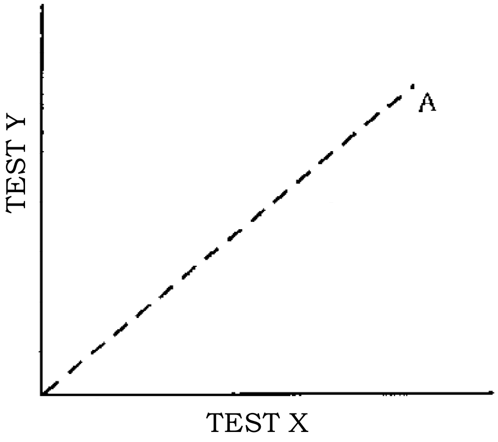


Fig. 2.4.10.1

the ranking of scores on the other variable if such as association continues among all the scores, the correlation is said to be positive and perfect. The value of co-efficient of correlation in such a case is +1. The following diagram (16.1) illustrates the scatter of scores on two tests in such situations.

High positive correlation exists when there is only slight deviation in the scores in one variable in comparison to scores in the other variable. In this situation also, the individual who is high on a test X tends to be high in test Y and the individual who is low in X tends to be low in Y. However, there will be some variation in the scores of test X or test Y from the mean of the value. The scatter of the scores of individual will take the following form (Fig. 2.4.10.2)

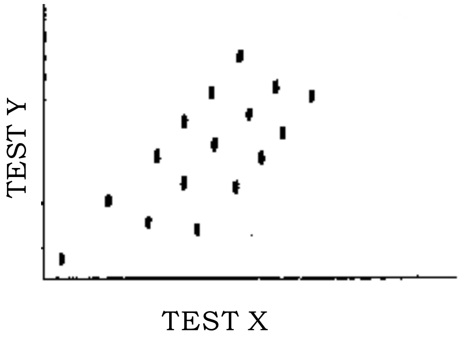


Fig. 2.4.10.2

In case of low positive correlation, the spread or scatter of scores becomes greater, An individual who is high on the test X is likely to be almost anywhere within the range in terms of the score on the higher direction on the test Y. The correspondence between the variation in the scores on the test X and Y becomes inconsistent and may take the following form (Fig. 2.4.10.3)

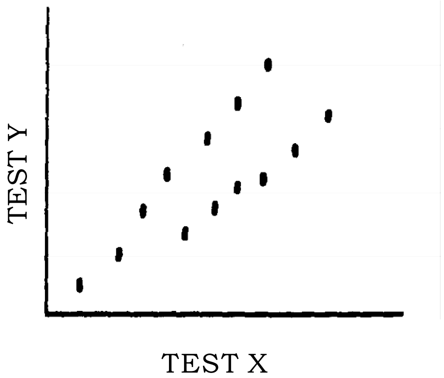
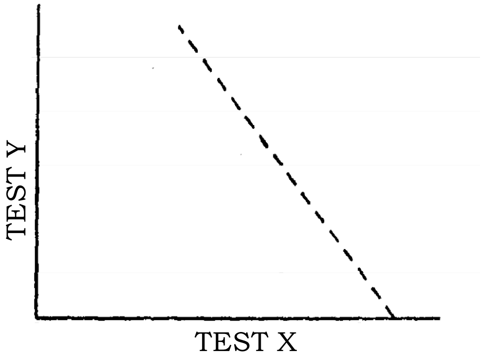


Fig. 2.4.10.3

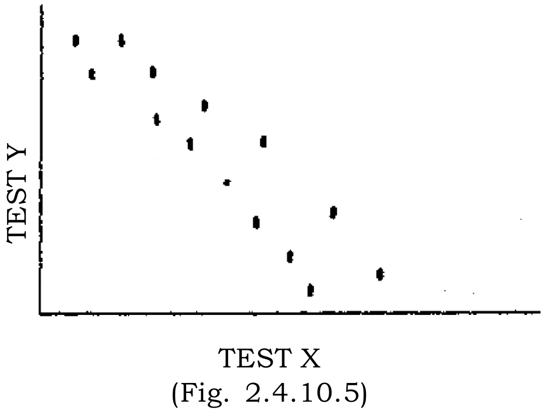
Sometimes students making high scores in one variable are likely to make low scores in another variable and vice-versa. Suppose that in a class of 15 students, the student who stood first in a test of Physics was ranked lowest on a test of Civics, and the students who stood second in Physics ranked next to the bottom (Fourteenth) in a test of civics, and that each student stood just as far from the top in Physics test as from the bottom in Civics test. Here the correspondence between ranks in Physics test and Civics test is regular and definite enough, but the direction of relationship is inverse (negative). This type of relationship can also be described perfect negative correlation, high negative correlation or low negative correlation.

Perfect negative linear correlation exists when high degree of one attribute or characteristics may be associated with low degree of another attribute or characteristic. In other words , and individual making high score in a trait X is likely to make low score on trait Y and similarly an individual making low score in trait X is likely to make a high score in trait Y. If such an association exists the correlation is said to be negative and perfect. The value of co-efficient of correlation in such situations is - 1.00 and the scatter of scores takes the following form (Fig. 2.4.10.4).



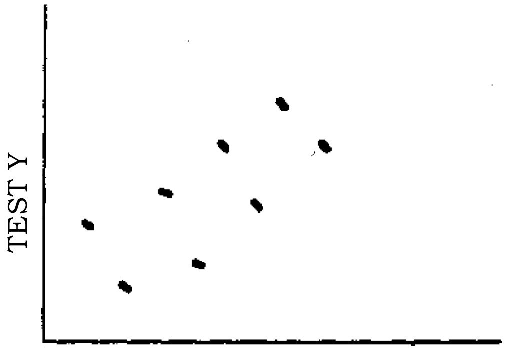
(Fig. 2.4.10.4)

When an indiviual who is high on a test X tends to be low on test Y and an individual who is low on test X tends to be high on test Y, and there is some variation in the scores of test X or of test Y from the mean values, high negative correlation is said to exist between the scores on test X and Y. The scatter of the scores on the two tests may take the following form (Fig. 2.4.10.5).



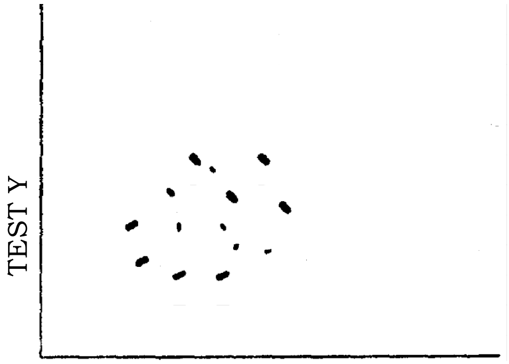
(Fig. 2.4.10.5)

The spread of the scores becomes greater when a low negative linear correlation exists between the two variables. An individual who is high on the test X is likely to be almost anywhere in terms of the score in the lower direction on the text Y. The correspondence between the variation of scores on test X and Y becomes inconsistent and the spread of the scores takes the following form (Fig. 2.4.10.6)



(Fig. 2.4.10.6)

When the relationship between two sets of variables is purely a chance relationship, we say there is no correlation. For example, in a class of 20 students, a student with a high score in Hindi test is likely to be anywhere within the total range in terms of his score in a test of personality. The four students scoring first four positions in the Hindi test may score seventh, tenth, eighth, and second positions on the personality test. The four lowest students in Hindi test may score fourth, seventh, sixth and tenth positions in the test of personality. In this type of situation the spread of scores is random. The co-efficient of correlation is 0.00 and the spread of scores in the graphic distribution takes the following form (Fig. 2.4.10.7).



(Fig. 2.4.10.7)

From the earlier discussion, it is evident that the intensity or degree of linear correlation is represented quantitatively by co-efficient. Its value ranges from-1.00 to +1.00 through 0.00. A. value of -1.00 describes perfect negative correlation and +1.00 describes perfect positive correlation. A Zero value describes complete lack of correlation between the two variables. The sign of co-efficient indicates the direction of relationship and the numerical value in magnitude.

2.4.4 Product Moment Correlation :

There are various methods of finding the degree of correlation. Their use is relative to the situation and type of data. There are many situations in which we have data expressed in terms of measurements on interval (numerical) scale. In such situations, product moment method is used to compute correlation between two variables.

This method was developed by Karl Pearson. It is denoted by the symbol r and is used for computing correlation when :

(i) The data for the two variables X and Y are expressed in interval or ratio level of measurement.

(ii) The distribution of the variable X and Y have a linear relationship.

(iii) The distribution of the variables X and Y are uni-model and their variances are approximately equal.

(iv) The sample size is fairly large.

2.4.4.1 Uses of Pearson's Correlation :

1. It is extensively used in the field of measurement.
   (a) In finding reliability
   (b) In estimating Validity
   (c) The cut of score is determined expirically with the help of scatter plot.
   (d) Item discrimination power is calculated by using r.
2. Higher technique of correlations are the extension of Pearson's r
   (a) multiple Correlations are based on Pearson r.
   (b) partial correlations employs r.
   (c) Factor analysis is the extension of Pearson r.
3. For prediction, r is used
4. Theories of intelligence and personality and developed using r.
5. Correlational studies in behavioural sciences employ r.

2.4.4.2 Product-Moment Correlation as a Ratio :

The product-moment co-efficient of correlation may be thought of essentially as that ratio which express the extent to which changes in one variable are accompanied by or are independent upon changes in a second variable. As an illustration consider the following example in which data are availabe in terms of scores on a Mathematics test and a test on Physics for a group of students.

1	2	3	4	5	6	7	8
Students	Scores in Mathematics Test X	Scores in Physics Test Y	(X-Mx) x	(Y-My) y	x^2 –	y^2 –	xy –
A	22	42	-4	-2	16	4	8
B	29	46	3	2	9	4	6
C	24	45	-2	1	4	1	-2
D	25	40	-1	-4	1	16	4
E	30	47	4	3	16	9	12
			$\Sigma x=0$	$\Sigma y=0\Sigma x^2$	$=46$	$\Sigma y^2=34$	$\Sigma xy=28$

Mx=26

My=44

$$\text{Correlation (r)} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 - \Sigma y^2}} = \frac{28}{\sqrt{46 \times 34}} = \frac{28}{39.55}$$

r=0.71

Computational Steps

Step 1. Find the means for the variables X scores and Y. These are Mx = 26, My = 44.

Step 2. Find the deviations of X scores and Y scores from their respective means and enter them in column (4) and (5) respectively. These are x and y columns.

Step 3. Find the sum of squares of deviations of x and y in column 6 and 7 which are denoted as Σx_2, and Σy_2. Also the sum of the product of x and y i.e. Σxy.

Step 4. Divide Σxy by square root of the product of Σx^2 and Σy^2 i.e., 46×34 and we are able to find the coefficient of correlation.

2.4.4.3 Computation of Product Moment Correlation from ungrouped Data Using Deviation from Assumed Means :

When the N (size of sample of group) is small or the raw scores are in small number, we make use of the following formula.

$$y = \frac{N\Sigma xy - \Sigma x\Sigma y}{\sqrt{\left[N\Sigma x^2 - (\Sigma x)^2\right]\left[N\Sigma y^2 - (\Sigma y)^2\right]}}$$

in which

x = Deviation of X measure from the assumed means.

Y= Deviation of Y measures from the assumed means.

To illustrate the use of this formula, let us compute product-moment r from the following data in the two variables X and Y for 11 Students.

X	Y	x	y	x^2	y^2	xy
45	56	-20	-9	400	81	180
55	50	-10	-15	100	225	150
56	48	-9 = -51	-17 = -50	81	289	153
58	60	-7	-5	49	25	35
60	62	-5	-3	25	9	15
65 (AM)	64	0	-1	00	1	0
68	65 (AM)	3	0	9	0	0
70	70	5	5	25	25	25
75	74	10 = +53	9 = +56	100	81	90
80	82	15	17	225	289	255
85	90	20	25	400	625	500
		Σx=2	Σy=6	Σx^2=1414	Σy^2=1650	xy=1403

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\left[N\Sigma x^2 - (\Sigma x)^2\right]\left[N\Sigma y^2 - (\Sigma y)^2\right]}}$$

$$= \frac{11 \times 1403 - 2 \times 6}{\sqrt{\left[11 \times 1414 - (2)^2\right]\left[11 \times 1650 - (6)^2\right]}} = 0.92$$

Computational Steps

Step 1. Choose any raw score in X measures as the assumed mean. Find the deviations of the raw scores from the assumed mean and enter them in column x and y respectively. Total column to obtain Σx and Σy.

Step 2. Find the square of all x's and y's enter these squares in columns x^2 and y^2 respectively, Total the columns to obtain Σx^2 and Σy^2.

Step 3. Multiply the x's and y's in the same rows and enter these product (with due regard for sign) in the xy column. Total the xy column, taking account of sign, to get Σxy.

Step 4. Substitute 1403 for Σxy : 2 for Σx, 6 for Σy, 1650 for Σy^2, 1414 for Σx^2 and 11 for N, in the formula and solve for r to obtain its value as 0.92.

2.4.4.4 Computation of Product Moment Correlation from ungrouped Data Using Raw Scores :

When a calculating machine is available, the value of product moment r can be obtained from original measures (raw scores) with the help of the following formula :

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\left[N\Sigma x^2 - (\Sigma x)^2\right]\left[N\Sigma y^2 - (\Sigma y)^2\right]}}$$

To illustrate the use of this formula, let us consider the following data of 12 students on two tests.

Students	Test 1	Test 2			
	x	y	x^2	y^2	xy
1	54	25	2916	625	1350
2	60	26	3600	676	1560
3	61	32	3721	1024	1952
4	71	34	5041	1156	2414
5	74	40	5476	1600	2960
6	59	28	3481	784	1652
7	67	28	4489	784	1876
8	65	30	4225	900	1950
9	71	36	5041	1296	2556
10	62	30	3844	900	1850
11	56	34	3136	1156	2904
12	50	22	2500	484	1100
	Σx=750	Σy=365	Σx^2=47470	Σy^2	Σxy=23134

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\left[N\Sigma x^2 - (\Sigma x)^2\right]\left[N\Sigma y^2 - (\Sigma y)^2\right]}}$$

$$r = \frac{12 \times 23134 - (750) + (365)}{\sqrt{\left[12 \times 47470 - (750)^2\right]\left[12 \times 11385 - (365)^3\right]}}$$

= 0.78

Computational Steps

Step 1. Find the totals of the columns X and Y separately to obtain ΣX and ΣY.

Step 2. Find the squares of all X's and Y's enter these squares in columns x^2 and y^2 respectively. Total these columns to obtain ΣX^2 and ΣY^2.

Step 3. Multiply the X's and Y's in the same rows, and enter products in the XY columns to get Σxy.

Step 4. Substitue for ΣXY, 23134; ΣX, 750; ΣY, 365; ΣX^2, 11385 and N, 12 in the formula and solve for r to obtain its value = 0.78.

10.4.5Computation of Product Moment Correlation from Grouped Data :

When N is large or even moderate in size, and when no calculating machine is available, the best procedure is to group data in the both variables X and Y and form a scattergram.

The scattergram represents in bivariate distribution by taking one variable X along the horizontal line and other variable Y along vertical line, suppose that we have records for 100 students, giving the height (X) and weight (Y) of students : then to each student corresponds a pair a value x,y. Each of the 100 students can be represented on the diagram with respect to height and weight. Suppose that student weighs 120 pounds and his Height is 60 inches. His weight locates him in the fifth column from top. Accordingly, a "tally" is placed in second cell of the fifth column.

Similarly each of 100 students is represented by a tally in cell of the table is accordance with the two characteristics of height and weight. Along the bottom of the diagram in fx row, is tabulated the number of students who fall in each height interval. The fy column and fx rows must total 100, (each) the total number of students. After listing all the frequencies in each cell is added and entered on the diagram. The scattergram is also called correlation table.

Let us consider the following scattergram showing paired scores of 12 students on the test of mathematics and physics.

Scattergram showing paired scores of 12 students on the test of Maths and Physics.

Calculation of Coefficient of Correlation

X / Y		Marks in Mathematics (X)								
		6-10	11-15	16-20	21-25	26-30	*fy*	*f*x	fy^2	*f*xy
	x	-2	-1	0	1	2				
Marks of Physics 6-10	$-2^{(x)}$	$1^{(4)}$	$1^{(2)}$	$0^{(0)}$	$0^{(0)}$	$0^{(0)}$	2	-4	8	6
11-15	-1	$0^{(0)}$	$1^{(1)}$	$1^{(0)}$	$0^{(0)}$	$0^{(0)}$	2	-2	2	1
16-20	0	$1^{(0)}$	$1^{(0)}$	$2^{(0)}$	$1^{(0)}$	$0^{(0)}$	5	0	0	0
21-25	1	$0^{(0)}$	$0^{(0)}$	$1^{(0)}$	$1^{(1)}$	$1^{(1)}$	3	3	3	2
f_y		2	3	4	2	1	N=12	Σfy= -3	Σfy^2=13	Σf×4=9
f_x		-4	-3	0	2	2	Σfx= -3			
fx^2		8	3	0	2	4	Σfx^2=17			
*f*xy		4	3	0	1	1	Σfxy=9			

The following formula is used for computing the product moment correlation between score onMathematics test (X) and scores on Physics test (Y).

$$y = \frac{N\Sigma fxy - (\Sigma\, fx)(\Sigma\, fy)}{\sqrt{\left[N\Sigma\, fx^2 - (\Sigma\, x)^2\right]\left[N\Sigma\, fy^2 - (\Sigma\, fy)^2\right]}}$$

The computation for the values of fx, fy, fx^2 and fxy are done by following steps :

Step 1. The distribu cores for the12 students is found in the fy olu attergram. Assume a mean for the distribution of scores of Physics. In the present example the mean for the Physics scores has been taken at 18 (mid-point of interval 16-20) and y's (deviation from the assumed mean) have taken from the point. Now fill in the fy and fy^2 columns.

Step 2. The distribution of the mathematics test scores for 12 students in the fx rows at the bottom of the scattergram. Assume a mean for this distribution. The mean for the Mathematics scores is taken at 18 (mid-point of interval 16-20), and x's (deviations from the assumed mean) are taken from this point. Fill in the fx and fx^2 rows.

Step 3. The fxy for a cell is computed by multiplying the frequency given in the particular cell with the correspondence x and y. For example, there is frequency I and the cell correspondence to Mathematics score interval 06-10 and Physics score interval, 6-10. The corresponding x for this cell frequency is-2 and the corresponding y is -2. Thus fxy for this cell is 1×2×2=4. Similarly the value for fxy is computed for all the cells and their sum fxy is calculated row-wise as well as columnwise. The two sums should be equal to each other. In the present example it has come equal to 09.

Step 4. Subsituting the values for Σfx, Σfx^2 Σfy^2 and xy in the formula to obtain the value of r :

$$y = \frac{N\Sigma fxy - (\Sigma\, fx)(\Sigma\, fy)}{\sqrt{\left[N\Sigma\, fx^2 - (\Sigma\, fx)^2\right]\left[N\Sigma\, fy^2 - (\Sigma\, fy)^2\right]}}$$

$$= \frac{12 \times 9 - (-3)(-3)}{\sqrt{(12 \times 17) - (-3)^2 (12 \times 13) - (-3)^2}}$$

$$= \frac{99}{\sqrt{195 \times 147}} = .58$$

10.5 Spearmen's Rho-Correlation :

This is the simplest method of finding coefficient of correlation. It is also called Rank Order coefficient of correlation and denoted by ρ (rho). This method is used when the data are available in ordinal form of measurement or if number of paired variables is more than 30 fewer than 9. To compute Spearman Rank order coefficient of correlation, the following formula is used.

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

Where D = The difference between paired ranks.

D^2 = the sumo f the squared differences between ranks.

N = number of paired ranks.

Illustration :

Two judges ranked 10 students reading ability test. The ranksgiven by judge X and judge Y have given below :

Students	Judge X	Judge Y	D	D^2
1	1	2	-1	1
2	5	3	2	4
3	4	5	-1	1
4	6	4	2	4
5	3	6	-3	9
6	2	1	1	1
7	7	7	0	0
8N N	9	8	1	1
9	10	9	1	1
10	8	10	-2	4
				$\Sigma D^2 = 26$

$$\rho = 1 - \frac{6x26}{10(100-1)}$$

Find Spearman’s Rank correlation between the scores of 10 students in first and second test as given below :

S.No.	X	Y	Rank on X R_1	Rank on Y R_2	Rank difference $D = R_1–R_2$	D^2
1	65	69	7	2	5.0	25
2	63	66	9	6.5	2.5	6.25
3	67	68	4.5	3.5	1.0	1.00
4	64	65	8	8.5	0.5	0.25
5	68	65	2.5	8.5	6.0	36.00
6	62	66	10	6.5	3.5	12.25
7	70	68	1	3.5	2.5	6.25
8	66	64	6	10	4.0	16.00
9	68	71	2.5	1	+1.5	2.25
10	67	67	4.5	5	.5	.25
n = 10						$\Sigma D^2 = 105.5$

In ranking the scores on Ist test (X), we give rank 1 to highest score of 70. Then, there are two next highest score of 68 ranked with mean rank (2+3)/2 = 2.5 each.

Again there are two next lower tied scores of 67 ranked with mean rank (4+5) / 2 =4.5 each. The next highest scores of 66 is then ranked as 6 (because we have already used ranks from one to five). Other lower united scores are ranked as usual. The same procedures is applied in ranking the scores on 2nd test (y). In the process, we observe that there are 2 tied ranks for score 68 and 2 for 67 on the first test. Similarly, these are 2 tied ranks for score 68 and 2 for score 66 on 2nd test andso on.

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

$$= \frac{6\times 105.5}{10\times 99} = .64$$

This is showing a positive correlation between two sets of scores.

Spearman's Rho provides a quick and convenient way of estimating the correlation when, N is small or when we have only ranks.

2.4.6 Summary :

At the end we can say that the tendency of simultaneous variation between two variables is called correlation or covariation correlation means the relationship between two variables. Co-efficient of correlation is numerical index that tells us to what extent the two variables are related and to what extent the variations in one variable changes with the variations in the other. The coefficient correlation is always symbolized either by r or ρ. The motion 'r' is known as product moment correlation co-efficient. The symbol 'P' is known as Rank Difference Correlation. The value of 'r' varies from +1 to -1. In a bivariate distribution, the correlation may be (1) Positive, Negative or Zero correlation and linear or Currilinear. Spearman's Rho provides estimating correlation when, N is small or when we have only ranks.

2.4.7 Suggested Questions :

1. Find the correlation coefficient between the following two sets of scores, using the Ratio Method.

Subjects	X	Y
1	16	41
2	18	42
3	20	50
4	17	46
5	19	49
6	26	44
7	14	40
8	20	40

2. Compute the Coefficient of correlation in the following table :

Achievement Test Score

	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	Total
130-139				1		1		1	3
120-129					1	3	1		6
110-119	1	2	5	6	14	6	3	2	3
100-109	3	7	9	17	13	6	2	2	59
90-99	4	10	10	12	5	26			
80-89	9	9	8	2	8		26		
Total	12	28	39	38	32	15	6	5	174

3. Compute the correlation between the following two sets of scores, where deviations are taken from the means of the two series :

Test 1	Test 2
150	60
126	40
135	45
176	51
138	56
142	43

142	57
163	38
137	41
178	55

4. Seven instructors are rated by urban and rural students on “clarity of presentation”. The results are tabulated in this manner.

Instructor	Urban	Rural
1	39	58
2	39	42
3	36	18
4	35	22
5	33	21
6	29	38
7	22	38

What is the Spearman’s rho (ρ) for these data? Interpret the result.

2.4.8 Suggested Readings :

1. Garrett, H.E. (1985) Statistics is Psychology and Education, Bombay : Vakils, Feffer and simons.
2. Guilford, J.P. (1965) fundamental Statistics in Psychology and Education. New York, McGraw-Hill Book Company.