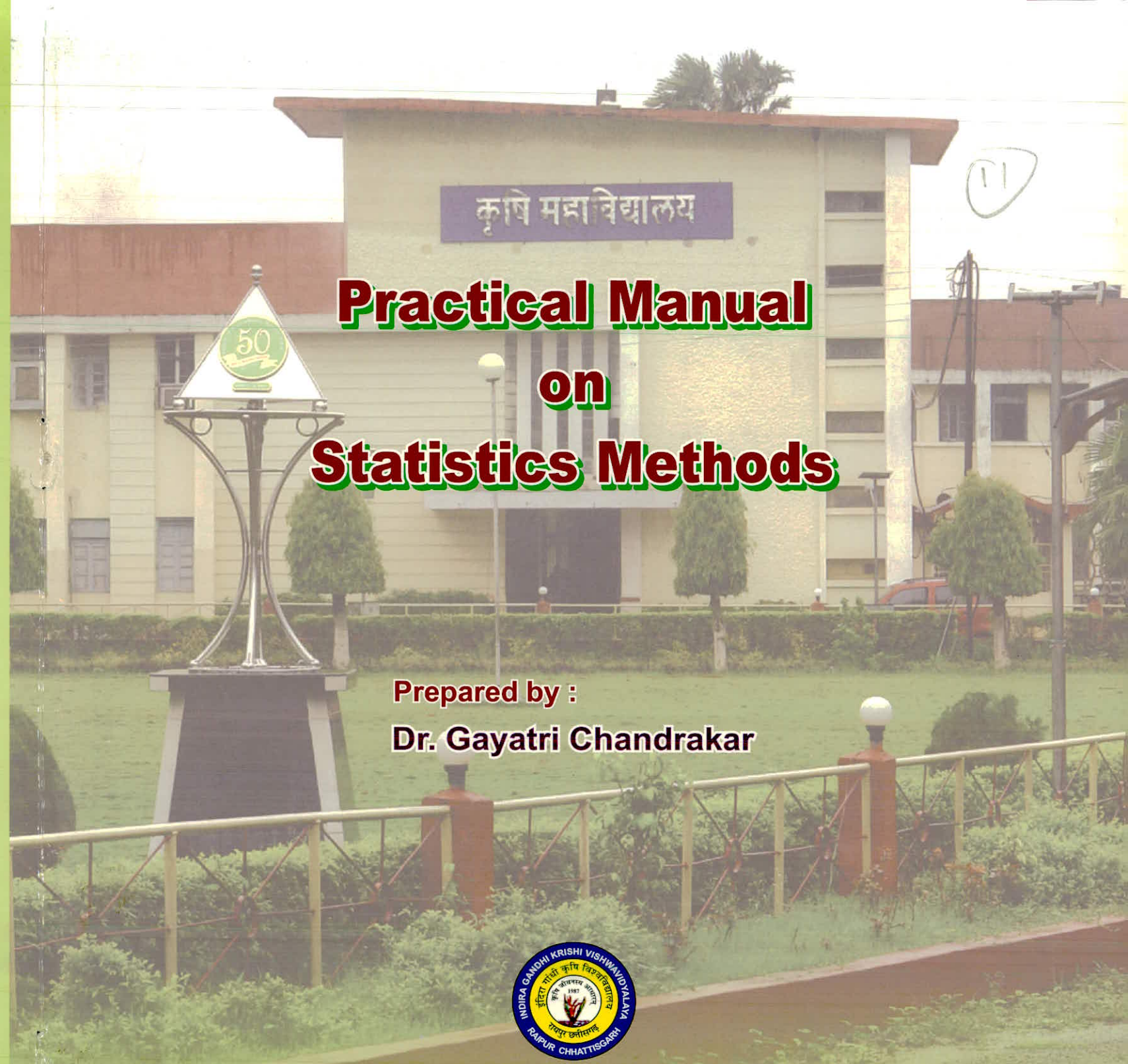
Chhattisgarh Samvad



Practical Manual on Statistics Methods

Prepared by :

Dr. Gayatri Chandrakar

Department of Agricultural Statistics and Social Science
College of Agriculture
INDIRA GANDHI KRISHI VISHWAVIDYALAYA
Raipur (Chhattisgarh) 492 012

Practical Manual
on
Statistics Methods

Prepared by:
Dr. Gayatri Chandrakar

Department of Agricultural Statistics and Social Science
College of Agriculture
INDIRA GANDHI KRISHI VISHWAVIDYALAYA
Raipur (Chhattisgarh) 492 012

Citation: Chandrakar, Gayatri 2013. Practical Manual on Statistics Methods
Pages : 55

Inspired by

Dr. S.K. Patil
Hon'ble Vice Chancellor
IGKV, Raipur (C.G.)

Guidance by

Dr. O.P. Kashyap
Dean, Faculty of Agriculture
College of Agriculture, Raipur (C.G.)

Prepared by

Dr. Gayatri Chandrakar

Course No. AST-211

Course title: Statistics Methods

Published by :
College of Agriculture, IGKV, Raipur (C.G.)

Publication year: 2013

No. of copies printed: 2500

Course No. : ..... **Credit .....**

Course Name : .....

Name of Students .....

Roll No. ..... **Batch .....**

Session ..... **Semester .....**

College of Agriculture .....

CERTIFICATE

This is to certify that Shri./Ku. ..... Enrolment No./ID No..... has completed the practical of Course No. ..... as per the syllabus of B.Sc. (Ag.) ..... year..... semester in the respective lab/field of College.

Date: **Course Teacher**

CONTENTS

Practical No. 1

Title : Construction of frequency distribution Tables and Frequency curves.

Objective : To prepare frequency distribution table and draw Histogram, frequency polygon & frequency curve.

Frequency distribution Table: frequency may be defined as he number of individuals having the same measurement or enumeration count or lies in the same measurement group. Frequency distribution is the distribution of frequencies over different measurements. A table showing the distribution of the frequencies in the different classes is called a frequency distribution table

Histogram : it consists of rectangles erected with bases equal to class intervals of frequency distribution and heights of rectangles are proportional to the frequencies of the respective classes in such a way that the areas of rectangles are directly proportional to the corresponding frequencies.

Frequency Polygon: If the points are plotted with mid values of the class intervals on the X-axis and the corresponding frequencies on the Y-axis, the figure obtained by joining these points with the help of a scale is known as frequency polygon.

Frequency Curve: If the points are plotted with mid values of the class intervals on X-axis and the corresponding frequencies on Y-axis, the figure formed by joining these points with a free hand is known as frequency curve.

Cumulative Frequency Curve (Ogive) : If the points are plotted with upper limits of classes on X-axis and the corresponding cumulative frequencies (less than) on Y- axis, the figure formed by joining these points with a free hand is known as cumulative frequency curve (less than). If the lower limits of classes are taken on X- axis and the corresponding cumulative frequencies on Y-axis, the curve so obtained is called cumulative frequency curve (greater than).

Example 1: Prepare frequency distribution table with suitable class interval from the data given below are the heights (in inches) of 75 plants in a field of a paddy crop.

17,	8,	23,	24,	26,	13,	31,	16,	14,	35,	6,	11,
12,	11,	15,	21,	10,	4,	3,	19,	35,	36,	19,	40,
28,	17,	12,	2,	27,	31,	11,	21,	16,	34,	39,	1,
7,	12,	13,	10,	6,	21,	24,	22,	26,	28,	17,	6,
5,	15,	11,	16,	28,	4,	3,	19,	27,	35,	37,	14,
2,	9,	8,	16,	13,	22,	8,	26,	13	12,	16,	14,
27,	31,	6,									

Solution : The difference between highest and lowest heights is 40-1=39 . Supposing that 10 groups are to be formed, the class interval for each class would be 39/10=3.9. The groups (or classes) will be formed with a class interval of 4 starting from 1 continuing upto 40. The number of plants will be accounted in each class with the help of vertical line called 'tally mark' .

Class	Tally marks	Frequency
1-4	𝍸 |	7
5-8	𝍸 ||||	9
9-12	𝍸 𝍸 |	11
13-16	𝍸 𝍸 ||||	14
17-20	𝍸 |	6
21-24	𝍸 |||	8
25-28	𝍸 ||||	9
29-32	|||	3
33-36	𝍸	5
37-40	|||	3

Example 2: Represent the following frequency distribution of farms according to area in a particular village by a histogram, frequency polygon, frequency.

Area (hectares)	0-2	2-4	4-6	6-8	8-10	10-12
No. of farms	40	48	25	18	12	7

From Fig. 1.1 one can infer that the maximum number of farms are laying in the group (2-4) and the minimum number in the group (10-12). The total area under the histogram is equal to the total frequency.

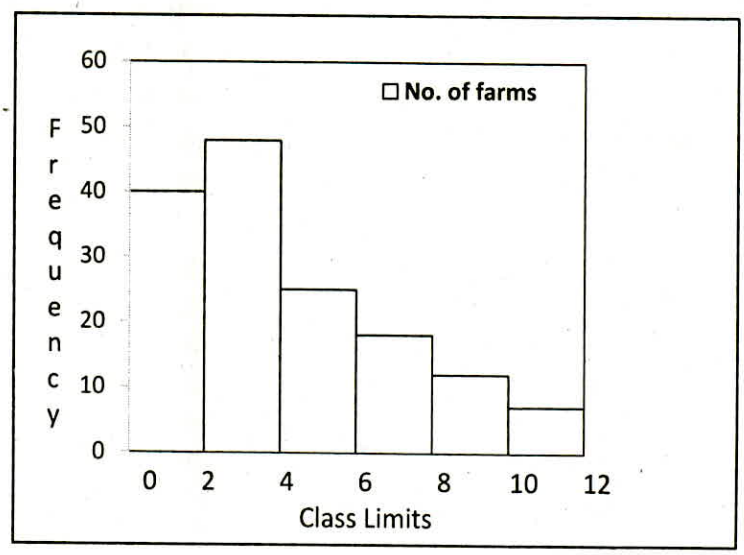
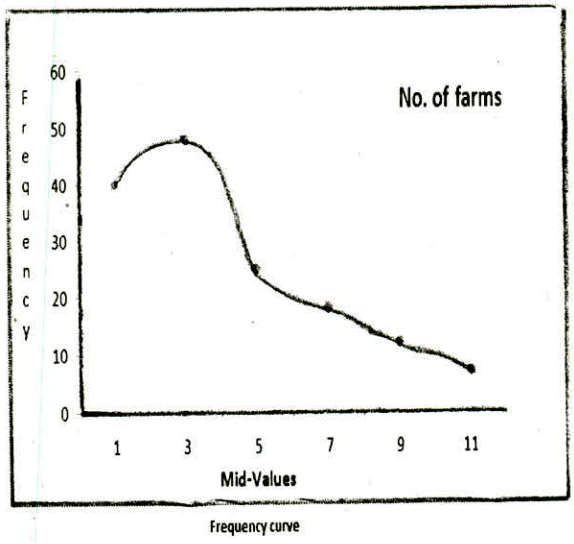


Fig. 1.1. Histogram



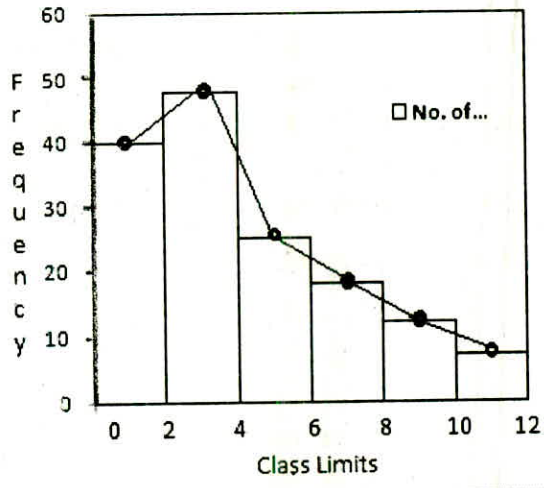


Fig. 1.2. Frequency Curve

Fig. 1.3. Frequency polygon

Exercise 1 : In a uniformity trial the weight of 50 onion bulbs were recorded in grams as follows:

200,	150,	75,	120,	120,	135,	180,	240,	80,	90,	130,
180,	95,	130,	150,	150,	165,	220,	85,	110,	170,	160,
210,	180,	170,	160,	160,	150,	130,	120,	150,	160,	180,
200,	220,	180,	190,	190,	130,	210,	180,	175,	165,	170,
190,	210,	170,	160	160,	140,	150,	210,	220,	190	180,

from a suitable frequency distribution in 7 classes.

Exercise 2: The following is the distribution of heights of plants of a particular crop.

Height (inches)	35-39	40-44	45-49	50-54	55-59	60-64
No. of plants	12	20	15	29	9	3

Draw the (i) Histogram (ii) Frequency polygon (iii) Frequency curve.

Practical No. 2

Title : **Computation of Arithmetic mean for grouped and un-grouped data**

Objective : **Find out Arithmetic mean for grouped and un-grouped data**

"The arithmetic mean is the amount secured by dividing the sum of values of the items in a series by their number."

If $x_1, x_2, x_3 \ldots\ldots, x_n$ be the values of x, then the arithmetic mean $\bar{x}$ is given by.

$$\bar{x} = \frac{x_1 x_2 x_3 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

If $x_1, x_2, \ldots\ldots x_n$ be the value of series and corresponding frequencies be $f_1, f_2, \ldots\ldots f_n$, then the arithmetic mean is given by

$$\bar{x} = \frac{f_1 x_2 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \cdots\ldots + f_n} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i}$$

In case of grouped data, we compute arithmetic mean $\bar{x} = A + \frac{\Sigma fd}{N} \times i$, where A is assumed mean $d = \frac{X - A}{i}$, i= class internal

Properties of Arithmetic mean

1 : The algebraic sum of the values from arithmetic mean is always zero.

2. The sum of the squares of the deviations of all the values from their mean is minimum.

3. If x and y are two independent variables and there is a relation y = a + bx, then $\bar{y} = a + b\,\bar{x}$, where $\bar{x}$ & $\bar{y}$ be the means of x and y.

4. A.M. is dependent with the change of origin and scale.

Example 1 : The rainfall of a certain town in centimeters for the first six months of the year are 102 , 103 , 100, 98 , 95 , 105. Find the average rainfall of that town for these six months.

Solution: average $\bar{x} = \frac{x_1 x_2 x_3 + \ldots + x_n}{n} = \frac{102 + 103 + 100 + 98 + 95 + 105}{6}$ =100.5 cms.

Example 2 : Calculate the arithmetic mean for the following frequency table :

X :	20-30	30-40	40-50	50-60	60-70
f :	8	26	30	20	16

Solution:

X	f	Min-value (x)	f.x
20-30	8	25	8*25=200
30-40	26	35	26*35=910
40-50	30	45	30*45=1350
50-60	20	55	20*55=1100
60-70	16	65	16*65=1040
	Σf =100		Σfx=4600

The required arithmetic mean = $\frac{\Sigma fx}{\Sigma f} = \frac{4600}{100} = 46$

Example 3. Calculate the arithmetic mean from the following Data :-

Age (years) :	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
No. of employees	8	10	28	30	22	10	4	3

Solution:

Age (years)	No. of employees f	Mid-value x	Step deviation d = (x-A)/5	fd
20-25	8	22.5	-3	-24
25-30	10	27.5	-2	-20
30-35	28	32.5	-1	-28
35-40	30	37.5	0	0
40-45	22	42.5	1	22
45-50	10	47.5	2	20
50-55	4	52.5	3	12
55-60	3	57.5	4	12
	∑ f=120			∑ fd = - 6

The required arithmetic mean = $A + \frac{\Sigma f d}{\Sigma f} \times i = 37.5 + \frac{(-6)}{(120)} \times 5 = 37.5 - 0.25 = 37.25$ years .

Exercise 1 : The weekly income of ten families (in rupees)in a certain locality are given below :

Family :	A	B	C	D	E	F	G	H	I	J
Income :	30	70	10	75	500	8	42	250	40	36

Calculate the arithmetic average.

Exercise 2 : From the following data of marks obtained by 60 students of a class. Calculate the arithmetic mean:

Marks :	20	30	40	50	60	70
No. of students:	8	12	20	10	6	4

Exercise 3 : Calculate arithmetic mean from the following data:

Marks less than :	80	70	60	50	40	30	20	10
No. of students :	100	90	80	60	32	20	13	5

Exercise 4: Given the following frequency distribution ,calculate the arithmetic mean:

Weekly Wages (in Rs.)	No .of Workers	Weekly Wages (in Rs.)	No. of Workers
12.5 - 17.5	2	37.5 – 42.5	4
17.5 – 22.5	22	42.5 – 47.5	6
22.5 – 27.5	10	47.5 – 52.5	1
27.5 – 32.5	14	52.5 -57.5	1
32.5 – 37.5	3		

Practical No. 3

Title : Computation of median for grouped and ungrouped data
Objective : To find out median for grouped & ungrouped data.

Median: If the variables of a series are arranged in ascending or descending order, then the value of the middle variable is defined as the median. If the number of variables is odd say and 2m+1, then the value of the (m+1) th variables is the median. If the number of variables is even say 2m, then the mean of the values of the mth and (m+1)th items is the median. According to Connor, "the median is that value of the variable which divides the group into two equal parts, one part comprising all values greater and the other all values less than the median.

Median for ungrouped data: If n be odd, then $\frac{1}{2}$ (n + 1)th item of series is median. and if n be even then there are two middle terms . The average of their values is the median.

In case of **discrete frequency distribution** median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

(i) Find N/2, where N=Σf
(ii) See the cumulative frequency (C.f.) just greater than N/2.
(iii) The corresponding value of x is median.

In the case of **continuous frequency distribution**, median is obtained by following formula:

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

Where l is the lower limit of the median class
f is the frequency of the median class
h is the magnitude of the median class,
C is the c.f. of the pre median class, and N=Σf.

Uses of the Median : Median is the only average to be used while dealing with qualitative data which cannot be measured quantitatively but still can be arranged in ascending or decending order of magnitude, for e,g, to find the average intelligence or average honesty among a group of people.

Example 1 : The consumption of electricity in units of first eleven months of a persons are:-

9,10,15,7,11,9,8,11,7,9,10 . Find the median.

Solution: Arranging the terms in ascending order, we get

7, 8, 9, 9, 9, 10, 10, 11, 11, 15

The number of items in this series is 11 which is odd. Hence the required median = value of $\frac{1}{2}$ (11 + 1)th term = value of 6^{th} term=9.

Example 2. The number of books issued on the consecutive days from a library are 75, 80, 96, 92, 89 ,94,100, 82 ,63, 105. Calculate the median.

Solution: Arranging the terms in ascending order, we get

63,75,80,82,89,92,94,96,100,105

Here the number of terms is 10 i.e. even There will be two middle terms viz. $\frac{1}{2}$ (10)th and $(\frac{10}{2} + 1)$th i.e. 5^{th} and 6^{th} .

The required median= $\frac{1}{2}$ (89 + 92)= 90.5

Example 3 : Obtain the median for the following frequency distribution:

x:	1	2	3	4	5	6	7	8	9
f:	8	10	11	16	20	25	15	9	6

Solution: For median first prepare cumulative frequency as

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6
c.f.	8	18	29	45	65	90	105	114	120

Here N= 120. Cumulative frequency (c.f.) just greater than N/2=60, is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

Example 4 : Find the median wages of the following distribution:

Wages (in Rs.) :	20-30	30-40	40-50	50-60	60-70
No. of Laboures:	3	5	20	10	5

Solution: Prepare cumulative frequency as

Wages (Rs.)	No. of laboures	C.F.
20-30	3	3
30-40	5	8
40-50	20	28
50-60	10	38
60-70	5	43

Here N/2 = 43/2 = 21.5, just greater than 21.5 in c.f. is 28 and the corresponding class is 40-50. Thus median class is 40-50.

Median= $1+\frac{h}{f}\left(\frac{N}{2}-C\right)=$ $40+\frac{10}{20}(21.5-8)=40+6.75=46.75$

Thus median wages is Rs. 46.75.

Exercise 1: (a) find the median height of hocky eleven, the heights in inches of whose player are:-

65, 67, 69, 61, 60, 65, 66,70, 71, 62, 72

(b) The daily wages is Rupees of ten labourers of a factoy are 4, 6, 9,12,11, 8,5,10, 11, 8. Find the median.

Exercise 2: Locate the median in the following distribution

Size :	8	10	12	14	16	18	20
Frequency:	3	7	12	28	10	9	6

Exercise 3 : Calculate median from the following data:-

Classes:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency:	5	8	7	12	28	20	10	10

Practical No. 4

Title : Computation of mode for grouped and ungrouped data
Objective : To compute mode for grouped & ungrouped data.

Mode : It is defined as the value of the variate having the maximum frequency. According to Kenney and Keeping, " The value of the variable which occure most frequently in a distribution is called the mode."

In case of discrete frequency distribution, mode is the value of x corresponding to maximum frequency. For example given below :

X	1	2	3	4	5	6	7	8
f	4	9	16	25	22	15	7	3

the value of x corresponding to the maximum frequency, viz., 25 is 4. Hence mode is 4 .

In case of for continuous frequency distribution, mode is given by the formula:

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

Where l is the lower limit, h the magnitude, f_1 is the frequency of the modal class. f_0 and f_2 are the frequencies of the preceding and succeeding the model class respectively.

Sometimes mode is estimated from the mean and the median. For a symmetrical distribution, mean, median and mode are coincide. If the distribution is moderately asymmetrical, the mean, median and mode obey the following empirical relationship (due to Karl Pearson):

$$\text{mean-median} = \frac{1}{3}(\text{Mean} - \text{mode})$$

Mode=3 Median-2Mean

But in any one (or more) of the following cases:

(i) If the maximum frequency is repeated.

(ii) If the maximum frequency occurs in the very beginning or at the end of the distribution, and.

(iii) If there are irregularities in the distribution.

The value of mode is determined by the method of grouping, which is illustrated below by an example.

Find the mode of the following frequency distribution:

Size (x)	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f)	3	8	15	23	35	40	32	28	20	45	14	6

Solution: Here we see that the distribution is not regular since the frequencies are increasing steadily up to 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. Here we cannot say that since maximum frequency is 45, mode is 10. Here we shall locate mode by the method of grouping as explained below:

Size (x)	frequency						Analysis Table		
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	Col. No.	max. frequency	value of x giving max. frequency
1	3	11		26			(i)	45	10
2	8		23		46		(ii)	75	5,6
3	15	38		98			(iii)	72	6,7
4	23		58			73	(iv)	98	4,5,6
5	35	75			107		(v)	107	5,6,7
6	40		72			100	(vi)	100	6,7,8
7	32	60		80					
8	28		48		93				
9	20	65							
10	45		59	65		79			
11	14	20							
12	6								

On examining the values in the last column we find that the value 6 is repeated maximum number of times and hence the value of mode is 6 and not 10 which is an irregular item.

Example1: Find the mode of the following numbers:

4, 5, 8, 6, 9, 8, 6, 5, 11, 9, 8

Solution : By inspection of the given numbers, we find that the frequency of the number 8 is greatest in the given data. Hence the required mode is 8.

Example 2: Calculate the mode from the following data:-

Monthly Rent (in Rupees)	No. of families	Monthly Rent (in Rupees)	No. of families
20-40	6	120-140	15
40-60	9	140-160	10
60-80	11	160-180	8
80-100	14	180-200	7
100-120	20		

Maximum frequency is 20 i.e. model class is 100-120.

Mode= $l + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2}$ = 100+$\frac{20-14}{2(20)-14-15}$ x 20 = 100 + $\frac{120}{11}$ = 100+ 10.9 = 110.9

Exercise 1: Locate mode in the following data :

8, 5, 7, 12, 9, 6, 4, 0, 8, 9, 7, 9

Exercise 2: Compute the mode for the following frequency distribution.

Class	:	2-4	4-6	6-8	8-10	10-12	12-14	14-16
Frequency	:	9	27	45	54	42	30	9

Practical No.5

Title : **Computation of Standard Deviation, Variance and Coefficient of Variation for grouped and ungrouped data**

Objective : **Calculate standard deviation, mean deviation and coefficient of variance for a given frequency distribution.**

The degree to which numerical data tend to spread about an average value is called the variation or dispersion of data. There are different measures of dispersion as:

(i) Range (ii) Quartile deviation (iii) Mean deviation (iv) Standard deviation

Range: The range is the difference between two extreme observations of the distribution. If A and B are the greatest and the smallest observations respectively in a distribution, then its range is $A - B$. Range is the simplest but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to chance fluctuations, it is not at all a reliable measure of dispersion.

Quartile deviation : Quartile deviation is the half of the difference between first and third quartiles and is given by $Q = \frac{1}{2}(Q_3 - Q_1)$,

Where Q_1 *and* Q_3 *are the* first and third quartiles of distribution respectively. Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data it cannot be regarded as a reliable measure.

Mean Deviation : If x_i / f_i , $i = 1, 2, \ldots\ldots, n$ is the frequency distribution ,then mean deviation from the average A(Usually mean , median or mode) is given by

Mean deviation $= \frac{1}{N} \sum f_i \mid x_i - A \mid$, $\sum f_i = N$, Where $\mid X_i - A \mid$ represents the absolute value of the deviation $(x_i - A)$, when the-ve sign is ignored.

Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviation $(x_i - A)$ creates artificiality and renders it useless for further mathematical treatment.

Standard deviation. : Standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean . For the frequency distribution x_i / f_i , $i = 1,2, \ldots\ldots, n$,

$\sigma = \sqrt{\frac{1}{N} \sum f_i (x_i - \bar{x})}\, 2$, where $\bar{x}$ is the arithmetic mean $\sum f_i = N$.

The Square of standard deviation is called the **variance** and is given by

$\sigma^2 = \frac{1}{N} \sum f_i (x_i - \bar{x})\, 2$

Properties : (i) Standard deviation is independent with the change of origin but not of seal .

(ii) If n_1, n_2 *are the sizes* ; $\bar{x}_1$, $\bar{x}_2$ the means and σ_1, σ_2 the standard deviations of series , then the standard deviation σ of the combined series is given by

$$\sigma^2 = \frac{1}{n_1+n_2}\left[n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)\right] \quad , \quad \text{Where} \quad d_1 = \overline{x_1} - \bar{x},\ d_2 = \overline{x_2} - \bar{x}$$

and $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1+n_2}$, is the mean of the combined series .

Co-efficient of Dispersion : If we want to compare the variability of the two series which differ widely in their average or which are measured in different units , we do not merely calculate the measures of dispersion but we calculate the co-efficient of dispersion which are pure numbers independent of the units of measurement. The co-efficient of dispersion (C.D.) based on different measures of dispersion are as follows:

1. C.D. based upon range $= \frac{A-B}{A+B}$, Where A and B are the greatest and smallest items in the series.
2. Based upon quartile deviation: $C.D. = \frac{(Q_3-Q_1)/2}{(Q_3+Q_1)/2} = \frac{Q_3-Q_1}{Q_3+Q_1} =$
3. Based upon mean deviation: $\text{C. D.} = \frac{\text{Mean deviation}}{\text{Average from which it is calculated}}$
4. Based upon standard deviation: $C.D. = \frac{S.D.}{Mean} = \frac{\sigma}{\bar{x}}$

Coefficient of Variation- 100 times the co-efficient of dispersion based upon standard deviation is called co-efficient of variation (C.V.), i.e. C.V.= $100 \times \frac{\sigma}{\bar{x}}$

According to Professor Karl Pearson who suggested this measure, C.V. is the percentage variation for each series. The series having greater C.V. is said to be more variable than the other and the series having lesser C.V. is said to be more consistent than the other.

Characteristic for an Ideal Measure of Dispersion: The criteria for an ideal measure of dispersion are the same as those for an ideal measure of central tendency, viz.:

(i) It should be rigidly defined.

(ii) It should be easy to calculate and easy to understand.

(iii) It should be based on all the observations.

(iv) It should suitable for further mathematical treatment.

(v) It should be affected as little as possible by fluctuation of sampling.

Example 1: Calculate the Standard deviation from the following data:

46, 51, 48, 62, 54, 51, 58, 60, 71, 75, 47, 73, 62, 65, 53, 57, 65, 72, 49, 51

Solution:

Masses (x)	x-$\bar{x}$	$(x-\bar{x})^2$
46	-12.5	156.25
51	-7.5	56.25
48	-10.5	110.25
62	3.5	12.25
54	-4.5	20.25
51	-7.5	56.25
58	-0.5	0.25
60	1.5	2.25
71	12.5	156.25
75	16.5	272.25
47	-11.5	132.25
73	14.5	210.25
62	3.5	42.25
65	6.5	12.25
53	-5.5	30.25
57	-1.5	2.25
65	6.5	42.25
72	13.5	182.25
49	-9.5	90.25
51	-7.5	56.25
$\Sigma x = 1170$		$\Sigma (x-\bar{x})^2 = 1643$

arithmetic mean $\bar{x} = \frac{\Sigma x}{N} = \frac{1170}{20} = 58.5$ and $\sigma = \sqrt{\left[\frac{\Sigma (x-\bar{x})^2}{N}\right]} = \sqrt{\left(\frac{1643}{20}\right)} = 9.064$

Example 2: Calculate the Standard deviation and Coefficient of variance of the following data:

Class :	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
Frequency :	6	5	15	10	5	4	3	2

Solution:

Class	Mid-value(x)	Frequency (f)	x-$\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
5-10	7.5	6	-13.7	187.69	1126.14
10-15	12.5	5	-8.7	75.69	378.45
15-20	17.5	15	-3.7	13.69	205.35
20-25	22.5	10	1.3	1.69	16.90
25-30	27.5	5	6.3	39.69	198.45
30-35	32.5	4	11.3	4127.69	510.76
35-40	37.5	3	16.3	265.69	797.07
40-45	42.5	2	21.3	453.69	907.38
		$\Sigma f = 50$			$\Sigma f\,(x-\bar{x})^2 = 4140.50$

Here $\bar{x} = \frac{\Sigma fx}{N} = \frac{(7.5)6 + (12.5)5 + (17.5)15 + (22.5)10 + (27.5)5 + (32.5)4 + (37.5)3 + (42.5)2}{50} = \frac{1060}{50} = 21.2$

and $\sigma = \sqrt{\left[\frac{\Sigma f_i\,(x-\bar{x})^2}{N}\right]} = \sqrt{\left[\frac{4140.50}{50}\right]} = \sqrt{(82.81)} = 9.1$

Coefficient of variance $= \frac{\sigma}{\bar{x}} \times 100 = 42.9$

Exercise 1 : Calculate the Standard deviation. of the following data:-

x :	1	2	3	4	5	6
f :	2	6	12	7	2	1

Exercise 2 : Calculate the Standard deviation and coefficient of variance from the following table.

Class :	0-10	10-20	20-30	30-40	40-50	50-60
Frequency :	15	17	19	27	19	12

Exercise 3 : Calculate the Standard deviation. for the following distribution of marks in a class of 50 students :

Marks Scored :	0-10	10-20	20-30	30-40	40-50
No. of Students :	2	4	6	9	15

Practical No. 6

Title : Large samples : SND Test for means, single sample.

Objective : To test the difference between sample mean and population mean.

Standard Normal Deviate tests: If X follows normal distribution with mean, μ and S.D., σ then $\bar{X}$ also follows normal distribution with mean μ and S.D. ., $\frac{\sigma}{\sqrt{n}}$. This can be denoted by

$\bar{X} \sim \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ i.e., $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

The expression on the left hand side is called as standard normal variate (or standard normal deviate) which follows normal distribution with mean zero and standard deviation unity. The test of hypothesis based on this deviate is called standard normal deviate test. The confidence limits for the population mean can be obtained as $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ i.e. the population mean, lies between $\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}$ and $\bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$, where 1.96 is the tabulated value of normal distribution at 5 per cent level of significance.

Under the null hypothesis: H_0 that the sample has been drawn from a population with mean μ and variance σ^2 .i.e., there is no. difference between the sample mean ($\bar{x}$) and population mean (μ),

the test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

If |Z| >1.96, H_0 is rejected at 5% level of significance otherwise accepted. If |z| >2.58, H_0 is rejected at 1 % level of significance.

Null Hypothesis: Null hypothesis is the statements about parameters which is likely to be rejected after testing. We start with the hypothesis that the two items are equal.

One Sample Test:

Assumptions : 1. Population is normal.

2. The sample is drawn at random.

Null hypothesis : $\mu = \mu_0$,

the statistic $Z = \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}$ follows normal distribution with zero mean and unit S.D. If Z (calculated)< Z (tabulated), 1.96 at 5 per cent level of significance, the null hypothesis is

accepted i.e., there is no significant difference between sample mean and population mean. Otherwise, the null hypothesis is rejected means there is significant difference between sample and population means.

Example 1: The average number of mango fruits per tree in a particular region was known from a considerable experience as 520 with a standard deviation 4.0. A sample of 20 trees given an average number of fruits 450 per tree. Test whether the average number of fruits per tree selected in the sample is in agreement with the average production in that region?

Null hypothesis = $\mu_0 = 520$,

The statistic $Z = \frac{|450-520|}{4/\sqrt{20}} = 78.26$

The Z (calculated)> Z (tabulated), 1.96 at 5 per cent level of significance. Therefore, we conclude that there is significant difference between sample mean and population.

Exercise 1: A sample of 900 members has a mean 3.5 cms. and s.d. 2.61 cms. Is the sample from a large population of mean 3.25 cms. and s.d. 2.61 cms.?

Exercise 2 : An insurance agent has claimed that the average age of policy holders who insure through him is less than the average for all agents which is 30.5 years.

A random sample of 100 policy holders who had insured through him gave the following age distribution:

Age last birthday:	16-20	21-25	26-30	31-35	36-40
No. of persons :	12	22	20	30	16

Calculate the arithmetic mean and standard deviation of this distribution and use these values to test his claim at the 5% level of significance . You are given that Z (1.645) =0.95.

Practical No. 7

Title : Large samples : SND test for means, two samples
Objective : To test the difference between two sample means.

Let $\bar{x}_1$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2 and $\bar{x}_2$ be the mean of an independent random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then,

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{and } \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Also $\bar{x}_1 - \bar{x}_2$, being the difference of two independent normal variates is also a normal variate. The Z (S.N.V.) corresponding to $\bar{x}_1 - \bar{x}_2$ is given by

Under the null hypothesis $H_0 : \mu_1 = \mu_2$ i.e. there is no difference between the sample means, we get

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\sigma_1^2/n_1) + \sigma_2^2/n_2)}} = N(0,1)$$

If Z (calculated) < Z (tabulated), 1.96 at 5 per cent level of significance, the null hypothesis is accepted i.e., there is no significant difference between the sample means. Otherwise, the null hypothesis is rejected i.e. there is significant difference between the sample means.

Example 1 : The mean of two single large samples of 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.5 inches?

Solution: We are given:

$n_1 = 1000, n_2 = 2000$; $\bar{x}_1 = 67.5$ inches, $\bar{x}_2 = 68.0$ inches.

Null hypothesis. H_0: $\mu_1 = \mu_2$ and σ = 2.5 inches, i.e., the samples have been drawn from the same population of standard deviation 2.5 inches.

Alternative Hypothesis, H_0: $\mu_1 \neq \mu_2$,

The Statistic $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$

$$Z = \frac{67.5 - 68.0}{2.5 \times \sqrt{\frac{1}{1000} + \frac{1}{2000}}} = \frac{-0.5}{2.5 \times 0.0387} = 5.1$$

Since |Z| > 3, the value is highly significant and we reject the null hypothesis and conclude that samples are not from the same population with standard deviation 2.5.

Exercise 1 : The following table presents data on the values of a harvested crop stored in the open and inside a godown:

	Sample size	Mean	$\Sigma (x - \bar{x})^2$
Out side	40	117	8.685
Inside	100	132	27.315

Assuming that the two samples are random and they have been drawn from normal populations with equal variances, examine if the mean value of the harvested crop is affected by weather conditions.

Exercise 2 : A random sample of 90 poultry farms of one variety gave an average production of 240 eggs per bird/year with a S.D. of 18 eggs. Another random sample of 60 poultry farms of another variety gave an average production of 195 eggs per bird/year with a S.D. 15 eggs. Distinguish between two varieties of birds with respect to their egg production.

Practical No. 8

Title : Students t-test for Two samples (paired and independent)
Objective : To test the difference between sample means

Students 't': Let x_i (i = 1, 2,....n) be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defiend as the statistic

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}}, \qquad \text{where} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is an unbiased estimate of the population variance σ^2, and it follows student's t-distribution with n-1 d.f.

Application of t-distribution.

(i) To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean.

(ii) To test the significance of the difference between two samples means.

(iii) To test the significance of an observed sample correlation co-efficient and sample regression coefficient.

(iv) To test the significance of observed partial and multiple correlation coefficients.

t-test for Single Mean : Suppose we want to test if a random sample x_i (i = 1, 2,....n) of size n has been drawn from a normal population with a specified mean, say μ_0, *or* if the sample mean differs significantly from the hypothetical value μ_0, of the population mean.

Under the null hypothesis H_0: (i) The sample has been drawn from the population with mean μ or (ii) there is no significant difference between the sample mean $\bar{x}$ and the population mean μ.

The statistic $t = \frac{\bar{x}-\mu_0}{S/\sqrt{n}}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Follows t distribution with n-1 d.f.

Compare the calculated value of t with the tabulated value at certain level of significance. If calculated |t| > tabulated t, null hypothesis is rejected otherwise H_0 may be accepted at the level of significance.

t- test for Difference of Means

Suppose we want to test if : (a) two independent samples $x_i (i = 1, 2, \ldots, n_1)$ and $y_i (i = 1,2, \ldots \ldots, n_2)$, have been drawn from the populations with same means or (b) the two sample means $\bar{x}$ and $\bar{y}$ differ significantly or not.

Under the null hypothesis H_0 that (a) samples have been drawn from the populations with the same means, i.e.., $\mu_x = \mu_y$ or (b) the sample means $\bar{x}$ and $\bar{y}$ do not differ significantly, *the statstic*:

$t = -\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, follows student's t- distribution with $(n_1 + n_2 - 2)\ d.f.$

where $S^2 = \frac{1}{n_1+n_2-2} \quad [\Sigma (x - \bar{x})^2 + \Sigma (y - \bar{y})^2]$

Paired t- test for Difference of Means: Let us now consider the case when (i) the sample sizes are equal *i. e.*, $n_1 = n_2 = n$ (say), and (*ii*) the two samples are not independent but the sample observations are paired together , i.e., the pair of observations (x_i, y_i), (i=1, 2,….,n) corresponds to the same (ith) smaple unit. The Problem is to test if the sample means differ significantly or not.

Here we consider the increments, $d_i = x_i - y_i$ (*i*=1, 2,….,n).

Under the null hypothesis H_0 that increments are due to fluctuations of sampling ,

the statistic $t = \frac{\bar{d}}{S/\sqrt{n}}$, where $\bar{d} = \frac{1}{n}\Sigma_{i=1}^{n} d_i$ and $S^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(d_i - \bar{d})^2$

follows student's t- distribution with (n-1) d. f.

Example 1 : A random sample of 10 boys had the following I.Q.'S : 70, 120, 110, 101, 88, 83, 95, 98,107, 100. Do these data support the assumption of a population mean I. Q. of 100 ?

Solution : Null hypothesis , H_0: the data are consistent with the assumption of a mean I.Q. of 100 in the population , *i. e.* $\mu = 100$.

The Statistic $t = \frac{\bar{x}-\mu}{\sqrt{s^2/n}} \sim t\ (n\text{-}1)$

Where $\bar{x}$ and s^2 are to be computed from the sample values of I.Q.S. as

	X	$(X-\bar{x})$	$(X-\bar{x})^2$
	70	-27.2	739.84
	120	22.8	519.84
	110	12.8	163.84
	101	3.8	14.44
	88	-9.2	84.64
	83	-14.2	201.64
	95	-2.2	4.84
	98	0.8	0.64
	107	9.8	96.04
	100	2.8	7.84
Total	972	0	1833.60

Here $n = 10, \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{972}{10} = 97.2$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ $s^2 = \frac{1833.60}{9} = 203.73$

$|t| = \frac{|97.2-100|}{\sqrt{203.73/10}} = \frac{2.8}{\sqrt{20.37}} = \frac{2.8}{4.514} = 0.62$

Tabulated $t_{0.05}$ for 9 d. f. is 2.262.

Since calculated t is less than tabulated $t_{0.05}$ for 9 d. f., H_o may be accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population .

Example 2 : The height of six randomly chosen sailors are in inches: 63, 65, 68, 69, 71 and 72. Those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss,, the light that these data throw on the suggestion that sailors are the average taller than soldiers.

Solution: The heights of sailors and soldiers be represented by the variables X and Y respectively then under the null hypothesis $\mu_X = \mu_Y$ i.e., the sailors are not on the average taller than the soldiers,

the test statistic t = $\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{D_1}+\frac{1}{n_2}}}$, where $S^2 = \frac{1}{n_1+n_2-2} [\Sigma (x-\bar{x})^2 + \Sigma (y-\bar{y})^2]$

X	d=X-68	d^2	Y	D=Y-66	D^2
63	-5	25	61	-5	25
65	-3	9	62	-4	16
68	0	0	65	-1	1
69	1	1	66	0	0
71	3	9	69	3	9
72	4	16	69	3	9
Total	0	60	70	4	16
			71	5	25
			72	6	36
			73	7	49
			Total	18	186

$\bar{x} = A + \Sigma d/n_1 = 68+0 = 68$ $\quad \bar{y} = B + \Sigma D/n_2 = 66 + \frac{18}{10} = 67.8$

$\Sigma (x - \bar{x})^2 = \Sigma d^2 - \frac{(\Sigma d)^2}{n_1} = 60\text{-}0 = 60$ $\quad \Sigma (y - \bar{y})^2 = \Sigma D^2 - \frac{(\Sigma D)^2}{n_2} = 186 - \frac{324}{10} = 153.6$

$S^2 = \frac{1}{n_1+n_2-2} [\Sigma (x - \bar{x})^2 + \Sigma (y - \bar{y})^2] = \frac{1}{14} (60 + 153.6) = 15.2571$

$t = \frac{68-67.8}{\sqrt{15.2571}\left(\frac{1}{6}+\frac{1}{10}\right)^{1/2}} = \frac{0.2}{\sqrt{15.2571 \times 0.2667}} = 0.99$

Since calculated t is much less than tabulated t (1.76). Hence null hypothesis may be retained at 5 % level of significance and we conclude that the data are inconsistent with the suggestion that the sailors are on the average taller than soldiers

Example 3 : A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure :

5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, and 6

Can it be concluded that stimulus will, in general, be accompanied by an increase in blood pressure ?

Solution: Here we are given the increments in blood pressure *i.e.* $d_i = (x_i - y_i)$.

Null Hypothesis : $H_0 : \mu_x = \mu_y$, *i.e.* there is no significant difference in the blood pressure reading of the patients before and after the drug.

The test statistic $t = \bar{d}/(s/\sqrt{n}) \sim t\,(n-1)$

d	5	2	8	-1	3	0	-2	1	5	0	4	6	31
d^2	25	4	64	1	9	0	4	1	25	0	16	36	185

$\bar{d} = \frac{\Sigma d}{n} = \frac{31}{12} = 2.58$ and

$$s^2 = \frac{1}{n-1}\Sigma(d - \bar{d})^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\Sigma d)^2}{n}\right]$$

$$= \frac{1}{11}\left[185 - \frac{(31)^2}{12}\right] = \frac{1}{11}(185 - 80.08) = 9.5382$$

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2.58 \times \sqrt{12}}{\sqrt{9.5382}} = \frac{2.58 \times 3.464}{3.09} = 2.89$$

Tabulated $t_{0\,05}$ for 11 d. f. is 1.80. Since calculated $t > t_{0.05,}$ Ho is rejected at 5% level of significance, we concluded that stimulus will, in general, be accompanied by an increase in blood pressure .

Exercise 1 : A random sample of 8 envelopes is taken from letter box of a post office and their weights in grams are found to be 12.1, 11.9,12.4, 12.3, 11.9, 12.1, 12.4, 12.1. Does this sample indicate at 1% level that the average weight of envelopes received at that post office is 12.35 gms.

Exercise 2 : Two independent samples of 8 and 7 items respectively had the following values :

Sample I	9	11	13	11	15	9	12	14
Sample II	10	12	10	14	9	8	10	

Is the difference between the means of samples significant?

Exercise 3 : Show how you would use student's t- test to decide whether the two sets of observations: [17,27,18,25,27,29,27,23,17] and [16,16,20,16,20,17,15,21] indicate samples drawn from the same universe.

Practical No. 9

Title : F- test for comparing two sample variances
Objective : To compare the sample variances.

If χ_1^2 and χ_2^2 a are two independent chi-square variants with v_1 and v_2 d. f. respectively , then F- Statistic is defined by $F = \frac{\chi_1^2/v_1}{\chi_2^2/v_2}$, In other words , F is defined as the ratio of two independent chi-*square variates* divided by their respective degrees of freedom and it follows snedecor' s F- distribution with (v_1, v_2) d.f.

F-statistic is also define as the ratio of two independent 'unbiased estimator' of the population variance. It is also known as variance ratio test.

$$F = \frac{large\ variance}{smaller\ variance}$$

Suppose we want to test (i) whether two independent samples x_i, $(i = 1, 2, \ldots., n_1)$ and y_i, $(i = 1, 2, \ldots., n_2)$ have been drawn from the normal populations with the same variance σ^2 or (ii) whether the two independent estimates of the population variance are homogeneous or not .

Setup the null hypothesis (H_o) that $\sigma x^2 = \sigma y^2 = \sigma^2\ i.e., the$ population variances are equal , the statistic F is given by $F=\frac{S_x^2}{S_y^2}$, $S_x^2 > S_y^2$

where $S_x^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1}(x_i - \bar{x})^2$ *and* $S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2}(y_i - \bar{Y})^2$

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows snedecor's F- distribution with $(n_1 - 1, n_2 - 1)\ d.f.$

Applications : (i) F- test for equality of population variances

(ii) F- test for testing the significance of an observed multiple correlation coefficient.

(iii) F- test for testing the linearity of regression

(iv) F- test for equality of several means

Example 1: Two random samples gave the following results:

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population.

Solution : Null Hypothesis H_0: The two samples have been drawn from the same normal population $\sigma_1^2 = \sigma_2^2$

Apply F-test , $F=\frac{S_x^2}{S_y^2}$, $S_x^2 > S_y^2$

where $S_x^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1}(x_i - \bar{x})^2$ and $S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2}(y_i - \bar{Y})^2$

given that $n_1=10$, $n_2=12$, $\bar{x}_1 = 15$, $\bar{x}_2 =14$, $\sum_{i=1}^{n_1}(x_i - \bar{x})^2 = 90$ and $\sum_{i=1}^{n_2}(y_i - \bar{Y})^2 = 108$

therefore $S_x^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1}(x_i - \bar{x})^2 = 10$ and $S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2}(y_i - \bar{Y})^2 = 9.82$

So $F = \frac{S_x^2}{S_y^2} = \frac{10}{9.82} = 1.018$, $S_x^2 > S_y^2$

Tabulated $F_{0.05}(9,11) = 2.90$

Since calculated F is less than tabulated F, it is not significant. Hence null hypothesis of equality of population variances may be accepted.

Example 2 : Two random samples drawn from two normal populations are:
20, 16, 26, 27, 23, 22, 18, 24, 25, 19 and 27, 33, 42, 35, 32, 34, 38, 28, 41, 43, 30, 37
Obtain estimates of the variances of the populations and test whether the populations have same variances. [Given $F_{0.05} = 3.11$ for 11 and 9 degrees of freedom]

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
20	27	-2	-8	4	64
16	33	-6	-2	36	4
26	42	4	7	16	49
27	35	5	0	25	0
23	32	1	-3	1	9
22	34	0	-1	0	1
18	38	-4	3	16	9
24	28	2	-7	4	49
25	41	3	6	9	36
19	43	-3	8	9	64
	30		-5		25
	37		2		4
$\Sigma x = 220$	$\Sigma y = 420$			$\Sigma (x-\bar{x})^2 = 120$	$\Sigma (y-\bar{y})^2 = 314$

Solution: $\bar{x} = 22, \quad \bar{y} = 35 \qquad S_x^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1}(x_i - \bar{x})^2 = \frac{120}{9} = 13.33$

and $S_y^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2}(y_i - \bar{Y})^2 = \frac{314}{11} = 28.545$

$$F = \frac{S_y^2}{S_x^2} = \frac{28.545}{13.33} = 2.141 , \quad S_y^2 > S_x^2$$

Tabulated $F_{0.05}$ (11, 9) = 3.11. Since calculated F is less than tabulated F, it is not significant. Hence null hypothesis of equality of population variances may be accepted.

Exercise 1 : Two random samples of sizes 8 and 11, drawn from two normal populations, are characterized as follows:

Sample is drawn	Size of sample	Sum of observation	Sum of squares of observations
I	8	9.6	61.52
II	11	16.5	73.26

You are to decide if the two populations can be taken to have the same variance.

Exercise 2: The nicotine content (in milligrams) of two samples of tobacco were found to be as follows:

Sample A : 24 27 26 21 25

Sample B : 27 30 28 31 22 36

Can it be said that the two samples come from the same normal population?

Practical No. 10

Title : **Chi square test in 2 x 2 Contingency Table and Yate's correction for continuity**

Objective : **To test the independence of attributes in 2 x 2 Contingency table.**

The square of standard normal variate is a chi square variate with 1 d.f. It is used to measure the deviation of observed frequencies in an experiment from the expected frequencies obtained from some hypothetical universe.

Conditions for validity of x^2 test :

For the validity of chi-square test of 'goodness of fit' between theory and experiment, the following conditions must be satisfied:

(i) The sample observations should be independent.

(ii) Constraints on the cell frequencies, should be linear i.e., $\Sigma O_i = \Sigma E_i$.

(iii) N, the total frequency should be reasonably large, say greater than 50,

(iv) No theoretical cell frequency should be less than 5. If any theoretical cell frequency is less that 5, then for the application of x^2 test , it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

Applications:

I. To test if the hypothetical value of the population variance is $\sigma^2=\sigma_0^2$ (say)

II. To test the 'goodness of fit'.

III. To test the independence of attributes.

IV. To test the homogeneity of independent estimates of the population variance.

V. To combine various probabilities obtained from independent experiment to give a single test of significance.

VI. To test the homogeneity of independent estimates of the population correlation coefficient.

2 x 2 Contingency Table

For the 2 x 2 table,

			Total
	a	b	a+b
	c	d	c+d
Total	a+c	b+d	N

$$x^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad \text{.....(i)}$$

where N= a + b + c + d

follows χ^2 distribution with (2-1) (2-1)= 1 d.f.

Yate's Correlation : In 2 x2 contingency table, if any one of the cell frequency is less than 5 then the use of pooling method of x^2 results in x^2 with o d.f. which is meaningless. In this case we apply a correction due to Yate's which is known as Yate's correction for continuity. This consists in adding 0.5 to cell frequency which less than 5 and then adjusting for the remaining cell frequency accordingly and then use formula for x^2 as (i) . The x^2 test is also applied without pooling method by formula.

$$x^2 = \frac{N\left[|ad-bc|-\frac{N}{2}\right]^2}{(a+c)(b+d)(a+b)(c+d)}$$, where N = a + b + c + d, with 1 d. f.

Example 1 : From the following table, test the hypothesis that flower colour is independent of flatness of leaf

	Flat leaves	Curled leaves	Total
White flowers	99	36	135
Red flowers	20	5	25
Total	119	41	160

Solution: Set the hypothesis that the flower colour is independent of the flatness of leaves.

The excepted frequency of the plants having white flowers and flat leaves i.e. corresponding to the frequency 99 is

$$\frac{(A_i)\,x\,(B_j)}{N} = \frac{135\,x\,119}{160} = 100$$

Other excepted frequency are given below:

	Flat leaves	Curled leaves
White flowers	100	35
Red flowers	19	06

$$\chi^2 = \Sigma\left[\frac{(O_i-E_i)^2}{E_i}\right] = \frac{(99-100)^2}{100} + \frac{(36-35)^2}{35} + \frac{(20-19)^2}{19} + \frac{(5-6)^2}{6} = 0.26$$

Tabulated value of χ^2 at 5% for one degree of freedom is 3.841. The calculated value of χ^2 is much less than this value. Hence we conclude that the colour of flower is independent of the flatness of leaf.

Example 2 : The following were the data of 40 individuals classified according to smoking and residential background. Test whether the smoking habit is inherent with the residential background.

	Smokers	Non-smokers	Total
Rural	15	3	18
Urban	15	7	22
Total	30	10	40

Yate's method: Here one cell frequency is less than 5, hence Yate's correction for continuity is applied, subtraction ½ from 15 and 7 and adding ½ to 15 and 3, we have

	Smokers	Non-smokers	Total
Rural	14.5	3.5	18
Urban	15.5	6.5	22
Total	30.0	10.0	40

Set the null hypothesis H_0 : Smoking and residential background are independent.

$\chi^2 = \frac{(14.5\ x\ 6.5 - 15.5\ x\ 3.5)^2}{30x\ 10x18x22}$ x 40 = 0.5387 *or*

$\chi^2 = \frac{N\ [|ad-bc|-\frac{N}{2}]^2}{(a+c)(b+d)(a+b)(c+d)}$, where N = a + b + c + d,

$= \frac{40\ [|15x7 - 15\ x\ 3|-\frac{40}{2}]^2}{(30)(10)(18)(22)}$ = 0.5387

The calculated value of square is less than the tabulated value 3.81 at 5 per cent level of significance of therefore there is evidence to say that the smoking habit is independent of the residential background.

Exercise 1 : 100 individuals of a particular race were tested with an intelligence test and classified into two classes. Another group of 120 individuals belong to another race were administered the same intelligence test and classified into the same two classes. The following are the observed frequencies of the two races:

	Intelligence	
Race	Intelligent	Non-intelligent
Race I	42	58
Race II	55	65

Test whether the intelligence is anything to do with the race.

Exercise 2 : In an experiment on the immunization of goats from anothrax the results were obtained as summarized below:

	Died of anthrax	Survived
Inoculated	2	10
Not inoculated	22	6

Test whether the survival from anthrax and inoculations are dependent or not?

Practical No. 11

Title : Computation of Correlation coefficient (r) and its testing
Objective : To find out correlation coefficient and test their significance.

Correlation coefficient is a measure of degree or extent of linear relationship between two variables x and y. The population correlation coefficient is denoted by ρ and its estimate by r and is defined as.

$$r = \frac{COV(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}, \text{ where cov (x, y)} = \frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})$$

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

Properties of correlation coefficient

(i) Correlation coefficient is a pure number i.e. it has no unit.

(ii) The correlation coefficient ρ ($or\ r$)ranges from -1 to 1.

(iii) The relation between the correlation coefficient 'r' and the two regression coefficients b_{YX} and b_{XY} is r= $\sqrt{b_{YX}.b_{XY}}$

(iv) The sign of r will be the same as that of b_{YX} and of b_{XY}.

(v) If the two variables are independent, the correlation coefficient between them is zero but the converse is not true.

Test of Significance : Under the null hypothesis H_o : $\rho = 0$ ie population correlation coefficient is zero, the statistic t= $\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ follows students t-distribution with (n-2) d.f. If the value of t comes out to be significant, we reject H_o and conclude that $\rho \neq 0$ i.e. correlation is significant in the population.

Example 1 : Find out the coefficient of correlation between the sales and expenses of the following 10 firms (figures in ' 000 Rs.).

Firms:		1	2	3	4	5	6	7	8	9	10
Sales	X :	50	50	55	60	65	65	65	60	60	50
Expenses	Y :	11	13	14	16	16	15	15	14	13	13

Sales X	Dev .from mean(58) x	Square of Dev x^2	Expenses Y	Dev .from mean (14) y	Square of Dev. y^2	Product of deviations xy
50	-8	64	11	-3	9	+24
50	-8	64	13	-1	1	+8
55	-3	9	14	0	0	0
60	+2	4	16	+2	4	+4
65	+7	49	16	+2	4	+14
65	+7	49	15	+1	1	+7
65	+7	49	15	+1	1	+7
60	+2	4	14	0		0
60	+2	4	13	-1	1	-2
50	-8	64	13	-1	1	+8
$\Sigma X = 580$	$\Sigma x = 0$	$\Sigma x^2 = 360$	$\Sigma Y = 140$	$\Sigma y = 0$	$\Sigma y^2 = 22$	$\Sigma xy = 70$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{580}{10} = 58 \text{ and } \bar{Y} = \frac{\Sigma Y}{N} = \frac{140}{10} = 4$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \; X \; \Sigma y^2}} = \frac{+70}{\sqrt{360 \; X \; 22}} = 0.786$$

to test the significance of correlation, use statistic $t = \frac{r}{\sqrt{1-r^2}} . \sqrt{n-2}$

$$= \frac{0.786}{\sqrt{1-(0.786)^2}} \sqrt{10-2} = 3.59$$

Tabulated value of t at 5% level of significance = 2.306 which is less then 3.59 therefore we conclude that correlation coefficient is significant.

Exercise1: Find the coefficient of correlation between the experience & performance of persons given below and also test its significance.

Experience (X):	16	12	18	4	3	10	5	12
Performance (Y):	23	22	24	17	19	20	18	21

Exercise 2: Yield of food grains in kg/ha in Kharif and Rabi seasons from 1975 -76 to 1984-85 were as follows:

Year	Kharif (X)	Rabi (Y)
1975-76	89	104
1976-77	94	109
1977-78	94	116
1978-79	78	105
1979-80	93	120
1980-81	95	119
1981-82	92	130
1982-83	106	131
1983-84	104	134
1984-85	105	142

Find the coefficient of correlation between yield of food grain in both seasons and also test its significance.

Practical No. 12

Title : Fitting of regression equation Y on X and X on Y.
Objective : To compute two lines of regression and correlation coefficient.

The line of regression is the straight line which in the least square sense gives the best fit to the given frequency .The two lines of regression are regression line of y on x is $y-\bar{y} = b_{yx}(x-\bar{x}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$ and regression line x on y is $x-\bar{x} = b_{xy}(y-\bar{y}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$ where b_{yx} and b_{xy} are regression coefficients of y on x and x on y respectively. The regression coefficient is a measure of change in dependent (response) variable corresponding to an unit change in independent variable .

Properties of Regression Coefficient

(i) Geometric mean of regression coefficients is correlation coefficient.
(ii) If one regression coefficients is greater than unity, the other must be less than unity.
(iii) Arithmetic mean of the regression coefficients is greater than the correlation coefficient.
(iv) Regression coefficients are independent with the change of origin but not of scale.

Example 1 : From the data given below compute regression equation of Y on X and X on Y

X	18	19	20	21	22	23	24	25	26	27
Y	17	17	18	18	18	19	19	20	21	22

Solution:

X	X-22=U	U^2	Y	Y-19=V	V^2	UV
18	-4	16	17	-2	4	8
19	-3	9	17	-2	4	6
20	-2	4	18	-1	1	2
21	-1	1	18	-1	1	1
22	0	0	18	-1	1	0
23	1	1	19	0	0	0
24	2	4	19	0	0	0
25	3	9	20	1	1	3
26	4	16	21	2	4	8
27	5	25	22	3	9	15
Total	$\Sigma U = 5$	$\Sigma U^2 = 85$		$\Sigma V = -1$	$\Sigma V^2 = 25$	$\Sigma UV = 43$

$$\bar{X} = 22 + \frac{5}{10} = 22.5, \quad \bar{Y} = 19 + \frac{-1}{10} = 18.9$$

$$b_{YX} = \frac{n\Sigma UV - \Sigma U.\Sigma V}{n\Sigma U^2 - (\Sigma U)^2} = \frac{10 \times 43 - 5x(-1)}{10 \times 85 - 25} = \frac{430+5}{850-25} = 0.527$$

$$b_{XY} = \frac{n\Sigma UV - \Sigma U.\Sigma V}{n\Sigma V^2 - (\Sigma V)^2} = \frac{10 \times 43 - 5x(-1)}{10 \times 25 - 1} = \frac{430+5}{250-1} = 1.747$$

Regression line of Y on X is $Y - 18.9 = 0.527\ (X - 22.5)$ and

Regression line of X on Y is $X - 22.5 = 1.747\ (Y - 18.9)$

Example 2 : Given that

$n = 10, \Sigma X = 666, \Sigma Y = 663,\ \Sigma X^2 = 44490,\ \ \Sigma Y^2 = 44061, \Sigma XY = 44224.$

Compute both lines of regression X on Y and Y on X and also estimate the value of Y when X = 70

$\bar{X} = \frac{\Sigma X}{n}$ 66.6 $\bar{Y} = \frac{\Sigma Y}{n} = 66.3,\ \Sigma u_i^2 = 44490 - \frac{(666)^2}{10} = 134.4$

$\Sigma v_i^2 = 44061 - \frac{(663)^2}{10} = 104.1$

$\Sigma u_i v_i = 44224 - \frac{666 \times 663}{10} = 68.2$

The regression coefficient $b_{YX} = \frac{68.2}{134.4} = 0.507$

Regression line of Y on X is $(Y - 66.3) = 0.507\ (X - 66.6))$

or $Y = 0.507\ X + 32.53$

estimate of Y when X=70 is $\hat{Y} = 0.507 \times 70 + 32.53 = 68.02$

For the regression line of X on Y, $b_{XY} = \frac{68.2}{104.1} = 0.655$ and

$$X - 66.6 = 0.655\ (Y - 66.3)$$

Exercise 1: You are given the data relating the purchases and sales. Obtain the two regression equations and estimate the likely sales when the purchases equal 100.

Purchases:	62	72	98	76	81	56	76	92	88	49
Sales:	112	124	131	117	132	96	120	136	97	85

Exercise 2: The average yield per acre of maize and the annual rainfall were 985 lbs. and 12.8 inches respectively, the standard deviations 70.1 lbs and 1.6 inches respectively. The coefficient of correlation between yield and rainfall is 0.52.

Calculate:- (i) the probable yield of maize when rainfall is 9.2 inches,

(ii) The probable annual rainfall for a yield of 1400 lbs/acre.

Exercise 3:- The equations of two regression lines obtained in a correlation analysis of 60 observations are $5x = 6y + 24$ and $1000\ y = 768\ x - 3608$. What is the correlation coefficient and find the ratio of variance of x and y ?

Practical No. 13

Title : **Analysis of completely Randomized Design (CRD) for equal and unequal number of replication.**

Objective : **To analyze data based CRD for equal & unequal number of replication.**

The Completely randomized design is the simplest of all the designs, based on principle of randomization and replication. In this design treatments are allocated at random to the experimental units over the entire experimental material.

Let us suppose that we have v treatments, the ith treatment being replicated r_i times, i= 1, 2....., v. Then the whole experimental material is divided into n= Σr_i experimental units and the treatments are distributed completely at random over the units subject to the condition that the ith treatment occurs r_i times.

Advantages :- (i) C.R.D. results in the maximum use of the experimental units since all the experimental material can be used.

(ii) The design is very flexible: Any number of treatments can be used and different treatments can be used unequal number of times without unduly complicating the statistical analysis in most of cases.

(iii) The statistical analysis remains simple if some or all the observation for any treatment are rejected or lost or missing.

(iv) It provides the maximum number of degrees of freedom for the estimation of the error variance, which increases the sensitivity on the precision of the experiment for small experiments.

Disadvantages:- (i) The design is not suitable for a small number of treatments.

(ii) It is difficult to find homogenous experimental units in all respects.

Applications :- (i) Completely randomized design is most useful in laboratory technique and methodological studies, like, in physics, chemistry or cookery, in chemical and biological experiments, in some green house studies, etc., where either the experimental material is homogeneous or the intrinsic variability between units can be reduced.

(iii) C.R.D. is also recommended in situations where an appreciable fraction of units is likely to be destroyed or to fail or respond.

Analysis :

I For equal number of replication :

Let yij be the jth observation of ith treatment, i= 1,2.....t and j= 1,2.....r.

Total observations N= rt and Grand total = $\Sigma\Sigma yij$, Correction factor = $\frac{(GT)^2}{N}$

Total Sum of square (TSS) = $\sum_i \sum_j y_{ij}^2 - C.F.$, Treatment sum of square (SSTr) = $\frac{\Sigma T_i^2}{r}$ - C.F.

Sum of square due to Error SSE= TSS – TrSS

ANOVA

Source of variation	d.f.	SS	MS	F
Treatment	t-1	SSTr	$\frac{SSTr}{t-1} = V_T$	V_T/V_E
Error	t (r-1)	SSE	$\frac{SSTE}{t(r-1)} = V_E$	
Total	r t-1	TSS		

If $F_{cal} \geq F_{tab\,\alpha\%}\,(t-1), t\,(r-1)$ then H_o is rejected at $\alpha\%$ level of significance otherwise accepted. For comparison between treatment means, we want to calculate critical difference CD = SE (d) x $t_{\alpha\%}$ (with error d. f.), where SE (d) = $\sqrt{\frac{2\,V_E}{r}}$

II For Unequal no. of replications

Total sum of squares TSS = $\Sigma\,\Sigma y_{ij}^2 - Cf.$ Sum of squares due to treatment SSTr=$\sum \frac{T_i^2}{r_i} - Cf.$

Error Sum of Squares = TSS - SSTr

ANOVA is same as I ie for equal number of replications.

If H_o is rejected, then we want to calculate CD for comparisons between treatment means as

CD= SE (d) x $t_{\alpha\%}$ at error d. f. , where SE (d) = $\sqrt{V_E\,(\frac{1}{r_i} + \frac{1}{r_j})}$

r_i & r_j are the no. of replication in ith and jth treatments respectively.

Example for equal no. of replications

Grain yield per plant (gms) of maize of nine varieties in completely a randomized design were as tabulated below:

Varieties	Rep. I	Rep. II	Rep.III
V_1	21.0	20.0	19.5
V_2	19.0	18.0	18.5
V_3	18.5	18.0	18.9
V_4	27.5	26.9	27.0
V_5	31.0	32.5	32.6
V_6	31.5	30.5	32.0
V_7	25.3	25.0	26.6
V_8	39.0	40.0	38.5
V_9	39.0	38.5	40.0

Analyse the experimental data and interpret the result.

Solution: Variety totals, replication totals and grand total are first calculated which are as follows:

$V_1 = 605, V_2 = 555, V_3 = 55.4, V_4 = 81.4,\ V_5 = 96.1, V_6 = 94.0, V_7 = 76.9, V_8 = 117.5,$

$V_9 = 117.5;\ C.F. = \frac{(GT)^2}{N} = \frac{(754.8)^2}{27} = 21100.85$, where $GT = 754.8$

Total SS $= \Sigma\,\Sigma y_{ij}^2 - C.F. =$ 22686.48 – 21100.85 = 1585.63

Treatment sum of squares SSTr = $\sum \frac{T_i^2}{r_i} - C.F.$ = 22677.65-21100.85 = 1576.80

Error S.S. = 1585.63 - 1576.80 = 8.83

Analysis of variance table

Source of variation	d.f.	SS	MS	F
Treatment	8	1576.03	197.00	401.63**
Error	18	8.83	0.4905	
Total	26	1585.63		

**Significant at 1 per cent level of significance between two-treatment means is

SE (d) = $\sqrt{\frac{2\,VE}{r}} = \sqrt{\frac{2\,X\,0.4905}{3}} = 0.5718$

CD = SE (d) x $t_{0.05\ at\ 15}\ d.f. = 0.5718$ x 2.131 = 1.218

There is a highly significant difference between varieties since the calculated value of F is greater than the tabulated value of F at 1 percent level of significance. Now we find which of

the paris of treatment means differ significantly. The variety means are arranged in ascending order and displayed below. Now comparing the difference between variety means by the appropriate critical difference (C.D.) value, the varieties which do not differ significantly are underlined.

$\bar{V}_2$	$\bar{V}_3$	$\bar{V}_1$	$\bar{V}_7$	$\bar{V}_4$	$\bar{V}_6$	$\bar{V}_5$	$\bar{V}_8$	$\bar{V}_9$
18.5	18.5	20.17	25.63	27.13	31.33	32.03	39.17	39.17

Examples for unequal number of replication

Given the following data obtained from a completely randomized design with four treatments, analyse the given data and draw conclusion about the equality of treatment effects.

Treatments			
T1	T2	T3	T4
20.9	23.7	13.2	5.8
12.4	14.4	10.2	6.1
10.1	9.0	5.1	4.8
4.2			1.5

Solution: Here we test null hypothesis: H_0: $\tau_1 = \tau_{2=}\tau_3 = \tau_4$

Treatment totals are $T_1 = 47.6,\ T_2 = 47.1, T_3 = 28.5,\ T_4 = 18.2\ amd\ G.T. = 141.4$

Number of replications for treatments $T_1, T_2, T_3\ and\ T_4$ are $4, 3, 4\ respectively\ and\ n = 14.$

$$C.F. \frac{(141.4)^2}{14} = 1428.14$$

Total S.S.= $(20.9^2 + 12.4^2 + \ldots\ldots 1.5^2) - C.F. = 2670.92 - 1428.14 = 1242.78$

Treatment S.S. = $\frac{47.6^2}{4} + \frac{47.1^2}{3} + \frac{28.5^2}{3} + \frac{18.2^2}{4} - C.F. = 326.54$

Error S.S. = 1242.78-326.54 = 916.24

Analysis of variance table

Source of variation	d.f.	SS	MS	F
Treatment	3	326.54	108.85	1.19
Error	10	916.24	91.62	
Total	13	1242.78		

Tabulated value of $F_{.05\ (3,10)} = 3.71$ which is greater than the calculated value of F= 1.19.
Hence, we accept null hypotheses which means that the four treatments are equally effective.

Exercise 1 : Carry out analysis of variance for the data of yield of 7 varieties, 5 observations being taken on each variety. Given that $F_{6,28}$ at 5% level of significance is 2.45.

V_1	V_2	V_3	V_4	V_5	V_6	V_7
13	5	14	14	17	15	16
11	11	10	10	15	9	12
10	13	12	15	14	13	13
16	18	13	17	19	14	15
12	12	11	10	12	10	12

Exercise 2 : Three processes A, B and C are tested to see whether their output are equivalent. The following observations of output are made:

A	10	12	13	11	10	14	15	13
B	9	11	10	12	13			
C	11	10	15	14	12	13		

Carry out the analysis of variance and state your conclusions.

Practical No. 14

Title : Analysis of Randomized block design (RBD)
Objective : To analyze experimental data on RBD and prepare ANOVA

In field experimentation, if the whole of the experimental area is not homogeneous and the fertility gradient is only in one direction, then a simple method of controlling the variability of the experimental material consists in stratifying or grouping the whole area into relatively homogeneous strata or sub-groups (or blocks or replicates, as they are called), perpendicular to the direction of the fertility gradient and the treatments are applied at random to relatively homogeneous units within each strata or block and replicated over all the blocks

Advantages (i) Accuracy: This design has been shown to be more efficient or accurate than C.R.D. for most types of experimental work. The elimination of between S.S. from residual S.S., usually in a decrease of error mean S.S.

(ii) Flexibility: In R.B.B. no restrictions are placed on the number of treatments or the number of replicates. In general, at least two replicates are required to carry out the test of significance (Factorial design is an exception). In addition, control (check) or some other treatments may be include more than once without complication in the analysis.

(iii) Easy of Analysis: Statistical analysis is simple and rapid. Moreover the error of any treatment can be isolated and any number of treatment may be omitted from the analysis without complicating it.

Disadvantages : R.B.D. is not suitable for large number of treatments or for cases in which complete block contains considerable variability.

Statistical Analysis

Let yij be the yield from ith treatments and jth replication, then under the null Hypothesis H_0: $\tau_1 = \tau_2 = \cdots \ldots \ldots \ldots$ i.e. there is no difference between treatment means.

ANOVA FOR RBD

Source of variation	d.f.	S.S.	MS	F
Replication	r-1	SSR	$\frac{SSR}{r-1} = V_R$	$\frac{V_R}{V_E}$
Treatment	t-1)	SSTr	$\frac{SSTr}{t-1} = V_T$	$\frac{V_T}{V_E}$
Error	(r-1) (t-1)	SSE.	$\frac{SSE}{(r-1)(t-1)} = V_E$	
Total	rt-1	TSS		

Correction factor = $\frac{(GT)^2}{rt}$ and G.T. = $\Sigma\ \Sigma y_{ij}$

Total sum of square (TSS) = $\Sigma\ \Sigma y_{ij}^2$ - CF, Sum of square due to treatment SSTr = $\frac{\Sigma Ti^2}{r} - CF$

Sum of square due to replication = = $\frac{\Sigma Ri^2}{t} - CF$ and Error sum of square = TSS -SSTr – SSR

(i) If $F_{cal} \geq F_{tab\cdot\alpha\%}$ $((t-1),(r-1)(t-1))$,then Ho is rejected at α% level of significance and concluded that treatments differ significantly.

Next step to compute critical difference CD for comparisons means as

CD= SE(d) x $t_{\alpha\%}$ (error d.f.), Where SE (d) = $\sqrt{\frac{2VE}{r}}$

(ii) If $F_{cal} < F_{tab\cdot\alpha\%}$ $((t-1),(r-1)(t-1))$, then we accept our hypothesis and conclude that treatments do not differ significantly.

Example 1. Consider the results given in the following table for an experiment involving six treatments in four randomized blocks. The treatments are indicated by numbers within parentheses.

Block	Yield for a randomized block experiment Treatment and yield					
1	(1) 24.7	(3) 27.7	(2) 20.6	(4) 16.2	(5) 16.2	(6) 24.9
2	(3) 22.7	(2) 28.8	(1) 27.3	(4) 15.0	(6) 22.5	(5) 17.0
3	(6) 26.5	(4) 19.6	(1) 38.5	(3) 36.8	(2) 39.5	(5) 15.4
4	(5) 17.7	(2) 31.0	(1) 28.5	(4) 14.1	(3) 34.9	(6) 22.6

Test whether the treatments differ significantly. Also (i) determine the critical difference between the means of any two treatments.

Solution : Under the null hypotheses H_0: $\tau_1 = \tau_{2=} = \tau_4$ *and* H_0: $b_1 = b_{2=} b_3 = b_4$

i.e. the treatments as well as blocks are homogeneous.

For finding the various S.S. we rearrange the above table as follows:

Blocks	Treatments						Totals (B_i)
	(1)	(2)	(3)	(4)	(5)	(6)	
1	24.7	20.6	27.7	16.2	16.2	24.9	130.0
2	27.3	28.8	22.9	15.0	17.0	22.5	133.3
3	38.5	39.5	36.9	19.6	15.4	15.4	176.1
4	28.5	31.0	34.9	14.1	17.7	17.7	148.8
Totals (T_i)	119.0	119.9	122.1	64.9	66.3	96.3	388.5=G

Correction Factor $= \frac{G^2}{N} = \frac{346332.25}{24} = 14{,}430.51$

Total S.S. $= \Sigma\Sigma\, y_{ij}^2$ - c.f. = 15,780.76 - 14,430.51 = 1,350.25

S.S. due to treatments (S.S.T.) $= \frac{1}{4}\ \sum_{i=1}^{6} T_i^2 - C.F. = \frac{61{,}326.81}{4}$ - 14,430.51= 901.19

S.S. due to blocks (S.S.B.) $= \frac{1}{6}\ \sum_{i=1}^{4} B_j^2 - C.F. = \frac{87899.63}{6} - 14430.51 = 219.43$

Error S.S.= T.S.S. - S.S.T. - S.S.B.= 1,350.25 -901.19 -219.43 = 229.63

ANALYSIS OF VARIANCE TABLE

Source of variation	d.f.	S.S.	MSS	Variance ratio F
Treatment	5	901.19	180.24	$F_t = \frac{180.24}{15.31} = 11.8^*$
Block	3	219.343	73.14	$F_b = \frac{73.14}{15.31} = 4.7$
Error	15	229.63	15.31	
Total	23	1,350.25		

Tabulated $F_{0.05\,(3.15)} = 5.42$ *and* $F_{0.05,5.15} = 4.5$

From the above table, we see that F_t is significant while F_b is not significant at 5% level of significance. Hence, F_t is rejected at 5% level of significance and we conclude that treatment effects are not alike. On the other hand, H_b may be retained at 5% level of significance and we may conclude that the blocks are homogenous.

Since H_0: $\tau_1 = \tau_{2=} \ldots\ldots \tau_6$ is rejected, we are interested to find out which treatment means differ significantly.

C.D. for any two treatment means = $t_{0.05}$ for error d.f. x $\sqrt{\frac{2VE}{r}}$ = 2.131 x $\sqrt{\frac{2\ X\ 15.31}{4}}$ = 2.131 x 2. = 5.97

By comparing the difference between the mean yields for different treatments with the critical difference, we find that the treatments 3, 2 and 1 are alike in giving significantly high yields while treatments 4 and 5 are alike in giving significantly low yields.

Exercise 1 . A varietal trial was conducted at a Research Station. The design adopted for the same was five randomized blocks of 6 plots each. The yield in 1b. per plot obtained from the experiment are as under.

	Varieties					
Blocks	V_1	V_2	V_3	V_4	V_5	V_6
I	30	23	34	25	20	13
II	39	22	28	25	23	32
III	56	43	43	31	49	17
IV	38	45	36	35	32	20
V	44	51	23	58	40	30

Analyse the data and comment on your findings.

Practical No. 15

Title : Analysis of Latin Square Design (LSD)

Objective : Perform ANOVA and calculate critical difference for treatment means in Latin Square Design.

A useful method of eliminating fertility variations consists in an experimental layout which will control variation in two perpendicular directions. Such a layout is Latin Square Design (L.S.D.)

Advantages (i) With two way grouping or stratification L.S.D. controls more of the variation than C.R.D. or R.B.D..

(ii) L.S.D. is an incomplete 3-way layout. Its advantage over the complete 3-way layout is that instead of m^3 experimental units only m^2units are needed.

(iii) The statistical analysis is simple though slightly complicated than for R.B.D. Even with missing data the analysis remains relatively simple.

(iv) More than one factor can be investigated simultaneously and with fewer trials than more complicated designs.

Disadvantages (i) Unlike R.B.D. , in L.S.D. the number of treatments is restricted to number of replications and this limits its field of application. L.S.D. is suitable for the number of treatments between 5 and 10 and for more than 10 to 12 treatments the design is seldom used since in that case the square becomes too large and does not remain homogenous.

(ii) In case of missing plots, when several units are missing that statistical analysis becomes quite complex.

Statistical Analysis

Let $y_{ijk}(i,j,k = 1,2,\ldots\ldots,m)$ be the yield from the unit in the ith row, jth coloumn and receiving the kth treatment, then total no. of observations m x m = m^2.

Let us set up the null hyposthesis.

For row effects $H_o : \alpha_1 = \alpha_2 = \ldots\ldots\ldots \alpha_m$, coloum effects $H_o : \beta_1 = \beta_2 = \ldots\ldots\ldots \alpha_m$,

treatment effects $H_o : \tau_1 = \tau_2 = \ldots\ldots\ldots \tau_m$

ANOVA FOR m x m LSD

Source of variation	d.f.	S.S.	MS	F
rows	m-1	SSR	$\frac{SSR}{m-1} = V_R$	$\frac{V_R}{V_E}$
columns	m-1	SSTC	$\frac{SSC}{m-1} = V_c$	$\frac{V_C}{V_E}$
treatments	m-1	SSTr	$\frac{SSTr}{m-1} = V_T$	$\frac{V_T}{V_E}$
error	(m-1) (m-2)	SSE	$\frac{SSE}{(m-1)(m-2)} = V_E$	
Total	m^2-1	TSS		

Where SSR $= \frac{\Sigma R_i^2}{m} - CF$, SSC $= \frac{\Sigma C_i^2}{m} - CF$, SSTr $= \frac{\Sigma T_k^2}{m} - CF$ and C.F.$= \frac{(GT)^2}{m^2}$

and GT is total of all the m^2 observations.

Total sum of squares TSS= $\Sigma \Sigma y_{ijk}^2 - Cf.$

Error SS = TSS - SSR - SSC - SSTr

Mean sum of squares and F-test are computed as given in the above table.

If $F_{cal} \geq F_{\alpha\%}\,(m-1)(m-2)$, then Ho is rejected at α%. Level and conclude that treatment means differ significantly and then we proceed to compute critical difference.

Critical difference CD = SE (d) x $t_{\alpha\%}$ (error d.f.) = $\sqrt{\frac{2V_E}{m}}$ x $t_{\alpha\%}$ (error d.f.)

If $F_{cal} < F_{\alpha\%}(m-1), (m-1)(m-2)$, then H_o may be accepted.

Example 1: An experiment was carried out to determine the effect of claying the ground on the field of barley grains; amount of clay used were as follows:

A : No clay , B: Clay at 100 per acre, C : Clay at 200 per acre, D: Clay at 300 per acre.

The yields were in plots of 8 meters by 8 meters by 8 meters and layout were:

Row \ Column	I	II	III	IV	Row Total (R_i)
I	D 29.1	B 18.9	C 29.4	A 5.7	83.1
II	C 16.4	A 10.2	D 21.2	B 19.1	66.9
III	A 5.4	D 38.8	B 24.0	C 37.0	105.2
IV	B 24.9	C 41.7	A 9.5	D 28.9	105.0
Column Total (C_i)	75.8	109.6	84.1	90.7	360.2

Perform the ANOVA and calculate the critical difference for the treatment mean yields.

Solution . The four treatment totals are :

A: 30. 8, B : 86. 9, C : 124.5, D :118.0, Grand total = 360.2, N=16.

$$\text{C. F .} = \left(\frac{360.2)^2}{16}\right) == 8109.0025$$

Total S.S.= $[(29.1)^2] + (18.9)^2 + ----+(28.9)^2$ -C.F. = 10052.08-8109.0025=1943.0775

SS.R.= $\frac{1}{4}[(83.1)^2] + (66.9)^2 + (105.2)^2 + (105.0)^2] - 8109.0025 = 259.3125$

S.S.C. = $\frac{1}{4}[(75.8)^2 + (109.6)^2 + (84.1)^2 + (90.7)^2] - 8109.0025 = 155.2725$

S.S.T. = $\frac{1}{4}[(30.8)^2 + (56.9)^2 + (124.5)^2 + (118.0)^2] - 8109.0025 = 1372.1225$

Error S.S. = T.S.S. = S.S.R.-S.S.C. – S.S.T. = 156.3700

ANOVA TABLE FOR L.S.D.

Source of variation	d.f.	S.S.	MS	F
rows	3	259.5375	86.4375	3.32<4.76
columns	3	155.2725	51.7575	1.98<4.76
treatments	3	1372.1225	457.3742	17.55>4.76
error	6	156.3700	26.0616	
Total	15	1943.0775		

where tabulated $F_{0.05}$ (3,6) = 4.76

From the above table we conclude that the variation due to rows and columns is not significant but the treatments, i.e., different levels of clay, have significant effect on the yield. To determine which of the treatment paris differ significantly, we have to calculate the critical difference (C.D.).

S.E. of difference between any two treatments means $= \sqrt{\frac{2V_E}{m}} = \sqrt{\frac{2x26.0616}{4}} = 3.609$

C.D.= 3.609 x $t_{0.05}$ (for error d. f.) = 3.609 x2.447= 8.83

The difference between mean yields of C and D is not significant and they may therefore be regarded alike as regards their effect on yield. Similar argument holds for the pair D and B. The treatments C and B are significantly different from each other as regards their effect on yield, since the difference between their mean yields, viz., 9.4 exceeds the C.D. As such treatment C is to be preferred to treatment B. Similar argument holds for any other pair left.

Exercise 1: Prepare analysis of variance for the following data of a Latin Square Design and interpret the results

A 12	C 19	B 10	D 8
C 12	B 12	D 6	A 7
B 22	D 10	A 5	C 21
D 12	A 7	C 27	B 17

APPENDICES

A1: t-table

Degrees of freedom	Level of significance	
	1%	5%
1	63.657	12.706
2	9.925	4.303
3	5.841	3.182
4	4.604	2.776
5	4.032	2.571
6	3.707	2.447
7	3.499	2.365
8	3.355	2.306
9	3.250	2.262
10	3.169	2.228
11	3.106	2.201
12	3.055	2.179
13	3.012	2.160
14	2.977	2.145
15	2.947	2.131
16	2.921	2.120
17	2.898	2.110
18	2.878	2.101
19	2.861	2.093
20	2.845	2.086
21	2.831	2.080
22	2.819	2.074
23	2.807	2.069
24	2.797	2.064
25	2.787	2.060
26	2.779	2.056
27	2.771	2.052
28	2.763	2.048
29	2.756	2.045
30	2.750	2.042

A2: χ^2 table

Degrees of freedom	Level of significance	
	1%	5%
1	6.635	3.841
2	9.210	5.991
3	11.345	7.815
4	13.277	9.488
5	15.086	11.070
6	16.812	12.592
7	18.475	14.067
8	20.090	15.507
9	21.666	16.919
10	23.209	18.307
11	24.725	19.675
12	26.217	21.026
13	27.688	22.362
14	29.141	23.685
15	30.578	24.996
16	32.000	26.296
17	33.409	27.587
18	34.805	28.869
19	36.191	30.144
20	37.566	31.410
21	38.932	32.671
22	40.289	33.924
23	41.638	35.172
24	42.980	36.415
25	44.314	37.652
26	45.642	38.885
27	46.963	40.113
28	48.278	41.337
29	49.588	42.557
30	50.892	43.773

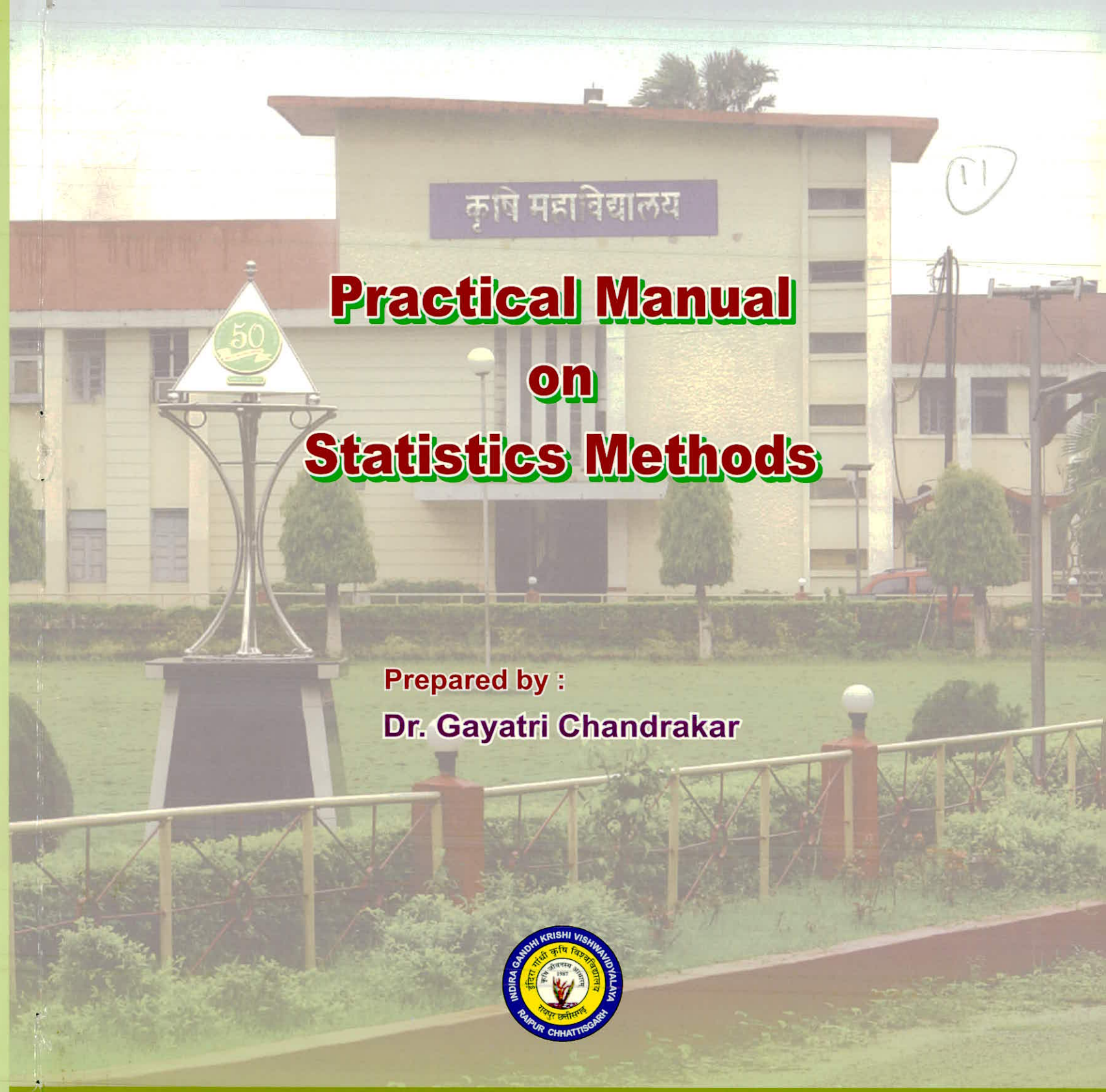
A3: F table values at 1% level of significance

Df for Smaller variance	Degrees of freedom for greater variance (numerator)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6083.32	6106.32	6123.86	6142.67	6157.28
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.42	99.43	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.98	26.92	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70
31	7.53	5.36	4.48	3.99	3.67	3.45	3.28	3.15	3.04	2.96	2.88	2.82	2.77	2.72	2.68
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.86	2.80	2.74	2.70	2.65
33	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91	2.84	2.78	2.72	2.68	2.63
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.82	2.76	2.70	2.66	2.61
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.80	2.74	2.69	2.64	2.60

A.3: F table values at 5% level of significance

Df for Smaller variance	Degrees of freedom for greater variance (numerator)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91	244.69	245.36	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.23	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15	2.11	2.08	2.05	2.03	2.00
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.04	2.02	1.99
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	2.09	2.06	2.03	2.01	1.98
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05	2.02	1.99	1.97
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04	2.01	1.99	1.96

Chhattisgarh Samvad



Practical Manual on Statistics Methods

Prepared by :

Dr. Gayatri Chandrakar

Department of Agricultural Statistics and Social Science
College of Agriculture
INDIRA GANDHI KRISHI VISHWAVIDYALAYA
Raipur (Chhattisgarh) 492 012