PRACTICAL MANUAL

ASS 212

STATISTICAL METHODS

2 (1+1)

B.Sc. (Hons.) Agriculture

Prepared by
Gaurav Shukla
Umesh Chandra
Annu

Department of Statistics & Computer Science
College of Agriculture
Banda University of Agriculture and Technology

Practical Manual on
Statistical Methods

Year
October, 2022

Publication No.: BUAT/M/2022/16

Prepared by:
Gaurav Shukla
Umesh Chandra
Annu

Published by:
College of Agriculture
Banda University of Agriculture and Technology
Banda-210001 (Uttar Pradesh)

FOREWORD

I am pleased to learn that the Department of Statistics & Computer Science is bringing out the practical (lab) manual of ASS 212: Statistical Methods for the students of B.Sc. (Hons.) Agriculture. The university has always been supportive for providing all sorts of help in facilitating the best teaching and learning environment. This practical (lab) manual will be helpful to improve students' understanding of the subject and easily accessible all the time. With this lab manual, the students will be able to develop their skills for better performance in academics and in the professional field as well.

I appreciate the tireless efforts of the faculty members of the Department of Statistics & Computer Science in developing and designing this manual. I am sure that this lab manual will be very useful to the students registered for the course of 'Statistical Methods'. This manual will work as a ready reckoner for the students to help them in preparation of competitive examination for higher studies.

With best wishes,

(G.S. Panwar)
Dean
College of Agriculture
Banda University of Agriculture and Technology
Banda-210001 (UP)

PREFACE

The practical manual entitled “Statistical methods” has been prepared keeping in view the syllabus of this course as per guidelines of ICAR’s 5th Dean’s Committee offered in B. Sc. (Honours) Agriculture degree programme and is intended to be used in conducting practicals of this course. This manual includes the basic knowledge and exercises related to frequency distribution, diagrammatic and graphical representation of data, measures of central tendency, measures of dispersion, skewness & kurtosis, correlation & regression, probability distributions, small and large sample tests, sampling techniques and design of experiments. The new chapters were added as per the revision and incorporated in such a way that made it easily understandable to the students to make it more clear and attractive. Pictures, graphs, figures, etc. are used at appropriate places. This manual is a combined effort of all the faculty members of the Department of Statistics & Computer Science, for which I am thankful for my teammates. The material included in this manual is taken from the different books, manuals, papers and internet facilities related to the subject.

On behalf of authors and as I/C Head of Department of Statistics & Computer Science, I acknowledge with thanks to Dr. N.P. Singh, Hon’ble Vice Chancellor, BUAT, Banda and Dr. G.S. Panwar, Dean, College of Agriculture, BUAT, Banda for encouraging us to bring out this practical manual.

With best wishes,

Gaurav

(Gaurav Shukla)
I/C, Head
Department of Statistics & Computer Science
College of Agriculture
BUAT, Banda

Abstract:

The practical manual entitled “Statistical methods” has been prepared keeping in view the syllabus of this course as per guidelines of ICAR’s 5th Dean’s Committee offered in B. Sc. (Honours) Agriculture degree programme and is intended to be used in conducting practicals of this course. This manual includes the basic knowledge and exercises related to frequency distribution, diagrammatic and graphical representation of data, measures of central tendency, measures of dispersion, skewness & kurtosis, correlation & regression, probability distributions, small and large sample tests, sampling techniques and design of experiments. The new chapters were added as per the revision and incorporated in such a way that made it easily understandable to the students to make it more clear and attractive. Pictures, graphs, figures, etc. are used at appropriate places. This manual is a combined effort of all the faculty members of the Department of Statistics & Computer Science, for which I am thankful for my teammates. The material included in this manual is taken from the different books, manuals, papers and internet facilities related to the subject.

Syllabus-ASS-212
Statistical Methods
Credit Hours: 2(1+1)

Theory

Introduction to Statistics and its Applications in Agriculture, Graphical Representation of Data, Measures of Central Tendency & Dispersion, Definition of Probability, Addition and Multiplication Theorem (without proof). Simple Problems Based on Probability. Binomial & Poisson distributions. Definition of Correlation, Scatter Diagram. Karl Pearson's Coefficient of Correlation. Linear Regression Equations. Introduction to Test of Significance, One sample & two sample test t for Means, Chi-Square Test of Independence of Attributes in 2 ×2 Contingency Table. Introduction to Analysis of Variance, Analysis of One Way Classification. Introduction to Sampling Methods, Sampling versus Complete Enumeration, Simple Random Sampling with and without replacement, Use of Random Number Tables for selection of Simple Random Sample.

Practical

Graphical Representation of Data. Measures of Central Tendency (Ungrouped data) with Calculation of Quartiles, Deciles & Percentiles. Measures of Central Tendency (Grouped data) with Calculation of Quartiles, Deciles & Percentiles. Measures of Dispersion (Ungrouped Data). Measures of Dispersion (Grouped Data). Moments, Measures of Skewness & Kurtosis (Ungrouped Data). Moments, Measures of Skewness & Kurtosis (Grouped Data). Correlation & Regression Analysis. Application of One Sample t-test. Application of Two Sample Fisher's t-test. Chi-Square test of Goodness of Fit. Chi-Square test of Independence of Attributes for 2 ×2 contingency table. Analysis of Variance One Way Classification. Analysis of Variance Two Way Classification. Selection of random sample using Simple Random Sampling.

List of Contents

Exercise – 1

FREQUENCY DISTRIBUTIONS

When observations, discrete or continuous are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. One of the aspects of classification is to prepare a frequency distribution.

In frequency distribution a number of classes are formed and number of observations belonging to each class is determined. The number of observations of a class is called as its frequency. A class may contain either a single value of the variable or a group of values of the variable. The group of values of the variable is commonly referred to as class-interval.

The class interval is characterized by two limits called as lower limit and upper limit. Average of two limits is called as mid value and is denoted by x. mid value is taken as representative value of that class- interval. In a frequency distribution, in general, number of classes should be between six and fifteen.

Preparation of a frequency distribution consists of following steps.

Step 1 – Form the classes. In case, class is in the form of class-interval calculate its mid value.

Step 2 --Read the observations one by one and make a tally mark against the class to which the observation belongs. When the class is in the form of class-interval, a particular class will include those observations whose value is equal to or greater than the lower limit but less than the upper limit.

Step 3 – Count the number of tally marks of each class. This number is called the frequency of that class.

- **Primary Data:** The data collected directly from the source is primary data. Suppose when you need to collect data for the favorite game of your classmates. You ask them directly. This is the primary data.

- **Secondary Data:** The collection of data indirectly or from some external source is the secondary data. These sources can be newspapers, magazines, television, internet etc. Suppose if you have to collect the data on the number of branches of your favourite restaurant in different cities. You may collect the data from newspapers or the internet. Such data is the secondary data.

Organization of the Data

If you have your data, what can be done with them? These data are raw data. We have to organize it in some meaningful way to take out the information. Consider the example of a collection of data of your favourite sport. From this raw data, you can count the number of the people who like a particular sport.

The representation of the various observations and tally marks in a form of table is the frequency distribution. The frequency is the number of the times an observation occurs. It is the number of repetitions. Consider in a class of 30 students, 5 like badminton. 10 students like cricket, 3 like tennis, 4 like football, 7 like volleyball and 1 likes hockey.

Sports	Number of Students = frequency	Tally Marks
Badminton	5	卌
Cricket	10	卌 卌
Football	4	||||
Hockey	1	|
Tennis	3	|||
Volleyball	7	卌 ||

Cumulative Frequency

Sometimes for additional analysis, we need to calculate, what is called as cumulative frequency. There are two types of cumulative frequencies; less than type cumulative frequency and more than type cumulative frequency. To calculate less than type cumulative frequency add the ordinary frequencies of successive classes one by one starting with class one. Less than type cumulative frequency of a particular class tells us how many observations are "less than" upper limit of that class.
To calculate more than type cumulative frequency, add the ordinary frequencies of successive classes one by one starting with last class. More than type cumulative frequency of a particular class tells us how many observations are "more than or equal to" lower limit of that class. It may be observed that knowing any type of cumulative frequency one can calculate other type of cumulative frequency or ordinary frequency.

For Example: The marks of 130 students in mathematics is given below

Marks	No. of Students	Less than Type C.F.	More than Type C.F.
21	3	3	130
22	7	10	127
23	10	20	120
24	17	37	110
25	23	60	93
26	25	85	70
27	18	103	45
28	12	115	27
29	10	125	15
30	5	130	5

Yule's Formula for appropriate number of classes: Yule's Formula for appropriate number of classes is 2.5 $(n^{1/4})$ or $1+3.322 \log_{10}$ n and the class interval can be calculated as (maximum value in data set – minimum value in data set) / number of classes.

For Example: The height of 84 plants selected at random from an experiment field from BUAT is given below. Construct a frequency distribution for the given data.

3.36	5.00	5.50	5.83	6.16	6.90
3.93	5.06	5.50	5.83	6.23	7.00
4.00	5.06	5.50	5.83	6.34	7.03
4.34	5.20	5.50	5.86	6.40	7.10

4.34	5.26	5.50	5.90	6.40	7.16
4.63	5.34	5.53	5.93	6.50	7.20
4.67	5.34	5.63	5.96	6.53	7.26
4.67	5.34	5.67	6.00	6.53	7.33
4.73	5.34	5.67	6.00	6.56	7.34
4.80	5.34	5.69	6.03	6.60	7.43
4.90	5.34	5.80	6.08	6.67	7.50
4.96	5.43	5.80	6.10	6.76	7.50
4.96	5.46	5.80	6.10	6.83	7.76
5.00	5.46	5.80	6.16	6.83	8.55

For appropriate number of classes, we use $2.5\,(n^{1/4}) = 2.5(84^{(1/4)}) = 7.56 \approx 8$

For class interval, $CI = \dfrac{(Max.Value - Min\,Value)}{Number\,of\ classes} = \dfrac{8.55 - 3.36}{8} = 0.65$

C.I.	**Frequency**
3.36-4.01	3
4.01-4.66	3
4.66-5.31	13
5.31-5.96	29
5.96-6.61	23
6.61-7.26	5
7.26-7.91	7
7.91-8.56	1
Total	**84**

Problem 1: The numbers of casual laborers, working on different fields during a harvest season are given below. Prepare a frequency distribution.
23, 26, 24, 26, 25, 26, 23, 26, 25, 25, 27,24, 28, 22, 25, 25, 26, 25, 26, 24, 27, 26, 24, 24, 26, 25, 26, 22, 26, 25, 27, 27, 25, 26, 24, 26, 25, 24, 26, 28, 23, 27, 25, 26, 25, 22, 26, 26, 22, 25,26, 24, 26, 28, 29, 22, 25, 24, 25, 27, 28, 27,27, 22, 29, 29, 29, 24,27, 23.
Solution:

X	**Frequency**	**Tally Marks**

Problem 2: The age (in years) at marriage of 80 women as recorded in the marriage register is given below. Prepare a frequency distribution and calculate cumulative frequencies.
20, 24, 19, 24, 18, 19, 24, 23, 22, 19, 19, 20, 24, 22, 21, 22, 23, 25, 19, 18, 23, 20 ,25, 24 , 21, 22, 21, 20, 20, 19, 23, 20, 22, 21, 20, 25, 23, 22, 21, 20, 21, 23, 20, 19, 23, 21, 22, 23, 24, 25, 22, 21 ,22, 19, 21, 23, 20, 21, 22, 24, 21, 22, 23, 24, 21, 19, 18, 23, 22, 21, 22, 23, 24, 25, 21, 22, 23, 23, 24, 22.

X	Tally Marks	Frequency	Less than type C.F.	More than type C.F.

Problem 3: Marks obtained by 60 students in a subject are given below. Prepare a frequency distribution by using Yule’s formula for appropriate number of classes. Also calculate less than type cumulative frequency and more than type cumulative frequency.

28, 63, 84, 92, 26, 66, 43, 39, 75, 53, 21, 86,
36, 49, 58, 79, 88, 31, 46, 52, 63, 73, 37, 48,
87, 71, 66, 52, 33, 22, 60, 57, 47, 32, 66, 79,
81, 37, 81, 62, 58, 46, 48, 63, 88, 96, 29, 92,
37, 46, 44, 76, 76, 66, 32, 39, 83, 54, 47, 43

Solution:

Class Interval	Tally Marks	Frequency	Less than type C.F.	More than type C.F.

Problem 4: Following data relate to the grain yield (in g per plot) of a sorghum variety from 100 experimental plots of equal area. Prepare a frequency distribution by using Yule's formula for appropriate number of classes and also calculate less than and more than type cumulative frequencies.

196	169	126	181	174	164	209	143	85	165	194	129
166	164	154	139	128	120	80	168	161	170	195	136
91	197	152	145	98	166	189	156	175	150	148	144
152	190	182	180	118	142	191	148	152	187	129	119
139	177	191	214	167	165	186	111	155	164	125	99
86	170	111	169	141	164	89	180	225	139	127	136
144	165	154	94	156	142	162	160	189	156	176	150
142	144	153	190	183	180	161	65	170	136	90	125
98	166	187	74								

Solution:

Class Interval	Tally Marks	Frequency	Less than type C.F.	More than type C.F.

Problem 5: The information about monthly income of a number of formers of a village is given below (minimum is Rs 6000). Prepare a suitable frequency distribution from the present information of monthly income of the formers.

Income	No. of formers
Less than Rs. 8000	25
Less than Rs. 10000	45
Less than Rs. 12,000	62
Less than Rs. 14,000	78
Less than Rs. 16,000	85
Less than Rs. 18,000	91
Less than Rs. 20,000	97
Less than Rs. 22,000	100

Solution:

Problem 6: The information about the seed yield of pegion pee from 100 plots is given below (minimum is 55 g). Prepare a suitable frequency distribution from the present information of monthly income of the formers.

Seed yield of sesamum	No of Plots
More than 55 g	100
More than 75 g	97
More than 95 g	92
More than 115 g	85
More than 135 g	65
More than 155 g	41
More than 175 g	15
More than 195 g	3
More than 215 g	1

Exercise – 2

DIGRAMATIC REPRESENTATION OF DATA

Following diagrams are used to depict a frequency distribution.
(a) Simple Bar Diagram
(b) Sub-divided Bar Diagram
(c) Multiple Bar Diagram
(d) Percentage Bar Diagram
(e) Pie Diagram

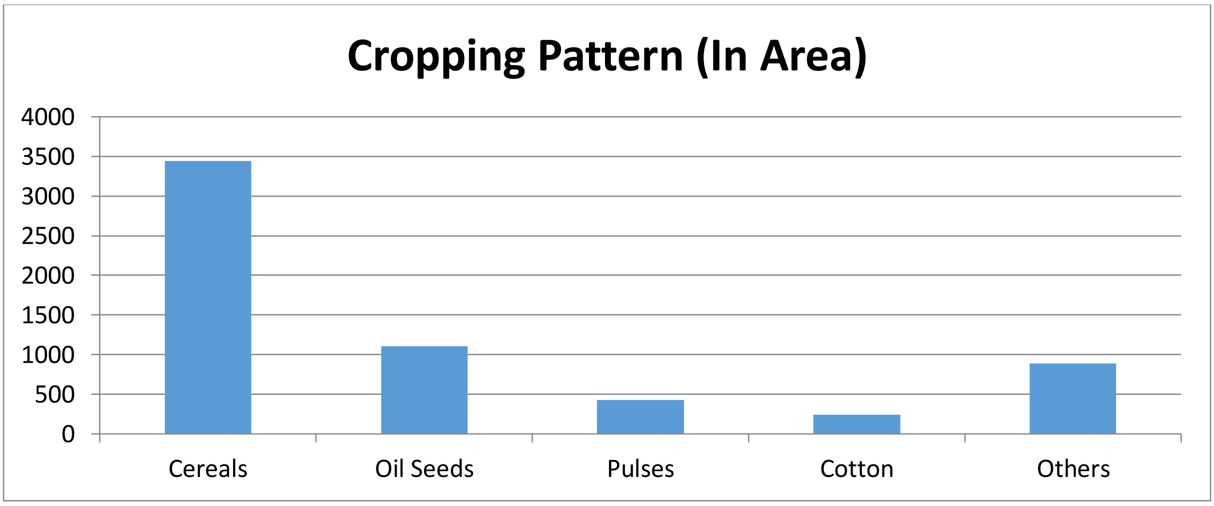
Simple Bar Diagram: If the classification is based on attributes and if attributes are to be compares with respect to the single characteristics, we use simple bar diagram. Simple bar diagram consist of vertical bars of equal width. The heights of the bars are proportional to the volume or magnitude of the attributes. All bars stand on same base line and separated each-others by equal width.

For Example: The cropping pattern in Kerala in the year 2010-2011 was as follows:

Crop	Area (in 1000 hec.)
Cereals	3440
Oil Seeds	1105
Pulses	425
Cotton	245
Others	885
Total	**6100**

Draw a simple bar diagram from the given data.

Solution:

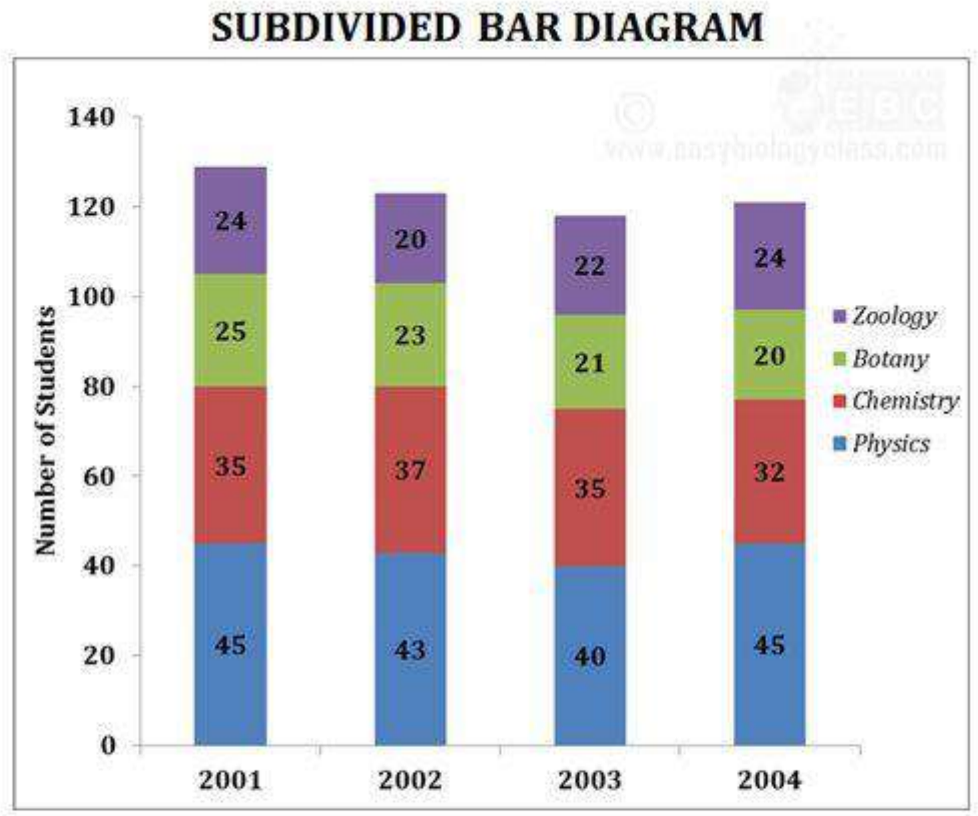


Sub-divided Bars Diagrams: Sometimes multi-character data may possess additive features; in this case the bars are sub-divided in to different parts whose heights are proportional to the volume or magnitude of the different sub-divisions of the attributes. Here, also all bars stand on same base line and separated each-others by equal width.

For Example: The following table gives the M.Sc. students enrolled in different subjects of a college. Draw a sub-divided bar diagram for the given data.

Sujects	Years			
	2001	**2002**	**2003**	**2004**
Physics	45	43	40	45
Chemistry	35	37	35	32
Botany	25	23	21	20
Zoology	24	20	22	24

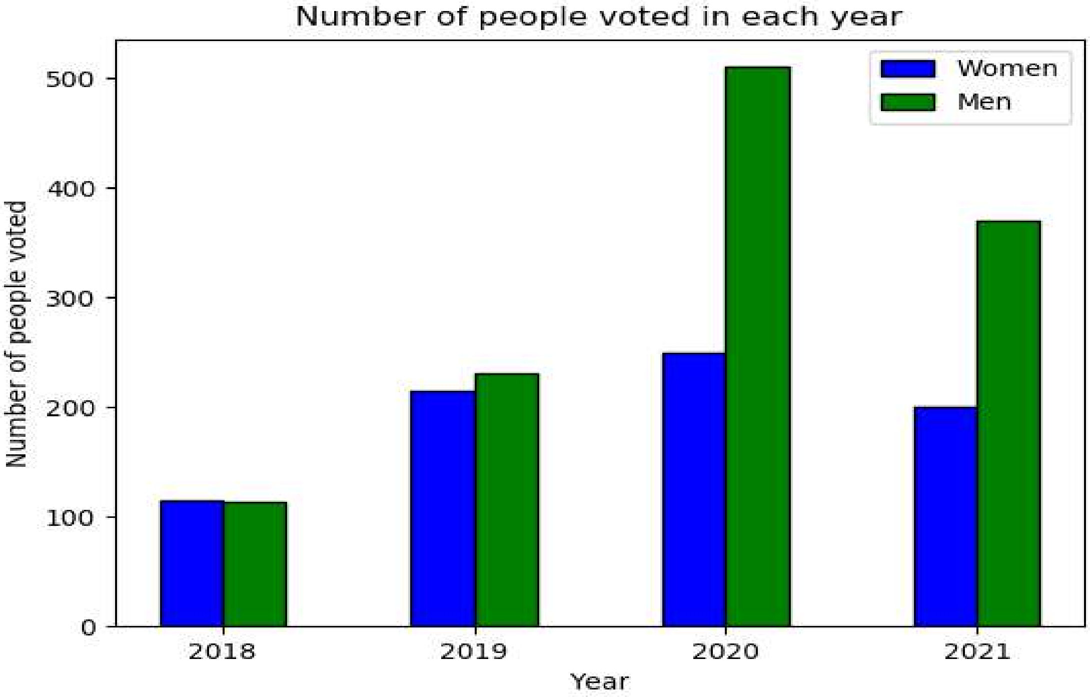
Solution:



Multiple Bar Diagram: If the data is classified by attributes and if two or more characters or groups are to be compared within each attribute, we use multiple bar diagram. Multiple bar diagram is simply extension of simple bar diagram. The bars which are to be compared (representing separate characteristic or group) are drown side by side. Each bar within an attribute will be marked or coloured differently in order to distinguish them. All bars stand on same base line.

For Example: The following table gives the numbers of people voted in each year is given below. Draw a multiple bar diagram for the given data.

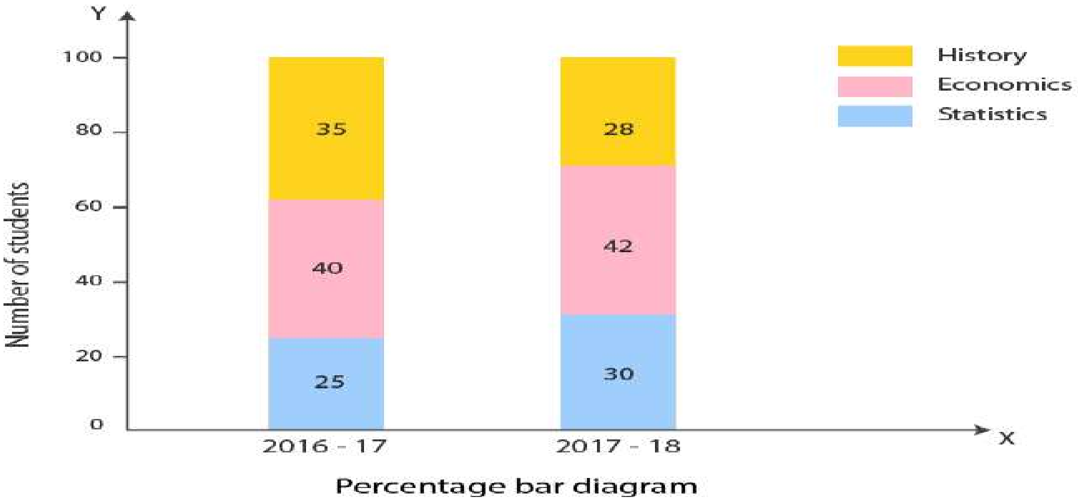
Voter	Voted in year			
	2018	**2019**	**2020**	**2021**
Women	100	210	230	200
Men	100	220	500	370



Percentage Bar Diagram: Sometimes the volumes of different attributes may be greatly different. For making meaningful comparisons, the components of the attributes are reduced to percentage. In that case each attribute will have 100 as its maximum volume. All bars stand on same base line and separated each-others by equal width with equal heights. Here we use a formula $P_i = \frac{v_i}{v} x100$ to calculate percentage of each attribute.

For Example: The following table gives the M.Sc. students enrolled in different subjects of a college in two sessions. Draw a sub-divided bar diagram for the given data.

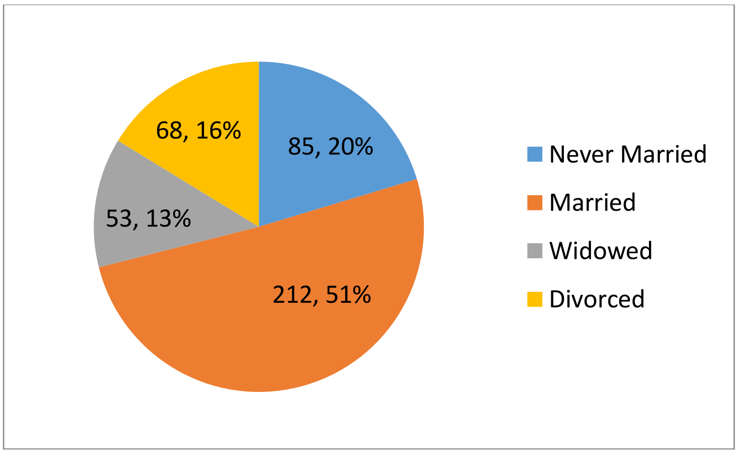
Session	**Subject**		
	Statistics	**Economics**	**History**
2016-17	20	32	28
2017-18	30	42	28



Pie Diagram or Pie Chart: Pie chart is a circular diagram. The circle is divided in to segments which are in proportion to the size of the component. These segments are marked or coloured differently in order to make this diagram attractive. Here we use a formula $A_i = \frac{v_i}{v} x 360$ to calculate angle of each segment.

For Example: Create the pie chart for the following data related to marital status of women.

Marital Status	**Women (in Lacs)**
Never Married	85
Married	212
Widowed	53
Divorced	68
Total	418



Problem 1: The cropping pattern in Karnataka in the year 2019-20 was as follows:

Crop	**Area (in 100 hec.)**
Cereals	2445
Oil Seeds	1005
Pulses	317
Cotton	183
Others	650
Total	**4600**

Problem 2: The following table gives the expenditure on the different segments of two families A and B. Draw a sub-divided bar diagram for the given data.

Items of Expenditure	**Family A**	**Family B**
	Expenditure in Rs.	**Expenditure in Rs.**
Fooding	140	240
Clothing	80	160
House rent	100	120
Education	30	80
Fual	40	40
Miscellaneous	40	80
Saving	70	80
Total	**500**	**800**

Problem 3: The distribution of bacterial population (10^5/g-soil) in the rhizosphere of certain ornamental plants in different physiological stage is given below:

Rhizosphere	Physiological Stages			Total
	Vegetable Stage	Flowering Stage	Seed Setting Stage	
Cock’s Comb	160	130	30	320
Salvia	144	60	56	260
Zinnia	95	40	35	170
Aster	60	75	10	145

Draw a sub-divided bar diagram for the given data.

Problem 4: The following table gives the expenditure on the different segments of three families A, B and C. Draw a multiple bar diagram for the given data.

Items of Expenditure	**Family A**	**Family B**	**Family B**
	Expenditure in Rs.	**Expenditure in Rs.**	**Expenditure in Rs.**
Fooding	50	45	60
Clothing	20	25	20
House rent	10	10	10
Education	5	10	5
Miscellaneous	15	10	5
Total	100	100	100

Problem 5: The distribution of bacterial population (10^5/g-soil) in the rhizosphere of certain ornamental plants in different physiological stage is given below:

Rhizosphere	Physiological Stages			Total
	Vegetable Stage	Flowering Stage	Seed Setting Stage	
Cock's Comb	167	130	33	330
Salvia	144	60	53	257
Zinnia	97	33	34	164
Aster	56	78	10	144

Draw a percentage bar diagram for the given data.

Problem 6: The cropping pattern in Kerala in the year 2005-2006 was as follows:

Crop	**Area (in 1000 hec.)**
Cereals	3950
Oil Seeds	1155
Pulses	464
Cotton	239
Others	892
Total	**6700**

Create the pie chart for the given data.

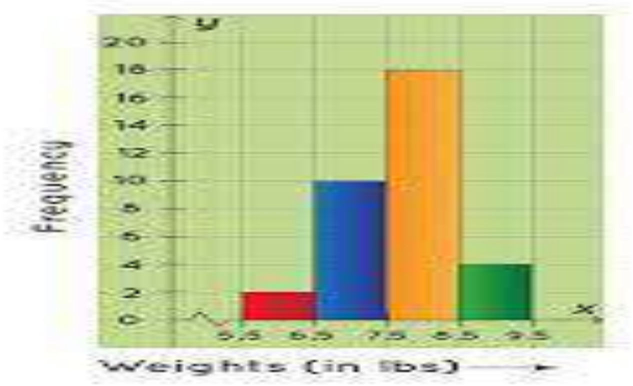
Exercise – 3

GRAPHICAL REPRESENTATION OF DATA

Following graphs are used to depict a frequency distribution.
(a) Histogram
(b) Frequency polygon
(c) Frequency curve
(d) Less than type cumulative frequency curve (ogive)
(e) More than type cumulative frequency curve (ogive).

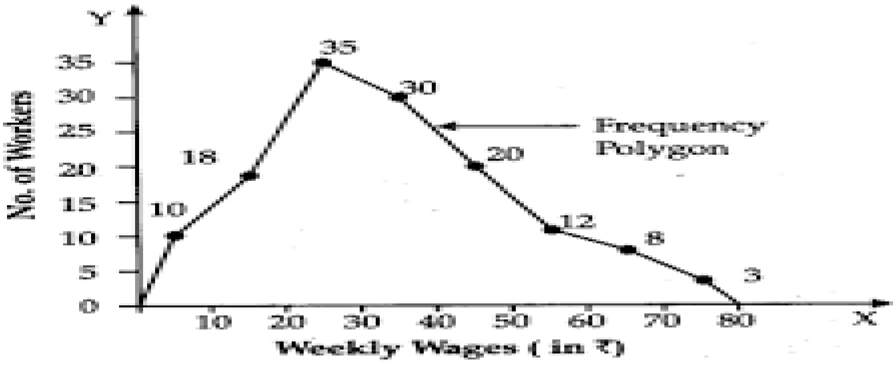
Histogram: When the data are classified based on the class interval, it can be represented by a histogram. Histogram is just like a simple bar diagram with only difference, there is no gap between the bars since the classes are continuous. To draw histogram takes class interval on X-axis and frequency on Y-axis. Draw rectangle with class-interval as base and frequency as height. When a class is in the form of single value the base will be of width unity with the value of the class at the centre.
For example: From the following data draw a histogram.

Weight (lbs)	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5
F	02	10	16	4



Frequency polygon: The type of data required for frequency polygon is same as for histogram. To draw frequency polygon takes mid value of each class interval on X-axis and corresponding frequencies on Y-axis. Plot the points and join them successively by straight line. First point is joined to lower limit of class one and last point is joined to upper limit of last class. When two or more frequency distributions are to be compared, frequency polygon is more useful.
For Example: From the following data, draw a frequency polygon.

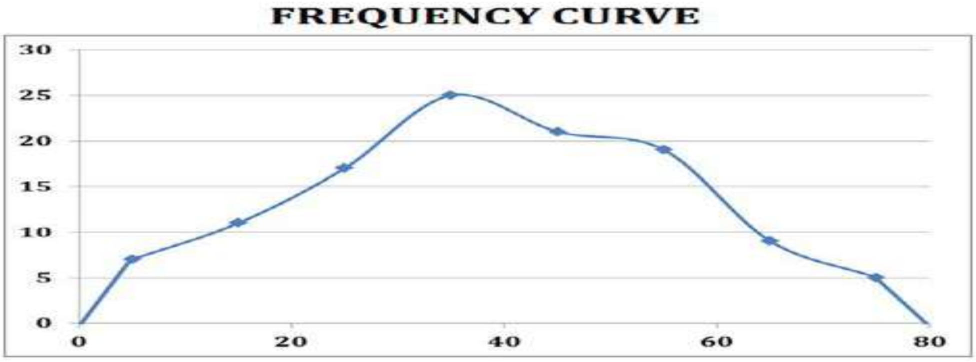
Weekly Wages (in R)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Workers	10	18	35	30	20	12	8	3



Frequency curve: The procedure and data for drawing a frequency curve is same as for frequency polygon but the points are joined by smooth and free hand curve. To draw frequency curve take mid value of each class interval on X-axis and corresponding frequencies on Y-axis. Plot the points and draw a freehand smooth curve such that points are near the curve as far as possible.

For example: From the following data draw a histogram.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Students	7	11	17	25	21	19	9	5



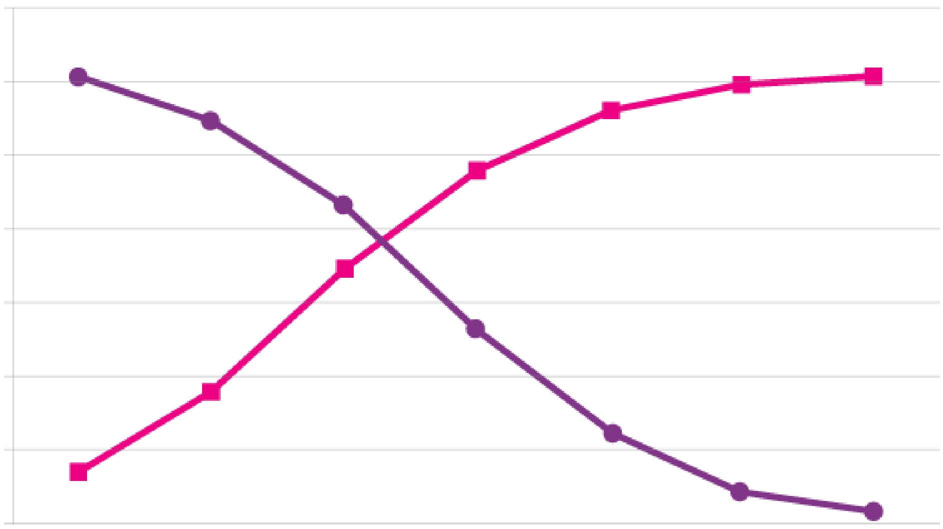
Cumulative frequency curve (ogive): It is a graph plotted for the variate values and their corresponding cumulative frequencies of a frequency distribution. Its shape is just like elongated S. This curve is prepared either for more than type or less than type.

Less than type cumulative frequency curve: To draw this curve takes upper limit on X-axis and less than type corresponding cumulative frequency (top to bottom) on Y-axis. Plot the points and draw a freehand smooth curve.

More than type cumulative frequency curve: To draw this curve takes lower limit on X-axis and more than type corresponding cumulative frequency (bottom to top) on Y-axis. Plot the points and draw a freehand smooth curve.

Note that if the both the ogives (mote than type and less than type) are drawn on same graph the two will intersect each other at a point. The intersection point corresponding to X coordinate would be median and corresponding to Y coordinate will show n / 2, where median is that value which divides total frequency into two equal parts and n is total frequency.

Note that when a class is in the form of a single value that value itself will be taken on X-axis to draw both the ogives.



© Byjus.com

(Exercise)

Problem 1: From the following data draw a histogram.

C.I.	0-10	10-20	20-30	30-40	40-50
F	18	32	45	28	17

Problem 2: From the following data, draw a histogram.

C.I.	1-10	11-20	21-30	31-40	41-50	51-60	61-70
F	5	14	17	23	16	10	5

Problem 3: From the following data, draw a frequency polygon.

C.I.	0-10	10-20	20-30	30-40	40-50
F	8	12	15	8	7

Problem 4: The rainfall distribution in a city, for the period 1996-2005 is given below:

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Rainfall(mm)	558	676	1035	713	1126	480	553	596	754	780

Draw a frequency polygon for the given data.

Problem 5: Frequency distribution of seed yield of 100 sesamum plants is given below:

Seed yields	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5	9.5-10.5
No. of Plants	4	6	10	26	24	15	10	5

Draw a frequency curve for the given data.

Problem 6: From the following data, draw frequency curve.

C.I.	0-5	5-10	10-15	15-20	20-25	25-30	30-35
F	8	12	20	30	18	12	5

Problem 7: From the following, draw ogive.

C.I.	0-10	10-20	20-30	30-40	40-50
F	8	12	15	8	7

Problem 8: From the following, draw ogive.

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70
F	3	7	13	17	12	6	2

Exercise - 4
MEASURES OF CENTRAL TENDENCY

One basic characteristic of a set of observations is their tendency to cluster around a value. Measurement of this tendency is called as measure of central tendency. This value is considered to be typical value and is taken as representative value of the data. Commonly used measures of central tendency are (a) arithmetic mean (b) median and (c) mode. But there are some other measures of central tendency known as geometric mean, harmonic mean and partition values.

Arithmetic mean (A.M.): It is obtained by dividing sum of all observations by total number of observations. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data, then

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

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N = \sum_{i=1}^{n} f_i$, then the arithmetic mean for such data is given as

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{f_1 + f_2 + ...f_n} = \frac{\sum_{i=1}^{n} f_i\, x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i\, x_i}{N}$$

Sometimes to avoid big calculations we can also use Step Deviation Method to find Arithmetic Mean, which is given as

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i\, d_i}{\sum_{i=1}^{n} f_i} \times h \qquad = A + \frac{\sum_{i=1}^{n} f_i\, d_i}{N} \times h$$

Geometric Mean (GM): It is defined as the positive nth root of the product of all n observations. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data, then

$$G = \sqrt[n]{x_1\, x_2 ....x_n} = (x_1\, x_2 ....x_n)^{1/n} \quad \text{Or} \quad \log G = \frac{1}{n} \log\, (x_1\, x_2 ....x_n)$$

$$G = \; anti\log \; [\frac{1}{n} \sum_{i=1}^{n} \log x_i]$$

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N = \sum_{i=1}^{n} f_i$, then the geometic mean for such data is given as

$$G = \; anti\log \; [\frac{1}{N} \sum_{i=1}^{n} f_i \log x_i]$$

Harmonic Mean (HM): It is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observations. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data, then

$$H = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N = \sum_{i=1}^{n} f_i$, then the harmonic mean for such data is given as

$$H = \frac{N}{\sum_{i=1}^{n}\left(\frac{f_i}{x_i}\right)}$$

Median (Me): It is value of the middle most observation when observations are arranged in increasing or decreasing order of magnitude.

Thus in case of ungrouped data, arrange the n observations in increasing or decreasing order. If n is odd then median is the value of the unit at (n+1)/2th position. If n is even then median is arithmetic mean of values of units at (n/2)th and [(n/2)+1]th positions.

For continuous data, the following formula is used to determine median…

Median (Me) = $L_1 + \frac{(N/2) - C}{f} \times h$

Where L_1 = Lower limit of median class, h = Width of median class
f = Frequency of median class, C = Cumulative frequency of pre median class
N = Total number of frequencies

Mode (Mo): It is that value which occurs maximum number of times in a data set. Thus, in ungrouped data determine the observation that has maximum frequency. This observation is the mode.

For continuous data, the following formula is used to determine mode…

Mode (Mo) = $L_1 + \frac{f - f_1}{2f - f_1 - f_2} \times h$

Where L_1 = Lower limit of model class, h = Width of model class
f = Frequency of model class, f_1 = Frequency of pre model class
f_2= Frequency of post model class

Partition Values: When we require decomposing a series in to more parts of equal size, the dividing places are known as partition values. These values are quartile, decile, percentile etc.

Quartile (Q_i): Three values which divide the whole series in to four equal parts after arranging them in ascending or descending order of magnitude are known as quartiles. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data and the data is arranged in some order of magnitude. Then

Q_i = Value of i(n+1)/4 th position in the series

For continuous data, the following formula is used to determine quartiles…

Quartile (Q$_i$) = $L_1 + \dfrac{\left(\dfrac{i \times N}{4}\right) - C}{f} \times h$ (i=1,2,3)

Where L$_1$ = Lower limit of ith quartile class, h = Width of ith quartile class
f = Frequency of ith quartile class, C = Cumulative frequency of pre ith quartile class
N = Total number of frequencies

Decile (D$_i$): Nine values which divide the whole series in to ten equal parts after arranging them in ascending or descending order of magnitude are known as deciles. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data and the data is arranged in some order of magnitude. Then
D$_i$ = Value of i(n+1)/10 th position in the series

For continuous data, the following formula is used to determine deciles…

Decile (D$_i$) = $L_1 + \dfrac{\left(\dfrac{i \times N}{10}\right) - C}{f} \times h$ (i=1,2,3,……,9)

Where L$_1$ = Lower limit of ith decile class, h = Width of ith decile class
f = Frequency of ith decile class, C = Cumulative frequency of pre ith decile class
N = Total number of frequencies

Percentile (P$_i$): Ninety nine values which divide the whole series in to hundred equal parts after arranging them in ascending or descending order of magnitude are known as percentiles. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data and the data is arranged in some order of magnitude. Then
P$_i$ = Value of i(n+1)/100 th position in the series

For continuous data, the following formula is used to determine percentile…

Percentile (P$_i$) = $L_1 + \dfrac{\left(\dfrac{i \times N}{100}\right) - C}{f} \times h$ (i=1,2,3,……,99)

Where L$_1$ = Lower limit of ith percentile class, h = Width of ith percentile class
C = Cumulative frequency of pre ith percentile class, f = Frequency of ith percentile class,
N = Total number of frequencies

For Example: The plant height of 11 plants of lentil at BUAT form is recorded as 33.4, 35.0, 34.6, 37.2, 36.8, 34.8, 36.2, 37.0, 36.7, 34.3, 34.0, calculate all the measures of central tendency.

$$AM = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{390}{11} = 35.45$$

$$GM = anti\log\left[\frac{1}{n}\sum_{i=1}^{n}\log x_i\right] = anti\log\left[\frac{17.04321}{11}\right] = anti\log(1.549383) = 35.43$$

$$HM = \frac{n}{\sum_{i=1}^{n}(1/x_i)} = \frac{11}{0.31067} = 35.40$$

Median = 35

For Example: The following is the distribution of body weights of 100 calves at the Ist lactation

Body weight (kg)	30-40	40-50	50-60	60-70	70-80
Calves	10	26	36	22	6

Find Mean, Median, Mode, GM., H.M., Q_1, and Q_3 of body weight of calves.

CI	**Mid Value**	**Frequency**	**fx**	**log x**	**f logx**	**f/x**	**cf**
30-40	35	10	350	1.544068	15.44068	0.285714	10
40-50	45	26	1170	1.653213	42.98353	0.577778	36
50-60	55	36	1980	1.740363	62.65306	0.654545	72
60-70	65	22	1430	1.812913	39.88409	0.338462	94
70-80	75	6	450	1.875061	11.25037	0.08000	100
Total			**5380**		**172.2117**	**1.936499**	

$$AM = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{5380}{100} = 53.80$$

$$GM = anti\log\left[\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right] = anti\log\left[\frac{172.2117}{100}\right] = anti\log(1.722117) = 52.73$$

$$HM = \frac{N}{\sum_{i=1}^{n}(f_i / x_i)} = \frac{100}{1.9364} = 51.63$$

$$Median = L_1 + \frac{(N/2) - C}{f} \times h = 50 + \frac{50-36}{36} \times 10 = 53.88$$

$$Mode = L_1 + \frac{f - f_1}{2f - f_1 - f_2} \times h = 50 + \frac{36-26}{72-26-22} \times 10 = 54.16$$

$$Q_1 = L_1 + \frac{\left(\frac{N}{4}\right) - C}{f} \times h = 40 + \frac{25-10}{26} \times 10 = 45.76$$

$$Q_3 = L_1 + \frac{\left(\frac{3N}{4}\right) - C}{f} \times h = 60 + \frac{75-72}{22} \times 10 = 61.36$$

(Exercise)

Problem 1: Daily expenditure on eggs, bread and butter by a family during a week was recorded as 32, 46, 16, 25, 29, 25 and 37. Calculate average daily expenditure using all the measures of central tendency.

Problem 2: The price (in Rs) of a fish product at 15 retailers was reported to be 44, 38, 40, 42, 36, 41, 39, 38, 41, 38, 40, 37, 38, 47 and 41. Calculate average price of the product and various other measures of central tendency.

Problem 3: The yields of paddy from 75 fields are given below:

Yields(ton)	4.8	5.0	5.2	5.4	5.6	6.0	6.2	6.4
No. of fields	4	6	10	15	20	12	10	8

Calculate various measures of central tendency.

Problem 4: The frequency distribution of weights of 200 sorghum ear heads is given below:

Weight	40-60	60-80	80-100	100-120	120-140	140-160	160-180	180-200
No. of ear heads	16	28	35	55	30	15	12	9

Calculate mean, median and mode for the given distribution.

Problem 5: Wages (in Rs) paid to workers of an organization are given below. Calculate arithmetic mean, geometric mean and harmonic mean from the given data.

Wages (in 100 Rs)	40-60	60-80	80-100	100-120	120-140	140-160	160-180
No. of workers	5	14	27	43	16	10	5

Problem 6: Determine first quartile, fourth decile and sixty sixth percentile from the given data.

4	5	7	6	8	10	18	21	15

Problem 7: From the given data find first and third quartiles, 46 th percentile, 7 th decile.

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70
F	7	13	20	30	12	6	2

Exercise - 5
MEASURES OF DISPERSION

Any measure of central tendency has its own limitation and they are only a representative of a frequency distribution but fail to give a complete picture of the distribution. They do not give any idea of dispersion or scatter ness of observations within distribution. Scatter ness or variation of observations from their average is known dispersion. The commonly used measures of dispersion are (a) Range (b) Quartile deviation (c) Mean deviation (d) Standard deviation and (e) Coefficient of variation.

Range: It is the difference between the highest and smallest observations of a data set. It is simple to calculate but is not based on all observations.

Coefficient of Range: The relative measure of range, called Coefficient of range is defined as

$$Coefficient\,of\;Range\;(\%)=\frac{X_{\max}-X_{\min}}{X_{\max}+X_{\min}}\times 100$$

Quartile Deviation: It is half of the difference between the upper quartile (Q_3) and lower quartile (Q_1).

$$QD=\frac{Q_3-Q_1}{2}$$

Coefficient of Quartile Deviation: The relative measure of quartile deviation is known as Coefficient of quartile deviation and it is defined as

$$Coefficient\,of\;QD\;\;(\%)=\frac{Q_3-Q_1}{Q_3+Q_1}\times 100$$

Mean deviation: Mean deviation from any average is defined as arithmetic mean of absolute deviations of the observations from that average. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data then

$$MD(\bar{x})=\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}$$

$$MD(Me)=\frac{\sum_{i=1}^{n}|x_i-Me|}{n}$$

$$MD(Mo)=\frac{\sum_{i=1}^{n}|x_i-Mo|}{n}$$

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N=\sum_{i=1}^{n} f_i$, then mean deviation about mean for such data is given as

$$MD(\bar{x})=\frac{\sum_{i=1}^{n} f_i|x_i-\bar{x}|}{N}$$

Mean deviation about median for such data is given as

$$MD(Me) = \frac{\sum_{i=1}^{n} f_i |x_i - Me|}{N}$$

Mean deviation about mode for such data is given as

$$MD(Mo) = \frac{\sum_{i=1}^{n} f_i |x_i - Mo|}{N}$$

Coefficient of Mean Deviation: The relative measure of mean deviation is known as Coefficient of mean deviation. Coefficient of mean deviation about mean, median and mode are defined as

$$Coefficient\,of\ MD\,(\bar{x})\ (\%) = \frac{MD(\bar{x})}{\bar{x}} \times 100$$

$$Coefficient\,of\ MD\,(Me)\ (\%) = \frac{MD(Me)}{Me} \times 100$$

$$Coefficient\,of\ MD\,(Mo)\ (\%) = \frac{MD(Mo)}{Mo} \times 100$$

Standard Deviation: It is defined as the positive square root of average of the squares of deviations of the values of the variable from their respective mean. Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data then

$$SD\ or\ \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \quad = \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^{\ 2} - (\bar{x})^2}$$

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N = \sum_{i=1}^{n} f_i$, then standard deviation for such data is given as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i\ (x_i - \bar{x})^2}{N}} \quad = \quad \sqrt{\frac{1}{N}\sum f_i\ x_i^{\ 2} - (\bar{x})^2}$$

Variance: Square of SD is known as variance.

Coefficient of Variation: Coefficient of variation is a relative measure of dispersion. It is denoted by CV and defined as

$$Coefficient\,of\ \text{var}\,iation\,(\%) = \frac{SD}{\bar{x}} x100 \quad = \frac{\sigma}{\bar{x}} x100$$

For Example: Monthly consumption of cooking oil (in kg) by 200 families is given below. Calculate range, mean deviation from mean, mean deviation from median, mean deviation from mode, standard deviation and coefficient of variation.

Oil (Kg)	2	4	6	8	10	12	14	Total
No. of Families	11	21	42	61	37	20	8	200

Solution

x	**f**	**CF**	**fx**	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$	$f_i \lvert x_i - \bar{x} \rvert$	$\lvert x_i - M_e \rvert$	$f_i \lvert x_i - M_e \rvert$	$\lvert x_i - M_o \rvert$	$f_i \lvert x_i - M_o \rvert$
2	11	11	22	-5.84	34.1056	375.1616	64.24	-6	66	-6	66
4	21	32	84	-3.84	14.7456	309.6576	80.64	-4	84	-4	84
6	42	74	252	-1.84	3.3856	142.1952	77.28	-2	84	-2	84
8	61	135	488	0.16	0.0256	1.5616	9.76	0	0	0	0
10	37	172	370	2.16	4.6656	172.6272	79.92	2	74	2	74
12	20	192	240	4.16	17.3056	346.112	83.2	4	80	4	80
14	8	200	112	6.16	37.9456	303.5648	49.28	6	48	6	48
total	200		1568			1650.88	444.32		436		436

Range= 14-2= 12

$$AM = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{1568}{200} = 7.84$$

Median= 8

Mode = 8

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\,(x_i - \bar{x})^2} = \sqrt{\frac{1650.88}{200}} = \sqrt{8.2544} = 2.873$$

$$MD(\bar{x}) = \frac{\sum_{i=1}^{n} f_i \lvert x_i - \bar{x} \rvert}{N} = \frac{444.32}{200} = 2.221$$

$$MD(Median) = \frac{\sum_{i=1}^{n} f_i \lvert x_i - M_e \rvert}{N} = \frac{436}{200} = 2.18$$

$$MD(Mode) = \frac{\sum_{i=1}^{n} f_i \lvert x_i - M_o \rvert}{N} = \frac{436}{200} = 2.18$$

$$CV\,(\%) = \frac{\sigma}{\bar{x}} x100 = \frac{2.873}{7.84} \times 100 = 36.64\%$$

$$Q_1 = \left[\frac{N+1}{4}\right]^{th} observation\ value = 6$$

$$Q_3 = 3\left[\frac{N+1}{4}\right]^{th} observation\ value = 10$$

$$QD = \left[\frac{Q_3 - Q_1}{2}\right] = 2$$

(Exercise)

Problem 1: Frequency distribution of heights of 100 plants of sunflower is given below:

Height (cm)	75	80	85	90	95	100	105
No. of Plants	5	8	13	26	24	14	10

From the following data find quartile deviation and coefficient of quartile deviation.

Problem 2: Frequency distribution of seed yield of 100 sesamum plants is given below:

Seed Yield (g)	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5
No. of Plants	5	8	13	26	24	16	8

From the following data find quartile deviation and coefficient of quartile deviation.

Problem 3: Frequency distribution of seed yield of 100 sesamum plants is given below:

Seed Yield (g)	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5	7.5-8.5	8.5-9.5
No. of Plants	5	8	13	26	22	16	10

From the following data find mean deviation from mean, mean deviation from median, mean deviation from mode, standard deviation and coefficient of variation.

Problem 4: Frequency distribution of height of 100 plants is given below:

Height (cm)	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No. of Plants	8	12	22	30	28	15	5

From the following data find mean deviation from mean, mean deviation from median, mean deviation from mode, standard deviation and coefficient of variation.

Exercise – 6

MOMENT, SKEWNESS AND KURTOSIS

Moment: Moment are just like a statistical tools which are useful to define all the charecteristics of a frequency distribution. The idea of moment comes from Mechanics, where moment is define as fxd (f= force and d = perpendicular distance).

Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data, then the rth moment about any arbitrary point "a", which is denoted by μ_r' (a) and defined as

$$\mu_r^{'}(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r \qquad ;(r = 1,2,3,4)$$

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N = \sum_{i=1}^{n} f_i$, then the rth moment about any arbitrary point "a" is $\mu_r^{'}(a) = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - a)^r \quad ; (r = 1,2,3,4)$

Moment about origin (Raw Moment): Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data, then the rth moment about origin, which is denoted by μ_r' and defined as

$$\mu_r^{'} = \frac{1}{n}\sum_{i=1}^{n}(x_i)^r \qquad ; (r = 1,2,3,4)$$

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N = \sum_{i=1}^{n} f_i$, then the rth moment about origin is $\mu_r^{'} = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)^r \quad ; (r = 1,2,3,4)$

Moment about mean (Central Moment): Let $x_1, x_2, ...x_n$ be the values of n observations of a series or data, then the rth moment about mean, which is denoted by μ_r and defined as

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r \qquad ; (r = 1,2,3,4)$$

Let there be k classes or groups whose values are $x_1, x_2, ...x_n$ and frequencies are $f_1, f_2, ...f_n$ respectively and the total number of frequencies is $N = \sum_{i=1}^{n} f_i$, then the rth moment about mean is $\mu_r = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^r \quad ; (r = 1,2,3,4)$

Relation between central moment and raw moment:

$$\mu_1 = 0$$

$$\mu_2 = \mu_2^{'} - (\mu_1^{'})^2$$

$$\mu_3 = \mu_3^{'} - 3\mu_2^{'}\mu_1^{'} + 2\left(\mu_1^{'}\right)^3$$

$$\mu_4 = \mu_4^{'} - 4\mu_3^{'}\mu_1^{'} + 6\mu_2^{'}\left(\mu_1^{'}\right)^2 - 3\left(\mu_1^{'}\right)^4$$

Sheppard Correction: W.F. Sheppard pointed out that in case of continuous frequency distributions, at the time of calculating moments,It is presumed that the frequencies are centered at the mid points of the class intervals. Such a presumption introduced some error in the calculation of moments. Hence, he suggested some correction in various moments. These corrections are known as Sheppard Correction. They are as follows:

$$Corrected\ \mu_2 = \mu_2 - \frac{h^2}{12}$$

$$Corrected\ \mu_3 = \mu_3$$

$$Corrected\ \mu_4 = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4$$

Where h is the class interval.

Skewness: By skewness of data, we mean the departure of symmetry. A frequency distribution of a set of values or data, which is not symmetrical, is called asymmetrical or skewed. In a skewed distribution, extreme values in a data set move towards one side or tail of a distribution. If the longer tails of the frequency curve, of the variable moves towards the higher values of the variable, this property is known as positive skewness and if the longer tails of the frequency curve of the variable moves towards the lower values of the variable, this property is known as negative skewness.

The measures of skewness are called coefficients of skewness. Some useful coefficients are:

Karl Pearson's measure of skewness (I) = $\frac{Mean - Mode}{S\,D}$

Karl Pearson's measure of skewness (II) = $\frac{3(Mean - Median)}{S\,D}$

Bowley's measure of skewness = $\frac{Q_3 + Q_1 - 2\,Median}{Q_3 - Q_1}$

Kelly's measure of skewness = $\frac{P_{90} + P_{10} - 2\,Median}{P_{90} - P_{10}} = \frac{D_9 + D_1 - 2\,Median}{D_9 - D_1}$

Measure of skewness based on moments:

$$\beta_1 = \frac{\mu_3^{\ 2}}{\mu_2^{\ 3}} \quad or \quad \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{\ 3/2}} = \frac{\mu_3}{S\,D^{\ 3}}$$

$$If \quad \gamma_1 = 0 \qquad Symmetric$$

$$If \quad \gamma_1 > 0 \qquad Positive\,Skewness$$

$$If \quad \gamma_1 < 0 \qquad Negative\,Skewness$$

Kurtosis: By kurtosis we mean the flatness or peakedness of the frequency curve of a data set. If the peakedness of the curve is large then the height of the curve from the X axis is large and the concentration of the values around their measure of central tendency is large. It is measured by the Coefficients β_2 or γ_2 . A curve having a standard height from X axis is called mesokurtic curve. A curve having more height than mesokurtic curve from X axis is called leptokurtic and a curve having lower height than mesokurtic curve from X axis is called platykurtic curve.

Measure of kurtosis based on moments:

$$\beta_2 = \frac{\mu_4}{\mu_2^{\ 2}} \qquad or \qquad \gamma_2 = \beta_2 - 3$$

$If \quad \gamma_2 = 0 \qquad Mesokurtic$

$If \quad \gamma_2 > 0 \qquad Leptokurtic$

$If \quad \gamma_2 < 0 \qquad Platykurtic$

For Example: Calculate Karl Pearson's measure of skewness from the following data.

Value	12	14	16	18	20
Freq.	12	28	35	15	10

Solution:

x	f	x²	fx	fx²
12	12	144	144	1728
14	28	196	392	5488
16	35	256	560	8960
18	15	324	270	4860
20	10	400	200	4000
Total	100		1566	25036

$$Mean \quad (\bar{x}) = \frac{\sum_i f_i x_i}{N} = \frac{1566}{100} = 15.66$$

Mode = 16

$$SD \ \ (\sigma) = \sqrt{\frac{1}{N}\sum f_i\, x_i^{\ 2} - (\bar{x})^2} = \sqrt{\frac{25036}{100} - (15.66)^2} \ = 2.263$$

$$KP(I) = \frac{Mean - Mode}{S\,D} = \frac{15.66 - 16.00}{2.263} = -0.15$$

For Example: Compute first four central moment from the given data. Also find β_1 and β_2.

Value	2	4	6	8	10
Freq.	12	28	35	15	10

Solution:

x	f	x^2	x^3	x^4	fx	fx^2	fx^3	fx^4
2	12	4	8	16	24	48	96	192
4	28	16	64	256	112	448	1792	7168
6	35	36	216	1296	210	1260	7560	45360
8	15	64	512	4096	120	960	7680	61440
10	10	100	1000	10000	100	1000	10000	100000
Total	100				566	3716	27128	214160

Raw Moments

$$\mu_1' = \frac{\sum_i f_i x_i}{N} = \frac{566}{100} = 5.66$$

$$\mu_2' = \frac{\sum_i f_i x_i^2}{N} = \frac{3716}{100} = 37.16$$

$$\mu_3' = \frac{\sum_i f_i x_i^3}{N} = \frac{27128}{100} = 271.28$$

$$\mu_4' = \frac{\sum_i f_i x_i^4}{N} = \frac{214160}{100} = 2141.60$$

Central Moments

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \left(\mu_1'\right)^2 = 5.12$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\left(\mu_1'\right)^3 = 2.94$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\left(\mu_1'\right)^2 - 3\left(\mu_1'\right)^4 = 63.63$$

$$\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3} = \frac{(2.93)^2}{(5.12)^3} = 0.06$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2{}^{3/2}} = 0.2593$$

$$\beta_2 = \frac{\mu_4}{\mu_2{}^2} = \frac{63.63}{(5.12)^2} = 2.42$$

$$\gamma_2 = \beta_2 - 3 = -0.58$$

Problem 1: Calculate Karl Pearson's and Bowley measure of skewness from the following data.

Value	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Freq.	6	8	17	21	15	11	2

Solution:

C.I.	**f**	**X**	**Fx**	**cf**

Problem 2: Calculate Kaley's and Bowley measure of skewness from the following data.

Value	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Freq.	2	6	8	11	7	4	2

Solution:

C.I.	**f**	**X**	**Fx**	**cf**

Problem 3: Compute first four central moment from the given data. Also find β_1 and β_2.

Value	2-4	4-6	6-8	8-10	10-12
Freq.	2	4	8	6	1

Exercise – 7

CORRELATION AND REGRESSION ANALYSIS

Many a times we have observations for two or more than two characteristics or variables. The objective, here is to study interrelationship among these variables. For example, we may have data for following variables: yield of a crop and fertilizer input; quantity of a commodity purchased, its price and number of members in the family; yield and protein contents in rice; age and blood pressure of a person. We will like to quantify this association and study some more aspect of this relationship. It is done through correlation and regression analysis.

Correlation is statistical analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. There are two types of correlation (1) Positive and (2) Negative. in positive correlation as value of one variable increases on an average value of other variable also increases and vice versa while in case negative correlation the variables change in opposite direction.

Scatter Diagram: Scatter diagram gives some idea about nature of relationship between two variables; whether there is positive correlation or negative correlation or poor correlation. Let $(x_1, y_1), (x_2, y_2), \ldots\ldots\ldots\ldots (x_n, y_n)$ be n pair of observations of a series or data. In scatter diagram, taking x variable on X-axis and y variable on Y- axis, then n observations are plotted on graph paper. If the points are clustered around a straight line with upward trend, correlation is said to be positive; if the points are clustered around a straight line with downward trend, correlation is said to be negative. If the points are haphazardly scattered without having trend with respect to a straight line correlation is said to be zero. Now we study how to quantify this extent of association. Recall that in case of one variable its variability is measured by variance. In case of two variables their joint variation is measured by a quantity called as covariance. Its formula is given by

$$Cov(x, y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = \frac{1}{n}\sum xy - (\bar{x}\,\bar{y})$$

Karl Pearson correlation coefficient: The correlation coefficient between x and y is defined as a measure of linear relationship between them. It is denoted by ‘r’ and is obtained as

$$r_{x,y} = \frac{Cov(x, y)}{\sqrt{V(x)V(y)}}$$

Where

$$V(x) = \frac{\sum(x-\bar{x})^2}{n} = \frac{1}{n}\sum x^2 - \left(\bar{x}^2\right)$$

$$V(y) = \frac{\sum(y-\bar{y})^2}{n} = \frac{1}{n}\sum y^2 - \left(\bar{y}^2\right)$$

Properties of correlation coefficient:

(a) It is free from units of measurement.

(b) Its value lies between -1 and +1. Value of r= + 1 is interpreted as perfect positive correlation and in this case all the points in the scatter diagram lie on a straight line. Similar is the case when r = - 1. When value of r goes away from +1 or -1 the spread in the scatter diagram increases and maximum spread occurs when r =0.

Probable Error: Probable error is an important measure to determine the limits of correlation coefficient and to assess the reliability of the value of coefficient. It can be defined as:

$$P.E.=\frac{0.6745(1-r^2)}{\sqrt{N}}$$

The limits of the correlation coefficient are defined as: $r \pm P.E.$

Spearman's Rank Correlation Coefficient: The Pearson correlation coefficient between the ranks of X and Y is called the rank correlation coefficient between the characteristics A and B for the group of individuals. It is known as Spearman's rank correlation coefficient. Spearman's rank correlation coefficient, usually denoted by "r" is given by the equation

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs. In tied observations case

$$r = 1 - \frac{6\left[\sum_{i=1}^{n} d_i^2 + \sum \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

Regression analysis: In regression analysis we express relationship between x and y by an equation called as regression equation. In this equation one variable is treated as dependent variable and the other as independent variable. When y is taken as dependent variable and x as independent variable the equation is called as regression equation of y on x. Since we interested in linear relationship between x and y. The equation is written as y = a + b x. Where a and b are constant to be determined. The constants a and b are called regression parameters "a" is the intercept term and b which is generally denoted by b_{yx} is called regression coefficient of y on x and it reveals the change in the dependent variable, y for unit increase in the independent variable, x. This regression equation is used to predict value of y for a given value of x. The formulae of regression parameters are obtained by using a method of least squares. The formulae are

$$b_{yx}=\frac{Cov(x,y)}{V(x)} \quad ; \quad a = \bar{y} - b_{yx}\,\bar{x}$$

Note that V(x) and V(y) are always positive while Cov(x,y) can be positive or negative. Therefore, when is Cov(x,y) positive, both r and b_{yx} are positive and when Cov(x,y) is negative both are negative.

We can also write the equation of regression line of Y on X as

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

Similarly the equation of regression line of X on Y as

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

Calculations in correlation and regression analysis are based on following steps:

Step 1 – Calculate ∑x, ∑y, $\sum x^2$, $\sum y^2$, and ∑xy.

Step 2 – Calculate V(x), V(y) and Cov (x,y).

Step 3 – Calculate correlation coefficient, r

Step 4 – Calculate regression coefficient of y on x, $b_{y\,x}$ and a

Step 5 – In the equation $y-\bar{y} = b_{yx}\,(x-\bar{x})$ substitute the numerical values of b_{yx} and mean values of x and y. This is the required regression equation of y on x. (Note that calculating regression equation is also called as fitting the regression equation).

Step 6 – In the equation obtained in previous step substitute the given value of x. It will give the corresponding predicted (estimated) value of y.

For Example: Find Karl Pearson's coefficient of correlation from the following data between height of father (x) and son (y) and also find Regression equations.

X	64	65	66	67	68	69	70	71
Y	66	67	65	68	72	70	70	68

Solution:

x	**y**	**x^2**	**y^2**	**xy**
64	66	4096	4356	4224
65	67	4225	4489	4355
66	65	4356	4225	4290
67	68	4489	4624	4556
68	72	4624	5184	4896
69	70	4761	4900	4830
70	70	4900	4900	4900
71	68	5041	4624	4828
540	**546**	**36492**	**37302**	**36879**

$$(\bar{x}) = \frac{\sum_i x_i}{n} = \frac{540}{8} = 67.5$$

$$(\bar{y}) = \frac{\sum_i y_i}{n} = \frac{546}{8} = 68.25$$

$$V(x) = \frac{1}{n}\sum x^2 - \left(\bar{x}^2\right) = \frac{36492}{8} - (67.5)^2 = 5.25$$

$$V(y) = \frac{1}{n}\sum y^2 - \left(\bar{y}^2\right) = \frac{37302}{8} - (68.25)^2 = 4.6875$$

$$Cov(x,y) = \frac{1}{n}\sum xy - \left(\bar{x}\,\bar{y}\right) = \frac{36879}{8} - (67.5)*(68.25) = 3$$

$$r_{x,y} = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{3}{\sqrt{(4.6825)\,(5.25)}} = 0.604$$

$$b_{yx} = \frac{Cov(x,y)}{V(x)} = \frac{3}{5.25} = 0.571$$

$$b_{xy} = \frac{Cov(x,y)}{V(y)} = \frac{3}{4.6875} = 0.64$$

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad \Rightarrow \quad y - 68.25 = 0.571(x - 67.5)$$

$$\Rightarrow \quad y = 0.571x + 29.7075$$

$$x - \bar{x} = b_{xy}(y - \bar{y}) \quad \Rightarrow \quad x - 67.5 = 0.64(y - 68.25)$$

$$\Rightarrow \quad x = 0.64y + 23.82$$

(Exercise)

Problem 1: Protein intake (x) and fat intake (y) (in gm) for ten old women was calculated as given below.

(a) Calculate correlation coefficient

(b) Calculate regression equation of y on x

(c) Estimate fat intake of a woman whose protein intake is 44 gm.

X:	56	47	33	39	42	38	46	47	38	32
Y:	56	83	49	52	65	52	56	48	59	70

Solution:

X	Y	XY	X^2	Y^2

Calculation:

Problem 2: Geographical area (x) and area under paddy cultivation (y) (in hectares) for 15 villages of a tehsil are given below. Calculate correlation coefficient and regression equation of y on x. Estimate area under paddy cultivation of a village whose geographical area is 116 hectares.

X: 103 106 120 120 151 160 155 136 178 196 100
140 160 166 112

Y: 41 33 87 78 81 90 85 70 100 102 35 70
82 85 50

Solution:

X	Y	XY	X^2	Y^2

Calculation:

Problem 3: Calculate correlation coefficient between marks obtained in first pre-final and second pre-final examinations on the basis of the following data collected for a sample of 10 students. Also fit the regression equation of second pre-final examination on first pre-final examination. Estimate marks in second pre-final examination for a student who gets 9 marks in first pre-final examination.

First Pre-final: 12 14 10 10 8 11 10 14 11 12

Second Pre-final: 11 13 12 14 7 14 12 10 9 12

Solution:

X	Y	XY	X^2	Y^2

Calculation:

Problem 4: The following table gives the ages and blood pressure of 9 women.

Age (X) :	56	42	36	47	49	42	60	72	63
Blood Pressure(Y)	147	125	118	128	145	140	155	160	149

(a) Find the correlation coefficient between X and Y.

(b) Determine the least square regression equation of Y on X.

(c) Estimate the blood pressure of a woman whose age is 45 years.

Problem 5: Two judges gave the following ranks to eight competitors in a beauty contest. Examine the relationship between their judgements.

Judge A	4	5	1	2	3	6	7	8
Judge B	8	6	2	3	1	4	5	7

Problem 6: Apply spearman’s Rank difference method and calculate coefficient of correlation between x and y from the data given below. Find the rank correlation coefficients.

X	22	28	31	23	2	9	31	27	22	31	18
Y	18	25	25	37	31	35	31	29	18	20	50.

Exercise -8

SIMPLE RANDOM SAMPLING

Survey is conducted to get some information about a population under study. By population we mean a group of units having some common characteristics in which an investigator is interested. For example, (i) students of a college admitted in a particular year; (ii) all dairy farms of a district. The most common objective in a survey is to know population mean. For example, (i) average monthly hostel expenditure of a student; (ii) average number of workers per dairy farm.

To fulfill this objective we select a part (i.e. a few units) of the population. This part is called as a sample. From sample data calculate sample mean which is taken as an estimator of population mean. Observe that estimator value will, in general be different from population mean and estimator value will change from sample to sample. Thus estimator is subject to error. This error is measured by a quantity called as standard error. There are various sampling methods. Simple random sampling is one of them. If is defined as a method wherein units are selected one by one assigning equal probability of selection at each draw. The selection can be done either with replacement or without replacement.

SRSWOR is a method of selection of *n* units out of the *N* units one by one such that at any stage of selection, anyone of the remaining units have same chance of being selected, i.e., 1/*N*.

SRSWR is a method of selection of *n* units out of the *N* units one by one such that at each stage of selection each unit has equal chance of being selected, i.e., 1/*N*.

Here we consider without replacement sampling. A random number table is used for selection of the sample. The selection method consists of following steps.

Step 1 – Determine the number of digits in N (where N = total number of units in the population). Let it be k.

Step 2 – Choose any k consecutive columns in the random number table.

Step 3 – From the chosen columns write down the numbers one by one. While writing the numbers delete number zero and a number greater than N. As the sample is to be selected without replacement only distinct numbers are to be considered. In this manner, write down n numbers where n denotes the number of units to be selected in the sample. n is also called as sample size. The units of the population bearing these serial numbers constitute the sample.

To summarize, basically in a survey, we (a) select a random sample from the population (b) calculate the estimator and (c) calculate standard error of the estimator. In case of simple random sampling without replacement following formulae are used.

Estimator of population mean $\bar{y} = \frac{1}{n}\sum y_i$

Standard Error of the estimator $\sqrt{\left(\frac{N-n}{Nn}\right)s^2}$

Where $s^2 = \frac{1}{n-1}\left(\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}\right)$ is known as sample variance and y denotes character under study for which estimator is to be obtained.

Here we consider with replacement sampling. A random number table is used for

selection of the sample. The selection method consists of following steps.
Step 1 – Determine the number of digits in N (where N = total number of units in the population). Let it be k.
Step 2 – Choose any k consecutive columns in the random number table.
Step 3 – From the chosen columns write down the numbers one by one. While writing the numbers delete number zero and a number greater than N. As the sample is to be selected with replacement numbers can to be repeat twice or thrice or more. In this manner, write down n numbers where n denotes the number of units to be selected in the sample. n is also called as sample size. The units of the population bearing these serial numbers constitute the sample.
To summarize, basically in a survey, we (a) select a random sample from the population (b) calculate the estimator and (c) calculate standard error of the estimator. In case of simple random sampling without replacement following formulae are used.

Estimator of population mean $\bar{y} = \frac{1}{n}\sum y_i$

Standard Error of the estimator $\sqrt{\frac{s^2}{n}}$

Where $s^2 = \frac{1}{n-1}\left(\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}\right)$ is known as sample variance and y denotes character under study for which estimator is to be obtained.

Advantages of sample survey over complete enumeration:
(i) Sample survey requires less time.
(ii) Sample survey requires less resource.
(iii) The quality of data obtained from the sample survey is better.

Problem 1: Monthly expenditure on vegetables (y in Rs.) for 40 families of a locality is given below. Draw a simple random sample of 5 families and calculate estimate of average monthly expenditure per family in that locality. Also calculate standard error of this estimate.

S.N.	**Y**	**S.N.**	**Y**	**S.N.**	**Y**	**S.N.**	**Y**	**S.N.**	**Y**
1	230	9	295	17	518	25	189	33	264
2	165	10	240	18	482	26	161	34	287
3	150	11	380	19	425	27	184	35	233
4	545	12	540	20	355	28	176	36	158
5	325	13	123	21	178	29	210	37	312
6	310	14	167	22	162	30	225	38	358
7	175	15	194	23	413	31	265	39	192
8	165	16	206	24	177	32	134	40	200

Solution:

Problem 2: Cultivated area (y in hectares) of 60 villages of a tehsil is given below. Draw a simple random sample of 8 villages; calculate estimate of average cultivated area for the tehsil and standard error of this estimate.

S.N.	**Y**	**S.N.**	**Y**	**S.N.**	**Y**	**S.N.**	**Y**	**S.N.**	**Y**

1	140	13	145	25	89	37	76	49	64
2	165	14	125	26	61	38	89	50	87
3	180	15	103	27	84	39	111	51	133
4	124	16	107	28	76	40	104	52	58
5	69	17	93	29	110	41	75	53	112
6	60	18	83	30	105	42	69	54	58
7	155	19	65	31	46	43	61	55	92
8	125	20	84	32	44	44	85	56	100
9	78	21	86	33	87	45	93	57	101
10	98	22	74	34	58	46	91	58	56
11	56	23	69	35	92	47	72	59	53
12	48	24	71	36	100	48	102	60	87

Exercise -9

STRATIFIED RANDOM SAMPLING

An important objective in any estimation problem is to obtain an estimator of a population parameter which can take care of all salient features of the population. If the population is homogeneous with respect to the characteristic under study, then we prefer to use the method of simple random sampling and the sample mean will serve as a good estimator of population mean. If the population is heterogeneous with respect to the characteristic under study, then we use stratified sampling. In stratified sampling, we divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and heterogeneous with respect to the characteristic under study between/among the subpopulations. Such subpopulations are termed as **strata.** And sample is selected by using SRS from each stratum. We are using SRS in each stratum then it is known as stratified random sampling. . In case of stratified random sampling following formulae are used.

Estimator of population mean $\bar{y}_{st} = \frac{1}{N}\sum_i N_i \bar{y}_i$

$$\bar{y} = \frac{1}{n}\sum_i n_i \bar{y}_i$$

Variance of the estimate $\bar{y}_{st}$ $V(\bar{y}_{st}) = \frac{1}{N^2}\sum_i N_i(N_i - n_i)\frac{S_i^2}{n_i}$

Let we have N units of a population. Divide these N units into k strata. Let the i^th stratum have N_i number of units. i= 1,2,3…..k. Note that there are k independent samples drawn through SRS of sizes n_1, n_2,……n_k. So, one can have k estimators of a parameter based on sizes n_1, n_2,…..n_k. Our interest is not to have k different estimators of the parameters but ultimate goal is to have a single estimator. First of all our aim to find sample sizes from each stratum (n_1, n_2,……n_k). The sample size cannot be determined by minimizing the cost and variability simultaneously. The cost function is directly proportional to the sample size whereas variability is inversely proportional to the sample size. Based on different ideas, some allocation procedures are as follows:

Equal Allocation: In equal allocation, sample of equal sizes are drown from each stratum.

i.e. $n_i = \frac{n}{k}$

and $V(\bar{y}_{st})_{eq} = \frac{1}{N^2}\sum_i N_i^2\left(\frac{k}{n} - \frac{1}{N_i}\right)S_i^2$

Proportional Allocation: In proportional allocation, samples are selected according to size of the stratum. This procedure gives a self-weighting sample.

i.e. $n_i = \frac{n}{N}N_i$

$$and \quad V(\bar{y}_{st})_{prop} = \left(\frac{N-n}{Nn}\right)\sum_i\left(\frac{N_i}{N}\right)S_i^2 = \left(\frac{N-n}{Nn}\right)\sum_i W_i S_i^2$$

Optimum Allocation: There are some situation under which an optimum allocation is considered: (1) when the variance of the stratified sample mean is minimized. (2) When the total cost of the survey is fixed.

(1) $$n_i = \frac{nW_i S_i}{\sum (W_i S_i)} = \frac{nN_i S_i}{\sum (N_i S_i)}$$

(2) $$n_i = \frac{nW_i S_i / \sqrt{c_i}}{\sum (W_i S_i / \sqrt{c_i})} = \frac{nN_i S_i / \sqrt{c_i}}{\sum (N_i S_i / \sqrt{c_i})}$$

If $c_i = c$, i.e. if the cost per unit is same in all strata then (2) is converted into (1).

$$V(\bar{y}_{st})_{opt} = \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_i N_i S_i\right)^2 - \sum_i N_i S_i^2\right] = \frac{1}{n}\left(\sum_i W_i S_i\right)^2 - \frac{1}{N}\left(\sum_i W_i S_i^2\right)$$

Problem 1: A sample survey was conducted in a locality containing 600 farmers to assess the requirement of seed paddy. The formers were stratified into three groups. The sample selection was based on proportional allocation. The average and variance were found for the sample. The results are given below:

Group	**No. of Formers**	**No. of Formers sampled**	**Average Requirment**	**Variance**
(i)	**N_i**	**n_i**	$\bar{y}_i$	s_i^2
1	300	30	75	12.5
2	160	16	120	20.4
3	140	14	170	36.2
Total	600	60		

Solution:

$$\bar{y}_{st} = \frac{1}{N}\sum_i N_i \bar{y}_i =$$

$$V(\bar{y}_{st})_{prop} = \left(\frac{N-n}{Nn}\right)\sum_i\left(\frac{N_i}{N}\right)S_i^2 =$$

Problem 2: An arial photograph was taken of forest and it was seen that the tract was divided into three major forest types, pine, bottom-land hardwood and upland hardwood. Ten one acre plots were selected from each stratum of sizes 320, 140 and 320 respectively at random (with equal probability). Volume obtained on these units is given below:

Pine	570	510	600	590	780	480	670	700	560	640
BLH	520	630	810	710	760	580	770	890	860	840
UH	420	540	320	210	180	270	290	260	200	350

Estimate the mean cubic feet volume per acre and also find its standard error.

Solution:

Exsercise-10

Test of Significance-I
(Large Sample Test)

Testing of hypothesis is one of the aspects of statistical inference. In statistical inference we draw conclusions about a population on the basis of a sample selected from it. (Population is a group of units possessing some common characteristics, e.g. all the farms in a region growing a particular crop; farmers taking loan from a nationalized bank; malnourished children in a school. Sample is a part of the population.) In hypothesis testing we make a statement about a population and on the basis of a sample drawn from it decide whether to accept or reject the statement.

Procedure of testing hypothesis:

1. Set up a hypothesis
2. Set up a suitable significance level
3. Set a test criterion according to null hypothesis
4. Making decisions

Statistical hypothesis: A statistical hypothesis is a conjecture about a population parameter. This conjecture may or may not be true.

Null Hypothesis: According to Prof. RA. Fisher, A null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. It is denoted by H_0.
Alternative Hypothesis: Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1. For example, if we want to test the null hypothesis that the population has a specified mean μ_o. i.e. $\mu = \mu_o$, then alternative hypothesis would be

(i) $\mu > \mu_o$ or
(ii) $\mu < \mu_o$ or
(iii) $\mu \neq \mu_o$

Two Types of Error: The main objective in sampling theory is to draw valid inferences about the population parameters on the basis, of the sample results. In the hypothesis-testing situation, there are four possible outcomes. In reality, the null hypothesis may or may not be true, and a decision is made to reject or not reject it on the basis of the result obtained from a sample. The four possible outcomes are shown in Figure. Notice that there are two possibilities for a correct decision and two possibilities for an incorrect decision.

	Ho is True	Ho is False
Reject Ho	Type I Error	Correct Decision
Accept Ho	Correct Decision	Type II Error

In practice we decide to accept or reject the lot after examining a sample from it. As such we are liable to commit the following two types of errors:
Type I Error: Reject Ho when it is true.
Type II Error: Accept Ho when it is wrong, i.e., accept Ho when H_I is true.

Statistical Significance: When we examine a sample drawn from a population under study for testing the null hypothesis, the sample estimate of a certain characteristic is found differ from the population parameter. This difference can be attributed due to two causes:

(1) It is arises due to errors of sampling

(2) It is arises due to fact that the parent population of the sample is differ from the population considered.

If the difference between the sample estimate and the population parameter under the null hypothesis is due to error of sampling, it will have some limiting value. This limiting value (say δ) can be positive or negative. If the observed difference between the sample estimate and the population parameter $|\bar{x} - \mu|$ is less than δ then we can say it is only due to sampling error alone and we accept our null hypothesis. And if this difference$|\bar{x} - \mu| \geq \delta$, then we can say it is not only due to sampling error but also due to fact that the parent population of the sample is differ from the population considered. Hence we reject our null hypothesis.

Critical Region: A region (corresponding to a test statistic) in the sample space S, which amounts to rejection of null hypothesis is termed as critical region or region of rejection. The selection of critical region is based on the probabilities of two types of error. The probability of I type is fixed and the critical region is chosen which minimize type II error.

Level of Significance: The probability 'α' that a random value of the test statistic belongs to the critical region is known as the level of significance. In other word, we can say the level of significance is the maximum probability of committing a type I error or level of significance is the size of the type I error.

Critical Values or Significant Values: The value of test statistic which separates the critical or rejection region and the acceptance region is called the critical value or significant value. It depends upon:

(I) The level of significance used, and

(ii) The alternative hypothesis, whether it is two-tailed or single-tailed.

One Tailed Test: A one-tailed test indicates that the null hypothesis should be rejected when the test value is in the critical region on one side of the sampling distribution. A one-tailed test is either a right tailed test or left-tailed test, depending on the direction of the inequality of the alternative hypothesis.

Two Tailed Test: If the direction of the inequality of the alternative hypothesis is not known, the critical region will be on either end of the sampling distribution, therefore the test is known as two tailed.

Degree of Freedom: Degree of freedom is the numbers of observations that are free to vary after certain restriction have been placed on data. If there are n observations in a sample, for each restriction imposed upon the original observations the number of degrees of freedom is reduced by one.

P-Value: The P-value or calculated probability is the probability of finding the observed or more extreme results when the null hypothesis is true. The definition of extreme is depends upon the condition, how the hypothesis being tested.

Test of Significance for Large Samples: Here we discuss the tests of significance when samples are large (n > 30) and it is based on assumption that they are drawn from a normal population with unknown mean and known variance. In all large sample tests we use the test statistic Z under the null hypothesis, where Z follows a normal distribution with mean 0 and variance 1.

Critical Values of Z-statistic

Level of significance	**Right tailed test**	**Left tailed test**	**Two tailed test**
α = 0.10	Z > 1.28	Z < - 1.28	$\|Z\| > 1.645$
α = 0.05	Z > 1.645	Z < - 1.645	$\|Z\| > 1.96$
α = 0.01	Z > 2.33	Z < - 2.33	$\|Z\| > 2.58$

(i) Testing of significance of single mean for large sample

Let x_1, x_2,……..,x_n be a random sample from a large population or normal population with mean μ and variance σ^2. Samples are drawn independently.

Testing Hypothesis: We want to test a null hypothesis

$$\mathbf{H_0}: \mu = \mu_0$$

Against $\mathbf{H_1}: \mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$

Test Statistic: We use a test statistic

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \sim N(0,1)$$

If variance of population is not known then we use its estimate, then

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \quad \sim N(0,1)$$

We reject our null hypothesis at α% level of significance if		
For **H₁:** $\mu > \mu_0$	$Z > Z_{\alpha}$	Otherwise accept H_0
For **H₁:** $\mu < \mu_0$	$Z < Z_{1-\alpha}$	Otherwise accept H_0
For **H₁:** $\mu \neq \mu_0$	$\boldsymbol{\|Z\| > Z_{\alpha/2}}$	Otherwise accept H_0

(ii) Testing of significance of the difference between two sample means for large samples:

Let $x_{11}, x_{12}, \ldots\ldots\ldots\ldots x_{1n_1}$ and $x_{21}, x_{22}, \ldots\ldots\ldots\ldots x_{2n_2}$ be two independent random samples from large populations or normal populations with mean μ_1, μ_2 and variance σ_1^2, σ_2^2 respectively. Where variances are known.

Testing Hypothesis: We want to test a null hypothesis

$$\mathbf{H_0}: \mu_1 = \mu_2$$

Against $\mathbf{H_1}: \mu_1 > \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 \neq \mu_2$

Test Statistic: We use a test statistic

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

1. If $\sigma_1^{\ 2} = \sigma_2^{\ 2} = \sigma^2 \quad (say)$ both variances are equal. then

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \sim N(0,1)$$

2. When the population variances are unknown and we have no justification to assume that the two variances are same, we can estimate these by corresponding sample variances. then

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^{\ 2}}{n_1} + \frac{S_2^{\ 2}}{n_2}}} \qquad \sim N(0,1)$$

We reject our null hypothesis at α% level of significance if		
For **H₁:** $\mu > \mu_0$	If $Z > Z_{\alpha}$	Otherwise accept H_0
For **H₁:** $\mu < \mu_0$	If $Z < Z_{1-\alpha}$	Otherwise accept H_0
For **H₁:** $\mu \neq \mu_0$	$\boldsymbol{|Z| > Z_{\alpha/2}}$	Otherwise accept H_0

Problem 1: The following data give yield of soybean in quintals per hectare obtained from 42 plots.

16.7 20.3 22.5 23.5 23.8 21.4 15.7 15.3 18.0 19.5 19.9 19.1 18.5 20.5

14.5 17.5 18.6 18.7 17.2 17.5 21.5 15.6 16.0 16.9 16.8 16.1 16.5 22.8

14.0 16.4 16.4 18.0 16.6 16.2 22.2 14.7 17.4 18.0 18.3 17.6 16.7 20.4

On the basis of above data, can it be concluded that the mean yield of soyabean is 20.00 quintals per hectare. Test at 5% level of significance.

Solution:

Testing Hypothesis:

H₀: $\mu = 20.00$

Against **H₁:** $\mu \neq 20.00$

Test Statistic

$$Z = \frac{\bar{x} - 20}{\sqrt{s^2 / n}}$$

Calculation:

Problem 2: A random sample of 900 farms in a certain year gives an average yield 2000 lbs. of Barley per acre with standard deviation of 192 lbs. It is also known that the average yield of barley during the past 30 years was 2200 lbs. per acre. Is the average yield in this year lower than the average yield obtained in previous 30 tears?

Solution:

Testing Hypothesis:

H_0:

Against **H_1:**

Test Statistic

Calculation:

Problem 3: A random sample of 400 house-holds in a city showed an average annual consumption of 16 kg of brand A coffee with a s.d. of 2 kg. Another sample of 300 households showed an average consumption of 14 kg of brand B with a s.d. of 3 kg. For a significance level of 5%, a marketing research agency wishes to judge the claims of each brand being the market leader.

Solution:

Testing Hypothesis:

$\mathbf{H_0}$:

Against $\mathbf{H_1}$**:**

Test Statistic

Calculation:

Problem 4: A sample of heights of 500 Neem tree has a mean of 20 .2 ft. with standard deviation of 1.7 ft. while a sample of height of 400 Mango trees has a mean of 19.6ft with standard deviation of 0.75 ft. Do the data indicate that height of Neem trees are on an average more than Mango trees?

Solution:
Testing Hypothesis:

H_0:

Against **H_1:**

Test Statistic

Calculation:

Exercise 11

Test of Significance-II
(Small Sample Test)

Assumption for “t” test:

1. Sample should be drawn from a normal population.
2. Sample size should be smaller than 30.
3. Variance of the population should be unknown.
4. Sample should be drawn independently.

(i) Testing of significance of single mean for small sample

Let x_1, x_2,……..,x_n be a random sample from a normal population with mean μ and unknown variance σ^2. Samples are drawn independently.

Testing Hypothesis: We want to test a null hypothesis

H_0: $\mu = \mu_0$

Against **H_1:** $\mu > \mu_0$ or $\mu < \mu_0$ or $\mu \neq \mu_0$

Test Statistic: We use a test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \sim t_{n-1}$$

$$Where \quad s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \quad and \quad \bar{x} = \frac{1}{n}\sum_i x_i$$

We reject our null hypothesis at α% level of significance if		
For **H_1:** $\mu > \mu_0$	$t > t_{\alpha:\, n-1}$	Otherwise accept H_0
For **H_1:** $\mu < \mu_0$	$t < t_{1-\alpha:\, n-1}$	Otherwise accept H_0
For **H_1:** $\mu \neq \mu_0$	$\boldsymbol{\lvert t\rvert > t_{\frac{\alpha}{2}:n-1}}$	Otherwise accept H_0

(ii) Testing of significance of the difference between two sample means for small samples:

Let $x_{11}, x_{12}, ...x_{1n_1}$ and $x_{21}, x_{22}, ...x_{2n_2}$ be two independent random samples from normal populations with mean μ_1, μ_2 and unknown variance σ_1^2, σ_2^2 respectively. It is assumed that both the variances are equal.

Testing Hypothesis: We want to test a null hypothesis

H_0: $\mu_1 = \mu_2$

Against **H_1:** $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 \neq \mu_2$

Test Statistic: We use a test statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \sim t_{n_1+n_2-2}$$

$$Where \quad s^2 = \frac{\sum_i (x_{1i} - \overline{x}_1)^2 + \sum_i (x_{2i} - \overline{x}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{n_1 + n_2 - 2}$$

We reject our null hypothesis at α% level of significance if		
For **H₁:** $\mu_1 > \mu_2$	$t > t_{\alpha: n_1+n_2-2}$	Otherwise accept H_0
For **H₁:** $\mu_1 < \mu_2$	$t < t_{1-\alpha: n_1+n_2-2}$	Otherwise accept H_0
For **H₁:** $\mu_1 \neq \mu_2$	$\lvert t \rvert > t_{\alpha/2: n_1+n_2-2}$	Otherwise accept H_0

(iii) Testing of significance of the difference between two sample means for small samples when the samples are correlated (Paired t test):

Let (x_1, y_1), (x_2, y_2),……….., (x_n, y_n) be n pair of observations (samples) from a bivariate normal population whose all parameters are unknown and d be the difference between the corresponding pair.

Testing Hypothesis: We want to test a null hypothesis

H₀: $\mu_d = 0$

Against **H₁:** $\mu_d > 0$ or $\mu_d < 0$ or $\mu_d \neq 0$

Test Statistic: We use a test statistic

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \qquad \sim t_{n-1}$$

$$Where \quad s_d{}^2 = \frac{1}{n-1}\sum_i (d_i - \bar{d})^2 \quad and \quad \bar{d} = \frac{1}{n}\sum_i d_i$$

We reject our null hypothesis at α% level of significance if		
For **H₁:** $\mu_d > 0$	$t > t_{\alpha: n-1}$	Otherwise accept H_0
For **H₁:** $\mu_d < 0$	$t < t_{1-\alpha: n-1}$	Otherwise accept H_0
For **H₁:** $\mu_d \neq 0$	$\lvert \boldsymbol{t} \rvert > \boldsymbol{t}_{\frac{\alpha}{2}: n-1}$	Otherwise accept H_0

(iv) Test for correlation coefficient

Let (x_1, y_1), (x_2, y_2),……….., (x_n, y_n) be n pair of observations (samples) from a bivariate normal population whose all parameters are unknown

Testing Hypothesis: We want to test a null hypothesis

H_0: $\rho = 0$

Against **H_1: $\rho \neq 0$**

Test Statistic: We use a test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \sim t_{n-2}$$

$$Where \quad r = \frac{\frac{1}{n}\sum xy - (\bar{x}\,\bar{y})}{\sqrt{\left(\frac{1}{n}\sum x^2 - \left(\bar{x}^2\right)\right)\left(\frac{1}{n}\sum y^2 - \left(\bar{y}^2\right)\right)}}$$

We reject our H_0 at α % level of significance if $|t| > t_{\alpha/2:n-2}$, otherwise accept H_0.

Problem 1: Daily protein intake (in gm) by an adult during a fortnight was reported as 48.1, 42.1, 52.0, 48.4, 49.4, 52.0, 46.6, 49.8, 46.4, 50.2, 48.8, 46.5, 50.0 51.4 and 48.0. Can we say that on an average daily protein intake by that adult is 50 gm? (Given that table value= t(14, 0.05/2)=2.145).

Solution:

Testing Hypothesis:

H_0:

Against **H_1:**

Test Statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \sim t_{n-1}$$

X	$(X - \bar{X})$	$(X - \bar{X})^2$

Calculation:

Problem 2: Tensile strength of threads manufactured by firm A for a random sample of 12 pieces of threads were measured as 2.20, 2.15, 2.17, 2.24, 2.20, 2.20, 2.25, 2.10, 2.26, 2.24, 2.18, 2.24 while that for threads of firm B for a random sample of 10 pieces were measured as 2.46, 2.42, 2.38, 2.26, 2.46, 2.38, 2.37, 2.35, 2.43, 2.41. Test whether average tensile strength of threads of these firms is same. (Given that t(20,0.05/2) =2.086).

Solution:

Testing Hypothesis:

H_0:

Against **H_1:**

Test Statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \sim t_{n_1+n_2-2}$$

Observation Table:

X_1	**X_2**	$X_1 - \overline{X}_1$	$(X_1 - \overline{X}_1)^2$	$X_2 - \overline{X}_2$	$(X_2 - \overline{X}_2)^2$

Calculation:

Problem 3: A drug is given to ten patients and their blood pressure before drug and after drug was recorded. Is it reasonable to believe that the drug has reduced the blood pressure? (Given that t (9, 0.05/2) =2.26).

Before	120	135	135	140	146	119	135	147	152	137
After	117	141	133	145	140	119	135	140	145	130

Solution:

Testing Hypothesis:

H_0:

Against **H_1:**

Test Statistic:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \qquad \sim t_{n-1}$$

Observation Table:

Before (X)	120	135	135	140	146	119	135	147	152	137
After (Y)	117	141	133	145	140	119	135	140	145	130
Difference (d) = Y-X										
$d - \bar{d}$										
$(d - \bar{d})^2$										

Calculation:

Problem 4: From a pair of 18 values the value of the correlation coefficient was calculated as 0.60. Test its significance.

Solution:

Testing Hypothesis:

H$_0$:

Against **H$_1$:**

Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \sim t_{n-2}$$

Calculation:

Problem 5: From the following data calculate S.E. of and also test their significance.

$If\ \ n=10,\ \sum x=40, \sum y=50,\ \sum x^2=270,\ \sum y^2=300,\ \sum xy=250,\ b_{yx}=0.83\ and\ b_{xy}\ =1$

Solution:

Testing Hypothesis:

H₀:

Against **H₁:**

Test Statistic:

$$t=\frac{b_{yx}}{S.E.b_{yx}} \qquad \sim t_{n-2}$$

Calculation:

Exercise – 12

TEST OF SIGNIFICANCE-III

(CHI SQUARE TEST)

Application of Chi-square test:

1. To test the significance of sample variance
2. To compare theoretical and observed proportion
3. To test the independence of attributes in a contingency table
4. To test the goodness of fit

(i) Test for significance of sample variance / S.D.

Let x_1, x_2,……..,x_n be a random sample from a normal population with mean μ and variance σ^2.

Testing Hypothesis: We want to test a null hypothesis

H_0: $\sigma = \sigma_0$

Against **H_1:** $\sigma > \sigma_0$ or $\sigma < \sigma_0$ or $\sigma \neq \sigma_0$

Test Statistic: We use a test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \quad \sim \chi^2_{n-1}$$

$$\text{Where} \quad s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

OR

H_0: $\sigma^2 = \sigma_0^2$

Against **H_1:** $\sigma^2 > \sigma_0$ or $\sigma^2 < \sigma_0^2$ or $\sigma^2 \neq \sigma_0^2$

We reject our null hypothesis at α% level of significance if		
For **H_1:** $\sigma > \sigma_0$	$\chi^2 > \chi^2_{\alpha:n-1}$	Otherwise accept H_0
For **H_1:** $\sigma < \sigma_0$	$\chi^2 < \chi^2_{1-\alpha:n-1}$	Otherwise accept H_0
For **H_1:** $\sigma \neq \sigma_0$	*Either* $\chi^2_{1-\alpha/2:n-1} < \chi^2$ *Or* $\chi^2 > \chi^2_{\alpha/2:n-1}$	Otherwise accept H_0

(ii) Test to compare theoretical and observed proportion

Let $r_1, r_2, \ldots\ldots, r_n$ be theoretical proportion and $n_1, n_2, \ldots. n_n$ be the observations of sample regarding theoretical proportion taken from a population under study.

Testing Hypothesis: We want to test a null hypothesis

H_0: Observations are in given proportion.

Against **H_1:** Observations are not in given proportion.

Test Statistic: We use a test statistic

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \sim \chi^2{}_{n-1}$$

Where $E_i = \dfrac{r_i}{r} \times N$

$r = r_1 + r_2 + \ldots\ldots + r_n$ *and* $N = n_1 + n_2 + \ldots\ldots n_n$

We reject our null hypothesis at α% level of significance if $\chi^2 > \chi^2_{\alpha:\,(n-1)}$ otherwise accept H_0.

(iii) Test for independence of attributes in a contingency table

A population is classified with respect to two characters A and B, one character is having p number of classes and the other is having q number of classes. Objective is to check whether the two characters are independent or not. For this purpose draw a random sample of n units and determine how many units belong to different classes of two characters. Let p = 2 and q =2. The data will be presented as in the following table called as 2 x 2 Contingency table.

A/B	B1	B2	Total
A1	A	B	a+b
A2	C	D	c+d
Total	a+c	b+d	a+b+c+d

Testing Hypothesis: We want to test a null hypothesis

H_0: Two characters are independent.

Against **H_1:** Two characters are not independent.

Test Statistic: We use a test statistic

$$\chi^2 = \frac{(ad - bc)^2\,(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)} \sim \chi^2{}_{(p-1)(q-1)}$$

We reject our null hypothesis at α% level of significance if $\chi^2 > \chi^2_{\alpha:\,(p-1)(q-1)}$ otherwise accept H_0.

(iv) Test for goodness of fit

A population has k number of classes from which a random sample of n units is drawn and observes how many belong to different classes. Let $O_1, O_2, \ldots\ldots, O_k$ be the number of observations in different classes. It is obvious that $O_1 + O_2 + \ldots\ldots + O_k = n$. According to some theory it is proposed that number of observations in these classes should be E_1, E_2, ………,E_k respectively. Now objective is to check whether the observed frequencies are according to the theoretical or expected frequencies. This objective is fulfilled by this test procedure. Using this test procedure we are testing whether there is agreement between observed and expected frequencies or not.

Testing Hypothesis: We want to test a null hypothesis

H_0: The observed frequencies are according to expected frequencies.

Against **H_1:** The observed frequencies are not according to expected frequencies.

Test Statistic: We use a test statistic

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \quad \sim \chi^2_{k-1}$$

Where E_i can be calculated by based appropriate theory.

We reject our null hypothesis at α% level of significance if $\chi^2 > \chi^2_{\alpha:\,(k-1)}$ otherwise accept H_0.

Problem 1: For years an experimenter has been using rats with standard deviation in weight as 26 gms. A new supplier assures him to supply the rats at low price. But the experimenter will not purchase them if standard deviation is more than 26 gms. So he selects a sample of 20 new rats and observes that s = 35. Will the experimenter order the rats from new supplier? (At 5% level of significance).

Solution:

Testing Hypothesis: **H_0**:

Against **H_1:**

Test Statistic:

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2} \quad \sim \chi^2_{n-1}$$

Calculation:

Problem 2: The following data is related to the segregation of 2 genes for purple-red flower colour and long-round pollen shape in peas:

Purple-long	Red-long	Purple-round	Red-round
296	27	19	8

Test at 5% level of significance that the segregation are in the ratio 9:3:3:1.

Solution:

Testing Hypothesis: **H$_0$**:

Against **H$_1$:**

Test Statistic:

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \sim \chi^2_{n-1}$$

$$\textit{Where} \quad E_i = \frac{r_i}{r}\times N$$

Observation Table:

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$

Calculation:

Problem 3: From a survey conducted in different regions (rural and urban)(A) to know preference of persons to different television programs (educational and entertainment) (B) following data was obtained. Test whether preference to a program depends on region. (Given that χ_2 (1, 0.05)= 3.84).

A/B	**B1**	**B2**	**Total**
A1	35	45	80
A2	20	50	70
Total	55	95	150

Solution:

Testing Hypothesis:

H_0:

Against **H_1:**

Test Statistic:

$$\chi^2 = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2_{(p-1)(q-1)}$$

Calculation:

Problem 4: 200 individuals are classified according to their eye and hair colours as given below. Test at 5% level of significance whether eye and hair colours are independent.

Eye colour				
Hair colour	Black	Blue	Brown	Total(A_i)
Black	40	20	60	120
Grey	20	30	30	80
Total (B_j)	60	50	90	200

Solution:

Testing Hypothesis: **H_0**:

Against **H_1:**

Test Statistic:

$$\chi^2 = \sum\sum\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right] \sim \chi^2_{n-1}$$

$$\textit{Where} \quad E_i = \frac{(A_i \times B_j)}{N}$$

Observation Table:

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Calculation:

Problem 5: A die is thrown 200 times with the following results:

No.	1	2	3	4	5	6
Frequency	32	40	28	35	25	40

Is the die unbiased at 5% level of significance?

Solution:

Testing Hypothesis: **H_0**:

Against **H_1:**

Test Statistic:

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] \sim \chi^2_{n-1}$$

$Where \quad E_i = N \times p_i$

Observation Table:

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$

Calculation:

Exercise – 13

ANALYSIS OF VARIANCE: ONE-WAY CLASSIFICATION

Analysis of variance is the systematic algebraic procedure to decomposing the overall variation in the responses observed in an experiment in to different component. Each component is attributed to an identifiable cause or sources of variation. Our purpose is to test the hypothesis about these components but the main aim is to estimating and comparing pairwise treatments. This technique is developed by " Prof. R.A. Fisher" in 1923 with the use of F Test.

Assumptions of ANOVA:

1. The total variance of the various sources of variation should be additive.
2. The errors attached to each observation are independently and normally distributed with mean zero and variance σ_e^2.
3. Observations should be independent each other.
4. The variance of sub group should be homogeneous.

One Way Classification:

We have studied two-sample t test for mean which is used to compare means of two populations. Now, we consider a test procedure to compare means of more than two (say, k) populations. Let n random observations are classified in to k different classes or groups in such a manner that i^{th} class or group contain n_i (i = 1,2,…,k) observations.

Classes or Groups					
A_1	**A_2**	**-**	**A_i**	**-**	**A_k**

In a wider perspective, the above data is referred to as one-way classified data as the observations are classified with respect to one factor, the different samples now being referred to as groups. The main objective in such studies is to compare averages of different groups. This objective is formulated in the form of a hypothesis as given below.

H_0: All group means are equal.

H_1: Not all group means are equal.

This hypothesis is tested by using F test. To perform F test, we use analysis of variance (ANOVA) technique, in which, total variation present in data is divided into between group variation and error variation. The F test involves comparison of between group variation with error variation. The calculations of this technique are presented in a table called as ANOVA table.

Source of Variation	**Degree of freedom**	**Sum of square**	**Mean sum of square**	**F ratio**

Between Groups	k-1	SSB	MSB	MSB/MSE
Error	n-k	SSE	MSE	-
Total	n-1	TSS	-	-

Calculations of this table are performed using following steps.
Step 1 – Calculate d.f. as given in that column.
Step 2 – Calculate total for each group. These are denoted by T_1, T_2, ……, T_k.
Step 3 – Calculate grand total, G = $T_1 + T_2 +$ …………+ T_k.
Step 4 – Calculate correction factor, (c f) = $G^2/$ n
Step 5 – Calculate SSB = Between groups sum of squares

$$SSB \; = \left[\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + ...+ \frac{T_k^2}{n_k} \right] - C.F.$$

Step 6 – Calculate TSS= total sum of squares = RSS – C.F.

$$TSS \; = \left[y_{11}^2 + y_{12}^2 + ...+ y_{1n_1}^2 + y_{21}^2 ......+ y_{kn_k}^2 \right] - C.F.$$

Step 7 – Calculate SSE = TSS - SSB
Step 8 – Calculate MSB = SSB/ (k-1)
MSE= SSE/ (n-k)
Step 9 – Calculate the value of test statistic, F =MSB/ MSE
Step 10 – Draw Conclusion:

The test statistic, F has (k – 1), (n – k) d.f. . Refer to table of F distribution. The table value is denoted by F (α: k–1, n –k). Reject H_0 if F ≥ F (α: k–1, n –k) otherwise accept null hypothesis. Acceptance of H_0 means there are no differences among group means. Rejection of H_0 means not all group means are equal. Thus when H_0 is rejected, it is necessary to know which group means differ from each other and which do not. For this purpose calculate critical difference (CD). If difference between any two group means is greater than or equal to critical difference those means are said to be significantly different from each other. The critical difference to compare ith mean with jth mean is calculated using following formula.

$$CD = t_{\alpha/2:error d.f.} \times \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Present the comparison between group means in the following table.

Pairs	**Difference of Means**	**CD**	**Conclusion**

Problem 1: The percentage protein content in soymilk prepared from five varieties of soybean was determined for a number of samples as given below. Test whether average protein content is same. Given table value: F (4.25, 0.05) = 2.76 and t (25, 0.05/2) = 2.060).

Variety	Percentage protein content	Total
V_1	2.52, 2.36, 2.64, 2.52, 2.44, 2.48	
V_2	2.88 , 2.78, 2.84, 2.86, 2.84, 2.80	
V_3	2.66, 2.58, 2.62, 2.54, 2.48, 2.52	

V_4	3.02, 3.06, 3.12, 3.18, 2.98, 3.06	
V_5	3.44, 3.48, 3.40, 3.46, 3.48, 3.46	

Solution:

Problem 2: The following figures related to the production of three varieties of wheat (in Kg) named A, B and C used in 15 plot.

Variety of Wheat	Yield (Kg)
A (RR-21) B (K-68) C (Sonalika)	14, 17, 16, 16 15, 11, 13, 15, 13, 14 18, 16, 18, 19, 15

Test, whether there is any significance difference in the production of three varieties or not.

Solution:

Exercise – 14

ANALYSIS OF VARIANCE: TWO-WAY CLASSIFICATION

Two Way Classifications (One observation per cell):

Let n random observations are classified into p categories or classes such as A_1, A_2,……,Ap according to some criterion A (i=1,2,3…,p) and into q categories or classes such as B_1, B_2,……,Bq according to some other criterion B (j=1,2,3…,q). In all the data we have pxq combinations in which (i, j)th cell is known as the observation corresponding to ith level of criterion A and jth level of criterion B. Let the observation corresponding (i, j)th cell is y_{ij}. The observations can be represented as follows:

A/B	B_1	B_2	-	B_j	-	B_q
A_1	y_{11}	y_{12}	-	y_{1j}		y_{1q}
A_2	y_{21}	y_{22}	-	y_{2j}		y_{2q}
. .	. .	. .	. .	. .	. .	. .
A_i	y_{i1}	y_{i2}	-	y_{ij}		y_{iq}
.	-	-	-	-	-	-
. .	-	-	-	-	-	-
A_p	y_{p1}	y_{p2}		y_{pj}		y_{pq}

In a wider perspective, the above data is referred to as two-way classified data as the observations are classified with respect to two factors. The main objective in such studies is to check different levels of factor A as well as Factor B, is they are equally effective or not? This objective is formulated in the form of a hypothesis as given below.

H_{01}: All levels of factor A are equally effective.
H_{11}: At-least one level of factor A is differ significantly.

H_{02}: All levels of factor B are equally effective.
H_{12}: At-least one level of factor B is differ significantly.

This hypothesis is tested by using F test. To perform F test, we use analysis of variance (ANOVA) technique, in which, total variation present in data is divided into between group variation (between the levels of A as well as between the levels of B) and error variation. The F test involves comparison of between variation due to factor A as well as factor B with error variation. The calculations of this technique are presented in a table called as ANOVA table.

Source of Variation	Degree of freedom	Sum of square	Mean sum of square	F ratio
Between the levels of A	p-1	SSA	MSA	MSA/MSE
Between the levels of B	q-1	SSB	MSB	MSB/MSE
Error	(p-1)(q-1)	SSE	MSE	-
Total	pq-1	TSS	-	-

Calculations of this table are performed using following steps.

Step 1 – Calculate d.f. as given in that column.
Step 2 – Calculate total of each row (Different levels of Factor A ($R_1, R_2, \ldots, R_p$)) as well as each column (Different levels of Factor B ($C_1, C_2, \ldots, C_q$)).
Step 3 – Calculate grand total, G = $R_1 + R_2 + \ldots\ldots\ldots\ldots + R_p = C_1 + C_2 + \ldots\ldots\ldots\ldots + C_q$
Step 4 – Calculate correction factor, (c f) = G^2/ pq
Step 5 – Calculate SSA = Sum of squares due to factor A

$$SSA \;=\frac{1}{q}\left(R_1^2+R_2^2+.......+R_p^2\right)\;-\;C.F.$$

Step 6 — Calculate SSB = Sum of squares due to factor B

$$SSA \;=\frac{1}{p}\left(C_1^2+C_2^2+.......+C_q^2\right)\;-\;C.F.$$

Step 7- Calculate TSS= total sum of squares = RSS – C.F.

$$TSS \;=\;\left[y_{11}^2+y_{12}^2+............+y_{pq}^2\right]-C.F.$$

Step 8 – Calculate SSE = TSS – SSA - SSB
Step 9 – Calculate MSA = SSA/ (p-1)
MSB= SSB/ (q-1)
MSE= SSE/ (p-1)(q-1)
Step 10 – Calculate the value of test statistic,
F_A =MSA/ MSE
F_B =MSB/ MSE
Step 11 – Draw Conclusion:

The test statistic, F has (p – 1), (p-1xq-1) d.f for factor A and (q – 1), (p-1xq-1) d.f for factor B. Refer to table of F distribution. The table value is denoted by F (α: p–1, p-1 x q-1) and by F (α: q–1, p-1 x q-1). Reject H_0 if $F_A \geq$ F (α: p–1, p-1 x q-1) otherwise accept null hypothesis. When H_0 is rejected, it is necessary to know which level means of A differ from each other and which do not. For this purpose calculate critical difference (CD). If difference between any two level means is greater than or equal to critical difference those means are said to be significantly different from each other. The critical difference to compare rth mean with sth mean is calculated by using following formula.

$$CD = t_{\alpha/2:error d.f.} \times \sqrt{\frac{2MSE}{q}}$$

Present the comparison between group means in the following table.

Pairs	Difference of Means	CD	Conclusion

The same process is done for factor B. We reject H_0 if $F_B \geq$ F (α: q–1, p-1 x q-1) otherwise accept null hypothesis. When H_0 is rejected, it is necessary to know which level means of B differ from each other and which do not. For this purpose calculate critical difference (CD). If difference between any two level means is greater than or equal to critical difference those means are said to be significantly different from each other. The critical difference to compare kth mean with lth mean is calculated by using following formula.

$$CD = t_{\alpha/2:error d.f.} \times \sqrt{\frac{2MSE}{p}}$$

Present the comparison between group means in the following table.

Pairs	**Difference of Means**	**CD**	**Conclusion**

Problem 1: A company appointed four salesmen A, B, C and D and observes their sales in three seasons: summer, winter and monsoon. The figures (in Lacs) are given in the following table:

Seasons	**Salesmen**			
	A	**B**	**C**	**D**
Summer	36	36	21	35
Winter	28	29	31	32
Monsoon	26	28	29	29

Carry out an analysis of variance.

Solution:

Problem 2: A farmer applies three types of fertilizers on four separate plots. The figures on yield per acre are tabulated below:

Fertilizers	**Yield**			
	A	**B**	**C**	**D**
Nitrogen	60	40	80	60
Potash	70	60	60	90
Phosphates	80	50	100	90

Find out if the plots are materially different, and also, if the three fertilizers make any material difference in yields.

Exercise -15

PROBABILITY AND PROBABILITY DISTRIBUTIONS

A probability is a quantitative measure of uncertainty a number that conveys the strength of our belief in the occurrence of an uncertain event.

Experiment: An experiment is a process that leads to one of several possible outcomes. An outcome of an experiment is some observation or measurement.

Random Experiment: A random experiment is an experiment which is repeated a large numbers of time whose all possible outcomes are known in advance but result is not known until it is observed.

Trial: When we repeat a random experiment several times, we call each of them as trial thus any particular performance of a random experiment is known as a trial.

Sample Space: A sample space, **Ω**, is a set of all possible outcomes of a random experiment. Every possible outcome must be listed once and only once.

Sample Point: A sample point is an element of the sample space. For example, if the sample space is **Ω** = $\{s_{1,} s_{2,} s_3\}$, then each s_i is a sample point.

Event: An event is a subset of the sample space. That is, any collection of outcomes forms an event. eg Toss a coin twice. Sample space: **Ω** = {HH, HT, TH, TT}

Exhaustive Outcomes: The total number of possible outcomes in any trial is known as exhaustive outcomes. For example (i) in tossing of a coin there are two exhaustive cases, viz., head and tai1. (ii) In throwing of a die, there are six, exhaustive cases since anyone of the 6 faces 1,2, ... ,6 may come uppermost.

Mutually Exclusive Outcomes: If the happening of anyone outcome precludes the happening of all the others i.e., if no two or more of them can happen simultaneously in the same trial, then they are said to be mutually exclusive. For example: (i) In throwing a die all 6 faces numbered 1 to 6 are mutually exclusive since if anyone of these faces comes, ,the possibility of others, in the same trial, is ruled out (a) Similarly in tossing a coin, the events head and tail are mutually exclusive.

Equally likely Outcomes: Outcomes of a trial are said to be equally likely if taking into consideration all the relevant evidence there is no reason to expect one in preference to the others. For example (i) In tossing an unbiased or uniform coin, head or tail ate mutually likely events. (ii) In throwing an unbiased die, all the six faces are equally likely to come.

Mathematical or Classical or a priori probability Definition (Laplace)

If a trial results in “n” exhaustive, mutually exclusive and equally likely cases out of them “m” are favorable to the happening of an event E, then the probability 'P' of happening of E is given by

$$P(E)=\frac{m}{n} = \frac{Favourable\ \ number\ \ of\ \ cases}{Total\ \ number\ \ of\ \ cases}$$

Remarks.

1. Probability 'p' of the happening of an event is also known as the probability of success and the probability 'q' of the non- happening of the event as the probability of failure.

2. If P (E) = 1, E is called a certain event and if P (E) = 0, E is called an impossible event.

Statistical or Empirical Probability Definition (Von Mises).

If a trial is repeated a large number of times under essentially homogeneous and identical conditions; then the limiting value of the ratio of' the number of times the event happens to the number of trials, as the number of trials become indefinitely large, is called the probability of happening of the event. (It is assumed that the limit is finite and unique). Symbolically, if in n trials an event E happens m times, then the probability 'p' of the happening of E is given by

$$P(E) = \lim_{n\to\infty} \frac{m}{n}$$

Axiomatic Definition

Let S be a sample space, P(A) is the probability function defined on a sigma field D of events if the following properties or axioms hold :

1. For each $A \in D$, P (A) is defined, is real and $P(A) \geq 0$

2. P(S) = 1

3. If A_n is any finite or infinite sequence of disjoint events, in D then $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$

The above three axioms are termed as the axiom of positive ness, certainty and union (additivity), respectively.

Theorems on probabilities of events

(**i**). Probability of the impossible event is zero, i e P (ϕ) = 0.

(ii). Probability of the complementary event it of A is given by P(A) = $1 - P(\bar{A})$.

(iii). For any two events A and B, $P(\bar{A} \bigcap B) = P(B) - P(A \bigcap B)$.

(iv). If A is subset of B, then $P(B - A) = P(B) - P(A)$.

(v).). If A is subset of B, then $P(B) \geq P(A)$

Addition Theorem of Probability

(i). If A and B are any two events (subsets of sample space S) and are not disjoint, then

$$P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$$

(ii). If A, B and C are any three events (subsets of sample space S) and are not disjoint, then

$$P(A \bigcup B \bigcup C) = P(A) + P(B) + P(C) - P(A \bigcap B) - P(A \bigcap C) - P(B \bigcap C) + P(A \bigcap B \bigcap C)$$

Conditional Probability: If A be any event of a sample space S and B be another event associated with S, we are interested to find out the probability of event A when event B has been already happen. Such type of probability is known as conditional probability. It is denoted by P (A/B) and defined as

$$P(A/B) = \frac{P(A \bigcap B)}{P(B)}$$

Provided, P(B) >0

Independent events: Several events are said to be independent if the happening (or non-happening) of an event is not affected by the occurrence of any number of the remaining events. For example (i) In tossing an unbiased coin the event of obtaining head in the first toss is independent of getting a head in the second, third and subsequent throws. In other words, we can say

An event B is said to be independent (or statistically independent) of event A, if the conditional probability of B given A i.e., P (B/A) is equal to the unconditional probability of B, i.e.,

P (B / A) = P (B)

Hence, the events A and B are independent if and only if $P(A \bigcap B) = P(A)P(B)$

Bayes Theorem: If $E_{1,} E_2, ..., E_n$ are mutually disjoint events with $P(E_i) \neq 0$ (i = 1,2, ... , n) then for any arbitrary event A which is a subset of $\bigcup E_i$, such that P (A) > O. we have

$$P\left(E_i/A\right)=\frac{P(E_i)\,P\left(A/E_i\right)}{\sum_i P(E_i)\,P\left(A/E_i\right)} \qquad ; \qquad (i = 1,2, \dots , n).$$

Binomial Distribution:

Binomial distribution is widely used probability distribution in many fields like as agriculture, medicine, quality inspection etc. This distribution has the following conditions:

1. Each trial has only two possible outcomes “success and failure”.
2. The repeated trials are independent to each other.
3. The probability of success in each trial remains constant.
4. The number of trials is finite.

A random variable X is said to follow a binomial distribution if it assume only nonnegative values and its probability mass function is given as

$$P\left[X = x\right]=C_x^n\; p^x q^{n-x} \qquad where\;\; x = 0,1,2,.....,n$$

n = number of trials,
x = number of success in a trial,
n-x = number of failures in a trial
p = probability of success,
q = probability of failure,

By this probability distribution we can find the probability of success x number of times (desired outcome) in n independent trials, regardless of their order of occurrence.

Properties of Binomial Distribution:

1. Mean of the distribution is np.
2. Variance of the distribution is npq.
3. Skewness of the distribution can be obtained by the formula $(q-p)/\sqrt{npq}$

   If q > p distribution is positively skewed, if q< p distribution is negatively skewed and if q = p distribution is symmetric.
4. Kurtosis of the distribution can be obtained by the formula 1-6pq / (npq). If pq > 1/6 distribution is platykurtic, if pq < 1/6 distribution is leptokurtic and if pq = 1/6 distribution is mesokutic.
5. Mode of the distribution can be obtained by the formula (n+1) p. If (n+1)p is not an integer say (m+f) where m if an integer value and f is fraction value then mode will be m and if (n+1)p is an integer value say (m) then there would be two mode values m and m-1.

Poisson Distribution:

Poisson distribution is also widely used probability distribution in many fields like as entomology, medicine, quality control etc. Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials (unlike that in binomial) of an experiment but which occur at random points of time and space where in our interest lies only in the number of occurrences of the event, not in its non-occurrences. This distribution has the following assumptions:

1. The occurrence or nonoccurrence of an event does not influence the occurrence or nonoccurrence of any other event.

2. The probability of happening of more than one event is a very small interval is negligible.
3. The probability of success for a short time interval or a small region of space is proportional to the length of the time interval or space as the case may be.

There are some instances where Poisson distribution may be successfully employed

(1) Number of deaths from a disease (not in the form of an epidemic) Such as heart attack or cancer or due to snake bite.
(2) Number of suicides reported in a particular city.
(3) The number of defective material in a packing manufactured by a good concern.
(4) Number of faulty blades in a packet of 100.
(5) Number of air accidents in some unit of time.
(6) Number of printing mistakes at each page of the book.
(7) Number of telephone calls received at a particular telephone exchange in some unit of time or connections to wrong numbers in a telephone exchange.
(8) Number of cars passing a crossing per minute during the busy hours of a day.
(9) The number of fragments received by a surface area 't' from a fragment atom bomb.
(10) The emission of radioactive (alpha) particles.

.

A random variable X is said to follow a Poisson distribution if its probability mass function is given as

$$P[X = x] = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad where \;\; x = 0,1,2,.....,\infty$$

Properties of Poisson Distribution:

1. Mean of the distribution is λ.
2. Variance of the distribution is λ.
3. Skewness of the distribution can be obtained by the formula 1/ λ. Since λ is positive hence distribution has positive skewness.
4. Kurtosis of the distribution can be obtained by the formula 3+(1/ λ). Since λ is positive hence distribution has leptokurtic.
5. Mode of the distribution can be obtained by the formula. If it is not an integer say (m+f) where m if an integer value and f is fraction value then mode will be λ and if it is an integer value say (m) then there would be two mode values λ and λ-1.

Normal Distribution:

The normal distribution was first discovered in 1733 by English mathematician De-Moivre, who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problems arising in the game of chance. It was also known to Laplace but due to some historical error it was credited to Gauss, who first made reference to it in the beginning of 19th century (1809), as the distribution of errors in Astronomy. Gauss used the normal curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies. Throughout the eighteenth and nineteenth centuries, various efforts were made to establish the normal model as the underlying law ruling all continuous random variables. The normal model has nevertheless, become the most important probability model in statistical analysis.

A random variable X is said to have a normal distribution with parameters μ (called "mean") and σ^2 (called "variance") if its density function is given by the probability law:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty,\ -\infty < \mu < \infty,\ \sigma^2 > O$$

Properties of Normal Distribution:

1. Mean of the distribution is μ.
2. Variance of the distribution is σ^2.
3. The curve is bell shaped and symmetrical about the line x = μ.
4. Distribution is mesokurtic.
5. Mean, median and mode of the distribution coincide.
6. Since f(x) being the probability, it can never be negative because no portion of the curve lies below the x-axis.
7. As x increases numerically, f(x) decreases rapidly, the maximum probability occurring at the point x = μ.
8. Linear combination of independent normal variates is also a normal variate.
9. In this distribution, Q. D. : M.D.: S.D. :: 10: 12 : 15

Area Property

$P(\mu - 0.6745\sigma < X < \mu + 0.6745\sigma) = 0{\cdot}5000$
$P(\mu - 1.00\sigma < X < \mu + 1.00\sigma) = 0{\cdot}6826$
$P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0{\cdot}9500$
$P(\mu - 2.00\sigma < X < \mu + 2.00\sigma) = 0{\cdot}9544$
$P(\mu - 2.58\sigma < X < \mu + 2.58\sigma) = 0{\cdot}9900$
$P(\mu - 3.00\sigma < X < \mu + 3.00\sigma) = 0{\cdot}9973$

A random variable Z is said to have a standard normal distribution if its density function is given by the probability law:

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}(z)^2} \qquad -\infty < z < \infty,$$

(Exercise)

1. Four persons A, B, C and D stand in a row. (a) write the no. of all possible ways in which they can stand. (b) write the no. of all possible ways in which they can stand such that C is always at second position [ans (a) 24, (b) 6].
2. There are 4 agricultural officers and 3 veterinary officers. A committee of 3 officers is to be formed such that (a) it has 2 agricultural officers and 1 veterinery officer (b) it has at least 2 agricultural officers. Determine the number of ways in which the committee can be formed in both situations. [ans (a) 18 (b) 22].
3. A coin is thrown 2 times. What is the probability of getting one head? (ans ½)
4. A coin is thrown 3 times. What is the probability that at least one head occur? [ans7/8]
5. One number is selected at random from each of the two sets (1,2,3,4) and (2,3,4,5,6). What is the probability that sum of the two selected numbers is (a) 8 (b) at most 6? [ans (a) 3/20 (b) 10/20].
6. A bag contains 30 tickets numbered serially from 1 to 30. One ticket is drawn at random. What is the probability it bears a number which is divisible by (i) 4 or 9 (ii) 3 or 6 (iii) 2 and 6 (iv) 5 and 7? [(ans (i) 10/30 (ii) 10/30 (iii) 5/30 (iv) 0].
7. If two dice are thrown simultaneously what is the probability that sum of numbers appearing on upper faces will be more than 8 ? [ans 10/36].
8. From a well-shuffled pack of ordinary cards 3 cards are drawn (i) one by one without replacement (ii) one by one with replacement. In each case find the probability that first is an ace, second a king and third is a diamond queen. [ans (i) (4 x 4 x 1) / (52 x 51 x 50) (ii) (4 x 4 x 1) / (52 x 52 x 52)].
9. From a pack of cards two are drawn one by one without replacement. Calculate the probability that one is king and one is diamond queen. Also answer the question if cards are drawn with replacement. [ans (2x 4 x 1) / (52 x 51) , (2 x 4 x 1) / (52x52)].
10. In a bag there are 5 red, 3 white and 6 black balls, all of the same size and shape. 4 balls are drawn at a time from this bag. What is the probability that they contain 2 red, 1 white and 1 black ball. [ans (10 x 3 x 6) / { 14 ! / (4 ! x 10 !) }].
11. What is the probability of not rolling any 6's in four rolls of a balanced die? [ans (5/6)4].
12. Calculate the prob. of getting three 3's and then a 4 or a 5 in four rolls of die. [ans (1/6) x (1/6) x (1/6) x (2/6)].
13. If a die is thrown three times, what is the prob. that (a) all throws show 6 (b) all throws are alike (show same face)?[ans(a)(1/6)x (1/6)x (1/6), (b) 6x (1/6) x (1/6) x (1/6)].
14. It is known that probability of a male or female calf in a calving is ½ each. Calculate the probability that in two successive calving of a cow (a) both are male (b) both are female (c) at least one is a male. [ans. (a) ¼ (b) ¼ (c) ¾]
15. From a bag having 8 white and 4 black balls all of the same size and shape, 3 balls are drawn at random one by one without replacement. What is the probability that the balls drawn will alternately be of different colours. [ans (8 x 4 x 7) /(12 x 11 x 10) + (4 x 8 x 3) / (12 x 11 x 10)].
16.For n = 4 , p = 1/3 calculate probabilities for all possible valise of x and verify that (i) sum of all prob. is one (ii) mean and variance for the binomial distribution are 4/3 and 8/9 respectively.
17. It is known that 80% of the insects are killed by a certain insecticide. A lot of 5 insects are put to this insecticide. Find the probability that (i) at least 4 insects will be killed (ii) not more than one will be killed. [hint; (i) P(4) + P(5), (ii) P (0) + P (1)]
18. It was found that 10 % of boys in certain class are suffering from short sight. What is the probability that a random sample of 5 boys will contain (i) more than 3 boys suffering from short sight (ii) at most 4 boys suffering from short sight ? [hint (i) P(4) + P (5) , (ii) 1 – P (

5)]

19. Probability of finding an animal in Pantnagar suffering from a disease is 1/8. If a group of 10 animals is selected at random what is the probability that in that group at most 2 animals will be suffering from the disease. State mean and variance for this binomial distribution. [hint: P (0) + P (1) + P(2)]

20. There are 4 agricultural officers and 3 veterinary officers. 3 are randomly selected to form a committee. Calculate the probability that the committee has (a) 2 agricultural officers and 1 veterinery officer (b) at least 2 agricultural officers.[ans.(i) 18/35 (ii) 22/35]

References

1. **Agarwal, B.L. (2003).** Elementary Statistics, New age international publishers, New Delhi.
2. **Amdekar, S.J. (2008).** A handbook of elementary statistics.
3. **Chandel, S.P.S. (1997).** A handbook of agricultural statistics, Achal prakashan mandir, Kanpur.
4. **Gomez, K.A. and Gomez, A.A. (1984).** Statistical Procedures for agricultural research, John willy and sons, New York.
5. **Goon, A.M., Gupta, M.K. and Dasgupta, B. (2005).** Fundamentals of statistics Vol.-1, World Press, Culcutta.
6. **Goon, A.M., Gupta, M.K. and Dasgupta, B. (2005).** Fundamentals of statistics Vol.-2, World Press, Culcutta.
7. **Gupta, S.C. and Kapoor, V.K. (2002).** Fundamentals of mathematical statistics, Sultan chand and sons, New Delhi.
8. **Rangaswami, R. (2010).** A text book of agricultural statistics, New age international publishers, New Delhi.
9. **Shukla, R. K. and Kumar, K. (2011).** A text book of pharmaceutical biostatistics, Vigyan bodh prakashan, Agra.

F table

Degrees of Freedom for Denominator	Significance Level: 0.025 one-tailed, 0.05 two-tailed Degrees of Freedom for Numerator											
	1	2	3	4	5	6	7	8	9	10	15	20
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.6	963.3	968.6	984.9	993.1
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79	2.68
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72	2.62
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67	2.56
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62	2.51
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46

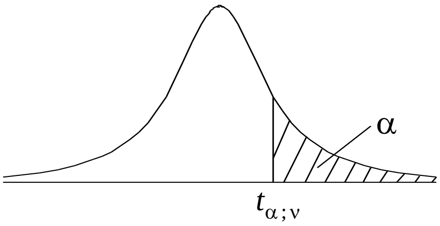
Chi-square table

	Significance Level			
Df	*0.1*	*0.05*	*0.025*	*0.01*
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.832	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.090
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209
11	17.275	19.675	21.920	24.725
12	18.549	21.026	23.337	26.217
13	19.812	22.362	24.736	27.688
14	21.064	23.685	26.119	29.141
15	22.307	24.996	27.488	30.578
16	23.542	26.296	28.845	32.000
17	24.769	27.587	30.191	33.409
18	25.989	28.869	31.526	34.805
19	27.204	30.144	32.852	36.191
20	28.412	31.410	34.170	37.566

Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where

$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



α ν	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table of the standard normal distribution values ($z \leq 0$)

–*z*	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45621	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42466
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10384	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08692	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03363	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00509	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00403	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00170	0.00164	0.00159	0.00154	0.00149	0.00144	0.00140
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
3.1	0.00097	0.00094	0.00090	0.00087	0.00085	0.00082	0.00079	0.00076	0.00074	0.00071
3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
3.4	0.00034	0.00033	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017

Table of the standard normal distribution values ($z \geq 0$)

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983