E- MATERIAL ON STATISTICAL METHODS

COURSE NO: ASAS-1101
COURSE CREDITS: 2(1+1)

By
Dr. Soumik Ray and Dr. Tufleuddin Biswas
Assistant Professor
Department of Agricultural statistics and computer applications
MSSSOA, CUTM

LECTURE OUTLINE

Course No. ASAS-1101 Credits: 2 (1+1)
Course Title: Statistical methods

THEORY

S. No.	Topic/Lesson
1	Introduction to Statistics, Definition, Advantages and Limitations.
2	Frequency distribution: Construction of Frequency Distribution table.
3	Measures of Central Tendency: Definition, Characteristics of Satisfactory Average.
4	Arithmetic Mean, Median, Mode for grouped and ungrouped data – Merits and Demerits of Arithmetic Mean.
5	Measures of Dispersion: Definition, standard deviation, variance and Coefficient of variation.
6	Normal Distribution and its properties. Introduction to Sampling: Random Sampling, concept of standard error of Mean.
7	Tests of Significance: Introduction, Types of errors, Null hypothesis, level of Significance and degrees of freedom, steps in testing of hypothesis.
8	Large sample tests: Test for Means – Z-test, One sample and Two samples with population S.D. known and Unknown.
9	Small sample tests: Test for Means – One sample t – test, Two samples t-test and Paired t-test.
10	Chi-Square test in 2x2 contingency table with Yate"s correction, F-test.
11	Correlation: Definition, types, properties, Scatter diagram, calculation and testing.
12	Regression: Definition, Fitting of two lines Y on X and X on Y, Properties, inter relation between correlation and regression.
13	Introduction to Experimental Designs, Basic Principles, ANOVA its assumptions.
14	Completely Randomized Design: Layout, Analysis with equal and unequal replications.
15	Randomized Block Design: Layout and Analysis.
16	Latin Square Design: Layout and Analysis.

PRACTICALS

S .No.	Topic
1	Construction of Frequency Distribution tables
2	Computation of Arithmetic Mean for Grouped and Un-grouped data
3	Computation of Median for Grouped and Un-grouped data
4	Computation of Mode for Grouped and Un-grouped data
5	Computation of Standard Deviation and variance for grouped and ungrouped data
6	Computation of coefficient of variation for grouped and ungrouped data
7	SND (Z) test for single sample, Population SD known and Unknown
8	SND (Z) test for two samples, Population SD known and Unknown
9	Student‟s t-test for single and two samples
10	Paired t-test and F-test
11	Chi-square test – 2x2 contingency table with Yate‟s correction
12	Computation of correlation coefficient and its testing
13	Fitting of simple regression equations Y on X and X on Y
14	Completely Randomized Design: Analysis with equal and unequal replications
15	Randomized Block Design: Analysis
16	Latin Square Design: Analysis

S T A T I S T I C S

Statistics has been defined differently by different authors from time to time. One can find more than hundred definitions in the literature of statistics.

Statistics can be used either as plural or singular. When it is used as plural, it is a systematic presentation of facts and figures. It is in this context that majority of people use the word statistics. They only meant mere facts and figures. These figures may be with regard to production of food grains in different years, area under cereal crops in different years, per capita income in a particular state at different times etc., and these are generally published in trade journals, economics and statistics bulletins, news papers, etc.,

When statistics is used as singular, it is a science which deals with collection, classification, tabulation, analysis and interpretation of data.

Important definition of statistics

Statistics is the branch of science which deals with the collection, classification and tabulation of numerical facts as the basis for explanations, description and comparison of phenomenon - Lovitt

The science which deals with the collection, analysis and interpretation of numerical data - Corxton & Cowden

The science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates - King

Statistics may be called the science of counting or science of averages or statistics is the science of the measurement of social organism, regarded as whole in all its manifestations - Bowley

Statistics is a science of estimates and probabilities -Boddington

Statistics is a branch of science, which provides tools (techniques) for decision making in the face of uncertainty (probability) - Wallis and Roberts

This is the modern definition of statistics which covers the entire body of statistics

All definitions clearly point out the four aspects of statistics collection of data, analysis of data, presentation of data and interpretation of data.

Importance: Statistics plays an important role in our daily life, it is useful in almost all sciences – social as well as physical – such as biology, psychology, education, economics, business management, agricultural sciences etc., . The statistical methods can be and are

being followed by both educated and uneducated people. In many instances we use sample data to make inferences about the entire population.

1. Planning is indispensable for better use of nation‟s resources. Statistics are indispensable in planning and in taking decisions regarding export, import, and production etc., Statistics serves as foundation of the super structure of planning.

2. Statistics helps the business man in the formulation of polices with regard to business. Statistical methods are applied in market and production research, quality control of manufactured products

3. Statistics is indispensable in economics. Any branch of economics that require comparison, correlation requires statistical data for salvation of problems

4. State. Statistics is helpful in administration in fact statistics are regarded as eyes of administration. In collecting the information about population, military strength etc., Administration is largely depends on facts and figures thud it needs statistics

5. Bankers, stock exchange brokers, insurance companies all make extensive use of statistical data. Insurance companies make use of statistics of mortality and life premium rates etc., for bankers, statistics help in deciding the amount required to meet day to day demands.

6. Problems relating to poverty, unemployment, food storage, deaths due to diseases, due to shortage of food etc., cannot be fully weighted without the statistical balance. Thus statistics is helpful in promoting human welfare

7. Statistics are a very important part of political campaigns as they lead up to elections. Every time a scientific poll is taken, statistics are used to calculate and illustrate the results in percentages and to calculate the margin for error.

In agricultural research, Statistical tools have played a significant role in the analysis and interpretation of data.

1. In making data about dry and wet lands, lands under tanks, lands under irrigation projects, rainfed areas etc.,

2. In determining and estimating the irrigation required by a crop per day, per base period.

3. In determining the required doses of fertilizer for a particular crop and crop land

In soil chemistry also statistics helps classifying the soils basing on their analysis results, which are analyzed with statistical methods.

4. In estimating the losses incurred by particular pest and the yield losses due to insect, bird, or rodent pests statistics is used in entomology.

5. Agricultural economists use forecasting procedures to determine the future demand and supply of food and also use regression analysis in the empirical estimation of function relationship between quantitative variables.

6. Animal scientists use statistical procedures to aid in analyzing data for decision purposes.

7. Agricultural engineers use statistical procedures in several areas, such as for irrigation research, modes of cultivation and design of harvesting and cultivating machinery and equipment.

Limitations of Statistics:

1. Statistics does not study qualitative phenomenon

2. Statistics does not study individuals

3. Statistics laws are not exact laws

4. Statistics does not reveal the entire information

5. Statistics is liable to be misused

6. Statistical conclusions are valid only on average base

Functions of statistics

Statistics simplifies complexity, presents facts in a definite form, helps in formulation of suitable policies, facilitates comparison and helps in forecasting.

Uses of statistics

Statistics has pervaded almost all spheres of human activities. Statistics is useful in the administration of various states, Industry, business, economics, research workers, banking, insurance companies etc.

Collection of data

Data can be collected by using sampling methods or experiments.

Data

The information collected through censuses and surveys or in a routine manner or other sources is called a raw data. When the raw data are grouped into groups or classes, they are known as grouped data.

There are two types of data

1. Primary data
2. Secondary data.

Primary data

The data which is collected by actual observation or measurement or count is called primary data.

Methods of collection of primary data

Primary data is collected in any one of the following methods

1. Direct personal interviews.
2. Indirect oral interviews
3. Information from correspondents.
4. Mailed questionnaire method.
5. Schedules sent through enumerators.

Secondary data

The data which are compiled from the records of others is called secondary data. The data collected by an individual or his agents is primary data for him and secondary data for all others. The secondary data are less expensive but it may not give all the necessary information.
Secondary data can be compiled either from published sources or from unpublished sources

Sources of published data

1. Official publications of the central, state and local governments.

2. Reports of committees and commissions.

3. Publications brought about by research workers and educational associations.

4. Trade and technical journals.

5. Report and publications of trade associations, chambers of commerce, bank etc.

6. Official publications of foreign governments or international bodies like U.N.O, UNESCO etc.

Sources of unpublished data

All statistical data are not published. For example, village level officials maintain records regarding area under crop, crop production etc. They collect details for

Variables

Variability is a common characteristic in biological Sciences. A quantitative or qualitative characteristic that varies from observation to observation in the same group is called a variable.

Quantitative data

The basis of classification is according to differences in quantity. In case of quantitative variables the observations are made in terms of kgs, Lt, cm etc. Example: weight of seeds, height of plants.

Qualitative data

When the observations are made with respect to quality is called qualitative data.

Eg: Crop varieties, Shape of seeds, soil type.

The qualitative variables are termed as attributes.

Classification of data

Classification is the process of arranging data into groups or classes according to the common characteristics possessed by the individual items.

Data can be classified on the basis of one or more of the following kinds namely

1. Geography
2. Chronology
3. Quality
4. Quantity

1. Geographical classification (or) Spatial Classification

Some data can be classified area-wise, such as states, towns etc.

Data on area under crop in India can be classified as shown below

Region	**Area (in hectares)**
Central India	-
West	-
North	-
East	-
South	-

2. Chronological or Temporal or Historical Classification

Some data can be classified on the basis of time and arranged chronologically or historically.
Data on Production of food grains in India can be classified as shown below

Year	**Tonnes**
1990-91	-
1991-92	-
1992-93	-
1993-94	-
1994-95	-

3. Qualitative Classification

Some data can be classified on the basis of attributes or characteristics. The number of farmers based on their land holdings can be given as follows

Type of farmers	**Number of farmers**
Marginal	907
Medium	1041
Large	1948
Total	**3896**

Qualitative classification can be of two types as follows

(i) Simple classification

(ii) Manifoldclassificatio

4. Quantitative classification

Some data can be classified in terms of magnitude. The data on land holdings by farmers in a block. Quantitative classification is based the land holding which is the variable in this example.

Land holding (hectare)	**Number of Farmers**
< 1	442
1-2	908
2-5	471
>5	124
Total	**1945**

Difference between Primary and secondary data

	Primary Data	**Secondary Data**
1. Original data	Primary data are original because investigation himself collects them.	Secondary data are not original since investigator makes use of the other agencies.
2. Suitability	If these data are collected accurately and systematically their suitability will be very positive.	These might or might not suit the objectives of enquiry.
3. Time and labour	These data involve large expenses in terms of money, time and manpower	These data are relatively less costly.
4. Precaution	don't need any great precaution while using these data.	These should be used with great care and caution.

Questions

1. A simple table contains data on

a) Two characteristics b) Several characteristics
c) One characteristic d) Three characteristics

Ans: One characteristic

2. When the collected data is grouped with reference to time, we have

a) Quantitative classification b) Qualitative classification
c) Geographical Classification d) Chorological Classification

Ans: Chorological Classification

3. Geographical classification means, classification of data according to Region.

Ans: True

4. An arrangement of data into rows and columns is known as Tabulation.

Ans: True

5. Data on yield is a quantitative variable

Ans: True

6. Qualitative variables are also called as attributes.

Ans: True

7. Define primary and secondary data

8. Give the advantages of tabulation.

9. Write a detail note on the types of classification

10. What are the essential characteristics of a good table?

11. Write the limitations of Statistics.

12. Difference between qualitative and quantitative data.

Lecture.2

Frequency distribution: Construction of Frequency Distribution table, Diagrammatic representation of data

Construction of Frequency Distribution Table:

In statistics, a **frequency distribution** is a tabulation of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way the table summarizes the distribution of values in the sample.

The following steps are used for construction of frequency table
Step-1: The number of classes are to be decided

The appropriate number of classes may be decided by Yule's formula, which is as follows:
Number of classes = $2.5 \times n^{1/4}$. Where " n" is the total number of observations

Step-2: The class interval is to be determined. It is obtain by using the relationship

$$C.I = \frac{\text{Maximum value in the given data} - \text{Minimum value in the given data}}{\text{Number of classes}}$$

Step-3: The frequencies are counted by using Tally marks

Step-4: The frequency table can be made by two methods

a) Exclusive method

b) Inclusive method

a) **Exclusive method:** In this method, the upper limit of any class interval is kept the same as the lower limit of the just higher class or there is no gap between upper limit of one class and lower limit of another class. It is continuous distribution.
Ex:

C.I.	Tally marks	Frequency (f)
0-10		
10-20		
20-30		

b) **Inclusive method:** There will be a gap between the upper limit of any class and the lower limit of the just higher class. It is discontinuous distribution

Ex:

C.I.	Tally marks	Frequency (f)
0-9		
10-19		
20-29		

To convert discontinuous distribution to continuous distribution by subtracting 0.5 from lower limit and by adding 0.5 to upper limit

Note: The arrangement of data into groups such that each group will have some numbers. These groups are called class and number of observations against these groups are called frequencies.

Each class interval has two limits 1. Lower limit and 2. Upper limit

The difference between upper limit and lower limit is called length of class interval. Length of class interval should be same for all the classes. The average of these two limits is called mid value of the class.

Example: Construct a frequency distribution table for the following data

25, 32, 45, 8, 24, 42, 22, 12, 9, 15, 26, 35, 23, 41, 47, 18, 44, 37, 27, 46, 38, 24, 43, 46, 10, 21, 36, 45, 22, 18.

Solution: Number of observations (n) = 30

Number of classes $= 2.5 \times n^{1/4}$

$= 2.5 \times 30^{1/4}$

$= 2.5 \times 2.3$

$= 5.8 - 6.0$

$$\text{Class interval} = \frac{\textit{Max.value - Min.value}}{\textit{No.of .classes}}$$

$$= \frac{47 - 8}{6}$$

$$= \frac{39}{6} = 6.5 \sim 6$$

Inclusive method:

C.I.	Tally marks	Frequency (f)
8-14	||||	4
15-21	||||	4
22-28	||||||	8
29-35	||	2
36-42	||||	5
42-49	|||| ||	7
Total		30

Exclusive method:

C.I.	Tally marks	Frequency (f)
7.5-14.5	||||	4
14.5-21.5	||||	4
21.5-28.5	|||| |||	8
28.5-35.5	||	2
35.5-42.5	||||	5
42.5-49.5	|||| ||	7
Total		30

Diagrams

Diagrams are various geometrical shape such as bars, circles etc. Diagrams are based on scale but are not confined to points or lines. They are more attractive and easier to understand than graphs.

Merits

1. Most of the people are attracted by diagrams.
2. Technical Knowledge or education is not necessary.
3. Time and effort required are less.
4. Diagrams show the data in proper perspective.
5. Diagrams leave a lasting impression.
6. Language is not a barrier.
7. Widely used tool.

Demerits (or) limitations

1. Diagrams are approximations.
2. Minute differences in values cannot be represented properly in diagrams.
3. Large differences in values spoil the look of the diagram.
4. Some of the diagrams can be drawn by experts only. eg. Pie chart.
5. Different scales portray different pictures to laymen.

Types of Diagrams

The important diagrams are

1. Simple Bar diagram.
2. Multiple Bar diagram.
3. Component Bar diagram.
4. Percentage Bar diagram.
5. Pie chart
6. Pictogram
7. Statistical maps or cartograms.

In all the diagrams and graphs, the groups or classes are represented on the x-axis and the volumes or frequencies are represented in the y-axis.

Simple Bar diagram

If the classification is based on attributes and if the attributes are to be compared with respect to a single character we use simple bar diagram.

Example

1. The area under different crops in a state.
2. The food grain production of different years.
3. The yield performance of different varieties of a crop.
4. The effect of different treatments etc.

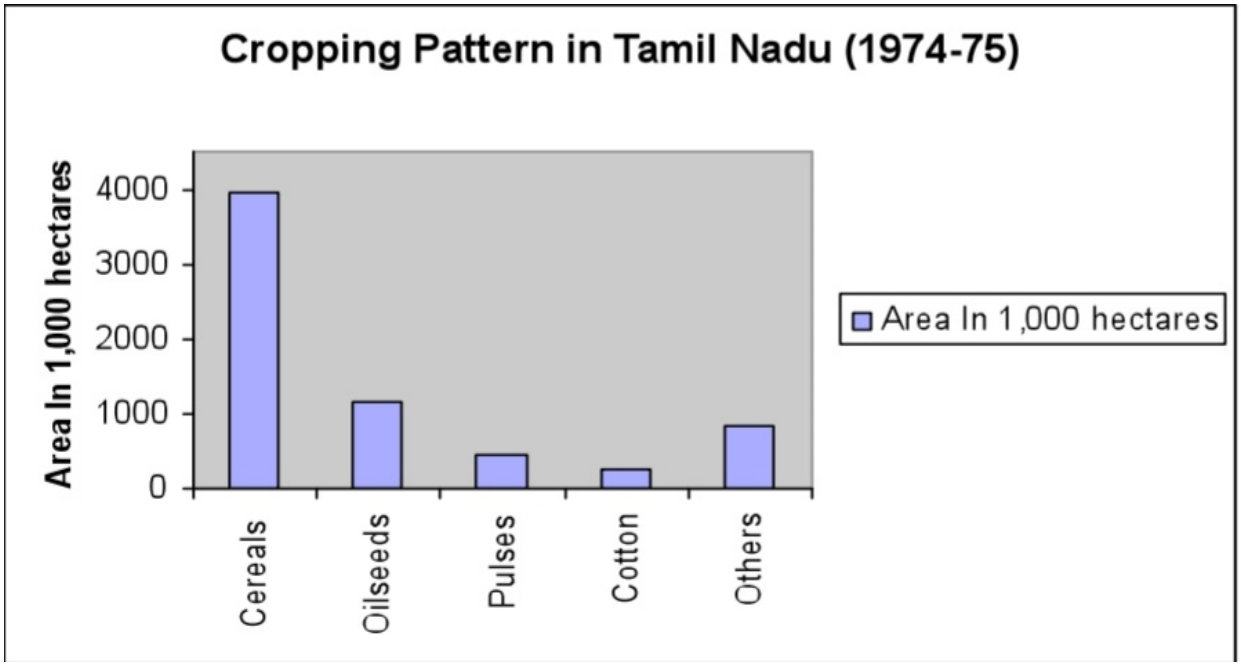
Simple bar diagrams Consists of vertical bars of equal width. The heights of these bars are proportional to the volume or magnitude of the attribute. All bars stand on the same baseline. The bars are separated from each others by equal intervals. The bars may be coloured or marked.

Example

The cropping pattern in Tamil Nadu in the year 1974-75 was as follows.

Crops	**Area In 1,000 hectares**
Cereals	3940
Oilseeds	1165
Pulses	464
Cotton	249
Others	822

The simple bar diagram for this data is given below.



Multiple bar diagram

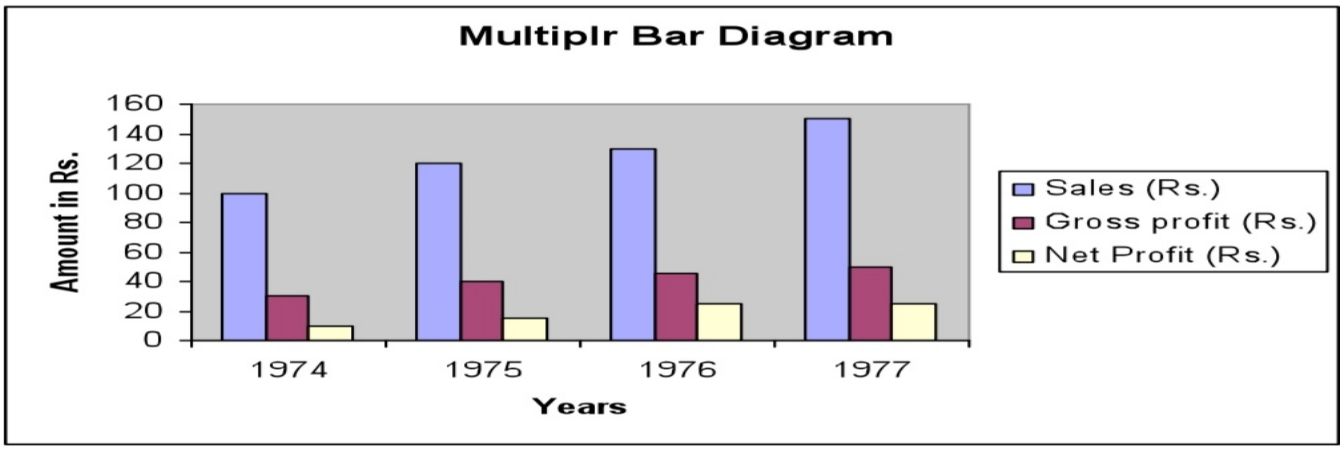
If the data is classified by attributes and if two or more characters or groups are to be compared within each attribute we use multiple bar diagrams. If only two characters are to be compared within each attribute, then the resultant bar diagram used is known as double bar diagram.

The multiple bar diagram is simply the extension of simple bar diagram. For each attribute two or more bars representing separate characters or groups are to be placed side by side. Each bar within an attribute will be marked or coloured differently in order to distinguish them. Same type of marking or colouring should be done under each attribute. A footnote has to be given explaining the markings or colourings.

Example

Draw a multiple bar diagram for the following data which represented agricultural production for the priod from 2000-2004

Year	Food grains (tones)	Vegetables (tones)	Others (tones)
2000	100	30	10
2001	120	40	15
2002	130	45	25
2003	150	50	25
2004			



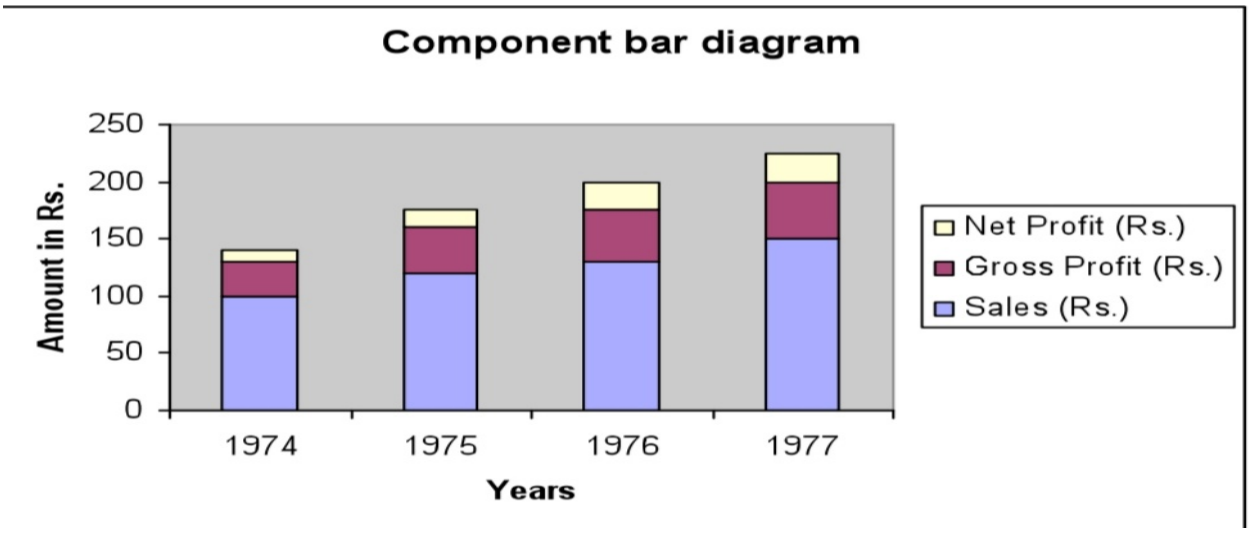
Component bar diagram

This is also called sub – divided bar diagram. Instead of placing the bars for each component side by side we may place these one on top of the other. This will result in a component bar diagram.

Example:

Draw a component bar diagram for the following data

Year	Sales (Rs.)	Gross Profit (Rs.)	Net Profit (Rs.)
1974	100	30	10
1975	120	40	15
1976	130	45	25
1977	150	50	25



Percentage bar diagram

Sometimes when the volumes of different attributes may be greatly different for making meaningful comparisons, the attributes are reduced to percentages. In that case each attribute will have 100 as its maximum volume. This sort of component bar chart is known as percentage bar diagram.

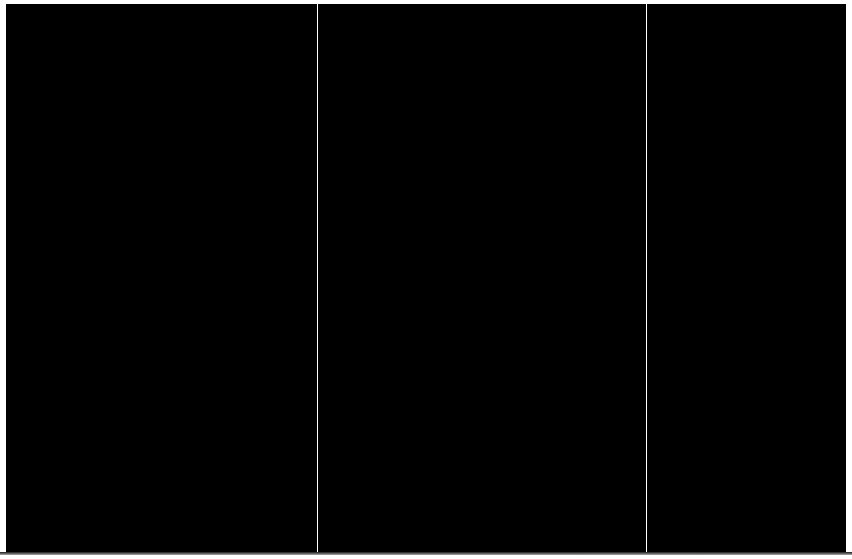
Percentage = $\frac{\text{Actual value}}{\text{Total of the actual value}} \text{x} 100,$

Example:

Draw a Percentage bar diagram for the following data

Using the formula Percentage = $\frac{\text{Actual value}}{\text{Total of the actual value}} \text{x} 100$, the above table is converted.

Year	Sales (Rs.)	Gross Profit (Rs.)	Net Profit (Rs.)
1974	71.43	21.43	7.14
1975	68.57	22.86	8.57
1976	65	22.5	12.5
1977	66.67	22.22	11.11



Pie chart / Pie Diagram

Pie diagram is a circular diagram. It may be used in place of bar diagrams. It consists of one or more circles which are divided into a number of sectors. In the construction of pie diagram the following steps are involved. Step 1:

Whenever one set of actual value or percentage are given, find the corresponding angles in degrees using the following formula

$$\text{Angle} = \frac{\text{Actual value}}{\text{Total of the actual value}} x 360^0$$

(or) $$\text{Angle} = \frac{\text{Percentage}}{100} x 360^0$$

Step 2:

Find the radius using the area of the circle πr^2 where value of π is 22/7 or 3.14 **Example**

Given the cultivable land area in four southern states of India. Construct a pie diagram for the following data.

State	Cultivable area(in hectares)
Andhra Pradesh	663
Karnataka	448
Kerala	290
Tamil Nadu	556
Total	**1957**

Using the formula

$$\text{Angle} = \frac{\text{Actual value}}{\text{Tota lof the actual value}} \text{x } 360^0$$

(or)

$$\text{Angle} = \frac{\text{Percentage}}{100} \text{x } 360^0$$

The table value becomes

State	Cultivable area
Andhra Pradesh	121.96
Karnataka	82.41
Kerala	53.35
Tamil Nadu	102.28

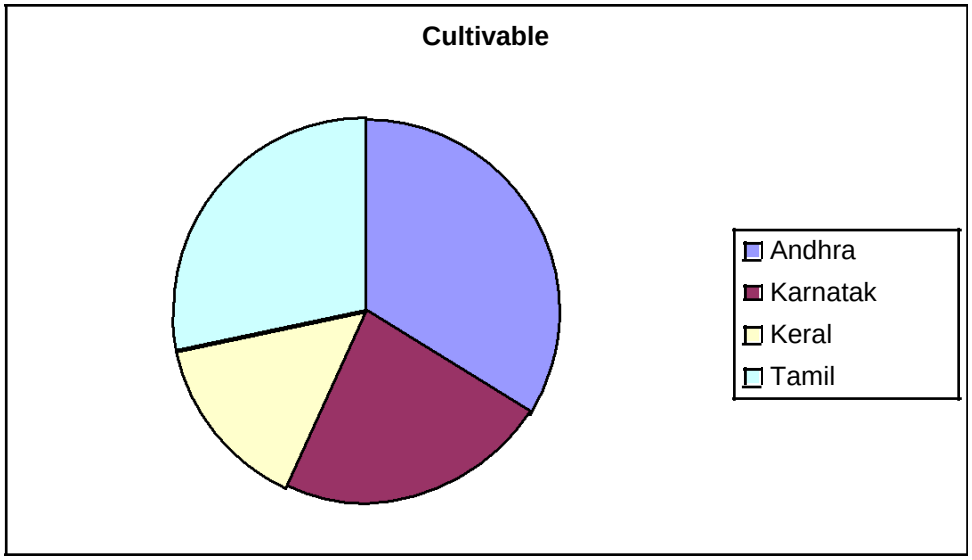
Radius = πr^2

Here πr^2 =1957

$r^2 = \frac{1957}{\pi} = \frac{1957}{3.14} = 623$

r = 24.96

r= 25 (approx)



Questions

1. In a component bar diagram the length of the bar

   a) Will be same for all c) will not be same

   b) Depends on the total d) none of these

   Ans: Depends on the total

2. The length of the bar will be same for all categories in

   a) Multiple bar diagram b) component bar diagram c)

   Percentage bar diagram d) none of these

   Ans: Percentage bar diagram

3. Sub-divided bar diagram are also called Component bar diagram.

   Ans: True

4. The multiple bar diagram is the extension of simple bar diagram.

   Ans: True

5. In a bar the width of the bars should be equal.

   Ans: True

6. In a percentage bar diagram the length of the bars will not be equal.

   Ans: False

7. How diagrams are useful in representing statistical data?
8. How to draw a pie chart?
9. Explain how to draw simple and multiple bar diagrams.
10. Explain how to draw Component and percentage bar diagrams.

Graphical representation – Histogram – Frequency polygon and Frequency curve

Graphs

Graphs are charts consisting of points, lines and curves. Charts are drawn on graph sheets. Suitable scales are to be chosen for both x and y axes, so that the entire data can be presented in the graph sheet. Graphical representations are used for grouped quantitative data.

Histogram

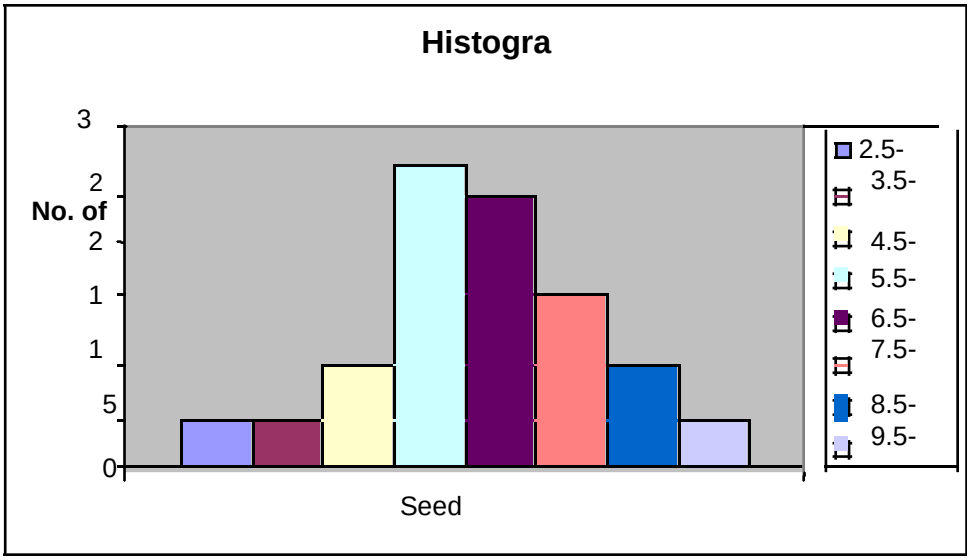
When the data are classified based on the class intervals it can be represented by a histogram. Histogram is just like a simple bar diagram with minor differences. There is no gap between the bars, since the classes are continuous. The bars are drawn only in outline without colouring or marking as in the case of simple bar diagrams. It is the suitable form to represent a frequency distribution.

Class intervals are to be presented in x axis and the bases of the bars are the respective class intervals. Frequencies are to be represented in y axis. The heights of the bars are equal to the corresponding frequencies.

Example

Draw a histogram for the following data

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



Frequency Polygon

The frequencies of the classes are plotted by dots against the mid-points of each class. The adjacent dots are then joined by straight lines. The resulting graph is known as frequency polygon.

Example

Draw frequency polygon for the following data

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

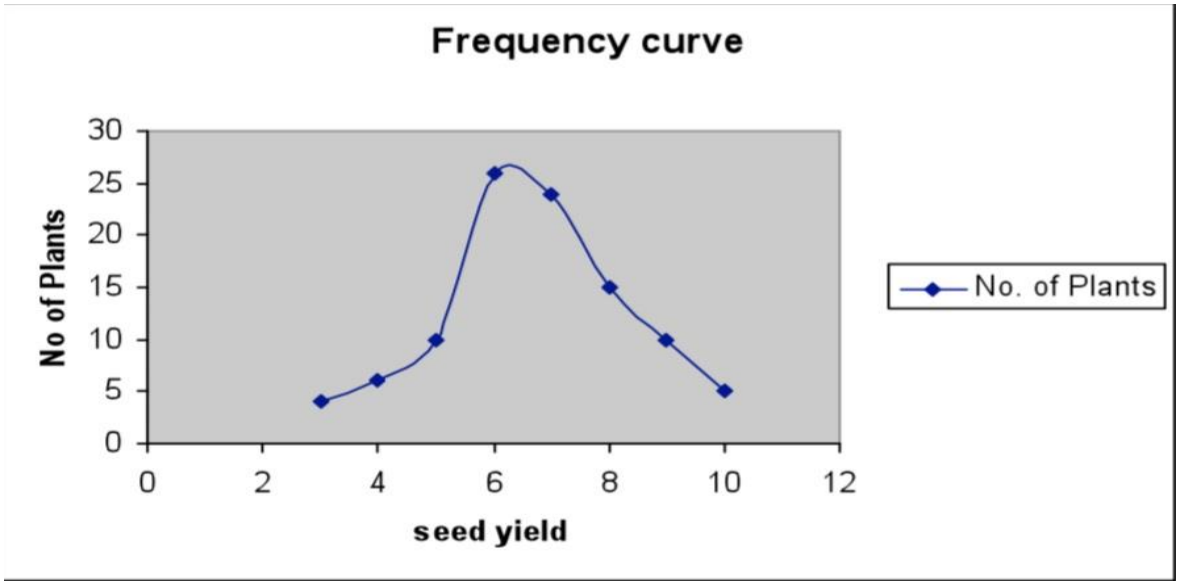
Frequency curve

The procedure for drawing a frequency curve is same as for frequency polygon. But the points are joined by smooth or free hand curve.

Example

Draw frequency curve for the following data

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



Ogives

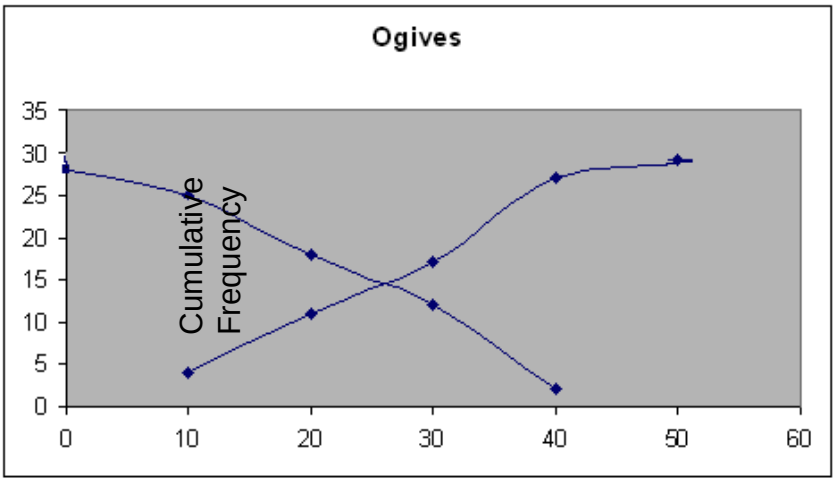
Ogives are known also as cumulative frequency curves and there are two kinds of ogives. One is less than ogive and the other is more than ogive.

Less than ogive: Here the cumulative frequencies are plotted against the upper boundary of respective class interval.

Greater than ogive: Here the cumulative frequencies are plotted against the lower boundaries of respective class intervals.

Example

Continuous Interval	Mid Point	Frequency	< cumulative Frequency	> cumulative frequency
0-10	5	4	4	29
10-20	15	7	11	25
20-30	25	6	17	18
30-40	35	10	27	12
40-50	45	2	29	2



Boundary values

Questions

1. With the help of histogram we can draw

   (a) Frequency polygon (b) frequency curve

   (c) Frequency distribution (d) all the above

   Ans: all the above

2. Ogives for more than type and less than type distribution intersect at

   (a) Mean (b) median

   (c) Mode (d) origin

   Ans: median

3. To draw the frequency polygon we take the mid values in the X axis.

4. To draw the frequency polygon we take the mid values in the X axis.

5. In a frequency curve the points are joined by bits of straight lines

   Ans: False

6. He stogram can be drawn for equal and unequal classes

   Ans: True

7. Explain how to draw frequency curve

8. Explain how to draw histogram.

9. Explain the diagrams that can be drawn for a frequency distribution table

10. Explain how to draw less than and more than Ogives.

Measures of central tendacy

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations. There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

Arithmetic mean or mean

Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. It is denoted by the symbol $\bar{x}$ If the variable x assumes n values x1, x2 … xn then the mean is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

This formula is for the ungrouped or raw data.

Example 1

Calculate the mean for pH levels of soil 6.8, 6.6, 5.2, 5.6, 5.8

Solution

$$\bar{x} = \frac{6.8 + 6.6 + 5.2 + 5.6 + 5.8}{5} = \frac{30}{5} = 6$$

Grouped Data

The mean for grouped data is obtained from the following formula:

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

Where x = the mid-point of individual class

f = the frequency of individual class

n = the sum of the frequencies or total frequencies in a sample.

Short-cut method

$$\bar{x} = A + \frac{\sum fd}{n} x\ c$$

Where $d = \frac{x - A}{c}$

A = any value in x

n = total frequency

c = width of the class interval

Example 2

Given the following frequency distribution, calculate the arithmetic mean

Marks : 64 63 62 61 60 59

Number of Students : 8 18 12 9 7 6

Solution

X	F	Fx	D=x-A	Fd
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	-1	-9
60	7	420	-2	-14
59	6	354	-3	-18
	60	3713		-7

Direct method

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

$$\bar{x} = \frac{3713}{60} = 61.88$$

Short-cut method

$$\bar{x} = A + \frac{\sum fd}{n} x\ c$$

Here A = 62

$$\bar{x} = 62 - \frac{7}{60} x\ 1 = 61.88$$

Example 3

For the frequency distribution of seed yield of seasamum given in table, calculate the mean yield per plot.

Yield per plot in(in g)	64.5-84.5	84.5-104.5	104.5-124.5	124.5-144.5
No of plots	3	5	7	20

Solution

Yield (in g)	**No of Plots (f)**	**Mid X**	$d = \frac{x - A}{c}$	**Fd**
64.5-84.5	3	74.5	-1	-3
84.5-104.5	5	94.5	0	0
104.5-124.5	7	114.5	1	7
124.5-144.5	20	134.5	2	40
Total	**35**			**44**

A=94.5

The mean yield per plot is

Direct method:

$$\bar{x} = \frac{\sum fx}{n} = \frac{(74.5 \times 3) + (94.5 \times 5) + (114.5 \times 7) + (134.5 \times 20)}{35}$$

$$= \frac{4187.5}{35} = 119.64 \text{ gms}$$

Shortcut method

$$\bar{x} = A + \frac{\sum fd}{n} x\ c$$

$$\bar{x} = 94.5 + \frac{44}{35} x\ 20 = 119.64\ g$$

Merits and demerits of Arithmetic mean

Merits

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

Median

The median is the middle most item that divides the group into two equal parts, one part comprising all values greater, and the other, all values less than that item.

Ungrouped or Raw data

Arrange the given values in the ascending order. If the number of values are odd, median is the middle value

If the number of values are even, median is the mean of middle two values.

By formula

When n is odd, Median = Md = $\left(\frac{n+1}{2}\right)^{th} value$

When n is even, Average of $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)^{th}$ $value$

Example 4

If the weights of sorghum ear heads are 45, 60,48,100,65 gms, calculate the median

Solution

Here n = 5

First arrange it in ascending order

45, 48, 60, 65, 100

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

$= \left(\frac{5+1}{2}\right) = 3^{rd}$ $value$ =60

Example 5

If the sorghum ear- heads are 5,48, 60, 65, 65, 100 gms, calculate the median.

Solution

Here n = 6

Meadian = Average of $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)^{th}$ $value$

$\left(\frac{n}{2}\right) = \frac{6}{2} = 3^{rd}\ value = 60$ and $\left(\frac{n}{2}+1\right) = \frac{6}{2}+1 = 4^{th}\ value = 65$

Median = $\frac{60+65}{2} = 62.5\,g$

Grouped data

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series

Step1: Find cumulative frequencies.

Step2 : Find $\left(\frac{n}{2}+1\right)$

Step3: See in the cumulative frequencies the value just greater than $\left(\frac{n}{2}+1\right)$

Step4: Then the corresponding value of x is median.

Example 6

The following data pertaining to the number of insects per plant. Find median number of insects per plant.

Number of insects per plant (x)	1	2	3	4	5	6	7	8	9	10	11	12
No. of plants(f)	1	3	5	6	10	13	9	5	3	2	2	1

Solution

Form the cumulative frequency table

x	f	cf
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
	60	

Median = size of $\left(\frac{n+1}{2}\right)^{th} item$

Here the number of observations is even. Therefore median = average of (n/2)th item and (n/2+1)th item.

= (30^{th} item +31^{st} item) / 2 = (6+6)/2 = 6

Hence the median size is 6 insects per plant.

Continuous Series

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{n}{2}\right)$

Step3: See in the cumulative frequency the value first greater than n/2, then the corresponding class interval is called the Median class. Then apply the formula

Median = $l + \frac{\frac{n}{2} - m}{f} \times c$

where l = Lower limit of the median class

m = cumulative frequency preceding the median class

c = width of the class

f =frequency in the median class.

n=Total frequency.

Example 7

For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the median.

Weights of ear heads (in g)	**No of ear heads (f)**	**Less than class**	**Cumulative frequency (m)**
60-80	22	<80	22
80-100	38	<100	60
100-120	45	<120	105
120-140	35	<140	140
140-160	24	<160	164
Total	164		

Solution

Median = $l + \frac{\frac{n}{2} - m}{f} \times c$

$\left(\frac{n}{2}\right) = \left(\frac{164}{2}\right) = 82$

It lies between 60 and 105. Corresponding to 60 the less than class is 100 and corresponding to 105 the less than class is 120. Therefore the median class is 100-120. Its lower limit is 100. Here l = 100, n=164 , f = 45 , c = 20, m =60

Median = $100 + \frac{82 - 60}{45} \times 20 = 109$

Merits of Median

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.

Demerits of Median

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in calculating mean deviation.
4. It does not take into account all the observations.

Mode

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it. It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in agriculture like to find typical height of a crop variety, maximum source of irrigation in a region, maximum disease prone paddy variety. Thus the mode is an important measure in case of qualitative data.

Computation of the mode

Ungrouped or Raw Data

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example 8

Find the mode for the following seed weight

2 , 7, 10, 15, 10, 17, 8, 10, 2 gms

Mode = 10

In some cases the mode may be absent while in some cases there may be more than one mode.

Example 9

(1) 12, 10, 15, 24, 30 (no mode)

(2) 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

the modal values are 7 and 10 as both occur 3 times each.

Grouped Data

For Discrete distribution, see the highest frequency and corresponding value of x is mode.

Example:

Find the mode for the following

Weight of sorghum in gms (x)	No. of ear head(f)
50	4
65	6
75	16
80	8
95	7
100	4

Solution

The maximum frequency is 16. The corresponding x value is 75.

mode = 75 gms.

Continuous distribution

Locate the highest frequency the class corresponding to that frequency is called the modal class. Then apply the formula.

Mode = $l + \frac{f_s}{f_p + f_s} \times c$

Where l = lower limit of the model class

f_p = the frequency of the class preceding the model class

f = the frequency of the class succeeding the model class

and c = class interval

Example 10

For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the mode

Weights of ear heads (g)	**No of ear heads (f)**	
60-80	22	
80-100	38	f_p
100-120	45	f
120-140	35	f_s
140-160	20	
Total	**160**	

Solution

Mode = $l + \frac{f_s}{f_p + f_s} \times c$

Here l=100, f = 45, c = 20, m =60, f_p=38, f_s=35

Mode = $100 + \frac{35_s}{38 + 35} \times 20$

$= 100 + \frac{35_s}{73} \times 20$

= 109.589

Geometric mean

The geometric mean of a series containing n observations is the nth root of the product of the values.

If x_1, x_2…, x_n are observations then

G.M= $\sqrt[n]{x_1, x_2 \ldots x_n}$

=

Log GM = $\frac{1}{n}\log(x_1, x_2 \ldots x_n)$

= $\frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n)$

= $\frac{\sum \log x_i}{n}$

GM = Antilog $\frac{\sum \log x_i}{n}$

For grouped data

GM = Antilog $\left[\frac{\sum f \log x_i}{n}\right]$

GM is used in studies like bacterial growth, cell division, etc.

Example 11

If the weights of sorghum ear heads are 45, 60, 48,100, 65 gms. Find the Geometric mean for the following data

Weight of ear head x (g)	**Log x**
45	1.653
60	1.778
48	1.681
100	2.000
65	1.813
Total	**8.925**

Solution

Here n = 5

GM = Antilog $\frac{\sum \log x_i}{n}$

= Antilog $\frac{8.925}{5}$

= Antilog 1.785 = 60.95

Grouped Data

Example 12

Find the Geometric mean for the following

Weight of sorghum (x)	No. of ear head(f)
50	4
65	6
75	16
80	8
95	7
100	4

Solution

Weight of sorghum (x)	No. of ear head(f)	Log x	f x log x
50	5	1.699	8.495
63	10	10.799	17.99
65	5	1.813	9.065
130	15	2.114	31.71
135	15	2.130	31.95
Total	**50**	**9.555**	**99.21**

Here n= 50

$$GM = \text{Antilog}\left[\frac{\left[\sum f \log x_i\right]}{n}\right]$$

$$= \text{Antilog}\left[\frac{99.21}{50}\right]$$

= Antilog 1.9842 = 96.43

Continuous distribution

Example 13

For the frequency distribution of weights of sorghum ear-heads given in table below. Calculate the Geometric mean

Weights of ear heads (in g)	**No of ear heads (f)**
60-80	22
80-100	38
100-120	45

120-140	35
140-160	20
Total	**160**

Solution

Weights of ear heads (in g)	**No of ear heads (f)**	**Mid x**	**Log x**	**f log x**
60-80	22	70	1.845	40 59
80-100	38	90	1.954	74.25
100-120	45	110	2.041	91.85
120-140	35	130	2.114	73.99
140-160	20	150	2.176	43.52
Total	**160**			**324.2**

Here $n = 160$

$$GM = \text{Antilog}\left[\frac{\sum f \log x_i}{n}\right]$$

$$= \text{Antilog}\left[\frac{324.2}{160}\right]$$

$$= \text{Antilog}\left[2.02625\right]$$

$$= 106.23$$

Harmonic mean (H.M)

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x1, x2…..xn are n observations,

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

For a frequency distribution

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

H.M is used when we are dealing with speed, rates, etc.

Example 13

From the given data 5, 10,17,24,30 calculate H.M.

X	$\frac{1}{x}$
5	0.2000
10	0.1000
17	0.0588
24	0.0417
30	0.4338

$$\text{H.M} = \frac{5}{0.4338} = 11.526$$

Example 14

Number of tomatoes per plant are given below. Calculate the harmonic mean.

Number of tomatoes per plant	20	21	22	23	24	25
Number of plants	4	2	7	1	3	1

Solution

tomatoes per plant (x)	No of plants(f)	$\frac{1}{x}$	$f\left(\frac{1}{x}\right)$
20	4	0.0500	0.2000
21	2	0.0476	0.0952
22	7	0.0454	0.3178
23	1	0.0435	0.0435
24	3	0.0417	0.1251
25	1	0.0400	0.0400
	18		0.8216

$$\text{H.M} = \frac{n}{\sum f\left(\frac{1}{x_i}\right)} = \frac{18}{0.1968} = 21.91$$

Merits of H.M

1. It is rigidly defined.

2. It is defined on all observations.

3. It is amenable to further algebraic treatment.

4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

Demerits of H.M

1. It is not easily understood.
2. It is difficult to compute.
3. It is only a summary figure and may not be the actual item in the series
4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.
5. It is rarely used in grouped data.

Percentiles

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The x^{th} percentile is that value below which x percent of values in the distribution fall. It may be noted that the median is the 50^{th} percentile.

For raw data, first arrange the n observations in increasing order. Then the x^{th} percentile is given by

$$P_x = \left(\frac{x(n+1)}{100}\right)^{th} \text{ item}$$

For a frequency distribution the x^{th} percentile is given by

$$P_x = l + \left(\frac{(x.n/100) - cf}{f} \times C\right)$$

Where

l= lower limit of the percentile calss which contains the x^{th} percentile value (x. n /100)

cf = cumulative frequency uotp l

f = frequency of the percentile class

C= class interval

N= total number of observations

Percentile for Raw Data or Ungrouped Data

Example 15

The following are the paddy yields (kg/plot) from 14 plots:

30,32,35,38,40.42,48,49,52,55,58,60,62,and 65 (after arranging in ascending order). The computation of 25^{th} percentile (Q_1) and 75^{th} percentile (Q_3) are given below:

$$P_{25}(or\ Q_1) = \left(\frac{25(14+1)}{100}\right)^{th} \text{ item}$$

$$= \left(3\frac{3}{4}\right)^{th} \text{ item}$$

$$= 3^{rd} \text{ item} + (4^{th} \text{ item} - 3^{rd} \text{ item}) \left(\frac{3}{4}\right)$$

$$= 35 + (38\text{-}35) \left(\frac{3}{4}\right)$$

$$= 35 + 3 \left(\frac{3}{4}\right) = 37.25 \text{ kg}$$

$$P_{75}(or\ Q_3) = \left(\frac{75(14+1)}{100}\right)^{th} \text{ item}$$

$$= \left(11\frac{1}{4}\right)^{th} \text{ item}$$

$$= 11^{th} \text{ item} + (12^{th} \text{ item} - 11^{th} \text{ item}) \left(\frac{1}{4}\right)$$

$$= 55 + (58\text{-}55) \left(\frac{1}{4}\right)$$

$$= 55 + 3 \left(\frac{1}{4}\right) = 55.75 \text{ kg}$$

Example 16

The frequency distribution of weights of 190 sorghum ear-heads are given below. Compute 25th percentile and 75th percentile.

Weight of ear-heads (in g)	**No of ear heads**
40-60	6
60-80	28
80-100	35
100-120	55
120-140	30
140-160	15
160-180	12
180-200	9
Total	**190**

Solution

Weight of ear-heads (in g)	No of ear heads	Less than class	Cumulative frequency
40-60	6	< 60	6
60-80	28	< 80	34
80-100	35	<100	69
100-120	55	<120	124
120-140	30	<140	154
140-160	15	<160	169
160-180	12	<180	181
180-200	9	<200	190
Total	**190**		

47.5

142.5

or P_{25}, first find out $\left(\frac{25(190)}{100}\right)$, and for P_{75}, $\left(\frac{75(190)}{100}\right)$, and proceed as in the case of median.

For P_{25}, we have $\left(\frac{25(190)}{100}\right) = 47.5$.

The value 47.5 lies between 34 and 69. Therefore, the percentile class is 80-100. Hence,

$$P_{25} = Q_1 = l + \left(\frac{(25.n/100) - cf}{f} \times C\right)$$

$$= 80 + \left(\frac{(47.5) - 34}{35} \times 20\right)$$

$$= 80 + \left(\frac{(13.5)}{35} \times 20\right)$$

= 80 +7.71 or 87.71 g.

Similarly,

$$P_{75} = l + \left(\frac{(75.n/100) - cf}{f} \times C\right) \text{Class}$$

$$= 120 + \left(\frac{(142.5) - 121}{30} \times 20\right)$$

$$= 120 + \left(\frac{(21.5)}{30} \times 20\right)$$

= 120 +14.33 =134.33 g.

Quartiles

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower).quartile (Q1) marks off the first one-fourth, the third (upper) quartile (Q3) marks off the three-fourth. It may be noted that the second quartile is the value of the median and 50^{th} percentile.

Raw or ungrouped data

First arrange the given data in the increasing order and use the formula for Q1 and Q3 then quartile deviation, Q.D is given by

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

Where $Q_1. = \left(\frac{n+1}{4}\right)^{th}$ item and $Q_3. = 3\left(\frac{n+1}{4}\right)^{th}$ item

Example 18

Compute quartiles for the data given below (grains/panicles) 25, 18, 30, 8, 15, 5, 10, 35, 40, 45

Solution

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

$$Q_1. = \left(\frac{n+1}{4}\right)^{th}$$

$$= \left(\frac{10+1}{4}\right)^{th}$$

$= (2.75)^{th}$ item

$= 2^{nd}$ item + $\left(\frac{3}{4}\right)$(3^{rd} item – 2^{nd} item)

$= 8 + \frac{3}{4}(10-8)$

$= 8 + \frac{3}{4} \times 2$

= 8+1.5

= 9.5

$Q_3 = 3\left(\frac{n+1}{4}\right)^{th}$

= 3 x (2.75) th item

= (8.75)th item

= 8th item + $\left(\frac{1}{4}\right)$(9th item – 8th item)

= 35+ $\frac{1}{4}$ (40-35)

= 35+1.25

= 36.25

Discrete Series

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{n+1}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{n+1}{4}\right)$, then the corresponding value of *x* is Q1

Step4: Find $3\left(\frac{n+1}{4}\right)$

Step5: See in the cumulative frequencies, the value just greater than $3\left(\frac{n+1}{4}\right)$,then the corresponding value of x is Q3

Example 19

Compute quartiles for the data given bellow (insects/plant).

X	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

Solution

x	f	cf
5	4	4
8	3	7
12	2	9
15	4	13
19	5	18
24	2	20

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} item = \left(\frac{24+1}{4}\right) = \left(\frac{25}{4}\right) = 6.25^{th} item$$

$$Q_3 = 3\left(\frac{n+1}{4}\right)^{th} item = 3\left(\frac{24+1}{4}\right) = 18.75\text{th item} \qquad Q1= 8;\ Q3=24$$

Continuous series

Step1: Find cumulative frequencies

Step2: Find $\left(\frac{n}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{n}{4}\right)$, then the corresponding class interval is called first quartile class.

Step4: Find $3\left(\frac{n}{4}\right)$ See in the cumulative frequencies the value just greater than $3\left(\frac{n}{4}\right)$ then the corresponding class interval is called 3rd quartile class. Then apply the respective formulae

$$Q_1 = l_1 + \frac{\frac{n}{4} - m_1}{f_1} \times c_1$$

$$Q_3 = l_3 + \frac{3\left(\frac{n}{4}\right) - m_3}{f_3} \times c_3$$

Where l_1 = lower limit of the first quartile class

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = c.f. preceding the first quartile class

l_3 = 1ower limit of the 3rd quartile class

f_3 = frequency of the 3rd quartile class

c_3 = width of the 3rd quartile class

m_3 = c.f. preceding the 3rd quartile class

Example 20: The following series relates to the marks secured by students in an examination.

Marks	No. of Students
0-10	11
10-20	18
20-30	25
30-40	28
40-50	30
50-60	33
60-70	22
70-80	15
80-90	12
90-100	10

Find the quartiles

Solution

C.I	f	cf
0-10	11	11
10-20	18	29
20-30	25	54
30-40	28	82
40-50	30	112
50-60	33	145
60-70	22	167
70-80	15	182
80-90	12	194
90-100	10	204
	204	

$$\left(\frac{n}{4}\right) = \left(\frac{204}{4}\right) = 51$$

$$3\left(\frac{n}{4}\right) = 153$$

$$Q_1 = l_1 + \frac{\frac{n}{4} - m_1}{f_1} \times c_1$$

$$= 20 + \frac{51 - 29_1}{25_1} \times 10 = 20 + 8.8 = 28.8$$

$$Q_3 = l_3 + \frac{3\left(\frac{n}{4}\right) - m_3}{f_3} \times c_3$$

$$= 60 + \frac{153 - 145}{22} \times 12 = 60 + 4.36 = 64.36$$

Questions

1. The middle value of an ordered series is called

a)2nd quartile b) 5th decile

c) 50th percentile d) all the above

Ans: all the above

2. For a set of values the model value can be

a) Unimodel b) bimodal

c) Trimodel d) All of these

d) Ans: all the above

3. Mode is suitable for qualitative data.

Ans: True

4. Decile divides the group in to ten equal parts.

Ans: True

5. Mean is affected by extreme values.

Ans: True

6. Geometric mean can be calculated for negative values.

Ans: False

7. Define mean and median

8. For what type of data mode can be calculated.

9. Explain how to calculate the arithmetic mean for raw and grouped data.

10. Explain how to calculate median and mode for grouped data.

Measures of Dispersion

The averages are representatives of a frequency distribution. But they fail to give a complete picture of the distribution. They do not tell anything about the scatterness of observations within the distribution.

Suppose that we have the distribution of the yields (kg per plot) of two paddy varieties from 5 plots each. The distribution may be as follows

Variety I	45	42	42	41	40
Variety II	54	48	42	33	30

It can be seen that the mean yield for both varieties is 42 kg but cannot say that the performances of the two varieties are same. There is greater uniformity of yields in the first variety whereas there is more variability in the yields of the second variety. The first variety may be preferred since it is more consistent in yield performance.

Form the above example it is obvious that a measure of central tendency alone is not sufficient to describe a frequency distribution. In addition to it we should have a measure of scatterness of observations. The scatterness or variation of observations from their average are called the dispersion. There are different measures of dispersion like the range, the quartile deviation, the mean deviation and the standard deviation.

Characteristics of a good measure of dispersion

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate

Range

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S.

Where L = Largest value.

S = Smallest value.

In individual observations and discrete series, L and S are easily identified.

In continuous series, the following two methods are followed.

Method 1

L = Upper boundary of the highest class

S = Lower boundary of the lowest class.

Method 2

L = Mid value of the highest class.

S = Mid value of the lowest class.

Example1

The yields (kg per plot) of a cotton variety from five plots are 8, 9, 8, 10 and 11. Find the range

Solution

L=11, S = 8.

Range = L – S = 11- 8 = 3

Example 2

Calculate range from the following distribution.

Size:	60-63	63-66	66-69	69-72	72-75
Number:	5	18	42	27	8

Solution

L = Upper boundary of the highest class = 75

S = Lower boundary of the lowest class = 60

Range = L – S = 75 – 60 = 15

Merits and Demerits of Range

Merits

1. It is simple to understand.

2. It is easy to calculate.

3. In certain types of problems like quality control, weather forecasts, share price analysis, etc.,

range is most widely used.

Demerits

1. It is very much affected by the extreme items.

2. It is based on only two extreme observations.

3. It cannot be calculated from open-end class intervals.

4. It is not suitable for mathematical treatment.

5. It is a very rarely used measure.

Standard Deviation

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean.

The standard deviation is denoted by s in case of sample and Greek letter σ (sigma) in case of population.

The formula for calculating standard deviation is as follows

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$ for raw data

And for grouped data the formulas are

$$s = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$ for discrete data

$$s = C \times \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$ for continuous data

Where $d = \frac{x - A}{C}$

C = class interval

Example 3

Raw Data

The weights of 5 ear-heads of sorghum are 100, 102,118,124,126 gms. Find the standard deviation.

Solution

x	x^2
100	10000
102	10404
118	13924
124	15376
126	15876
570	**65580**

Standard deviation $s = \sqrt{\dfrac{\sum x^2 - \dfrac{(\sum x)^2}{n}}{n-1}}$

$$= \sqrt{\frac{65580 - \dfrac{(570)^2}{5}}{5-1}} = \sqrt{150} = 12.25 \ gms$$

Example 4

Discrete distribution

The frequency distributions of seed yield of 50 seasamum plants are given below. Find the standard deviation.

Seed yield in gms (x)	3	4	5	6	7
Frequency (f)	4	6	15	15	10

Solution

Seed yield in gms (x)	**f**	**fx**	**fx^2**
3	4	12	36
4	6	24	96
5	15	75	375
6	15	90	540
7	10	70	490
Total	**50**	**271**	**1537**

Here $n = 50$

Standard deviation $s = \sqrt{\dfrac{\sum fx^2}{n} - \left(\dfrac{\sum fx}{n}\right)^2}$

$$= \sqrt{\frac{1537}{50} - \left(\frac{271}{50}\right)^2}$$

$$= \sqrt{30.74 - 29.3764}$$

= 1.1677 gms

Example 5

Continuous distribution

The Frequency distributions of seed yield of 50 seasamum plants are given below. Find the standard deviation.

Seed yield in gms (x)	2.5-35	3.5-4.5	4.5-5.5	5.5-6.5	6.5-7.5
No. of plants (f)	4	6	15	15	10

Solution

Seed yield in gms (x)	**No. of Plants f**	**Mid x**	**d= $\frac{x-A}{C}$**	**df**	**d^2 f**
2.5-3.5	4	3	-2	-8	16
3.5-4.5	6	4	-1	-6	6
4.5-5.5	15	5	0	0	0
5.5-6.5	15	6	1	15	15
6.5-7.5	10	7	2	20	40
Total	**50**	**25**	**0**	**21**	**77**

A=Assumed mean = 5

n=50, C=1

$$s = C \times \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$$

$$= 1 \times \sqrt{\frac{77}{50} - \left(\frac{21}{50}\right)^2}$$

$$= \sqrt{1.54 - 0.1764}$$

$$= \sqrt{1.3636} = 1.1677$$

Merits and Demerits of Standard Deviation

Merits

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling and hence stable.
6. It is the basis for measuring the coefficient of correlation and sampling.

Demerits

1. It is not easy to understand and it is difficult to calculate.
2. It gives more weight to extreme values because the values are squared up.

3. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.

Variance

The square of the standard deviation is called variance.

(i.e.) variance = $(SD)^2$.

Coefficient of Variation

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of plants cannot be compared with the standard deviation of weights of the grains, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation. The coefficient of variation is obtained by dividing the standard deviation by the mean and expressed in percentage. Symbolically, Coefficient of variation (C.V) = $\frac{SD}{mean} \times 100$

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable or more stable or more uniform or more consistent or more homogeneous.

Example 6

Consider the measurement on yield and plant height of a paddy variety. The mean and standard deviation for yield are 50 kg and 10 kg respectively. The mean and standard deviation for plant height are 55 am and 5 cm respectively.

Here the measurements for yield and plant height are in different units. Hence the variabilities can be compared only by using coefficient of variation.

For yield, CV= $\frac{10}{50} \times 100$ = 20%

For plant height, CV= $\frac{5}{55} \times 100$ = 9.1%

The yield is subject to more variation than the plant height.

Questions

1. Which measure is affected most by the presence of extreme values.

a) Range b) Standard Deviation

b) Quartile Deviation d) Mean deviation

Ans: Standard Deviation

2. Variance is square of _____

a) Range b) Standard Deviation

c) Quartile Deviation d) Mean deviation

Ans: Standard Deviation

3. If the CV of variety I is 30% and variety II is 25% then Variety II is more consistent.

Ans: True

4. For the set of data 5, 5, 5,5,5,5 the Standard deviation value is zero.

Ans: True

5. The absolute measures of dispersion will have the original units.

Ans: True

6. The mean deviation value for a set of data can take even negative value.

Ans: False

7. Define dispersion.

8. Define C.V. What are its uses?

9. What are the differences between absolute measure and relative measure of dispersion?

10. How to calculate the standard deviation for raw and grouped data?

Probability

Probability

The concept of probability is difficult to define in precise terms. In ordinary language, the word probable means likely (or) chance. Generally the word, probability, is used to denote the happening of a certain event, and the likelihood of the occurrence of that event, based on past experiences. By looking at the clear sky, one will say that there will not be any rain today. On the other hand, by looking at the cloudy sky or overcast sky, one will say that there will be rain today. In the earlier sentence, we aim that there will not be rain and in the latter we expect rain. On the other hand a mathematician says that the probability of rain is '0' in the first case and that the probability of rain is '1' in the second case. In between 0 and 1, there are fractions denoting the chance of the event occurring. In ordinary language, the word probability means uncertainty about happenings.In Mathematics and Statistics, a numerical measure of uncertainty is provided by the important branch of statistics – called theory of probability. Thus we can say, that the theory of probability describes certainty by 1 (one), impossibility by 0 (zero) and uncertainties by the co-efficient which lies between 0 and 1.

Trial and Event An experiment which, though repeated under essentially identical (or) same conditions does not give unique results but may result in any one of the several possible outcomes. Performing an experiment is known as a trial and the outcomes of the experiment are known as events.

Example 1: Seed germination – either germinates or does not germinates are events.

2. In a lot of 5 seeds none may germinate (0), 1 or 2 or 3 or 4 or all 5 may germinate.

Sample space (S)

A set of all possible outcomes from an experiment is called sample space. For example, a set of five seeds are sown in a plot, none may germinate, 1, 2, 3 ,4 or all five may germinate. i.e the possible outcomes are {0, 1, 2, 3, 4, 5. The set of numbers is called a sample space. Each possible outcome (or) element in a sample space is called sample point.

Exhaustive Events

The total number of possible outcomes in any trial is known as exhaustive events (or) exhaustive cases.

Example

1. When pesticide is applied a pest may survive or die. There are two exhaustive cases namely (survival, death)
2. In throwing of a die, there are six exhaustive cases, since anyone of the 6 faces

   1, 2, 3, 4, 5, 6 may come uppermost.
3. In drawing 2 cards from a pack of cards the exhaustive number of cases is $52C_2$, since 2 cards can be drawn out of 52 cards in $52C_2$ ways

Trial	**Random Experiment**	**Total number of trials**	**Sample Space**
(1)	One pest is exposed to pesticide	$2^1=2$	{S,D}
(2)	Two pests are exposed to pesticide	$2^2=4$	{SS, SD, DS, DD}
(3)	Three pests are exposed to pesticide	$2^3=8$	{SSS, SSD, SDS, DSS, SDD, DSD,DDS, DDD
(4)	One set of three seeds	$4^1= 4$	{0,1,2,3}
(5)	Two sets of three seeds	$4^2=16$	{0,1},{0,2},{0,3} etc

Favourable Events

The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event.

Example

1. When a seed is sown if we observe non germination of a seed, it is a favourable event. If we are interested in germination of the seed then germination is the favourable event.

Mutually Exclusive Events

Events are said to be mutually exclusive (or) incompatible if the happening of any one of the events excludes (or) precludes the happening of all the others i.e.) if no two or more of the events can happen simultaneously in the same trial. (i.e.) The joint occurrence is not possible.

Example

1. In observation of seed germination the seed may either germinate or it will not germinate. Germination and non germination are mutually exclusive events.

Equally Likely Events

Outcomes of a trial are said to be equally likely if taking in to consideration all the relevant evidences, there is no reason to expect one in preference to the others. (i.e.) Two or more events are said to be equally likely if each one of them has an **equal chance of occurring**.

Independent Events

Several events are said to be independent if the happening of an event is not affected by the happening of one or more events.

Example

1. When two seeds are sown in a pot, one seed germinates. It would not affect the germination or non germination of the second seed. One event does not affect the other event.

Dependent Events

If the happening of one event is affected by the happening of one or more events, then the events are called dependent events.

Example

If we draw a card from a pack of well shuffled cards, if the first card drawn is not replaced then the second draw is dependent on the first draw.

Note: In the case of independent (or) dependent events, the joint occurrence is possible.

Definition of Probability

Mathematical (or) Classical (or) a-priori Probability

If an experiment results in 'n' exhaustive cases which are mutually exclusive and equally likely cases out of which 'm' events are favourable to the happening of an event 'A', then the probability 'p' of happening of 'A' is given by

$$p = P(A) = \frac{Favourable\ number\ of\ cases}{Exhaustive\ number\ of\ cases} = \frac{m}{n}$$

Note

1. If m = 0 P(A) = 0, then 'A' is called an impossible event. (i.e.) also by P(ϕ) = 0.

2. If m = n P(A) = 1, then 'A' is called assure (or) certain event.
3. The probability is a non-negative real number and cannot exceed unity (i.e.) lies between 0 to 1.
4. The probability of non-happening of the event 'A' (i.e.) P($\bar{A}$). It is denoted by 'q'.

   $P(\bar{A}) = \frac{n - m}{n} = 1 - \frac{m}{n} =$

   — $1 - P(A)$

   $q = 1 - p$

   $p + q = 1$

   (or) $P(A) + P(\bar{A}) = 1$.

Statistical (or) Empirical Probability (or) a-posteriori Probability

If an experiment is repeated a number (n) of times, an event 'A' happens 'm' times then the statistical probability of 'A' is given by

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

Axioms for Probability

1. The probability of an event ranges from 0 to 1. If the event cannot take place its probability shall be '0' if it certain, its probability shall be '1'.

   Let $E_1, E_2, \ldots, E_n$ be any events, then $P(E_i) \geq 0$.

2. The probability of the entire sample space is '1'. (i.e.) P(S) = 1.

   Total Probability, $\sum_{i=1}^{n} P(E) = 1$

3. If A and B are mutually exclusive (or) disjoint events then the probability of occurrence of either A (or) B denoted by P(AUB) shall be given by

   $P(A \cup B) = P(A) + P(B)$

   $P(E_1 E_2 \ldots E_n) = P(E_1) + P(E_2) + \ldots\ldots + P(E_n)$ If

   $E_1, E_2, \ldots, E_n$ are mutually exclusive events.

Example 1: Two dice are tossed. What is the probability of getting (i) Sum 6 (ii) Sum 9?

Solution

When 2 dice are tossed. The exhaustive number of cases is 36 ways.

(i) Sum 6 = {(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)}

Favourable number of cases = 5

$$P(\text{Sum } 6) = \frac{5}{36}$$

(ii) Sum 9 = {(3, 6), (4, 5), (5, 4), (6, 3)}

Favourable number of cases = 4

$$P(\text{Sum } 9) = \frac{4}{36} = \frac{1}{9}$$

Example 2: A card is drawn from a pack of cards. What is a probability of getting (i) a king (ii) a spade (iii) a red card (iv) a numbered card?

Solution

There are 52 cards in a pack.

One can be selected in $52C_1$ ways.

Exhaustive number of cases is = $52C_1 = 52$.

(i) A king

There are 4 kings in a pack.

One king can be selected in $4C_1$ ways.

Favourable number of cases is = $4C_1 = 4$

Hence the probability of getting a king = $\frac{4}{52}$

(ii) A spade

There are 13 kings in a pack.

One spade can be selected in $13C_1$ ways.

Favourable number of cases is = $13C_1 = 13$

Hence the probability of getting a spade = $\frac{13}{52}$

(iii) A red card

There are 26 kings in a pack.

One red card can be selected in $26C_1$ ways.

Favourable number of cases is = $26C_1 = 26$

Hence the probability of getting a red card = $\frac{26}{52}$

(iv) A numbered card

There are 36 kings in a pack.

One numbered card can be selected in $36C_1$ ways.

Favourable number of cases is = $36C_1 = 36$

Hence the probability of getting a numbered card = $\frac{36}{52}$

Example 3: What is the probability of getting 53 Sundays when a leap year selected at random?

Solution

A leap year consists of 366 days.

This has 52 full weeks and 2 days remained.

The remaining 2 days have the following possibilities.

(i) Sun. Mon (ii) Mon, Tues (iii) Tues, Wed (iv) Wed, Thurs (v) Thurs, Fri (vi) Fri, Sat (vii) Sat, Sun.

In order that a lap year selected at random should contain 53 Sundays, one of the 2 over days must be Sunday.

Exhaustive number of cases is = 7

Favourable number of cases is = 2

Required Probability is = $\frac{2}{7}$

Conditional Probability

Two events A and B are said to be dependent, when B can occur only when A is known to have occurred (or vice versa). The probability attached to such an event is called the conditional probability and is denoted by P (A/B) (read it as: A given B) or, in other words, probability of A given that B has occurred.

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

If two events A and B are **dependent**, then the conditional probability of B given A is,

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

Theorems of Probability

There are two important theorems of probability namely,

1. The addition theorem on probability
2. The multiplication theorem on probability.

I. Addition Theorem on Probability

(i) Let A and B be any two events which are **not mutually exclusive**

$$P(A \text{ or } B) = P(A \cup B) = P(A + B) = P(A) + P(B) - P(A \cap B) \quad \text{(or)}$$

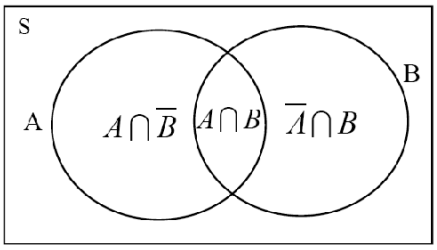
$$= P(A) + P(B) - P(AB)$$

Proof

Let us take a random experiment with a sample space S of N sample points.

Then by the definition of probability ,

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N}$$



From the diagram, using the axiom for the mutually exclusive events, we write

$$P(A \cup B) = \frac{n(A) + n(\overline{A} \cap B)}{N}$$

Adding and subtracting $n(A \cap B)$ in the numerator,

$$= \frac{n(A) + n(\overline{A} \cap B) + n(A \cap B) - n(A \cap B)}{N}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{N}$$

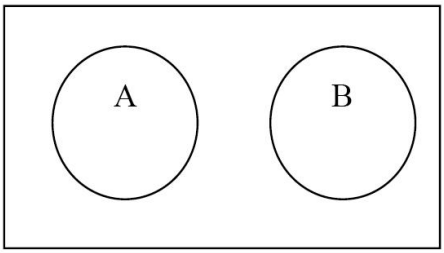
$$= \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(ii) Let A and B be any two events which are **mutually exclusive**

$$P(A \text{ or } B) = P(A \cup B) = P(A + B) = P(A) + P(B)$$

Proof



We know that, n (A B) = n (A) + n (B)

$$P(A \quad B) = \frac{n(A \cup B)}{n}$$

$$= \frac{n(A) + n(B)}{n}$$

$$= \frac{n(A)}{n} + \frac{n(B)}{n}$$

P (A B) = P (A) + P (B)

Note

(i) In the case of 3 events, **(not mutually exclusive events)**

P (A or B or C) = P (A B C) = P (A + B + C)

= P (A) + P (B) + P (C) – P (A ∩ B) – P (B ∩ C) – P (A ∩ C) + P (A ∩ B ∩ C)

(ii) In the case of 3 events, **(mutually exclusive events)**

P (A or B or C) = P (A B C) = P (A + B + C) = P (A) + P (B) + P (C)

Example

Using the additive law of probability we can find the probability that in one roll of a die, we will obtain either a one-spot or a six-spot. The probability of obtaining a one-spot is 1/6. The probability of obtaining a six-spot is also 1/6. The probability of rolling a die and getting a side that has both a one-spot with a six-spot is 0. There is no side on a

die that has both these events. So substituting these values into the equation gives the following result:

$$\frac{1}{6} + \frac{1}{6} - 0 = \frac{2}{6} = \frac{1}{3} = 0.3333$$

Finding the probability of drawing a 4 of hearts or a 6 or any suit using the additive law of probability would give the following:

$$\frac{1}{52} + \frac{4}{52} - 0 = \frac{5}{52} = 0.0962$$

There is only a single 4 of hearts, there are 4 sixes in the deck and there isn't a single card that is both the 4 of hearts and a six of any suit.

Now using the additive law of probability, you can find the probability of drawing either a king or any club from a deck of shuffled cards. The equation would be completed like this:

$$\frac{4}{52}+\frac{13}{52}-\frac{1}{52}=\frac{16}{52}=0.3077$$

There are 4 kings, 13 clubs, and obviously one card is both a king and a club. We don't want to count that card twice, so you must subtract one of it's occurrences away to obtain the result.

II. Multiplication Theorem on Probability

(i) If A and B be any two events which are **not independent**, then (i.e.) **dependent**.

P (A and B) = P (A ∩ B) = P (AB) = P (A). P (B/A) ⟶ (I)

= P (B). P (A/B) ⟶ (II)

Where P (B/A) and P (A/B) are the conditional probability of B given A and A given B respectively.

Proof

Let n is the total number of events

n (A) is the number of events in A

n (B) is the number of events in B

n (A B) is the number of events in (A B)

n (A ∩ B) is the number of events in (A ∩ B)

$$P(A\cap B)=\frac{n(A\cap B)}{n}$$

$$=\frac{n(A\cap B)}{n}\times\frac{n(A)}{n(A)}$$

$$=\frac{n(A)}{n}\times\frac{n(A\cap B)}{n(A)}$$

P (A ∩ B) = P (A). P (B/A) ⟶ (I)

$$P(A\cap B)=\frac{n(A\cap B)}{n}$$

$$=\frac{n(A\cap B)}{n}\times\frac{n(B)}{n(B)}$$

$$\frac{n(B)}{n}\times\frac{n(A\cap B)}{n(B)}$$

$$P(A \cap B) = P(B).\ P(A/B) \qquad \text{(II)}$$

(ii) If A and B be any two events which are **independent**, then,

$P(B/A) = P(B)$ and $P(A/B) = P(A)$

$P(A \text{ and } B) = P(A \cap B) = P(AB) = P(A)\ .\ P(B)$

Note

(i) In the case of 3 events, **(dependent)**

$P(A \cap B \cap C) = P(A).\ P(B/A).\ P(C/AB)$

(ii) In the case of 3 events, **(independent)**

$P(A \cap B \cap C) = P(A).\ P(B).\ P(C)$

Example

So in finding the probability of drawing a 4 and then a 7 from a well shuffled deck of cards, this law would state that we need to multiply those separate probabilities together. Completing the equation above gives:

$$p(4\ and\ 7) = \frac{4}{52} \times \frac{4}{52} = \frac{16}{2704} = 0.0059$$

Given a well shuffled deck of cards, what is the probability of drawing a Jack of Hearts, Queen of Hearts, King of Hearts, Ace of Hearts, and 10 of Hearts?

$$p(10, J, Q, K, A\ of\ hearts) = \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} \times \frac{1}{52} = 0.0000000026$$

In any case, given a well shuffled deck of cards, obtaining this assortment of cards, drawing one at a time and returning it to the deck would be highly unlikely (it has an exceedingly low probability).

Questions

1. Probability is expressed as

(a) Ratio (b) percentage
(c) Proportion (d) all the above

Ans: all the above

2. Probability can take values from

(a) - ∞ to + ∞ (b) - ∞ to 1
(c) 0 to +1 (d) –1 to +1

Ans: 0 to +1

3. The probability of a sure event is One.

Ans: True

4. If A and B are mutually exclusive events, then P (AUB) = **P (A) + P (B)**

5. An integer is chosen from 1 to 20. The probability that the number is divisible by 4 is ¼.

Ans: True

6. Mean of the Binomial Distribution is npq.

Ans: False

7. Define an independent event.

8. What is conditional probability?

9. State the addition and multiplication laws..

10. Additional theorem of probability.

Theoretical Distributions

Theoretical distributions are

1. Binomial distribution
2. Poisson distribution

} Discrete distribution

3. Normal distribution → Continuous distribution

Discrete Probability distribution

Bernoulli distribution

A random variable x takes two values 0 and 1, with probabilities q and p ie., p(x=1) = p and p(x=0)=q, q-1-p is called a Bernoulli variate and is said to be Bernoulli distribution where p and q are probability of success and failure. It was given by Swiss mathematician James Bernoulli (1654-1705)

Example

- Tossing a coin(head or tail)
- Germination of seed(germinate or not)

Binomial distribution

Binomial distribution was discovered by James Bernoulli (1654-1705). Let a random experiment be performed repeatedly and the occurrence of an event in a trial be called as success and its non-occurrence is failure. Consider a set of n independent trails (n being finite), in which the probability p of success in any trail is constant for each trial. Then q=1-p is the probability of failure in any trail..

The probability of x success and consequently n-x failures in n independent trails. But x successes in n trails can occur in nc_x ways. Probability for each of these ways is $p^x q^{n-x}$.

P(sss…ff…fsf…f)=p(s)p(s)….p(f)p(f)….

= p,p…q,q…

= (p,p…p)(q,q…q) (x

times) (n-x times)

Hence the probability of x success in n trials is given by

$nc_x\ p^x q^{n-x}$

Definition

A random variable x is said to follow binomial distribution if it assumes non-negative values and its probability mass function is given by

$$P(X=x) = p(x) = \begin{cases} nc_x\ p^x q^{n-x}, & x=0,1,2\ldots n \\ & q=1-p \\ 0, & \text{otherwise} \end{cases}$$

The two independent constants n and p in the distribution are known as the parameters of the distribution.

Condition for Binomial distribution

We get the binomial distribution under the following experimentation conditions

1. The number of trial n is finite
2. The trials are independent of each other.
3. The probability of success p is constant for each trial.
4. Each trial must result in a success or failure.
5. The events are discrete events.

Properties

1. If p and q are equal, the given binomial distribution will be symmetrical. If p and q are not equal, the distribution will be skewed distribution.
2. Mean = E(x) = np

3. Variance =V(x) = npq (mean>variance)

Application

1. Quality control measures and sampling process in industries to classify items as defectives or non-defective.
2. Medical applications such as success or failure, cure or no-cure.

Example 1

Eight coins are tossed simultaneously. Find the probability of getting atleast six heads.

Solution

Here number of trials, n = 8, p denotes the probability of getting a head.

$p = \frac{1}{2}$ and $q = \frac{1}{2}$

If the random variable X denotes the number of heads, then the probability of a success in n trials is given by

$P(X = x) = {}^{n}c_{x}\, p^{x}\, q^{n-x}$, $x = 0, 1, 2, ..., n$

$$= 8C_x \left(\frac{1}{2}\right)^{x} \left(\frac{1}{2}\right)^{8-x} = 8C_x \left(\frac{1}{2}\right)^{8}$$

$$= \frac{1}{2^8} 8C_x$$

Probability of getting atleast six heads is given by

$P(x \geq 6) = P(x = 6) + P(x = 7) + P(x = 8)$

$$= \frac{1}{2^8} 8C_6 + \frac{1}{2^8} 8C_7 + \frac{1}{2^8} 8C_8$$

$$= \frac{1}{2^8}\left[8C_6 + 8C_7 + 8C_8\right]$$

$$= \frac{1}{2^8}\left[28 + 8 + 1\right] = \frac{37}{256}$$

Example 2 Ten coins are tossed simultaneously. Find the probability of getting (i) atleast seven heads (ii) exactly seven heads (iii) atmost seven heads

Solution

p = Probability of getting a head = $\frac{1}{2}$ 2

q = Probability of not getting a head = $\frac{1}{2}$

The probability of getting x heads throwing 10 coins simultaneously is given by

$P(X = x) = nC_x\, p^x\, q^{n-x}$, x = 0, 1, 2, ..., n

$$= 10C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = 10C_x \left(\frac{1}{2}\right)^{10}$$

$$= \frac{1}{2^{10}} 10C_x$$

i) Probability of getting atleast seven heads

$P(x \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$

$$= \frac{1}{2^{10}}\left[10C_7 + 10C_8 + 10C_9 + 10C_{10}\right]$$

$$= \frac{1}{1024}\left[120 + 45 + 10 + 1\right] = \frac{176}{1024}$$

ii) Probability of getting exactly 7 heads

$$P(x = 7) = \frac{1}{2^{10}} 10C_7 = \frac{1}{2^{10}}(120) = \frac{120}{1024}$$

iii) Probability of getting almost 7 heads

$P(x \leq 7) = 1 - P(x > 7)$

= 1 symbol {P(x = 8) + P (x = 9) + P(x = 10)}

$$= 1 \quad \frac{1}{2^{10}}\left[10C_8 + 10C_9 + 10C_{10}\right]$$

$$= 1 - \frac{1}{2^{10}}\left[45 + 10 + 1\right]$$

$$= 1 - \frac{56}{1024}$$

$$= \frac{968}{1024}$$

Example 3:20 wrist watches in a box of 100 are defective. If 10 watches are selected at random, find the probability that (i) 10 are defective (ii) 10 are good (iii) at least one watch is defective (iv) at most 3 are defective.

Solution

20 out of 100 wrist watches are defective

Probability of defective wrist watch, p

$$q = 1 - p = \frac{4}{5}$$

Since 10 watches are selected at random, n =10

$P(X = x) = nC_x\, p^x\, q^{n-x},\ x = 0, 1, 2, ..., 10$

$$= 10C_x \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{10-x}$$

i) Probability of selecting 10 defective watches

$$P(x = 10) = = 10C_{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^{0} = 1.\frac{1}{5^{10}}.1 = \frac{1}{5^{10}}$$

ii) Probability of selecting 10 good watches (i.e. no defective)

$$P(x = 0) = \quad 10C_0 \left(\frac{1}{5}\right)^{0} \left(\frac{4}{5}\right)^{10}$$

$$= \ 1.1\left(\frac{4}{5}\right)^{10} = \left(\frac{4}{5}\right)^{10}$$

iii) Probability of selecting at least one defective watch

$P(x \geq 1) = 1 - P(x < 1)$

$= 1 - P(x = 0)$

$$= 1 \ - \ 10C_0 \left(\frac{1}{5}\right)^{0} \left(\frac{4}{5}\right)^{10}$$

$$=1 - \left(\frac{4}{5}\right)^{10}$$

iv) Probability of selecting at most 3 defective watches

P (x £ 3) = P (x = 0) + P(x =1) + P(x = 2) + P(x = 3)

$$= 10C_0 \left(\frac{1}{5}\right)^{0} \left(\frac{4}{5}\right)^{10} + 10C_1 \left(\frac{1}{5}\right)^{1} \left(\frac{4}{5}\right)^{9} + 10C_2 \left(\frac{1}{5}\right)^{2} \left(\frac{4}{5}\right)^{8} + \ 10C_3 \left(\frac{1}{5}\right)^{3} \left(\frac{4}{5}\right)^{7}$$

$$= \ 1.1\left(\frac{4}{5}\right)^{10} + 10\left(\frac{1}{5}\right)^{1} \left(\frac{4}{5}\right)^{9} + \frac{10.9}{1.2}\left(\frac{1}{5}\right)^{2} \left(\frac{4}{5}\right)^{8} + \frac{10.9.8}{1.2.3}\left(\frac{1}{5}\right)^{3} \left(\frac{4}{5}\right)^{7}$$

= 1. (0.107) + 10 (0.026) + 45 (0.0062) + 120 (0.0016)

= 0.859 (approx)

Poisson distribution

The Poisson distribution, named after Simeon Denis Poisson (1781-1840). Poisson distribution is a discrete distribution. It describes random events that occurs rarely over a unit of time or space.

It differs from the binomial distribution in the sense that we count the number of success and number of failures, while in Poisson distribution, the average number of success in given unit of time or space.

Definition

The probability that exactly x events will occur in a given time is as follows

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x=0,1,2\ldots$$

called as probability mass function of Poisson distribution.

where λ is the average number of occurrences per unit of time

$$\lambda = np$$

Condition for Poisson distribution

Poisson distribution is the limiting case of binomial distribution under the following assumptions.

1. The number of trials n should be indefinitely large ie., $n->\infty$
2. The probability of success p for each trial is indefinitely small.
3. $np= \lambda$, should be finite where λ is constant.

Properties

1. Poisson distribution is defined by single parameter λ.
2. Mean = λ
3. Variance = λ. Mean and Variance are equal.

Application

1. It is used in quality control statistics to count the number of defects of an item.
2. In biology, to count the number of bacteria.
3. In determining the number of deaths in a district in a given period, by rare disease.
4. The number of error per page in typed material.
5. The number of plants infected with a particular disease in a plot of field.
6. Number of weeds in particular species in different plots of a field.

Example 4: Suppose on an average 1 house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district, what is the probability that exactly 5 houses will have a fire during the year? [given that $e^{-2} = 0.13534$]

Solution:

Mean, $\bar{x}$ = np , n = 2000 and p = $\frac{1}{1000}$

$$= 2000 \times \frac{1}{1000}$$

$$\lambda = 2$$

The Poisson distribution is

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$P(X = 5) = \frac{e^{-2}2^{5}}{5!}$$

$$= \frac{(0.13534) \times 32}{120}$$

$$= \mathbf{0.036}$$

Example 5

If 2% of electric bulbs manufactured by a certain company are defective. Find the probability that in a sample of 200 bulbs i) less than 2 bulbs ii) more than 3 bulbs are defective.[e-4 = 0.0183]

Solution

The probability of a defective bulb $= p = \frac{2}{100} = 0.02$

Given that n = 200 since p is small and n is large

We use the Poisson distribution

mean, m = np = 200 · 0.02 = 4

Now, Poisson Probability function, $P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

i) Probability of less than 2 bulbs are defective

$= P(X<2)$

$= P(x = 0) + P(x = 1)$

$= e^{-4} + e^{-4}(4)$

$= e-4 (1 + 4) = 0.0183 \cdot 5$

$= 0.0915$

ii) Probability of getting more than 3 defective bulbs

$P(x > 3) = 1 - P(x \leq 3)$

$= 1 - \{P(x = 0) + P(x =1) + P(x=2) + P(x=3)\}$

$= 1 - e^{-4}\left\{1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\right\}$

$= 1 - \{0.0183 \cdot (1 + 4 + 8 + 10.67)\}$

$= 0.567$

Normal distribution

Continuous Probability distribution is normal distribution. It is also known as error law or Normal law or Laplacian law or Gaussian distribution. Many of the sampling distribution like student-t, f distribution and $\chi 2$ distribution.

Definition

A continuous random variable x is said to be a normal distribution with parameters μ and σ^2, if the density function is given by the probability law

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$; $-\infty < x < \infty$, $-\infty < \propto < \infty$, $\sigma > 0$

Note

The mean m and standard deviation s are called the parameters of Normal distribution.

The normal distribution is expressed by $X \sim N(m, s^2)$

Condition of Normal Distribution

i) Normal distribution is a limiting form of the binomial distribution under the following conditions.

a) n, the number of trials is indefinitely large ie., nà¥ and

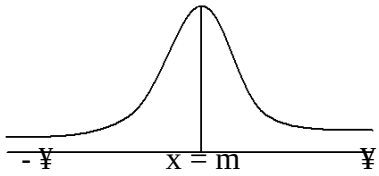
b) Neither p nor q is very small.

ii) Normal distribution can also be obtained as a limiting form of Poisson distribution with parameter mà¥

iii) Constants of normal distribution are mean = m, variation $=s^2$, Standard deviation = s.

Normal probability curve

The curve representing the normal distribution is called the normal probability curve. The curve is symmetrical about the mean (m), bell-shaped and the two tails on the right and left sides of the mean extends to the infinity. The shape of the curve is shown in the following figure.



Properties of normal distribution

1. The normal curve is bell shaped and is symmetric at x = m.
2. Mean, median, and mode of the distribution are coincide

   i.e., Mean = Median = Mode = m
3. It has only one mode at x = m (i.e., unimodal)
4. The points of inflection are at $x = m \pm s$
5. The maximum ordinate occurs at x = m and its value is $= \frac{1}{\sigma\sqrt{2\pi}}$
6. Area Property $P(m - s < ´ < m + s) = 0.6826$

   $P(m - 2s < ´ < m + 2s) = 0.9544$

P(m - 3s < ´ < m + 3s) = 0.9973

Standard Normal distribution

Let X be random variable which follows normal distribution with mean m and variance s2 .The standard normal variate is defined as $Z = \frac{X - \mu}{\sigma}$ which follows standard normal distribution with mean 0 and standard deviation 1 i.e., Z ~ N(0,1). The standard normal distribution is given by $\phi(Z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2}$; -¥< z< ¥

The advantage of the above function is that it doesn't contain any parameter. This enables us to compute the area under the normal probability curve.

Note

Property of $\phi(Z)$

1. $\phi(-Z) = 1 - \phi(Z)$
2. $P(a \le Z \le b) = \phi(b) - \phi(a)$

Example 6: In a normal distribution whose mean is 12 and standard deviation is 2. Find the probability for the interval from x = 9.6 to x = 13.8

Solution

Given that Z~ N (12, 4)

$$P(1.6 \le Z \le 13.8) = P\left(\frac{9.6 - 12}{2} \le Z \le \frac{13.8 - 12}{2}\right)$$

= P(-1.2 ≤ Z ≤ 0)+P(0 ≤ Z ≤ 0.9)

= P(0≤ Z ≤ 1.2)+P(0 ≤ Z ≤ 0.9) [by using symmetric property]

=0.3849 +0.3159

=0.7008

When it is converted to percentage (ie) 70% of the observations are covered between 9.6 to 13.8.

Example 7: For a normal distribution whose mean is 2 and standard deviation 3. Find the value of the variate such that the probability of the variate from the mean to the value is 0.4115

Solution:

Given that Z~ N (2, 9)

To find X_1:

We have P ($2 \le Z \le X_1$) =0.4115

$$P\left(\frac{2-2}{3} \le \frac{X-\mu}{\sigma} \le \frac{X_1-2}{3}\right) = 0.4115$$

P ($0 \le Z \le Z_1$) =0.4115 where $Z_1 = \frac{X_1 - 2}{3}$

[From the normal table where 0.4115 lies is rthe value of

Z_1] Form the normal table we have Z_1=1.35

$$\therefore 1.35 = \frac{X_1 - 2}{3}$$

$3(1.35)+2=X_1$

$=X_1=6.05$

(i.e) 41 % of the observation converged between 2 and 6.05

Questions

1. For a Poisson distribution

(a) mean > variance (b) mean = variance

(c) mean < variance (d) mean < variance

Ans: mean = variance

2. In normal distribution, skewness is

(a) one **(b) zero**

(c) greater than one (d) less than one

Ans: zero

3. Poisson distribution is a distribution for rare events

Ans: True

4. The total area under normal probability curve is one.

Ans: True

5. Poisson distribution is for continuous variable.

Ans: False

6. In a symmetrical curve mean, median and mode will coincide.

Ans: True

7. Give any two examples of Poisson distribution

8. The variance of a Poisson distribution is 0.5. Find P(x = 3).

[$e^{-0.5} = 0.6065$]

9. The customer accounts of a certain departmental store have an average balance of Rs.1200 and a standard deviation of Rs.400. Assuming that the account balances are normally distributed. (i) what percentage of the accounts is over Rs.1500? (ii) What percentage of the accounts is between Rs.1000 and Rs.1500? iii) What percentage of the accounts is below Rs.1500?

10. State the Properties of normal distribution

Sampling

Population (Universe)

Population means aggregate of all possible units. It need not be human population. It may be population of plants, population of insects, population of fruits, etc.

Parameter

A summary measure that describes any given characteristic of the population is known as parameter. Population are described in terms of certain measures like mean, standard deviation etc. These measures of the population are called parameter and are usually denoted by Greek letters. For example, population mean is denoted by $\propto$, standard deviation by σ and variance by σ^2.

Sample

A portion or small number of unit of the total population is known as sample.

- All the farmers in a village(population) and a few farmers(sample)
- All plants in a plot is a population of plants.
- A small number of plants selected out of that population is a sample of plants.

Statistic

A summary measure that describes the characteristic of the sample is known as statisitic. Thus sample mean, sample standard deviation etc is statistic. The statistic is usually denoted by roman letter.

$\bar{x}$ - sample mean

s – standard deviation

The statistic is a random variable because it varies from sample to sample.

Sampling

The method of selecting samples from a population is known as sampling.

Sampling technique

There are two ways in which the information is collected during statistical survey. They are

1. Census survey
2. Sampling survey

Census

It is also known as population survey and complete enumeration survey. Under census survey the information are collected from each and every unit of the population or universe.

Sample survey

A sample is a part of the population. Information are collected from only a few units of a population and not from all the units. Such a survey is known as sample survey.

Sampling technique is universal in nature, consciously or unconsciously it is adopted in every day life.

For eg.

1. A handful of rice is examined before buying a sack.
2. We taste one or two fruits before buying a bunch of grapes.
3. To measure root length of plants only a portion of plants are selected from a plot.

Need for sampling

The sampling methods have been extensively used for a variety of purposes and in great diversity of situations.

In practice it may not be possible to collected information on all units of a population due to various reasons such as

1. Lack of resources in terms of money, personnel and equipment.
2. The experimentation may be destructive in nature. Eg- finding out the germination percentage of seed material or in evaluating the efficiency of an insecticide the experimentation is destructive.

3. The data may be wasteful if they are not collected within a time limit. The census survey will take longer time as compared to the sample survey. Hence for getting quick results sampling is preferred. Moreover a sample survey will be less costly than complete enumeration.
4. Sampling remains the only way when population contains infinitely many number of units.
5. Greater accuracy.

Sampling methods

The various methods of sampling can be grouped under

1) Probability sampling or random sampling

2) Non-probability sampling or non random sampling

Random sampling

Under this method, every unit of the population at any stage has equal chance (or) each unit is drawn with known probability. It helps to estimate the mean, variance etc of the population.

Under probability sampling there are two procedures

1. Sampling with replacement (SWR)
2. Sampling without replacement (SWOR)

When the successive draws are made with placing back the units selected in the preceding draws, it is known as sampling with replacement. When such replacement is not made it is known as sampling without replacement.

When the population is finite sampling with replacement is adopted otherwise SWOR is adopted.

Mainly there are many kinds of random sampling. Some of them are.

1. Simple Random Sampling
2. Systematic Random Sampling
3. Stratified Random Sampling
4. Cluster Sampling

Simple Random sampling (SRS)

The basic probability sampling method is the simple random sampling. It is the simplest of all the probability sampling methods. It is used when the population is homogeneous.

Questions

1. If each and every unit of population has equal chance of being included in the sample, it is known as
   (a) Restricted sampling (b) Purposive sampling
   (c) Simple random sampling (d) None of the above
   Ans: Simple random sampling

2. In a population of size 10 the possible number of samples of size 2 will be

(a) 45 (b) 40 (c) 54 (d) None of the above

Ans: 45

3. A population consisting of an unlimited number of units is called an infinite population.

Ans: True

4. If all the units of a population are surveyed it is called census.

Ans: True

5. Random numbers are used for selecting the samples in simple random sampling method.

Ans: True

6. The list of all units in a population is called as Frame.

Ans: True

7. What is sampling?

8. Explain the Lottery method.

9. Explain the method of selection of samples in simple random sampling.

10. Explain the method of selection of samples in Stratified random sampling.

Lecture.9

Test of significance

Hypothesis

Hypothesis is a statement or assumption that is yet to be proved.

Statistical Hypothesis

When the assumption or statement that occurs under certain conditions is formulated as scientific hypothesis, we can construct criteria by which a scientific hypothesis is either rejected or provisionally accepted. For this purpose, the scientific hypothesis is translated into statistical language. If the hypothesis in given in a statistical language it is called a statistical hypothesis.

For eg:-

The yield of a new paddy variety will be 3500 kg per hectare – scientific hypothesis.

In Statistical language if may be stated as the random variable (yield of paddy) is distributed normally with mean 3500 kg/ha.

Simple Hypothesis

When a hypothesis specifies all the parameters of a probability distribution, it is known as simple hypothesis. The hypothesis specifies all the parameters, i.e μ and σ of a normal distribution.

Eg:-

The random variable x is distributed normally with mean $\mu=0$ & SD=1 is a simple hypothesis. The hypothesis specifies all the parameters (μ & σ) of a normal distributions.

Composite Hypothesis

If the hypothesis specific only some of the parameters of the probability distribution, it is known as composite hypothesis. In the above example if only the μ is specified or only the σ is specified it is a composite hypothesis.

Null Hypothesis - Ho

Consider for example, the hypothesis may be put in a form ‘paddy variety A will give the same yield per hectare as that of variety B’ or there is no difference between the average yields of paddy varieties A and B. These hypotheses are in definite terms. Thus these hypothesis form a basis to work with. Such a working hypothesis in known as null hypothesis. It is called null hypothesis because if nullities the original hypothesis, that variety A will give more yield than variety B.

The null hypothesis is stated as ‘there is no difference between the effect of two treatments or there is no association between two attributes (ie) the two attributes are

independent. Null hypothesis is denoted by Ho.

Eg:-

There is no significant difference between the yields of two paddy varieties (or) they give same yield per unit area. Symbolically, Ho: $\mu 1=\mu 2$.

Alternative Hypothesis

When the original hypothesis is $\mu 1>\mu 2$ stated as an alternative to the null hypothesis is known as alternative hypothesis. Any hypothesis which is complementary to null hypothesis is called alternative hypothesis, usually denoted by H_1.

Eg:-

There is a significance difference between the yields of two paddy varieties. Symbolically,

H_1: $\mu 1 \neq \mu 2$ (two sided or directionless alternative)

If the statement is that A gives significantly less yield than B (or) A gives significantly more yield than B. Symbolically,

H_1: $\mu 1 < \mu 2$ (one sided alternative-left tailed)

H_1: $\mu 1 > \mu 2$ (one sided alternative-right tailed)

Sampling Distribution

By drawing all possible samples of same size from a population we can calculate the statistic, for example, $\bar{x}$ for all samples. Based on this we can construct a frequency distribution and the probability distribution of $\bar{x}$. Such probability distribution of a statistic is known a sampling distribution of that statistic. In practice, the sampling distributions can be obtained theoretically from the properties of random samples.

Standard Error

As in the case of population distribution the characteristic of the sampling distributions are also described by some measurements like mean & standard deviation. Since a statistic is a random variable, the mean of the sampling distribution of a statistic is called the expected valued of the statistic. The SD of the sampling distributions of the statistic is called standard error of the Statistic. The square of the standard error is known as the variance of the statistic. It may be noted that the standard deviation is for units whereas the standard error is for the statistic.

Testing of Hypothesis

Once the hypothesis is formulated we have to make a decision on it. A statistical procedure by which we decide to accept or reject a statistical hypothesis is called testing of hypothesis.

Sampling Error

From sample data, the statistic is computed and the parameter is estimated through the statistic. The difference between the parameter and the statistic is known as the sampling error.

Test of Significance

Based on the sampling error the sampling distributions are derived. The observed results are then compared with the expected results on the basis of sampling distribution. If the difference between the observed and expected results is more than specified quantity of the standard error of the statistic, it is said to be significant at a specified probability level. The process up to this stage is known as test of significance.

Decision Errors

By performing a test we make a decision on the hypothesis by accepting or rejecting the null hypothesis Ho. In the process we may make a correct decision on Ho or commit one of two kinds of error.

- We may reject Ho based on sample data when in fact it is true. This error in decisions is known as Type I error.

- We may accept Ho based on sample data when in fact it is not true. It is known as Type II error.

	Accept Ho	**Reject Ho**
Ho is true	Correct Decision	Type I error
Ho is false	Type II error	Correct Decision

The relationship between type I & type II errors is that if one increases the other will decrease. The probability of type I error is denoted by α. The probability of type II error is denoted by β. The correct decision of rejecting the null hypothesis when it is false is known as the power of the test. The probability of the power is given by $1-\beta$.

Critical Region

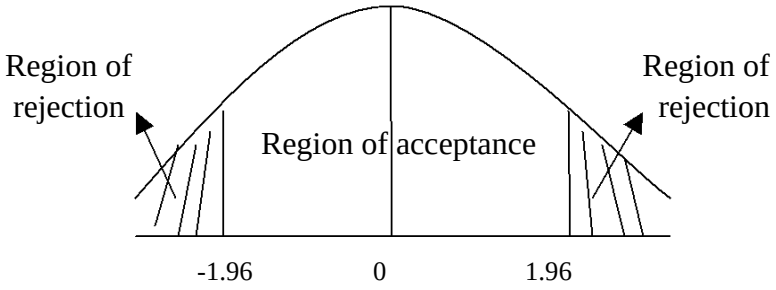
The testing of statistical hypothesis involves the choice of a region on the sampling distribution of statistic. If the statistic falls within this region, the null hypothesis is rejected: otherwise it is accepted. This region is called critical region.

Let the null hypothesis be Ho: $\mu 1 = \mu 2$ and its alternative be H_1: $\mu 1 \neq \mu 2$. Suppose Ho is

true. Based on sample data it may be observed that statistic $(\bar{x}_1 - \bar{x}_2)$ follows a normal distribution given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

We know that 95% values of the statistic from repeated samples will fall in the range $(\bar{x}_1 - \bar{x}_2)$ ± 1.96 times SE $(\bar{x}_1 - \bar{x}_2)$. This is represented by a diagram.



The border line value ±1.96 is the critical value or tabular value of Z. The area beyond the critical values (shaded area) is known as critical region or region of rejection. The remaining area is known as region of acceptance.

If the statistic falls in the critical region we reject the null hypothesis and, if it falls in the region of acceptance we accept the null hypothesis.

In other words if the calculated value of a test statistic (Z, t, χ^2 etc) is more than the critical value in magnitude it is said to be significant and we reject Ho and otherwise we accept Ho. The critical values for the t and χ^2 are given in the form of readymade tables. Since the criticval values are given in the form of table it is commonly referred as table value. The table value depends on the level of significance and degrees of freedom.

Example: Z cal < Z tab -We accept the Ho and conclude that there is no significant difference between the means.

Test Statistic

The sampling distribution of a statistic like Z, t, & χ^2 are known as test statistic. Generally, in case of quantitative data

$$\text{Test statistic} = \frac{Statistic - Parameter}{SandardError(Statistc)}$$

Level of Significance

The probability that the statistic will fall in the critical region is $\frac{\alpha}{2}+\frac{\alpha}{2}=\alpha$. This α is nothing but the probability of committing type I error. Technically the probability of committing type I error is known as level of Significance.

One and two tailed test

The nature of the alternative hypothesis determines the position of the critical region. For example, if H_1 is $\mu 1 \neq \mu 2$ it does not show the direction and hence the critical region falls on either end of the sampling distribution. If H_1 is $\mu 1 < \mu 2$ or $\mu 1 > \mu 2$ the direction is known. In the first case the critical region falls on the left of the distribution whereas in the second case it falls on the right side.

Degrees of freedom

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

The number of independent variables which make up the statistic is known as the degrees of freedom and is denoted by γ (Nu).

Steps in testing of hypothesis

The process of testing a hypothesis involves following steps.

1. Formulation of null & alternative hypothesis.
2. Specification of level of significance.
3. Selection of test statistic and its computation.
4. Finding out the critical value from tables using the level of significance, sampling distribution and its degrees of freedom.
5. Determination of the significance of the test statistic.
6. Decision about the null hypothesis based on the significance of the test statistic.
7. Writing the conclusion in such a way that it answers the question on hand.

Large sample theory

The sample size n is greater than 30 ($n \geq 30$) it is known as large sample. For large samples the sampling distributions of statistic are normal(Z test). A study of sampling distribution of statistic for large sample is known as large sample theory.

Small sample theory

If the sample size n ils less than 30 ($n < 30$), it is known as small sample. For small

samples the sampling distributions are t, F and χ^2 distribution. A study of sampling distribution for small samples is known as small sample theory.

Test of Significance

The theory of test of significance consists of various test statistic. The theory had been developed under two broad heading

1. Test of significance for large sample

   Large sample test or Asymptotic test or Z test ($n \geq 30$)
2. Test of significance for small samples($n<30$)

   Small sample test or Exact test-t, F and χ^2.

It may be noted that small sample tests can be used in case of large samples also.

Large sample test

Large sample test are

1. Sampling from attributes
2. Sampling from variables

Sampling from attributes

There are two types of test for attributes

1. Test for single proportion
2. Test for equality of two proportions

Test for single proportion

In a sample of large size n, we may examine whether the sample would have come from a population having a specified proportion P=Po. For testing

We may proceed as follows

1. **Null Hypothesis (Ho)**

   Ho: The given sample would have come from a population with specified proportion P=Po
2. **Alternative Hypothesis(H_1)**

   H_1 : The given sample may not be from a population with specified

   proportion $P \neq Po$ (Two Sided)

   $P>Po$(One sided-right sided)

   $P<Po$(One sided-left sided)
3. **Test statistic**

$$Z = \frac{|p - P|}{\sqrt{\frac{PQ}{n}}}$$

It follows a standard normal distribution with μ=0 and σ^2=1

4. Level of Significance

The level of significance may be fixed at either 5% or 1%

5. Expected vale or critical value

In case of test statistic Z, the expected value is

Z_e =1.96 at 5% level
2.58 at 1% level ⟶ Two tailed test

Z_e = 1.65 at 5% level
2.33 at 1% level ⟶ One tailed test

6. Inference

If the observed value of the test statistic Z_o exceeds the table value Z_e we reject the Null Hypothesis Ho otherwise accept it.

Test for equality of two proportions

Given two sets of sample data of large size n_1 and n_2 from attributes. We may examine whether the two samples come from the populations having the same proportion. We may proceed as follows:

1. Null Hypothesis (Ho)

Ho: The given two sample would have come from a population having the same proportion $P_1=P_2$

2. Alternative Hypothesis (H_1)

H_1 : The given two sample may not be from a population with specified

proportion $P_1 \neq P_2$ (Two Sided)

$P_1>P_2$(One sided-right sided)

$P_1<P_2$(One sided-left sided)

3. Test statistic

$$Z = \frac{\left| (p_1 - p_2) - (P_1 - P_2) \right|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

When $P_1$and P_2 are not known, then

$$Z = \frac{\left| p_1 - p_2 \right|}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$ for heterogeneous population

Where $q_1 = 1-p_1$ and $q_2 = 1-p_2$

$$Z = \frac{|p_1 - p_2|}{\sqrt{pq\left(\frac{1}{n} + \frac{1}{n2}\right)}}$$ for homogeneous population

p= combined or pooled estimate.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

4. Level of Significance

The level may be fixed at either 5% or 1%

5. Expected vale

The expected value is given by

Z_e = 1.96 at 5% level
2.58 at 1% level
→ Two tailed test

Z_e = 1.65 at 5% level
2.33 at 1% level
→ One tailed test

6. Inference

If the observed value of the test statistic Z exceeds the table value Z_e we may reject the Null Hypothesis Ho otherwise accept it.

Sampling from variable

In sampling for variables, the test are as follows

1. Test for single Mean
2. Test for single Standard Deviation
3. Test for equality of two Means
4. Test for equality of two Standard Deviation

Test for single Mean

In a sample of large size n, we examine whether the sample would have come from a population having a specified mean

1. Null Hypothesis (Ho)

Ho: There is no significance difference between the sample mean ie.,

$\mu=\mu_o$ or

The given sample would have come from a population having a specified mean ie., $\mu=\mu_o$

2. Alternative Hypothesis(H_1)

H_1 : There is significance difference between the sample mean ie., $\mu\neq\mu_o$ or $\mu>\mu_o$ or $\mu<\mu_o$

3. Test statistic

$$Z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}}$$

When population variance is not known, it may be replaced by its estimate

$$Z = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}}$$

where $s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$

4. Level of Significance

The level may be fixed at either 5% or 1%

5.Expected vale

The expected value is given by

Z_e = 1.96 at 5% level
2.58 at 1% level
→ Two tailed test

Z_e = 1.65 at 5% level
2.33 at 1% level
→ One tailed test

6. Inference

If the observed value of the test statistic Z exceeds the table value Z_e we may reject the Null Hypothesis Ho otherwise accept it.

Test for equality of two Means

Given two sets of sample data of large size n_1 and n_2 from variables. We may examine whether the two samples come from the populations having the same mean. We may proceed as follows

1. Null Hypothesis (Ho)

Ho: There is no significance difference between the sample mean ie.,

$\mu=\mu_o$ or

The given sample would have come from a population having a specified mean ie., $\mu_1=\mu_2$

2. Alternative Hypothesis (H_1)

H_1: There is significance difference between the sample mean ie.,

$\mu=\mu_o$ ie., $\mu_1\neq\mu_2$ or $\mu_1<\mu_2$ or $\mu_1>\mu_2$

3. Test statistic

When the population variances are known and unequal (i.e) $\sigma_1^2\neq\sigma^2{}_2$

$$Z = \frac{\left|(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)\right|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma^2{}_2}{n_2}}}$$

When $\sigma_1^2=\sigma^2{}_2$

$$Z = \frac{\left|(\bar{x}_1-\bar{x}_2)\right|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $\sigma = \frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2}$

The equality of variances can be tested by using F test.

When population variance is unknown, they may be replaced by their estimates s_1^2 and s_2^2

$$Z = \frac{\left|(\bar{x}_1-\bar{x}_2)\right|}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad \text{when } s_1^2\neq s_2^2$$

when $s_1^2 = s_2^2$

$$Z = \frac{\left|(\bar{x}_1 - \bar{x}_2)\right|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^2 = \frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1 + n_2}$$

Questions

1. A hypothesis may be classified as

(a)Simple (b) Composite

(c)Null (d) All the above

Ans: All the above

2. Area of the critical region depends on

(a) Size of type I error (b) Size of type II error

(c)Value of the statistics (d) Number of observations

Ans: Size of type I error

3. Large sample test can be applied when the sample size exceeds 30.

Ans: True

4. If the calculated test statistic is greater than the critical value, the null hypothesis is rejected.

Ans: True

5. The standard error of mean is given by $\frac{\sigma}{\sqrt{n}}$

Ans: True

6. If the alternative hypothesis is $\mu_1 \neq \mu_2$ then the test is known as one tailed test.

Ans: False

7. Define standard error.

8. Define Type I and Type II error.

9. Describe the procedure of comparing two group means.

10. Describe the procedure of comparing two proportions.

Small sample test

Student's t test

When the sample size is smaller, the ratio $Z = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}}$ will follow t distribution and not the standard normal distribution. Hence the test statistic is given as $t = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}}$ which follows normal distribution with mean 0 and unit standard deviation. This follows a t distribution with (n-1) degrees of freedom which can be written as $t_{(n-1)}$ d.f.

This fact was brought out by Sir William Gossest and Prof. R.A Fisher. Sir William Gossest published his discovery in 1905 under the pen name Student and later on developed and extended by Prof. R.A Fisher. He gave a test known as t-test.

Applications (or) uses

1. To test the single mean in single sample case.
2. To test the equality of two means in double sample case.
   (i) Independent samples(Independent t test)
   (ii) Dependent samples (Paired t test)
3. To test the significance of observed correlation coefficient.
4. To test the significance of observed partial correlation coefficient.
5. To test the significance of observed regression coefficient.

Test for single Mean

1. Form the null hypothesis

Ho: $\mu=\mu_o$

(i.e) There is no significance difference between the sample mean and the population mean

2. Form the Alternate hypothesis

H_1: $\mu\neq\mu_o$ (or $\mu>\mu_o$ or $\mu<\mu_o$)

ie., There is significance difference between the sample mean and the population mean

3. Level of Significance

The level may be fixed at either 5% or 1%

4. Test statistic

$$t = \left| \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \right|$$ which follows t distribution with (n-1) degrees of freedom

$$\text{where } \bar{x} = \frac{\sum x_i}{n} \text{ and } s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

6. Find the table value of t corresponding to (n-1) d.f. and the specified level of significance.
7. Inference

   If $t < t_{tab}$ we accept the null hypothesis H_0. We conclude that there is no significant difference sample mean and population mean

(or) if $t > t_{tab}$ we reject the null hypothesis H_0. (ie) we accept the alternative hypothesis and conclude that there is significant difference between the sample mean and the population mean.

Example 1

Based on field experiments, a new variety of green gram is expected to given a yield of 12.0 quintals per hectare. The variety was tested on 10 randomly selected farmer's fields. The yield (quintals/hectare) were recorded as 14.3,12.6,13.7,10.9,13.7,12.0,11.4,12.0,12.6,13.1. Do the results conform to the expectation?

Solution

Null hypothesis H_0: $\propto$=12.0

(i.e) the average yield of the new variety of green gram is 12.0 quintals/hectare.

Alternative Hypothesis: H_1:$\propto \neq$ 12.0

(i.e) the average yield is not 12.0 quintals/hectare, it may be less or more than 12 quintals / hectare

Level of significance: 5 %

Test statistic:

$$t = \left| \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \right|$$

From the given data

$$\sum x = 126.3 \qquad \sum x^2 = 1605.77$$

$$\bar{x} = \frac{\sum x}{n} = \frac{126.3}{10} = 12.63$$

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{1605.77 - 1595.169}{9}} = \sqrt{\frac{10.601}{9}}$$

$$= 1.0853$$

$$\frac{s}{\sqrt{n}} = \frac{1.0853}{\sqrt{10}} = 0.3432$$

Now $t = \left| \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \right|$

$$t = \frac{12.63 - 12}{0.3432} = 1.836$$

Table value for t corresponding to 5% level of significance and 9 d.f. is 2.262 (two tailed test)

Inference

$t < t_{tab}$

We accept the null hypothesis H_0

We conclude that the new variety of green gram will give an average yield of 12 quintals/hectare.

Note

Before applying t test in case of two samples the equality of their variances has to be tested by using F-test

$$F = \frac{s_1^2}{s_2^2} \sim F_{(n_1 - 1, n_2 - 1)} \; d.f \quad \text{if} s_1^2 > s_2^2$$

or

$$F = \frac{s_2^2}{s_1^2} \sim F_{(n_2 - 1, n_1 - 1)} \; d.f \quad \text{if} \quad s_2^2 > s_1^2$$

where s_1^2 is the variance of the first sample whose size is n_1.

s_2^2 is the variance of the second sample whose size is n_2.

It may be noted that the numerator is always the greater variance. The critical value for F is read from the F table corresponding to a specified d.f. and level of significance

Inference

$F < F_{tab}$

We accept the null hypothesis H_0.(i.e) the variances are equal otherwise the variances are unequal.

Test for equality of two Means (Independent Samples)

Given two sets of sample observation $x_{11}, x_{12}, x_{13} \ldots x_{1n}$, and $x_{21}, x_{22}, x_{23} \ldots x_{2n}$ of sizes n1 and n2 respectively from the normal population.

1. Using F-Test , test their variances

(i) **Variances are Equal**

Ho:., $\mu_1 = \mu_2$

H_1 $\mu_1 \neq \mu_2$ (or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$)

Test statistic

$$t = \frac{\left| (\bar{x}_1 - \bar{x}_2) \right|}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where the combined variance

$$s^2 = \frac{\left[\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n1}\right] + \left[\sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n2}\right]}{n_1 + n_2 - 2}$$

The test statistic t follows a t distribution with (n1+n2-2) d.f.

(ii) Variances are unequal and n1=n2

$$t = \frac{\left|(\bar{x}_1 - \bar{x}_2)\right|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

It follows a t distribution with $\left(\frac{n1 + n2}{2}\right)$ 1d.f.

(i) Variances are unequal and n1≠n2

$$t = \left|\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\right|$$

This statistic follows neither t nor normal distribution but it follows Behrens-Fisher d distribution. The Behrens – Fisher test is laborious one. An alternative simple method has been suggested by Cochran & Cox. In this method the critical value of t is altered as t_w (i.e) weighted t

$$t_w = \frac{t_1\left(\frac{s_1^2}{n1}\right) + t_2\left(\frac{s_2^2}{n2}\right)}{\frac{s_1^2}{n1} + \frac{s_2^2}{n2}}$$

where $t_1$is the critical value for t with (n1-1) d.f. at a dspecified level of significance

and t2 is the critical value for t with (n2-1) d.f. at a dspecified level of significance and

Example 2

In a fertilizer trial the grain yield of paddy (Kg/plot) was observed as follows

Under ammonium chloride 42,39,38,60 &41 kgs

Under urea 38, 42, 56, 64, 68, 69,& 62 kgs.

Find whether there is any difference between the sources of nitrogen?

Solution

Ho: $\mu_1=\mu_2$ (i.e) there is no significant difference in effect between the sources of nitrogen.

H_1: $\mu_1\neq\mu_2$ (i.e) there is a significant difference between the two sources

Level of significance = 5%

Before we go to test the means first we have to test their variances by using F-test.

F-test

Ho:., $\sigma_1^2=\sigma_2^2$

H1:., $\sigma_1^2\neq\sigma_2^2$

$$s_1^2=\frac{\sum x_1^2-\frac{(\sum x_1)^2}{n1}}{n1-1}=82.5$$

$$s_2^2=\frac{\sum x_2^2-\frac{(\sum x_2)^2}{n2}}{n2-1}=154.33$$

$$F=\frac{s_2^2}{s_1^2}\sim F_{(n_2-1,n_{1-}1)}\ d.f\ \text{ if } s_2^2>s_1^2$$

$$F=\frac{154.33}{32.5}=1.8707$$

$F_{tab}(6,4)$ d.f. = 6.16

$F < F_{tab}$

We accept the null hypothesis H_0. (i.e) the variances are equal.

Use the test statistic

$$t=\frac{\left|(\bar{x}_1-\bar{x}_2)\right|}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where

$$s^2=\frac{\left[\sum x_1^2-\frac{(\sum x_1)^2}{n1}\right]+\left[\sum x_2^2-\frac{(\sum x_2)^2}{n2}\right]}{n_1+n_2-2}=\frac{330+926}{10}=125.6$$

$$t=\frac{|44-57|}{\sqrt{125.6\left(\frac{1}{7}+\frac{1}{75}\right)}}=1.98$$

The degrees of freedom is 5+7-2= 10. For 5 % level of significance, table value of t is 2.228

Inference:

$t < t_{tab}$

We accept the null hypothesis H_0

We conclude that the two sources of nitrogen do not differ significantly with regard to the grain yield of padd

Example 3

The summary of the results of an yield trial on onion with two methods of propagation is given below. Determine whether the methods differ with regard to onion yield. The onion yield is given in Kg/plot.

Method I	**Method II**
n1=12	n2=12
SS1=186.25	SS2=737.6667
$\bar{x}_1 = 25.25$	$\bar{x}_2 = 28.83$
$s_1^2 = 16.9318$	$s_2^2 = 67.0606$

Solution

Ho:., $\mu_1=\mu_2$ (i.e) the two propagation methods do not differ with regard to onion yield.

H_1 $\mu_1 \neq \mu_2$ (i.e) the two propagation methods differ with regard to onion yield.

Level of significance = 5%

Before we go to test the means first we have to test their variability using F-test.

F-test

Ho: $\sigma_1^2 = \sigma_2^2$

H1: $\sigma_1^2 \neq \sigma_2^2$

$$s_1^2 = \frac{\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n1}}{n1-1} = 16.9318$$

$$s_2^2 = \frac{\sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n2}}{n2-1} = 67.0606$$

$$F = \frac{s_2^2}{s_1^2} \sim F_{(n_2-1, n_1-1)} \; d.f \qquad s_2^2 \quad s_1^2$$

$$F = \frac{67.0606}{16.9318} = 3.961$$

$F_{tab}(11,11)$ d.f. = 2.82

$F > F_{tab}$

We reject the null hypothesis H_0.we conclude that the variances are unequal.

Here the variances are unequal with equal sample size then the test statistic is

$$t = \frac{\left|(\bar{x}_1 - \bar{x}_2)\right|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$s^2 = \frac{\left[\sum x_1^2 - \frac{(\sum x_1)^2}{n1}\right] + \left[\sum x_2^2 - \frac{(\sum x_2)^2}{n2}\right]}{n_1 + n_2 - 2}$$

$$s^2 = \frac{SS1 + SS2}{n_1 + n_2 - 2} = \frac{186.25 + 737.6667}{12 + 12 - 2} = 41.9962$$

$$t = \frac{25.25 - 28.83}{\sqrt{41.9962\left(\frac{1}{12} + \frac{1}{12}\right)}} = \frac{3.58}{\sqrt{6.9994}} = 1.353$$

t =1.353

The table value for $\left(\frac{12+12}{2} - 1\right)$ =11 d.f. at 5% level of significance is 2.201

Inference:

$t < t_{tab}$

We accept the null hypothesis H_0

We conclude that the two propagation methods do not differ with regard to onion yield.

Example 4

The following data relate the rubber yield of two types of rubber plants, where the sample have been drawn independently. Test whether the two types of rubber plants differ in their yield.

Type I	6.21	5.70	6.04	4.47	5.22	4.45	4.84	5.84	5.88	5.82	6.09	5.59
	6.06	5.59	6.74	5.55								

Type II	4.28	7.71	6.48	7.71	7.37	7.20	7.06	6.40	8.93	5.91	5.51	6.36

Solution

Ho:., $\mu_1=\mu_2$ (i.e) there is no significant difference between the two rubber plants.

H_1 $\mu_1\neq\mu_2$ (i.e) there is a significant difference between the two rubber plants.

Level of significance = 5%

Here

n1=16	n2=12
$\sum x_1 = 90.09$	$\sum x_2 = 80.92$
$\bar{x}_1 = 5.63$	$\bar{x}_2 = 6.7431$
$\sum x_1^2 = 513.085$	$\sum x_2^2 = 561.64$

Before we go to test the means first we have to test their variability using F-test.

F-test

Ho:., $\sigma_1^2=\sigma_2^2$

H1:., $\sigma_1^2\neq\sigma_2^2$

$$s_1^2 = \frac{\sum x_1^2 - \frac{(\sum x_1)^2}{n1}}{n1-1} = 0.388$$

$$s_2^2 = \frac{\sum x_2^2 - \frac{(\sum x_2)^2}{n2}}{n2-1} = 1.452$$

$$F = \frac{s_2^2}{s_1^2} \sim F_{(n_2-1,n_1-1)}\ d.f.\ \text{if}\ s_2^2 > s_1^2$$

$$F = \frac{1.452}{0.388} = 3.742$$

$F_{tab}(11,15)$ d.f.=2.51

$F > F_{tab}$

We reject the null hypothesis H_0. Hence, the variances are unequal.

Here the variances are unequal with unequal sample size then the test statistic is

$$t = \left| \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right|$$

$$t = \frac{(5.63 - 6.7431_2)}{\sqrt{\frac{0.388}{16} + \frac{1.452}{12}}} = 2.912$$

$$t_w = \frac{t_1\left(\frac{S_1^2}{n1}\right) + t_2\left(\frac{S_2^2}{n2}\right)}{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}$$

$t_1 = t_{(16-1)}$ d.f.=2.131

$t_2 = t_{(12-1)}$ d.f .=2.201

$$t_w = \frac{2.131\left(\frac{0.388}{16}\right) + 2.201\left(\frac{1.452}{12}\right)}{\frac{0.388}{16} + \frac{1.425}{12}} = 2.187$$

Inference:

$t > t_w$

We reject the null hypothesis H_0. We conclude that the second type of rubber plant yields more rubber than that of first type.

Equality of two means (Dependant samples)

Paired t test

In the t-test for difference between two means, the two samples were independent of each other. Let us now take particular situations where the samples are not independent.

In agricultural experiments it may not be possible to get required number of homogeneous experimental units. For example, required number of plots which are similar in all; characteristics may not be available. In such cases each plot may be divided into two

equal parts and one treatment is applied to one part and second treatment to another part of the plot. The results of the experiment will result in two correlated samples. In some other situations two observations may be taken on the same experimental unit. For example, the soil properties before and after the application of industrial effluents may be observed on number of plots. This will result in paired observation. In such situations we apply paired t test.

Suppose the observation before treatment is denoted by x and the observation after treatment is denoted by y. for each experimental unit we get a pair of observation(x,y). In case of n experimental units we get n pairs of observations : (x1,y1), (x2,y2)…(xn,yn). In order to apply the paired t test we find out the differences (x1- y1), (x2-y2),..,(xn-yn) and denote them as d1, d2,…,dn. Now d1, d2…form a sample . we

apply the t test procedure for one sample (i.e) $t = \frac{|\bar{d}|}{\sqrt{s^2/n}}$

where $\bar{d} = \frac{\sum di}{n}$, $s^2 = \frac{\sum di^2 - \frac{(\sum di)^2}{n}}{n-1}$

the mean d may be positive or negative. Hence we take the absolute value as $|\bar{d}|$.
The test statistic t follows a t distribution with (n-1) d.f.

Example 5

In an experiment the plots where divided into two equal parts. One part received soil treatment A and the second part received soil treatment B. each plot was planted with sorghum. The sorghum yield (kg/plot) was absorbed. The results are given below. Test the effectiveness of soil treatments on sorghum yield.

Soil treatment A	49	53	51	52	47	50	52	53
Soil treatment B	52	55	52	53	50	54	54	53

Solution

H_0: $\propto_1 = \propto_2$, there is no significant difference between the effects of the two soil treatments

H_1: $\propto_1 \neq \propto_2$, there is significant difference between the effects of the two soil treatments Level of significance = 5%

Test statistic

$t = \frac{|\bar{d}|}{\sqrt{s^2/n}}$

x	y	d=x-y	d^2
49	52	-3	9
53	55	-2	4
51	52	-1	1
51	52	-1	1
47	50	-3	16
50	54	-4	16
52	54	-2	4
53	53	0	0
Total		-16	44

$$\bar{d} = \frac{\sum di}{n} = \frac{-16}{8} = -2,$$

$$s^2 = \frac{\sum di^2 - \frac{\left(\sum di\right)^2}{n}}{n-1} = 1.7143$$

$$t = \frac{|-2|}{\sqrt{1.7143/8}} = 4.32$$

Table value of t for 7 d.f. at 5% l.o.s is 2.365

Inference:

$t > t_{tab}$

We reject the null hypothesis H_0. We conclude that the is significant difference between the two soil treatments between A and B. Soil treatment B increases the yield of sorghum significantly,

Questions

1. The test statistic $F = \frac{s_1^2}{s_2^2}$ is used for testing

(a) H_0: $\propto_1 = \propto_2$ (b) H_0: $\sigma_{1^2} = \sigma_{2^2}$

(c) H0: $\sigma_1 = \sigma_2$ (d) H_0: $\sigma_2 = \sigma_{0^2}$

Ans: H_0: $\sigma_{1^2} = \sigma_{2^2}$

2. In paired t test with n observations in each group the degrees of freedom is

(a) n (b) n-1 (c) n-2 (d) n+1

Ans: n-1

3. Student t- test is applicable in case of small samples.

Ans: True

4.F test is also known as variance ratio test.

Ans: True

5. In case of comparing the equality of two variances the greater variance should be taking in the numerator.

Ans: True

6. While comparing the means of two independent samples the variances of the two samples will be always equal.

Ans: False

7. Define t statistic.

8. Define F statistic.

9. Explain the procedure of testing the equality of two variances.

10. How to compare the means of two independent small samples.

Attributes- Contingency table – 2x2 contingency table – Test for independence of attributes – test for goodness of fit of mendalian ratio Test based on χ^2 -distribution

In case of attributes we can not employ the parametric tests such as F and t. Instead we have to apply χ^2 test**.** When we want to test whether a set of observed values are in agreement with those expected on the basis of some theories or hypothesis. The χ^2 statistic provides a measure of agreement between such observed and expected frequencies.

The χ^2 test has a number of applications. It is used to

(1) Test the independence of attributes
(2) Test the goodness of fit
(3) Test the homogeneity of variances
(4) Test the homogeneity of correlation coefficients
(5) Test the equaslity of several proportions.

In genetics it is applied to detect linkage.

Applications

χ^2 – test for goodness of fit

A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as "chi-square test of goodness of fit ".

If 0_i, (i=1,2,…..,n) is a set of observed (experimental frequencies) and Ei (i=1,2,…..,n) is the corresponding set of expected (theoretical or hypothetical) frequencies, then,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

It follows a χ^2 distribution with n-1 d.f. In case of χ^2 only one tailed test is used.

Example

In plant genetics, our interest may be to test whether the observed segregation ratios deviate significantly from the mendelian ratios. In such situations we want to test the agreement between the observed and theoretical frequency, such test is called as test of goodness of fit.

Conditions for the validity of χ^2 -test:

χ^2 -test is an approximate test for large values of 'n' for the validity of χ^2 -test of goodness of fit between theory and experiment, the following conditions must be satisfied.

1. The sample observations should be independent.

2. Constraints on the cell freqrequency, if any, should be linear.

Example: $\sum O_i = \sum E_i$

3. N, the total frequency should be reasonably large, say greater then (>) 50.

4. No theoretical cell frequency should be less than (<)5. If any theoretical cell frequency is <5, then for the application of χ^2 - test, it is pooled with the preceding or scecceeding frequency so that the pooled frequency is more than 5 and finally adjust for degree's of freedom lost in pooling.

Example1

The number of yiest cells counted in a haemocytometer is compared to the theoretical value is given below. Does the experimental result support the theory?

No. of Yeast cells in the square	Obseved Frequency	Expected Frequency
0	103	106
1	143	141
2	98	93
3	42	41
4	8	14
5	6	5

Solution

H_0: the experimental results support the theory

H_1: the esperimental results does not support the theory.

Level of significance=5%

Test Statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i	E_i	O_i-E_i	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
103	106	-3	9	0.0849
143	141	2	4	0.0284
98	93	5	25	0.2688
42	41	1	1	0.0244
8	14	-6	36	2.5714
6	5	1	1	0.2000
400	400			**3.1779**

χ^2 =3.1779

Table value

χ^2 (6-1=5 at 5 % l.os)= 11.070

Inference

$\chi^2 < \chi^2_{tab}$

We accept the null hypothesis.

(i.e) there is a good correspondence between theory and experiment.

χ^2 test for independence of attributes

At times we may consider two charactertistics on attributes simultaneously. Our interest will be to test the association between these two attributes

Example:- An entomologist may be interested to know the effectiveness of different concentrations of the chemical in killing the insects. The concentrations of chemical form one attribute. The state of insects 'killed & not killed' forms another attribute. The result of this experiment can be arranged in the form of a contingency table. In general one attribute may be divided into m classes as $A_1, A_2, \ldots\ldots A_m$ and the other attribute may be divided into n classes as $B_1, B_2, \ldots\ldots B_n$. Then the contingency table will have m x n cells. It is termed as m x n contingency table

A	A1	A2…	Aj	…	Am	Row Total
B						

B1	O11	O12	…	O1j		O1m	r1
B2	O21	O22	…	O2j		O2m	r2
.							
.							
.							
Bi	Oij	Oi2	…	Oij		Oim	ri
.							
.							
.							
Bn	On1	On2	…	Onj		Onm	rk
Column Total	c1	c2	…	cj	…	cm	$n=\sum ri=\sum cj$

where Oij's are observed frequencies.

The expected frequencies corresponding to Oij is calculated as $\frac{ri.cj}{n}$. The χ^2 is computed as

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(o_{ij}-E_{ij}\right)}{E_{ij}}$$

where

O_{ij} – observed frequencies
E_{ij} – Expected frequencies
n= number of rows
m= number of columns
It can be verified that $\sum Oij = \sum Eij$
This χ^2 is distributed as χ^2 with (n-1) (m-1) d.f.

2x2 – contingency table

When the number of rows and numberof columns are equal to 2 it is termed as 2 x 2 contingency table .It will be in the following form

	B_1	B_2	Row Total	
A_1	a	b	a+b	r_1
A_2	c	d	c+d	r_2
Column Total	a+c	b+d	a+b+c+d =n	
	c_1	c_2		

Where a, b, c and d are cell frequancies c1 and c2 are column totals, r1 and r2 are row totals and n is the total number of observations.

In case of 2 x 2 contigency table χ^2 can be directly found using the short cut formula,

$$\chi^2 = \frac{n(ad-bc)^2}{c1.c2.r1.r2}$$

The d.f associated with χ^2 is (2-1) (2-1) =1

Yates correction for continuity

If anyone of the cell frequency is < 5, we use Yates correction to make χ^2 as continuous. The yares correction is made by adding 0.5 to the least cell frequency and adjusting the other cell frequencies so that the column and row totals remain same . suppose, the firat cell frequency is to be corrected then the consigency table will be as follows:

	B1	B2	Row Total
A1	a + 0.5	b – 0.5	a+b=r1
A2	c – 0.5	d+ 0.5	c+d =r2
Column Total	a+c=c1	b+d=c2	n = a+b+c+d

Then use the χ^2 - statistic as

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{c1.c2.r1.r2}$$

The d.f associated with χ^2 is (2-1) (2-1) =1

Exapmle 2

The severity of a disease and blood group were studied in a research projest. The findings sre given in the following table, knowmn as the m xn contingency table. Can this severity of the condition and blood group are associated.

Severity of a disease classified by blood group in 1500 patients.

Condition	Blood Groups				Total
	O	A	B	AB	

Severe	51	40	10	9	110
Moderate	105	103	25	17	250
Mild	384	527	125	104	1140
Total	540	670	160	130	1500

Solution

H_0: The severity of the disease is not associated with blood group.

H_1: The severity of the disease is associated with blood group.

Calculation of Expected frequencies

Condition	Blood Groups				Total
	O	A	B	AB	
Severe	39.6	49.1	11.7	9.5	110
Moderate	90.0	111.7	26.7	21.7	250
Mild	410.4	509.2	121.6	98.8	1140
Total	540	670	160	130	1500

Test statistic:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\left(o_{ij} - E_{ij}\right)}{E_{ij}}$$

The d.f. associated with the χ^2 is (3-1)(4-1) = 6

Calculations

O_i	E_i	O_i-E_i	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
51	39.6	11.4	129.96	3.2818
40	49.1	-9.1	82.81	1.6866
10	11.7	-1.7	2.89	0.2470
9	9.5	-0.5	0.25	0.0263
105	90.0	15	225.00	2.5000
103	111.7	-8.7	75.69	0.6776
25	26.7	-1.7	2.89	0.1082
17	21.7	-4.7	22.09	1.0180
384	410.4	-26.4	696.96	1.6982
527	509.2	17.8	316.84	0.6222
125	121.6	3.4	11.56	0.0951

104	98.8	5.2	27.04	0.2737
Total				12.2347

χ^2 =12.2347

Table value of χ^2 for 6 d.f. at 5% level of significance is

12.59 Inference

$\chi^2 < \chi^2_{tab}$

We accept the null hypothesis.

The severity of the disease has no association with blood group.

Example 3

In order to determine the possible effect of a chemical treatment on the rate of germination of cotton seeds a pot culture experiment was conducted. The results are given below

Chemical treatment and germination of cotton seeds

	Germinated	Not germinated	Total
Chemically Treated	118	22	140
Untreated	120	40	160
Total	238	62	300

Does the chemical treatrment improve the germination rate of cotton seeds?

Solution

H_0:The chemical treatment does not improve the germination rate of cotton seeds.

H_1: The chemical treatment improves the germination rate of cotton seeds.

Level of significance = 1%

Test statistic

$$= \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \text{ with 1 d.f.}$$

$$\chi^2 = \frac{300(118\times 40 - 22\times 120)^2}{140\times 160\times 62\times 238} = 3.927$$

Table value

χ^2 (1) d.f. at 1 % L.O.S = 6.635

Inference

$\chi^2 < \chi^2_{tab}$

We accept the null hypothesis.

The chemical treatmentwill not improve the germination rate of cotton seeds significantly.

Example 4

In an experiment on the effect of a growth regulator on fruit setting in muskmelon the following results were obtained. Test whether the fruit setting in muskmelon and the application of growth regulator are independent at 1% level.

	Fruit set	**Fruit not set**	**Total**
Treated	16	9	25
Control	4	21	25
Total	20	30	50

Solution

H_0:Fruit setting in muskmelon does not depend on the application of growth regulator.

H_1: Fruit setting in muskmelon depend on the application of growth regulator.

Level of significance = 1%

After Yates correction we have

	Fruit set	Fruit not set	Total
Treated	15.5	9.5	25
Control	4.5	20.5	25
Total	20	30	50

Tet statistic

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{50\left[|15.5\times 20.5 - 9.5\times 4.5|-\frac{50}{2}\right]^2}{25\times 25\times 20\times 30} = 8.33$$

Table value

χ^2 (1) d.f. at 1 % level of significance is 6.635

Inference

$\chi^2 > \chi^2_{tab}$

We reject the null hypothesis.

Fruit setting in muskmelon is influenced by the growth regulator. Application of growth regulator will increase fruit setting in musk melon.

Questions

1. The calculated value of χ^2 is
(a) always positive (b) always negative
(c) can be either positive or negative (d) none of these

Ans: always positive

2. Degrees of freedom for Chi-square in case of contingency table of order (4 ×3) are
(a) 12 (b) 9 (c) 8 (d) 6

Ans: 6

3.One condition for application of χ^2 test is that no cell frequency should be less than five.

Ans: True

4**.** The distribution of the χ^2 depends on the degrees of freedom.

Ans: True

5. The greater the discrepancy between the observed and expected Frequency lesser the value of χ^2.

Ans: False

6**.** When observed and expected frequencies completely coincide χ_2 will be zero.

Ans: True

7. What is a contignecy table?

8. When and how to apply Yates correction?

9. Explain the χ_2 test of goodness of fit?

10. Explain how to test the indepence of attributes?

Lecture.12

Correlation – definition – Scatter diagram -Pearson's correlation co-efficient – properties of correlation coefficient

Correlation

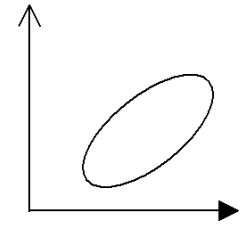
Correlation is the study of relationship between two or more variables. Whenever we conduct any experiment we gather information on more related variables. When there are two related variables their joint distribution is known as bivariate normal distribution and if there are more than two variables their joint distribution is known as multivariate normal distribution.

In case of bi-variate or multivariate normal distribution, we are interested in discovering and measuring the magnitude and direction of relationship between 2 or more variables. For this we use the tool known as correlation.

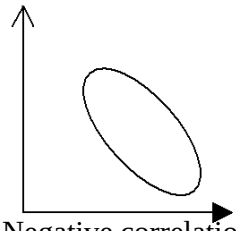
Suppose we have two continuous variables X and Y and if the change in X affects Y, the variables are said to be correlated. In other words, the systematic relationship between the variables is termed as correlation. When only 2 variables are involved the correlation is known as simple correlation and when more than 2 variables are involved the correlation is known as multiple correlation. When the variables move in the same direction, these variables are said to be correlated positively and if they move in the opposite direction they are said to be negatively correlated.

Scatter Diagram

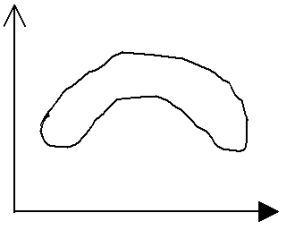
To investigate whether there is any relation between the variables X and Y we use scatter diagram. Let (x_1,y_1), (x_2,y_2)….(x_n,y_n) be n pairs of observations. If the variables X and Y are plotted along the X-axis and Y-axis respectively in the x-y plane of a graph sheet the resultant diagram of dots is known as scatter diagram. From the scatter diagram we can say whether there is any correlation between x and y and whether it is positive or negative or the correlation is linear or curvilinear.



Positive Correlation

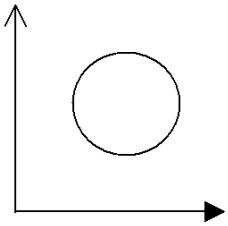


Negative correlation



Curvilinear

(or) non linear



no correlation

Pearsons Correlation coefficient

The measures of the degree of relationship between two continuous variables is called correlation coefficient. It is denoted by r.(in case of sample)and ρ (in case of population). The correlation coefficient r is known as Pearson's correlation coefficient as it was discovered by Karl Pearson. It is also called as product moment correlation.

The correlation coefficient r is given as the ratio of covariance of the variables X and Y to the product of the standard deviation of X and Y. Symbolically,

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

which can be simplified as

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

This correlation coefficient r is known as Pearson's Correlation coefficient. The numerator is termed as sum of product of X and Y and abbreviated as SP(XY). In the denominator the first term is called sum off squares of X (i.e) SS(X) and second term is called sum of squares of Y (i.e) SS(Y)

$$r = \frac{SP(XY)}{\sqrt{SS(X)}\sqrt{SS(Y)}}$$

The denominator in the above formula is always positive. The numerator may be positive or negative making r to be either positive or negative.

Assumptions in correlation analysis:

Correlation coefficient r is used under certain assumptions, they are

1. The variables under study are continuous random variables and they are normally distributed
2. The relationship between the variables is linear
3. Each pair of observations is unconnected with other pair (independent)

Properties

1. The correlation coefficient value ranges between –1 and +1.
2. The correlation coefficient is not affected by change of origin or scale or both.
3. If r > 0 it denotes positive correlation
   r< 0 it denotes negative correlation between the two variables x and y.
   r = 0 then the two variables x and y are not linearly correlated.(i.e)two variables are independent.
   r = +1 then the correlation is perfect positive
   r = -1 then the correlation is perfect negative.

Testing the significance of r

The significance of r can be tested by Student's t test. The test statistics is given by

$$t = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}$$

This t is distributed as Student's t distribution with (n-2) degrees of freedom.

The relationship between the variables is interpreted by the square of the correlation coefficient (r^2) which is called coefficient of determination. The value $1-r^2$ is called as coefficient of alienation. If r^2 is 0.72, it implies that on the basis of the samples 72% of the variation in one variable is caused by the variation of the other variable. The coefficient of determination is used to compare 2 correlation coefficients.

Problem

Compute Pearsons coefficient of correlation between plant height (cm) and yield (Kgs) as per the data given below:

Plant Height (cm)	39	65	62	90	82	75	25	98	36	78
Yield in Kgs	47	53	58	86	62	68	60	91	51	84

Solution

H_o: The correlation coefficient r is not significant

H_1: The correlation coefficient r is significant.

Level of significance 5%

From the data

n = 10

$$\sum x = 650 \quad \sum y = 660 \quad \sum xy = 45604 \quad \sum x^2 = 47648 \quad \sum y^2 = 45784$$

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

$$= \frac{45604 - \frac{(650)(660)}{10}}{\sqrt{47648 - \frac{(650)^2}{10}}\sqrt{45784 - \frac{(660)^2}{10}}}$$

$$= \frac{45604 - 42900}{(73.47)(47.1)} = 0.7804$$

Correlation coefficient is positively correlated.

Test Statistic

$$t = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}} \sim (n-2) \text{ d.f.}$$

$$t = \frac{0.7804}{\sqrt{\frac{1-(0.7804)^2}{10-2}}} = 3.530$$

$t_{tab} = t_{(10\text{-}2,\ 5\%los)} = 2.306$

Inference

$t > t_{tab}$, we reject null hypothesis.

The correlation coefficient r is significant. (i.e) there is a relation between plant height and yield.

Questions

1. Limits for correlation coefficient.

(a) $-1 \le r \le 1$ (b) $0 \le r \le 1$
(c) $-1 \le r \le 0$ (d) $1 \le r \le 2$

Ans: $-1 \le r \le 1$

2. The correlation coefficient is unaffected by change of

(a) Origin (b) scale
(c) Scale & origin (d) None of these

Ans: scale & origin

3. When r = +1, there is Perfect positive correlation.

Ans: True

4. Karl pearsons correlation coefficient is calculated only when the two variables are continuous.

Ans: True

5. The correlation between two variables is symmetric

Ans: True

6. The correlation between two variables is known as multiple correlation.

Ans: False

7. What is a scatter diagram? Mention its uses

8. Define correlation.

9. Explain the method how to calculate the Karl pearsons correlation coefficient?

10. Mention the properties of the correlation coefficient?

Regression

Regression is the functional relationship between two variables and of the two variables one may represent cause and the other may represent effect. The variable representing cause is known as independent variable and is denoted by X. The variable X is also known as predictor variable or repressor. The variable representing effect is known as dependent variable and is denoted by Y. Y is also known as predicted variable. The relationship between the dependent and the independent variable may be expressed as a function and such functional relationship is termed as regression. When there are only two variables the functional relationship is known as simple regression and if the relation between the two variables is a straight line I is known a simple linear regression. When there are more than two variables and one of the variables is dependent upon others, the functional relationship is known as multiple regression. The regression line is of the form y=a+bx where a is a constant or intercept and b is the regression coefficient or the slope. The values of 'a' and 'b' can be calculated by using the method of least squares. An alternate method of calculating the values of a and b are by using the formula:

The regression equation of y on x is given by y = a + bx

The regression coefficient of y on x is given by

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

and $a = \bar{y} - b\,\bar{x}$

The regression line indicates the average value of the dependent variable Y associated with a particular value of independent variable X.

Assumptions

1. The x's are non-random or fixed constants
2. At each fixed value of X the corresponding values of Y have a normal distribution about a mean.
3. For any given x, the variance of Y is same.
4. The values of y observed at different levels of x are completely independent.

Properties of Regression coefficients

1. The correlation coefficient is the geometric mean of the two regression coefficients
2. Regression coefficients are independent of change of origin but not of scale.
3. If one regression coefficient is greater than unit, then the other must be less than unit but not vice versa. ie. both the regression coefficients can be less than unity but both cannot be greater than unity, ie. if $b_1>1$ then $b_2<1$ and if $b_2>1$, then $b_1<1$.
4. Also if one regression coefficient is positive the other must be positive (in this case the correlation coefficient is the positive square root of the product of the two regression coefficients) and if one regression coefficient is negative the other must be negative (in this case the correlation coefficient is the negative square root of the product of the two regression coefficients). ie.if $b_1>0$, then $b_2>0$ and if $b_1<0$, then $b_2<0$.
5. If θ is the angle between the two regression lines then it is given by

$$\tan\theta = \frac{(1-r^2)\sigma_x\sigma_y}{r(\sigma_x^2+\sigma_y^2)}$$

Testing the significance of regression co-efficient

To test the significance of the regression coefficient we can apply either a t test or analysis of variance (F test). The ANOVA table for testing the regression coefficient will be as follows:

Sources of variation	**d.f.**	**SS**	**MS**	**F**
Due to regression	1	SS(b)	S_b^2	S_b^2 / S_e^2

Deviation from regression	n-2	SS(Y)-SS(b)	S_e^2	
Total	n-1	SS(Y)		

In case of t test the test statistic is given by

$$t = b / SE(b) \text{ where } SE(b) = s_e^2 / SS(X)$$

The regression analysis is useful in predicting the value of one variable from the given values of another variable. Another use of regression analysis is to find out the causal relationship between variables.

Uses of Regression

The regression analysis is useful in predicting the value of one variable from the given value of another variable. Such predictions are useful when it is very difficult or expensive to measure the dependent variable, Y. The other use of the regression analysis is to find out the causal relationship between variables. Suppose we manipulate the variable X and obtain a significant regression of variables Y on the variable X. Thus we can say that there is a causal relationship between the variable X and Y. The causal relationship between nitrogen content of soil and growth rate in a plant, or the dose of an insecticide and mortality of the insect population may be established in this way.

Example 1

From a paddy field, 36 plants were selected at random. The length of panicles(x) and the number of grains per panicle (y) of the selected plants were recorded. The results are given below. Fit a regression line y on x. Also test the significance (or) regression coefficient.

The length of panicles in cm (x) and the number of grains per panicle (y) of paddy plants.

S.No.	Y	X	S.No.	Y	X	S.No.	Y	X
1	95	22.4	13	143	24.5	25	112	22.9
2	109	23.3	14	127	23.6	26	131	23.9
3	133	24.1	15	92	21.1	27	147	24.8
4	132	24.3	16	88	21.4	28	90	21.2
5	136	23.5	17	99	23.4	29	110	22.2
6	116	22.3	18	129	23.4	30	106	22.7
7	126	23.9	19	91	21.6	31	127	23.0

8	124	24.0	20	103	21.4	32	145	24.0
9	137	24.9	21	114	23.3	33	85	20.6
10	90	20.0	22	124	24.4	34	94	21.0
11	107	19.8	23	143	24.4	35	142	24.0
12	108	22.0	24	108	22.5	36	111	23.1

Null Hypothesis H_o: regression coefficient is not significant.

Alternative Hypothesis H_1: regression coefficient is significant.

$$\sum y = 4174 \qquad \sum y^2 = 496258 \qquad \bar{y} = \frac{\sum y}{n} = 115.94$$

$$\sum x = 822.9 \qquad \sum x^2 = 18876.83 \qquad \bar{x} = \frac{\sum x}{n} = 22.86$$

$$\sum xy = 96183.4$$

$$\text{SS(Y)} = \sum y^2 - \frac{(\sum y)^2}{n} = 496258 - \frac{(4174)^2}{36} = 12305.8889$$

$$\text{SS(X)} = \sum x^2 - \frac{(\sum x)^2}{n} = 18876.83 - \frac{(822.9)^2}{36} = 66.7075$$

The regression line y on x is $\bar{y}$ =a+ b $\bar{x}$

$$b1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{96183.4 - \frac{(822.9)(4174)}{36}}{66.7075} = 11.5837$$

$\bar{y}$ =a+ b $\bar{x}$

115.94 = a + (11.5837)(22.86)

a=115.94-264.8034

a=-148.8633

The fitted regression line is y =-148.8633+11.5837x

$$SS(b) = \frac{\left(\sum xy - \frac{\sum x \sum y}{n}\right)^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} = \frac{(722.7167)^2}{66.7075} = 8950.8841$$

Anova Table

Sources of Variation	d.f.	SS	MSS	F
Regression	1	8950.8841	8950.8841	90.7093
Error	36-2=34	3355.0048	98.6766	
Total	35	12305.8889		

For t-test

$$t = \frac{b}{SE(b)} \sim t_{(n-2)} d.f$$

$$SE(b = \sqrt{\frac{Se^2}{SS(X)}} = \sqrt{\frac{98.6776}{66.7075}} = 1.2162$$

$$t = \frac{11.5837}{1.2162} = 9.5245$$

Table Value:

$t_{(n-2)}$ d.f.=t_{34} d.f at 5% level=2.032

$t > t_{tab}$. we reject H_o.

Hence t is significant.

Questions

1. When the correlation coefficient r = +1, then the two regression lines

a) are perpendicular to each other b) coincide

c) are parallel to each other d) none of these

Ans: coincide

2. If one regression coefficient is greater than unity then the other must be

a) greater than unity b) equal to unity

c) less than unity d) none of these

Ans: less than unity

3. If the correlation between the two variables is positive the regression coefficient will be positive.

Ans: True

4. The Dependent variable is also called as predicted variable.

Ans: True

5. Correlation coefficient is the geometric mean of two regression coefficients.

Ans: True

6. Regreesion gives the functional relationship between two variables.

Ans: True

7. What is meant by Cause and effect?

8. State the properties of regression coefficient.

9. From the following data, find the regression equation

$\sum X = 21, \sum Y = 20, \sum X^2 = 91, \sum XY = 74, n = 7$

10. Explain how to fit the regression equation of y on x and test the significance of the regression coefficient.

Design of experiments

Choice of treatments, method of assigning treatments to experimental units and arrangement of experimental units in different patterns are known as designing an experiment. We study the effect of changes in one variable on another variable. For example how the application of various doses of fertilizer affects the grain yield. Variable whose change we wish to study is known as **response variable**. Variable whose effect on the response variable we wish to study is known as **factor**.

Treatment: Objects of comparison in an experiment are defined as treatments.
Examples are Varieties tried in a trail and different chemicals.

Experimental unit: The object to which treatments are applied or basic objects on which the experiment is conducted is known as experimental unit.

Example: piece of land, an animal, etc

Experimental error: Response from all experimental units receiving the same treatment may not be same even under similar conditions. These variations in responses may be due to various reasons. Other factors like heterogeneity of soil, climatic factors and genetic differences, etc also may cause variations (known as extraneous factors). The variations in response caused by extraneous factors are known as **experimental error**.

Our aim of designing an experiment will be to minimize the experimental error.

Basic principles

To reduce the experimental error we adopt certain principles known as basic principles of experimental design.

The basic principles are 1) Replication, 2) Randomization and 3) Local control

Replication

Repeated application of the treatments is known as replication.

When the treatment is applied only once we have no means of knowing about the variation in the results of a treatment. Only when we repeat several times we can estimate the experimental error.

With the help of experimental error we can determine whether the obtained differences between treatment means are real or not. When the number of replications is increased, experimental error reduces.

Randomization

When all the treatments have equal chance of being allocated to different experimental units it is known as randomization.

If our conclusions are to be valid, treatment means and differences among treatment means should be estimated without any bias. For this purpose we use the technique of randomization.

Local Control

Experimental error is based on the variations from experimental unit to experimental unit. This suggests that if we group the homogenous experimental units into blocks, the experimental error will be reduced considerably. Grouping of homogenous experimental units into blocks is known as local control of error.

In order to have valid estimate of experimental error the principles of replication and randomization are used.

In order to reduce the experimental error, the principles of replication and local control are used.

In general to have precise, valid and accurate result we adopt the basic principles.

Questions

1. For valid conclusions we should have

(a) Unbiased estimate (b) biased estimate
(c) random estimate (d) none of these

Ans: Unbiased estimate

2. Response variable is also called as

(a) Independent variable (b) dependent variable
(c) treatment (d) error

Ans: dependent variable

3. The genetic differences of varieties are termed as extraneous factors.

Ans: True

4. Repetition of the treatment is known as replication.

Ans: True

5. Replicaion will increase the error.

Ans: False

6. Basic principles are adopted to reduce the experimental error.

Ans: True

7. What is experimental error?

8. Define treatment and experimental unit.

9. What is meant by designing an experiment?

10. Explain the basic principles and its uses?

Completely Randomized Design (CRD)

CRD is the basic single factor design. In this design the treatments are assigned completely at random so that each experimental unit has the same chance of receiving any one treatment. But CRD is appropriate only when the experimental material is homogeneous. As there is generally large variation among experimental plots due to many factors CRD is not preferred in field experiments.

In laboratory experiments and greenhouse studies it is easy to achieve homogeneity of experimental materials and therefore CRD is most useful in such experiments.

Layout of a CRD

Completely randomized Design is the one in which all the experimental units are taken in a single group which are homogeneous as far as possible.

The randomization procedure for allotting the treatments to various units will be as follows.

Step 1: Determine the total number of experimental units.

Step 2: Assign a plot number to each of the experimental units starting from left to right for all rows.

Step 3: Assign the treatments to the experimental units by using random numbers.

The statistical model for CRD with one observation per unit

$Y_{ij} = \propto + t_i + e_{ij}$

$\propto$ = overall mean effect

t_i = true effect of the i^{th} treatment

e_{ij} = error term of the j^{th} unit receiving i^{th} treatment

The arrangement of data in CRD is as follows:

	Treatments				
	T_1	T_2	T_i	T_K	
	y_{11}	y_{21}	y_{i1}	Y_{K1}	
	y_{12}	y_{22}	y_{i2}	Y_{K2}	
	y_{1r1}	y_{2r2}	y_{iri}	$Y_{k\,rk}$	
Total	Y_1	Y_2	Y_i	T_k	GT

(GT – Grand total)

The null hypothesis will be

$H_o : \propto_1 = \propto_2 = \ldots\ldots\ldots\ldots = \propto_k$ or There is no significant difference between the treatments

And the alternative hypothesis is

$H_1: \propto_1 \neq \propto_2 \neq \ldots\ldots\ldots\ldots \neq \propto_k$. There is significant difference between the treatments

The different steps in forming the analysis of variance table for a CRD are:

1. $C.F = \frac{(GT)^2}{n}$

   n= Total number of observations

2. $\text{Total SS} = \text{TSS} = \sum_{i=1}^{k}\sum_{j=1}^{v} y_{ij}^{\ 2} - C.F$

3. $\text{Treatment SS} = \text{TrSS} = \frac{Y_1^2}{r_1} + \frac{Y_2^2}{r_2} + \ldots\ldots\ldots + \frac{Y_k^2}{r_k} - C.F$

   $= \sum_{i=1}^{k} \frac{Y^2{}_i}{r_i} - C.F$

4. $\text{Error SS} = \text{ESS} = \sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}^{\ 2} - \sum_{i=1}^{k} \frac{Y_i^2}{r_i}$

   = TSS – TrSS

5. Form the following ANOVA table and calculate F value.

Source of variation	**d.f.**	**SS**	**MS**	**F**
Treatments	t-1	TrSS	TrMS= $\frac{TrSS}{t-1}$	$\frac{TrMS}{EMS}$
Error	n-t	ESS	EMS= $\frac{ESS}{n-t}$	
Total	n-1	TSS		

6. Compare the calculated F with the critical value of F corresponding to treatment degrees of freedom and error degrees of freedom so that acceptance or rejection of the null hypothesis can be determined.

7. If null hypothesis is rejected that indicates there is significant differences between the different treatments.

8. Calculate C D value.

C.D. = SE(d). t

$$\text{where S.E(d)} = \sqrt{EMS(\frac{1}{r_i} + \frac{1}{r_j})}$$

r_i = number of replications for treatment i

r_j = number of replications for treatment j and

t is the critical t value for error degrees of freedom at specified level of significance, either 5% or 1%.

Advantages of a CRD

1. Its layout is very easy.
2. There is complete flexibility in this design i.e. any number of treatments and replications for each treatment can be tried.
3. Whole experimental material can be utilized in this design.
4. This design yields maximum degrees of freedom for experimental error.
5. The analysis of data is simplest as compared to any other design.
6. Even if some values are missing the analysis can be done.

Disadvantages of a CRD

1. It is difficult to find homogeneous experimental units in all respects and hence CRD is seldom suitable for field experiments as compared to other experimental designs.
2. It is less accurate than other designs.

Questions

1. CRD can be used with
   (a) Equal replication (b) unequal replication
   (c)Equal and unequal replication (d) single replication

Ans: Equal and unequal replication

2. When there are 5 treatments each replicated 4 times the total number of experimental plots will be
(a)5 (b) 4 (c) 9 (d)20
Ans: 20

3. In CRD the error degrees of freedom is rt-1.
Ans: True

4. CRD can be adopted only when the experimental material is homogenous.
Ans: True

5. CRD is a single factor experiment.
Ans: True

6. In CRD the total sum of squares is divided into treatment sum of squares, Replication sum of squares and error sum of squares.
Ans: False

7. Mention any two advantages of CRD?

8. When the treatments are large in a CRD what will happen to the precision of the experiment?

9. Explain the Layout of the CRD?

10. Explain the Layout of the CRD?

Randomized Blocks Design (RBD)

When the experimental material is heterogeneous, the experimental material is grouped into homogenous sub-groups called blocks. As each block consists of the entire set of treatments a block is equivalent to a replication.

If the fertility gradient runs in one direction say from north to south or east to west then the blocks are formed in the opposite direction. Such an arrangement of grouping the heterogeneous units into homogenous blocks is known as randomized blocks design. Each block consists of as many experimental units as the number of treatments. The treatments are allocated randomly to the experimental units within each block independently such that each treatment occurs once. The number of blocks is chosen to be equal to the number of replications for the treatments.

The analysis of variance model for RBD is

$Yij = \propto + ti + rj + eij$

where

$\propto$ = the overall mean

t_i = the i^{th} treatment effect

r_j = the j^{th} replication effect

eij = the error term for i^{th} treatment and j^{th} replication

Analysis of RBD

The results of RBD can be arranged in a two way table according to the replications (blocks) and treatments.

There will be r x t observations in total where r stands for number of replications and t for number of treatments. .

The data are arranged in a two way table form by representing treatments in rows and replications in columns.

Treatment	Replication					Total
	1	2	3	……	r	
1	y11	y12	y13	……	y1r	T1
2	y21	y22	y23	……	y2r	T2
3	y31	y32	y33	……	y3r	T3
t	yt1	yt2	yt3	…….	ytr	Tt
Total	R1	R2	R3		Rr	G.T

In this design the total variance is divided into three sources of variation viz., between replications, between treatments and error

$$CF = \frac{(GT)^2}{rt}$$

Total SS=TSS= $\Sigma \Sigma$ y ij ^{2} – CF

Replication SS=RSS= = $\frac{1}{t} \Sigma Rj^2$ – CF

Treatments SS=T_rSS = $\frac{1}{r} \Sigma Ti^2$ - CF

Error SS=ESS = Total SS – Replication SS – Treatment SS

The skeleton ANOVA table for RBD with t treatments and r replications

Sources of variation	**d.f.**	**SS**	**MS**	**F Value**
Replication	r-1	RSS	RMS	RM S/ EM S
Treatment	t-1	TrSS	TrMS	TrMS/EMS
Error	(r-1) (t-1)	ESS	EMS	
Total	rt –1	TSS		

CD = SE(d) . t where S.E(d)= $\sqrt{\frac{2EMS}{r}}$

t = critical value of t for a specified level of significance and error degrees of freedom

Based on the CD value the bar chart can be drawn. From the bar chart conclusion can be written.

Advantages of RBD

The precision is more in RBD. The amount of information obtained in RBD is more as compared to CRD. RBD is more flexible. Statistical analysis is simple and easy. Even if some values are missing, still the analysis can be done by using missing plot technique.

Disadvantages of RBD

When the number of treatments is increased, the block size will increase. If the block size is large maintaining homogeneity is difficult and hence when more number of treatments is present this design may not be suitable.

Questions

1. RBD can be used with

(a) Equal replication (b) unequal replication
(c)Equal and unequal replication (d) single replication

Ans: Equal replication

2. When there are 5 treatments each replicated 4 times the total number of experimental plots will be

(a) 5 (b) 4 (c) 9 (d) 20

Ans: 20

3. In RBD the error degrees of freedom is (r-1)(t-1).

Ans: True

4. RBD can be adopted when the experimental material is heterogeneous.

Ans: True

5. In RBD the blocking is done in one direction.

Ans: True

6. In RBD the total sum of squares is divided into treatment sum of squares, Replication sum of squares and error sum of squares.

Ans: True

7. Mention any two advantages of RBD?

8. Furnish the ANOVA model for RBD

9.Explain the Layout of the RBD?

10. Explain the computational procedure of RBD?

Latin Square Design

When the experimental material is divided into rows and columns and the treatments are allocated such that each treatment occurs only once in each row and each column, the design is known as L S D.

In LSD the treatments are usually denoted by A B C D etc.

For a 5 x 5 LSD the arrangements may be

A	***B***	***C***	***D***	***E***
B	***A***	***E***	***C***	***D***
C	***D***	***A***	***E***	***B***
D	***E***	***B***	***A***	***C***
E	***C***	***D***	***B***	***A***
Square 1				

	B	***C***	***D***	***E***
B	***A***	***D***	***E***	***C***
C	***E***	***A***	***B***	***D***
D	***C***	***E***	***A***	***B***
E	***D***	***B***	***C***	***A***
Square 2				

A	***B***	***C***	***D***	***E***
B	***C***	***D***	***E***	***A***
C	***D***	***E***	***A***	***B***
D	***E***	***A***	***B***	***C***
E	***A***	***B***	***C***	***D***
Square 3				

Analysis

The ANOVA model for LSD is

$$Y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk}$$

r_i is the i^{th} row effect

c_j is the j^{th} column effect

t_k is the k^{th} treatment effect and

eijk is the error term

The analysis of variance table for LSD is as follows:

Sources of Variation	d.f.	S S	M S	F
Rows	t-1	RSS	RMS	RMS/EMS
Columns	t-1	CSS	CMS	CMS/EMS
Treatments	t-1	TrSS	TrMS	TrMS/EMS
Error	(t-1)(t-2)	ESS	EMS	
Total	t^2-1	TSS		

F table value

$F_{[t-1),(t-1)(t-2)]}$ degrees of freedom at 5% or 1% level of significance

Steps to calculate the above Sum of Squares are as follows:

Correction Factor $(\mathrm{CF}) = \frac{(GT)^2}{(t)^2}$

Total Sum of Squares $(\mathrm{TSS}) = \sum (y_{ijk})^2 - CF$

Row sum of squares $(\mathrm{RSS}) = \frac{1}{t}\sum_{i=1}^{t} (R_i)^2 - CF$

Column sum of squares $(\mathrm{CSS}) = \frac{1}{t}\sum_{j=1}^{t} (C_j)^2 - CF$

Treatment sum of squares $(\mathrm{TrSS}) = \frac{1}{t}\sum_{k=1}^{t} (T_k)^2 - CF$

Error Sum of Squares = TSS-RSS-CSS-TrSS

These results can be summarized in the form of analysis of variance table.

Calculation of SE, SE (d) and CD values

$$\mathrm{SE} = \sqrt{\frac{EMS}{r}}$$

where r is the number of rows

$\bar{\ }$.

CD= SE (d). t

where t = table value of t for a specified level of significance and error degrees of freedom

Using CD value the bar chart can be drawn and the conclusion may be written.

Advantages

- LSD is more efficient than RBD or CRD. This is because of double grouping that will result in small experimental error.
- When missing values are present, missing plot technique can be used and analysed.

Disadvantages

- This design is not as flexible as RBD or CRD as the number of treatments is limited to the number of rows and columns. LSD is seldom used when the number of treatments is more than 12. LSD is not suitable for treatments less than five.

Because of the limitations on the number of treatments, LSD is not widely used in agricultural experiments.

Note: The number of sources of variation is two for CRD, three for RBD and four for LSD.

Questions

1. In a Latin Square design the number of rows will be equal to
   a) No. of columns b) No. of Treatments
   c) No. of Replications d) No. of Columns & Number of Treatments

Ans: No. of Columns & Number of Treatments

2. In a Latin Square design with 5 treatments the number of experimental units will be equal to
   a) 25 b) 20 c) 24 d) 36

Ans: 25

3. If the number of experimental units is 36 then the number of rows will be equal to 6.
Ans: True

4. The error degrees of freedom in LSD with t treatments will be (t-1)(t-2).
Ans: True

5. If the experimental material is homogeneous then LSD can be adopted.
Ans: False

6. In a LSD each treatment should occur only once in each row and each column.
Ans: True

7. Furnish the ANOVA model for LSD.

8. What is a Latin Square Design?

9. State the advantages and disadvantages of LSD.

10. Explain the computational procedure of LSD?

References Books:

1. Statistics for Agricultural Sciences, G. Nageswara Rao, Second Edition, BS Publications, Hyderabad

2. A Text book of Agricultural Statistics, R. Rangaswamy, New Age International (P) Limited, publishers

3. Statistical Methods, K.P. Dhamu and K. Ramamoorthy, AGROBIOS (INDIA)

4. Fundamentals of Mathematical Statistics, S.C. Gupta and V.K. Kapoor, Sultan Chand & Sons Educational Publications

5. Fundamentals Applied Statistics, S.C. Gupta and V.K. Kapoor, Sultan Chand & Sons Educational Publications

6. Design Resources Server: www.iasri.res.in