Teaching Material

Course Title: Statistical Methods
Credit Hrs. 2(1+1)
Course No. ASM- 211
Semester-III

Compiled by:

Dr. S. N. Singh (Univ.Professor)
Dr. Fozia Homa (Asstt. Professor)
Mr. Subrat Keshori Behera (Asstt. Professor)

Department of Statistics, Mathematics & Computer Application

BIHAR AGRICULTURAL COLLEGE, SABOUR

BIHAR AGRICULTURAL UNIVERSITY, SABOUR
BHAGALPUR

PIN 813 210.

Course Title: Statistical Methods
Credit Hrs. 2(1+1)
Course Content

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

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

References:

1. Hand Book of Agricultural Statistics by S.R.S. Chandel.
2. Fundamentals of Mathematical Statistics (Vol.I&II) by S.C. Gupta and V.K. Kapoor.
3. Mathematical Statistics by J.N. Kapur and H.C. Saxena.
4. Elements of Statistics by B.N. Asthana.
5. Elements of Statistics by E.B. Mode.
6. Statistical Methods for Agricultural Workers by V.G.Panse & P.V. Sukhatme.
7. Design and Analysis of Experiments by M.N. Das & N.C. Giri.

Introduction

Origin and Development of Statistics: The Statistics seems to have been derived from the Latin word 'Status' or the Italian word 'Statista' or the German word 'Statistik' each of which means a "political state". In ancient times, the government used it to collect the information regarding the population and "property or wealth" of the country.

Sir, Ronald A. Fisher (1890-1962) known as the father of statistics who applied statistics into various field such as Genetics, Biometry, Education and Agriculture etc.

Definition of statistics: "These are the aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other". by Prof. Horace & Secrist.

When it is used in plural, it means the quantitative data.

When it is used in singular, it is defined as "science which deals with collection, presentation, analysis and interpretation of numerical data". by Croxton and Cowden.

Purpose or Function of statistics:

1. To summarise the large mass of data into a few representative value.
2. To establish a relation among the data sets or within each data set.

Importance and Scope:

It has wide applications in almost all sciences like social as well as physical: Planning, Economics, Business, Industry, Meteorology, Education, War, Agriculture, Psychometry etc.

Limitations of statistics:

- Statistics is not suited to the study of qualitative phenomenon.
- Statistics does not study individuals.
- Statistical laws are not exact.
- Statistics is liable to be misused.

Application of Statistics in Agriculture

In Agriculture it is used as collection, presentation, analysis and interpretation of numerical data. In Agriculture it is applied in design of experiments through Analysis of variance and various statistical tools are applied to find:

- Suitable fertilizer dose.
- Suitable varieties of different crops.
- Date of sowing,
- Method of transplanting
- In meteorology weather forecasting,
- Disease and insect pest forecasting,
- Weather parameters (temperature rainfall, sunshine, wind velocity, humidity etc.)

- Yield of the different crops,
- Yield attributes, morphological and biochemical traits,
- Chemical and physical studies of soil,
- Evaluation of pesticide efficacy,
- Cost of cultivation.
- Crop cutting experiment to estimate the yield of the different crops,
- Preharvest forecastof yield based on biometrical characters and farmers' appraisa.
- Forecasting of yield of different crops based on meteorological data,

Statisstical tools:

- Measures of central tendency,measures of dispersion, graphical representations
- Different Sampling techniques in sample survey.
- Different Test of significance
- Correlation,regression,multiple correlations and multiple regression.
- Rank correlation and more.

Frequency distribution

It is an arrangement of variate values along with their respective frequency.

Frequency : Frequency is derived from "how frequently a variable occurs"

Each class is defined by two boundaries Lower boundary is called lower limit and upper boundary is called upper limit.

Range = Maximum Value – Minimum Value.
Class Interval = Upper limit – Lower limit
Mid value = (Upper limit + Lower limit)/2; Frequency density = frequency/class width
Relative frequency = $\frac{frequence\ of\ each\ class}{Total\ frequency}$

The following points may be kept in mind for classification of data:

(i) The classes should be clearly defined and free from ambiguity.
(ii) The classes should be exhaustive, i.e. each of the given value should be included in one of the class.
(iii) The classes should be mutually exclusive and non-overlapping.
(iv) The classes should be of equal width.
(v) Indeterminate classes, open end classes: less than or greater than should be avoided as far as possible.
(vi) The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15. Struges used the formulae for determining the approximate number of classes $K = 1 + 3.322 \log_{10}N$, where N is the total frequency.

Graphical Representation

Graphical representations are represented by points plotted on a graph paper which makes the unwieldy data intelligible and conveys to the eye the general run of observations. Graphical representation also facilitates the comparison of two or more frequency distribution.

Some important type of graphical representation are:

(i) Histogram
(ii) Frequency Polygon
(iii) Frequency curve

Histogram: If the frequency distribution is not continuous first it is to be converted into continuous distribution by subtracting 0.5 from the lower limit and adding 0.5 to the upper limit of each classes. In drawing histogram of a continuous frequency distribution we first mark off class intervals on x-axis and corresponding frequency on y-axis by selecting a suitable scale. On each class interval we erect rectangles with heights proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If, however, the classes are of unequal width then the heights of the rectangle, will be proportional to the ratio of the frequency to the width of the class, the diagrams of continuous rectangles so obtained is called histogram.

Frequency polygon: For ungrouped distribution, the frequency polygon is obtained by plotting the points with abscissa as the variate values and the ordinate as the corresponding frequency and joins the points by means of straight line. For a grouped frequency distribution the abscissa of the points are mid values of the class intervals. The frequency polygon so obtained should be extended to the base line(x-axis) at both ends so that it meets the x-axis at the mid points of two hypothetical classes, the class before the first class and the class after the last class, each assumed to have zero frequency.

Frequency curve: If the class intervals are of small width, the frequency polygon can be approximated to frequency curve and we join the points with smooth hand. The frequency curve can also be obtained by drawing a smooth free hand curve through the vertices of the frequency polygon.

Measures of central Tendency

"Central tendency may be defined as a value of the variate which is thoroughly representative of the series or the distribution as a whole". They give us an idea about the concentration of the values in the central part of the distribution. The following are the measures of central tendency.

(i) Arithmetic mean or mean.
(ii) Median
(iii) Mode
(iv) Geometric Mean
(v) Harmonic Mean

Characteristics for an ideal measures of central tendency:

i. It should be rigidly defined.
ii. It should be readily comprehensible and easy to calculate.
iii. It should be based upon all the observations.
iv. It should be suitable for further mathematical treatment.
v. It should be affected as little as possible by fluctuation of sampling.
vi. It should not affected much by extreme values.

Arithmetic Mean: If $x_1, x_2, \ldots x_n$ are n observations, then Arithmetic mean is given by

$A.M. = (x_1 + x_2 + \ldots + x_n)/n$

In case of frequency distribution,

Mean $= (x_1 f_1 + x_2 f_2 + \ldots + x_n f_n)/N$ Where, $N = \sum_{i=1}^{n} f_i$

In case of grouped or continuous frequency distribution, x is taken as the mid value of the corresponding class.

Properties of Arithmetic Mean:

1. AM is independent of change of origin and scale both.

2.Algebraic sum of the deviations of a set of values from their arithmetic mean is zero.

3. The sum of the squares of the deviations of a set of values is minimum when taken about mean.

4. Combined mean, $\overline{x} = \frac{\sum_{i=1}^{n_1} n_i \overline{x_i}}{\sum_{i=1}^{n_1} n_i}$

Merits of Arithmetic mean:

(i) It is rigidly defined.
(ii) It is easy to understand and easy to calculate.
(iii) It is based upon all the observations.
(iv) It is amenable to algebraic treatment.
(v) It is affected least by of fluctuation of sampling. This property is some time described by saying that A.M. is **stable average**.

Demerits:

1. Arithmetic mean is affected very much by the extreme values.
2. It can not be determined by inspection.
3. It can not be used in qualitative characteristics like intelligence, honesty, beauty.
4. Arithmetic mean can not be accurately obtained if single observation is missing or lost.
5. Arithmetic mean can not be calculated if the extreme class is open.

Uses: It is generally used in all the subjects of studies like social and economic studies. Average cost of production, Average price, Average yield/ acre etc.

Median:

Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observation i.e. it is the value such that the number of observation above it is equal to the number of observation below it.

Step-I: In case of ungrouped data, if the number of observation is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude.

Step-II: In case of even number of observations there are two middle terms and median is obtained by taking the arithmetic mean of these middle terms after arranging the series in ascending or descending order.

Step-III: In case of discrete frequency distribution, median is obtained by:

(i) Construct cumulative frequencies.
(ii) Find N/2, Where, N = $\sum_{i=1}^{n} f_i$
(iii) See the cumulative frequency (c.f) just greater than N/2 and the corresponding value of x gives median.

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

Medain = $L + \frac{\frac{N}{2} - C}{f} \times h$

Where, L is the lower limit of the median class.
f is the frequency of the median class.
h is the magnitude of the median class.
c is the cumulative frequency preceding the median class.
N = $\sum f_i$

Merits of Median:

(i) It is rigidly defined.
(ii) It is easy to understand and to calculate.
(iii) It is not at all affected by extreme values.
(iv) It can be calculated for distribution with open the classes.

Demerits of median:

(i) It is not amenable to algebraic treatment.
(ii) It is affected much by fluctuation of sampling.
(iii) In case of even number of observation median can not be determined exactly.
(iv) It is not based on all the observations.

Uses: 1. Median is the only average to be used while dealing with qualitative data. e.g. to find the average intelligence or average honesty among a group of people.

2. It is to be used for determining the typical value in problems concerning distribution of wages etc.

Mode: This is that value of the variable which occurs most frequently or whose frequency is maximum.

In case of continuous distribution mode is given by:

Mode = $L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$

Where, L= Lower limit of the modal class.

f_m = maximum frequency of modal class.

f_1 & f_2 are the frequencies of preceding and following of the modal class respectively.

h = Magnitude of the modal class.

Merits:

1. It is readily comprehensible and easy to calculate.
2. It is not at all affected by the extreme values
3. It can be obtained simply by inspection.
4. It can be computed in case of open end class.

Demerits:

1. It is not rigidly defined. A distribution with two modes is called **bi-modal** and the distribution with more than two modes is called **multi-modal**.
2. It is not suitable for further mathematical treatment.
3. It is not based on all the observations.
4. It is affected to a great extent by fluctuation of sampling.

Uses: Mode is the average to be used in finding the ideal size e.g. in business forecasting, in manufacture of ready-made garments, shoes size etc.

For a symmetrical distribution; mean, median and mode coincide. If the distribution is moderately asymmetrical the mean ,median and mode obey the following empirical relations:

Mean – median = 1/3 (Mean – mode)
mode = 3 median – 2 mean

DISPERSION

"Dispersion is the measure of extent to which individual items vary by" L.R Connor.

Consider the series (i) 7, 8, 9, 10, 11 (ii) 3, 6, 9, 12, 15 (iii) 1, 5, 9, 13, 17

In all these cases we see that the number of observation is 5 and the mean is 9. We can not form an idea as to whether it is the average of 1st series or 2nd series or third series or any other series of 5 observation whose sum is 45. Thus we see that the measure of central tendency are inadequate to give us a complete idea of distribution. They must be supported and supplemented by some other measures. One such measure is dispersion.

Literal meaning of dispersion is 'Scatteredness'. In dispersion, we have an idea about the homogeneity or heterogeneity of the distribution. We say that series (i) is more homogeneous (less dispersed) than the series (ii) or (iii) or we say that series (iii) is more heterogeneous (more scattered) than the series (i) or (ii)

Characteristics for an ideal Measure of dispersion:

i. It should be rigidly defined.
ii. It should be easy to calculate and easy to understand.
iii. It should be based on all the observations.
iv. It should be amenable to further mathematical treatment.
v. It should be affected as little as possible by fluctuation of sampling.

Following are the measures of dispersion:

1. Range.
2. Quartile deviation or Semi- interquartile range .
3. Mean deviation.
4. Standard deviation.

1.Range: Range is the difference between two extreme observations of the distribution. If A and B are two extreme values then

Range = A - B

Where, A and B are the two extreme value

example: 2, 4, 6, 8, 25, 30

Range = 30 - 2 = 28

Range is not a reliable measure of dispersion as it is based upon only two extreme values.

2. Quartile deviation = $(Q_3 - Q_1)/2$

Where, Q_1 and Q_3 are the 1st and 3rd quartile respectively.

It is not a reliable measure of dispersion as it covers only 50% of the distribution.

(3) Mean Deviation :

If xi / f_i is the frequency distribution then mean deviation is given by

Mean Deviation = $\frac{1}{N}\Sigma fi\,|xi - A|$ Where , A = Mean
= Median
= Mode

- Mean deviation is also not a reliable measure of dispersion as it takes only positive value due modulus sign.
- Mean deviation is least when measured from median.

(4) Standard deviation

It is the positive square root of the arithmetic mean of the square of deviation from arithmetic mean.

Standard deviation is denoted by σ (sigma)

$\sigma = \sqrt{\frac{1}{n}\Sigma (x_i - \bar{x})^2}$ for ungrouped data.

$\sigma = \sqrt{\frac{1}{N}\Sigma fi\,(x_i - \bar{x})^2}$ for grouped data.

Shortcut method

$\sigma^2 x_i = \sigma^2 d_i = \frac{1}{n}\Sigma\, di^2 - \left(\frac{1}{n}\Sigma fdi\right)^2$ where, $d_i = x_i - A$

$\sigma^2 x_i = h^2\sigma^2 d_i = h^2\left[\frac{1}{N}\Sigma fi\, di^2 - \left(\frac{1}{N}\Sigma fid\right)^2\right]$ where, $d_i = \frac{x_i - A}{h}$

Where A= Arbitrary value

h = Class interval

It is a reliable measures of dispersion as it satisfies all characteristics for an ideal measures of dispersion.

Standard deviation or, variance is independent of change of origin but not of scale.

Coefficient of dispersion

Whenever, you want to compare the variability in two series, we compute coefficient of dispersion, not measure of dispersion. Coefficient of dispersion is independent of unit of measurement

3, 5, 7, 11, 15, 17 (cm)

4, 6, 8, 10, 12 (kg)

We can compare the above two series although they are measured in different units. Following are the measures of dispersion:

1. Coefficient of dispersion based upon range $= \frac{A - B}{A + B}$
2. Coefficient of dispersion based upon Quartile deviation $\frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$
3. Coefficient of dispersion based upon mean deviation $= \frac{\text{M. D.}}{\text{Av. from which it is calculated}}$
4. Coefficient of dispersion based upon S.D.$= \frac{\sigma}{\bar{x}}$

Coefficient of variation: It is 100 times co efficient of dispersion based upon standard deviation.

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

Whenever, we want to compare the variation in two series,we compute coefficient of variation each series separately. The series having more. C.V. in comparison to other is said to be more variable than others and the series having less. C.V. in comparison to other is said to be more consistent than others.

Moments.

The r^{th} moment of a variable X about the point x = A, usually denoted by μ_r' is given by:

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} fi\,(x_i - A)^r,\ \sum fi = N$$

$$= \frac{1}{N}\ \sum fi\, d^r where\ d_i = X_i - A$$

The r^{th} moment of a variable x about the mean, $\bar{x}$ usually denoted by μ_r is given by:

$$\mu_r = \frac{1}{N}\sum fi\,(x_i - \bar{x})^o = \frac{1}{N}\sum fi\ = 1$$

and $\mu_1 = \frac{1}{N}\sum fi\,(x_i - \bar{x}) = 0$ (being the algebraic sum of deviation from mean is zero)

Also, $\mu_2 = \frac{1}{N}\sum_i fi\ (x_{i-}\overline{x})^2 = \sigma^2$

i.e. $\mu_0 = 1$, $\mu_1 = 0$ and $\mu_2 = \sigma^2$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\beta_2 = \frac{\mu_4}{\mu_2^2}$

Skewness and kurtosis

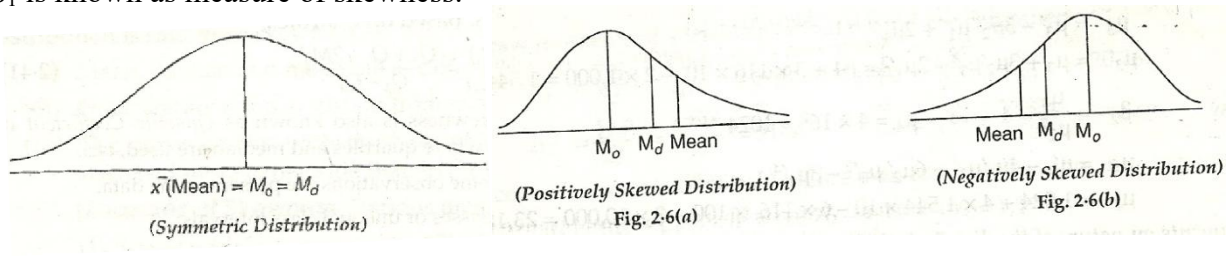
Skewness mean 'lack of symmetry'. It gives us an idea about the shape of the curve, in skewness.

(i) Mean ≠ median ≠ mode.

(ii) Quartiles are not equidistant from median

(iii) The curve is not symmetrical but stretched more to one side than the other.

β_1 is known as measure of skewness.



Kurtosis enables us to have an idea about the 'flatness or peakedness' of the frequency curve. It is measured by the coefficient β_2 or its derivation γ_2 given by:

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

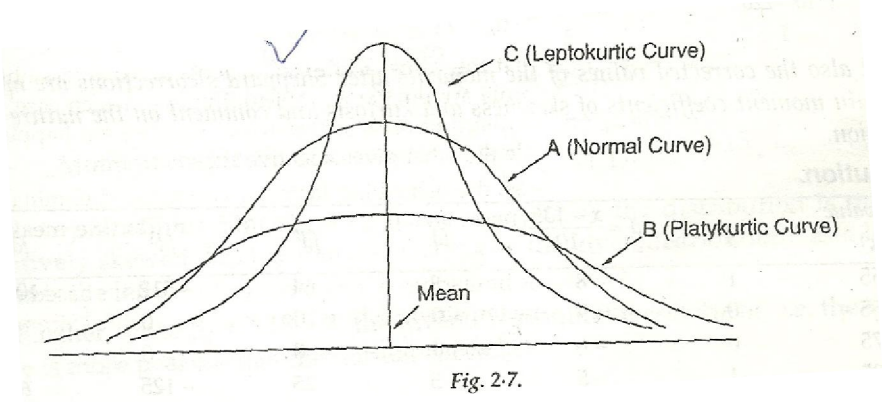


Fig. 2·7.

Theory of Probability

Introduction: if an experiment is conducted under essentially homogeneous and similar conditions we generally come across two types of situations.

Deterministic or Predictable phenomena	**Probabilistic or** unpredictable **phenomena**.
Result can be predicted with certainty eg. for a perfect gas V α $\frac{1}{P}$	Result can not be predicted with certainty eg. in tossing of a coin one may not be sure whether he will get head or tail.

Mathematical or classical or Priori Definition of Probability: If a trial results in n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to the happening of an event E, then the probability 'p'of happening of E is given by

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

Obviously, p as well as q are non- negative i.e . $0 \le p \le 1$, $0 \le q \le 1$
P(E) = 1, E is certain event, P(E) = O, E is impossible event.

p = probability of success or probability of happening of the event.
q = probability of failure or non happening of the event.
p + q = 1

Statistical or Empirical Definition of probability:
If a trial is repeated a number of times under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event happens to the number of trials, as the number of trials becomes indefinitely large, is called the probability of the happening of the event.

Symbolically, if n trials, an event E happens m times, then the probability of the happening of E is given by:

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

Definitions of various terms:

Trial and Event: Consider an experiment which, though repeated under essentially homogeneous and identical conditions, does not give unique results but may result in any one of the several possible outcomes. The experiments are known as a trial and the outcomes are known as events or cases. For example:

i. Throwing of a die is a trail and getting 1or 2 or 3 or 4 or 5 or 6 is an event.
ii. Tossing of coin is a trail and getting head (H) or tail (T) is an event.

Exhaustive Events:

(i) The total number of possible out comes in any trial is known as exhaustive events or exhaustive cases. In tossing of a coin there are two exhaustive cases head and tail (the possibility of the coin standing on its edge is being ignored)

(ii) In throwing of a die, there are 6 exhaustive cases since any one of six faces 1, 2, …,6 may come uppermost.

(iii) In drawing two cards from a pack of 52 cards, the exhaustive number of cases is $^{52}C_2$

(iv) In throwing of two dice, the exhaustive number of cases is $6^2 = 36$

(v) In general, in throwing of n dice, the exhaustive numbers of cases is 6^n.

Favourable Events:

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

i. In drawing a card from a pack of 52 cards the number of cases favourable to drawing of an ace is 4, for drawing a spade is 13 and for drawing a red card is 26.

Mutually Exclusive Events: Events are said to be mutually exclusive or incompatible if the happening of any one of them precludes the happening of all the others i.e., if no two or more of them can happen simultaneously in the same trial. For example,

1. In throwing a die all the 6 faces numbered 1 to 6 are mutually exclusive since if any one of these faces comes, the possibility of others, in the same trial, is ruled out.
2. Similarly in tossing a coin the events head and tail are mutually exclusive.

Equally Likely Events: Outcomes of trial are said to be equally likely if taking into consideration all the relevant evidences, there is no reason to expect one in preference to the others. For example,in a random toss of an unbiased or uniform coin, head and tail are equally likely events.

Independent Events. Several events are said to be independent if the happening (or non-happening) of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events. For example,in tossing an unbiased coin, the event of getting a head in the first toss is independent of getting a head in the second, third and in any subsequent throw.

Addition law of probability:

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

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

Multiplication law of probability:

For two events A and B,

$$P(A \cap B) = P(A).P(B \mid A), P(A) > 0$$
$$= P(B).P(A \mid B), P(B) > 0$$

Where P (B/A) represents conditional probability of occurrence of event B when the event A has already happened and P (A | B) is the conditional probability of happening of event A, given that B has already happened.

$$P(A) = n(A)/n(S) \quad P(B) = n(B)/n(S), P(A \cap B) = n(A \cap B)/n(S),$$

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

$A \cup B \longrightarrow$ At least one of the events A or B occurs.

$A \cap B \longrightarrow$ Both the events A and B occur.

Normal Distribution

The normal distribution was first discovered in 1733 by English mathematician De-Moivre.

Definition: A random variable X is said to follow normal distribution with parameters μ (called 'mean') and σ^2 (called 'variance') if its probability density function is given by:

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

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ is a standard normal variate with E(Z) = 0 and Var (Z) = 1 and we write $Z \sim N(0\ 1)$

The p.d.f. of standard normal variate Z is given by $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ $-\infty < z < \infty$

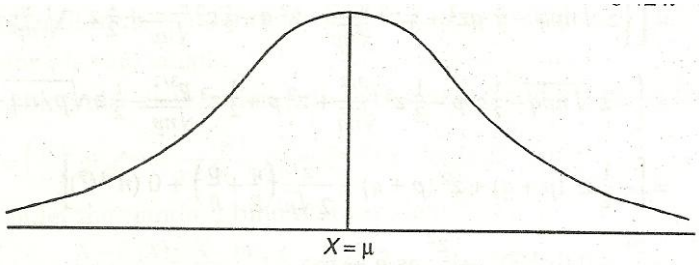
Chief Characteristics of the Normal Distribution and Normal Probability Curve.

The normal probability curve with mean μ and standard deviation σ is given by the equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},\ -\infty < x < \infty$$

and has the following properties:

1. The curve is bell- shaped and symmetrical about the line at x = μ
2. Mean, median and mode of the distribution coincide.
3. As x increases numerically, $f(x)$ decreases rapidly, the maximum probability occurring at the point x =μ, and is given by: $[p(x)]_{max} = \frac{1}{\sigma\sqrt{2\pi}}$
4. $\beta_1 = 0$ and $\beta_2 = 3$.
5. Liner combination of independent normal variants is also a normal variate.
6. X-axis is an asymptote to the curve.
7. The curve has two point of inflextion at $\mu + \sigma$ and $\mu + \sigma$



Probability curve

8. Mean deviation about mean = $\sqrt{\frac{2}{\pi}}\sigma \cong \frac{4}{5}\sigma$ (approx)

   Q:D.:M.D.: S.D. :: $\frac{2}{3}\sigma$: $\frac{4}{5}\sigma$: σ:: $\frac{2}{3}$: $\frac{4}{5}$: 1 ⇒ Q.D.: M.D.: S.D. :: 10:12:15

9. Area Property:

   $P(\mu - \sigma < X < \mu + \sigma) = 0.6826$

   $P((\mu - 2\sigma < X < \mu + 2\sigma) = 0.9544$

   $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

BINOMIAL DISTRIBUTION

Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700 and was first published posthumously in 1713.

Definition:

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

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

The two independent constants n and p in the distribution are known as the parameters of the distribution 'n' is also sometimes, known as the degree of the binomial distribution.

Binomial distribution is a discrete distribution as X can take only the integral value viz., 0,1,2,…,n. Any random variable which follows binomial distribution is known as binomial variate.

We shall use the notation X ~ B (n, p) to denote that the random variable X follows binomial distribution with parameters n and p.

Mean of binomial distribution = np

Variance of binomial distribution = npq

Physical conditions for Binomial Distribution. We get the binomial distribution under the following experimental conditions.

(i) Each trial results into exhaustive and mutually disjoint outcome, termed as success and failure.
(ii) The number of trials 'n' is finite.
(iii) The trails are independent of each other.
(iv) The probability of success 'p' is constant for each trial.

The trails satisfying the conditions (i), (iii) and (iv) are also called Bernoulli trials.

The problems relating to tossing of a coin or throwing of dice or drawing cards from a pack of cards with replacement lead to binomial probability distribution.

Binomial distribution is important not only because of its wide applicability, but because it gives rise to many other probability distributions.

Example: - Ten coins, are thrown simultaneously. Find the probability of getting at least seven heads.

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

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

The probability of getting x heads in a random throw of 10 coins is :

$$P(x) = \binom{10}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = \binom{10}{x} \left(\frac{1}{2}\right)^{10}; \text{ x = 0, 1,2,…, 10}$$

∴ Probability of getting at least seven heads is given by:

P(X ≥7) = p(7) + p(8) + p(9) + p(10)

$$= \left(\frac{1}{2}\right)^{10}\left\{\binom{10}{7}+\binom{10}{8}+\binom{10}{9}+\binom{10}{10}\right\} = \frac{120+45+10+1}{1024} = \frac{17\ 6}{1024}$$

POISSON DISTRIBUTION

Poisson distribution was discovered by the French mathematician and physicist Simeon Denis Poisson (1781-1840) who published it in 1837. Poisson distribution is a limiting case of the binomial distribution under the following conditions.

(i) n, the number of trials is indefinitely large, i.e., n → ∞

(ii) p, the constant probability of success for each trial is indefinitely small, i.e.,p →0.

(iii) np = λ, (say) is finite.

Thus p = λ/n, q = 1 - λ/n, where λ is a positive real number.

The probability of x successes in a series of n independent trials is:

$$B(x; n,p) = \binom{n}{x} p^x q^{n-x}; x = 0, 1, 2, \ldots., n$$

Definition. A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by:

$$P(x, \lambda) = P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}; x = 0\ 1, 2, \ldots;\ \lambda > 0 \\ 0\ , \text{otherwise} \end{cases}$$

Here λ is known as the parameter of the distribution. We shall use the notation
X ~ P(λ),denote that X is a Poisson variate with parameter λ.
Mean and variance of poisson distribution are equal and equal to λ.

Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials (unlike that in binomial distribution) of an experiment but which occur at random points of time and space wherein our interest lies only in the number of occurrences of the event, not in its non-occurrences.

Following are some instance where Poisson distribution may be successfully employed:

(i) Number of deaths from a disease (not in the form of an epidemic) such as heart attack or cancer or due to snake bite.

(ii) Number of suicides reported in a particular city.

(iii) The number of defective material in a packing manufactured by a good concern.

(iv) Number of faulty blades in a packet of 100.

(v) Number of air accidents in some unit of time.

(vi) Number of printing mistakes at each page of the book.

(vii) Number of telephone calls received at a particular telephone exchange in some unit of time or connections to wrong numbers in a telephone exchange.

(viii) Number of cars passing a crossing per minute during the busy hours of a day.

(ix) The number of fragments received by a surface area 'A' from a fragment atom bomb.

(x) The emission of radioactive (alpha) particles.

Introduction to sampling

Population: It is an aggregate of objects (animate or inanimate) under study is known as population, It may be finite or infinite.
Sample: A finite subset of statistical individuals in a population is known as sample and the number of individuals in the sample is known as sample size.

Random Sampling: Random Sample is one in which each unit of population has got an equal chances of selection and the technique of drawing random sample is termed as Random Sampling. Fairly good random samples can be obtained by the use of Tippet's random number tables, or by tossing of a coin or drowning a lottery etc.

Parameter: It is the characteristics of population values such as population mean (μ) and population variance (σ^2).

Statistic: It is an estimate of parameter obtained from the sample is the function of sample value only. Eg. sample mean ($\bar{x}$), sample variance (S^2)
Standard Error: The standard deviation of the sampling distribution of a statistic is known as standard error and denoted by S.E.
Standard error of mean: It is the positive square root of the variance of sampling distribution of mean

S.E. of Mean = $\sqrt{(\sigma 2/n)}$ Where, σ = population standard deviation and n = sample size

Utility of S.E.- S.E. plays every important role in large sample theory and forms the basis of the testing of hypothesis if t is any statistic, then for large samples.

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0,1)$$

Sampling vs complete enumeration

Sampling survey: A survey involving only a part of population is called sample survey. A sample is a subset of population.

Complete enumeration/ census survey: A survey in which each and every unit of the population is under consideration is known as complete enumeration. The money manpower and time required to carry out complete enumeration are generally larger then sample survey.

The main merits of sampling technique over the complete enumeration may be outlined as follows:

1. Less time. There is considerable saving in time and labour since only a part of the population has to be examined. The sampling results can be obtained more rapidly and the data can be analysed much faster since relatively fewer data have to be collected and processed.

2. Reduced cost of the survey. Sampling usually results in reduction in cost in terms of money and in terms of man hours. Although,the amount of labour and the expenses involved in collecting information are generally greater per unit of sample than in complete enumeration, the total cost of the sample survey is expected to be much smaller than that of the complete census.
3. Greater Accuracy of Results. The results of a sample survey are usually much more reliable than those obtained from a complete census.
4. Greater Scope. Sample survey has generally greater scope as compared with complete census. The complete enumeration is impracticable, rather inconceivable if the survey requires a highly trained personnel and more sophisticated equipment for the collection and analysis of the data. Since sample survey saves in time and money. It is possible to have a thorough and intensive enquiry because a more detailed information can be obtained from a small group of respondents.
5. If the population is too large, as for example, trees in a jungle, we are left with no way but to resort to sampling.
6. If testing is destructive, i.e., if the quality of an article can be determined only by destroying the article in the process of testing, as for example.
   (i) Testing the quality of milk or chemical salt by analysis,
   (ii) Testing the breaking strength of chalks,
   (iii) Testing of crackers and explosives,
   (iv) Testing the life of an electric tube or bulb, etc.

Simple random sampling

Simple random sampling is the most widely used simplest method of drawing smaple from a population such that each and every unit is the population has an equal probability of being included to sample.

From a population of N units, we select one unit by giving equal probability 1/N to all unit with the help of random numbers. A unit is selected, noted and returned to the population before drawing the second unit and the process is repeated 'n' times to fit a simple random sample of 'n' units. This procedure of selecting a sample is known as '**Simple Random Sampling with Replacement (SRSWR)**'. If, however, this procedure is continued till 'n' distinct units are selected ignoring all repetitions a '**Simple Random Sample Without Replacement (SRSWOR)**' is obtained. The latter procedure is exactly same as retaining the unit selected and selecting a further unit with equal probability from the units that remain in the population.

Use of Random Number Table for Selection of Simple Random Sample

The most common and inexpensive method of selecting a random sample consists in the use of Random Number tables, which have been so constructed that each of the digits 0,1,2,…9 appears with approximately the same frequency and independently of each other. If we have to select a sample from a population of size N (≤ 99) then the numbers can be combined two by two to given pairs from 00 to 99. Similarly if $N \leq 999$ or $N \leq 9999$, and so on, then combining the digits three by three (or four by four), and so on), we get numbers from 000 to 999 or (0000 to 9999), and so on. Since each of the digit 0,1, 2,...9 occurs with approximately

the same frequency and independently of each other, so does each of the pairs 00 to 99 or triplets 000 to 999 or quadruplets 0000 to 9999, and so on.

The method of drawing the random number consists in the following steps:

(i) To identify the N units in the population with the numbers from 1 to N.

(ii) To select at random, any page of the "random number table" and pick up the numbers in any row or column at random.

The population units corresponding to the numbers selected in step (ii) constitute the random sample.

Test of Significance

It is the statistical procedure for deciding whether the difference under study is significant or, not.Common test of significance are t-test, F-test, Chi-square (χ^2) test.

Null Hypothesis : It is the hypothesis of no difference and it is denoted by H_0.
Alternative Hypothesis: Any hypothesis which is complementary to null hypothesis is known as alternative hypothesis and it is denoted by H_1.

Under, $H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$,two tailed test
$H_1 : \mu > \mu_0$ right tailed test
$H_1 : \mu < \mu_0$ left tailed test.

Error in Sampling:

The main theory of sampling is to draw a valid inference about the population parameter on the basis of sample drawn from it and in this way, we are liable to commit two types of error: type-I error and type II error.

Type-I error: Reject H_0 when it is true = P {Reject H_0 when it is true}= P {Reject H_0/H_0}and denoted by "α". It is also known as "producer's risk". α is the size of type I error.

Type-II error: Accept H_0 when it is wrong = P {Accept H_0 when it is wrong}= P {Accept H_0/H_1 } denoted by "β". It is also known as "Consumer's risk". β is the size of type II error.
Power of test = 1 - β
Critical Region: A region in the sample space (S) which amounts the rejection of H_0.
Level of Significane: The probability (α) that a random value of statistic (t) belong to critical region. It is the size of Type-I error or It is the maximum probability with which we will be willing to risk an error. It is generally fixed in advance; like 5% level of significance and 1% level of significance.
Steps to solve the problems of test of significance:
1. Frame H_0 according to question and also frame H_1.
2. Apply suitable statistic according to question.
3. Calculate the value of right hand side of applied statistic.
4.Compare the calculated value of applied statistic with tabulated value (given value) at required degree of freedom and level of significance.

If cal. Value ≥ tab. value at given degree of freedom and level of significance. Result is significant and we reject H_0.i.e. we accept H_1. If cal. value < tab. The result is non-significant, we accept H_0 i.e. we reject H_1 and to draw conclusion accordingly.

Degree of freedom: Number of observations (n) - number of restrictions (k) imposed upon them, degree of freedom = n-k

Student's t-test

t- test was first given by W.S. Gosset in 1908 and modified by R.A. fisher in1926.

Definition:Let x_i (i = 1,2,..., n) be a random sample of size n drawn from a normal population with mean μ and variance σ^2. Then student's t is defined by the statistic:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \quad \text{with (n-1) d.f.}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

is an unbiased estimate of the population variance σ^2.

Assumption of t- test :

- Parent population from which sample is drawn should be normal.
- Sample observations are independent i.e. samples are random.
- Population standard deviation (**σ**) is unknown.

Fisher's't' (Definition). It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If ξ is a N(0,1) and χ^2 is an independent chi-square variate n d.f., then Fisher's t is given by:

$$t = \frac{\xi}{\sqrt{\chi^2/n}}$$

and it follows Student's 't' distribution with n degree of freedom.

Application of t-test:

- To test the significance of difference from the sample mean from the hypothetical value of population mean.
- To test the significance difference between two samples mean.
- To test the significance of observed sample correlation coefficient and regression coefficient.

First application, under H_0: (i) The sample has been drawn from the population with mean, μ_0 or (ii) There is no significant difference between the sample mean $\bar{x}$ and the population mean μ_0.

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \quad \text{with (n-1) d.f.} \quad \text{where, } \bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i \text{ is the sample mean and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Second application, under H_0: (i).$\mu_x = \mu_y$ (ii) The sample means $\bar{x}$ and $\bar{y}$ do not differ significantly.[$n_1 \neq n_2$; and $\sigma_x^2 = \sigma_y^2 = \sigma^2$ i.e. population variances are equal and unknown].

t = $\frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ with($n_1 + n_2 - 2$) and $S^2 = \frac{1}{n_1+n_2-2}[\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{j=1}^{n_2}(y_j - \bar{y})^2]$

is an unbiased estimate of the common population variances σ^2.

and $\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$, $\bar{y} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_j$

If $n_1 = n_2$ then samples are not independent but paired together and we apply paired t-test

t = $\frac{\bar{d}}{S/\sqrt{n}}$ with (n-1) d.f. ; where $d_i = x_i - y_i$ and $\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - \bar{d})^2$

F-test

Definition of F-test: If X and Y are two independent chi-square variates with $\upsilon_1$and υ_2 degree of freedom respectively, then F-test is given by:

F = (X/ υ_1)/(Y/ υ_2) with (υ_1- 1) and (υ_2 – 1) d.f.

Applications of F-test :

1. To test the equality of two population variances.
2. To test the significance of an observed sample correlation coefficient.
3. To test the significance of an observed multiple correlation co-efficient.
4. To test the significance of quality of several mean (design of experiment).
5. To test the linearity of regression.

To test the equality of two population variances:

$H_0 : \sigma_x^2 = \sigma_y^2 = \sigma^2$ (say)

$F = \frac{Sx^2}{Sy^2}$ with (n_1 – 1) and (n_2 -1) d.f.

where $S^2x = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$

and $Y^2x = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$

Chi-square test

If X ~N (μ, σ^2), then $Z = \frac{X-\mu}{\sigma}$ is a standard normal variate and square of standard normal variate is known as ch-square variate with 1 d.f.

Chi-square test was first discovered by Karl Pearson in 1900.

Definition of Chi-square test of Goodness of fit: If O_i (i = 1,2,3...,n) is a set of observed frequencies and E_i (i = 1,2,3...,n) is the corresponding set of expected frequencies, then chi-square is given by

$\chi^2 = \sum_{i=1} \frac{(Oi - Ei)^2}{E_i}$ with (n – 1) d.f.

Conditions for the validity of χ^2 test:

(i) The sample observations should be independent i.e samples are random.

(ii) No theoretical cell frequency should be less than 5.

(iii)Total number of frequencies should be reasonable large(>50).

(iv)Σ Oi = ΣEi.

Application of χ^2 test:

1. **Test of Goodness of fit**: It enables us to find the deviations in experiment from theory is just by chance.
2. **Test of independence of attributes**: We test whether two or more attributes are independent to each other.

Contingency Table.

We consider two attributes A and B, A divided into r classes $A_1, A_2 \ldots A_r$ and B divided into s classes $B_1, B_2, \ldots, B_s$. Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the tables known as r x s manifold contingency table where (A) is the number of persons possessing the attribute A_i,(i = 1, 2, 3,...r), (B_j) is the number of persons possessing the attribute B_j (j= 1, 2, ...,s) and ($A_i B_j$) is the number of persons possessing both the attributes A_i and B_j, (i = 1, 2, ..., r; j = 1, 2,..., s).

Also $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} (B_j) = N$, where N is the total frequency.

CONTINGENCY TABLE(r x s)

A / B	A_1	A_2	...	A_j	...	A_r	**Total**
B_1	(A_1B_1)	(A_2B_1)	...	(A_iB_1)	...	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	...	(A_iB_2)	...	(A_rB_2)	(B_2)
:	:	:	:	:	:	:	:
B_j	(A_1B_j)	(A_2B_j)	...	(A_iB_j)	…	(A_rB_j)	(B_j)
:	:	:	:	:	:	:	:
B_s	(A_1B_s)	(A_2B_s)	…	(A_iB_s)	…	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	…	(A_j)	…	(A_r)	**N**

Yate's correction for continuity: In a 2 x 2 contingency table, the number of d.f. is (2-1) (2-1)= 1. If any one of the theoretical cell frequencies is less than 5, then use of pooling method for χ^2- test results in χ^2 with 0 d.f. (since 1 d.f. is lost in pooling) which is meaningless. In this case we apply correction due to F. Yates (1934), which is usually known as "Yate's Correction for continuity" [as we know, χ^2 is a continuous distribution and it fails to maintain its character of continuity if any of the expected frequency is less than 5; hence the name 'Correction for continuity'].This consists in adding 0.5 to the cell frequency which is less than 5 and then adjusting for the remaining cell frequency accordingly. The χ^2- test of goodness of fit is then applied without pooling method.

For a 2 x 2 contingency table,

a	b
c	d

we have $\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$

According to Yate's correction, as explained above, we subtract (or add) ½ from 'a' and 'd' and add (subtract) ½ to 'b' and 'c' so that the marginal totals are not disturbed at all.

There, corrected value of χ^2 is given as: $\frac{N\left[\left(a\mp\frac{1}{2}\right)\left(d\mp\frac{1}{2}\right)-\left(b\pm\frac{1}{2}\right)\left(c\pm\frac{1}{2}\right)\right]^2}{(a+c)(b+d)(a+b)(c+d)}$

Numerator = $N\left[(ad-bc)\mp\frac{1}{2}(a+b+c+d)\right]^2 = N\left[|ad-bc| - \frac{N}{2}\right]^2$

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

CORRELATION

If the change in one variable affects a change in the other variables, the variables are said to be correlated. If the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive e.g., (i) height and weight of a group of persons (ii) income and expenditure.

If increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be indirect or negative. e.g., (i) volume and pressure of a perfect gas. (ii) price and demand of a commodity.

Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

Karl Pearson coefficient of correlation or correlation coefficient: Correlation coefficient between two variables X and Y is a numerical measure of linear relationship between them and is given by

$$r(X,Y) = \frac{Cov(X,\ Y)}{\sigma_x\,\sigma_y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}.\,\bar{Y})}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \overline{X^2}\right)\left(\frac{1}{n}\Sigma Y^2 - \overline{Y^2}\right)}} = \frac{\frac{1}{n}\Sigma UV - \bar{U}.\,\bar{V})}{\sqrt{\left(\frac{1}{n}\Sigma U^2 - \overline{U^2}\right)\left(\frac{1}{n}\Sigma V^2 - \overline{V^2}\right)}}$$

Where, U = (X- a) or (X – a)/ h and V = (X- b) or (X – b)/ k

where, a = arbitrary value in X variable and h is the magnitude of X variable

b = arbitrary value in Y variable and k is the magnitude of Y variable

Range of correlation coefficient: $-1 \le r \le 1$

Correlation coefficient (r) is independent of change of origin and scale both.

Scatter Diagram

In bivariate distribution, if the values of the variables X and Y are plotted along the x-axis and y-axis respectively in the (x,y) plane, the diagram of dots so obtained is known as scatter diagram. From the scatter diagram, we can form a fairly good idea whether the

variables are correlated or not. If the points are very dense, i.e. very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is fairly large.

a. Lines of Regression :-

If the variables in a bivariate distribution are related we will find that the points in Scatter diagram will cluster armed some curve called the curve of regrerssion. If the curves is straight line it is called the line of regression Line of

REGRESSION

Literal meaning of regression is "stepping back toward average". It was first used by Sir Francis Galton, Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data. In regression analysis there are two types of variables

I. **Dependent Variable** : The variable whose value is influenced or is to be predicted is known as dependent Variable. eg. Yield. Itis also known as regressed or explained variable.

II. **Independent Variable** : The Variable which influences the values or is used for prediction is known as independent variable. eg. Fertiliser, irrigation etc. It is also termed as regressed or predictor or explanatory variable.

Lines of Regression :If the variables in a bivariate distribution are related ,we will find that the points in scatter diagram are clustered around some curve called the curve of regression and if the curves is straight line it is called the line of regression. Line of regression is the line of "best fit". Line of regression of Y on X is Y = a + bx and Line of regression of X on Y is X = a+by

Regression Coefficient : 'b' the slope of the line of regression of Y on X is called coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable(X).

b_{yx}= Regression co-efficient of Y on X = $\frac{Cov\,(x,y)}{Var\,(x)} = \frac{\mu_{11}}{\sigma_{x^2}} = r\frac{\sigma_y}{\sigma_x}$

The regression co-efficient of X on Y represents the change in the value of independent variable corresponding to a unit change in the value of dependent variable and is given by

b_{xy}= Regression co-efficient of X on Y = $\frac{Cov\,(x,y)}{Var\,(y)} = \frac{\mu_{11}}{\sigma_{y^2}} = r\frac{\sigma_x}{\sigma_y}$

Properties of regression coefficient:

1. Regression co-efficient is independent of change of origin but not of scale.
2. $b_{xy} \neq b_{yx}$ where as , r(X,Y) = r(Y,X) .
3. Correlation co-efficient is the geometric mean between regression coefficients

$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$

$r = \pm\sqrt{b_{xy} X\ b_{yx}}$

4. If one regression coefficient is greater than unity the other must be less than unity.

Lines of Regression of X on Y is

$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

Lines of Regression of Y on X is

$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

Design of Experiment

Basic Principles of Experimental Designs:

The basic principles of experimental designs are randomization, replication and local control. These principles make a valid test of significance possible. Each of them is described briefly in the following subsections.

(1) Randomization: The first principle of an experimental design is randomization, which is a random process of assigning treatments to the experimental units. The principle of randomization asserts that each treatment has equal chance/probability of being allotted to the same plot.

(2) Replication: Repetition of the treatment is called Replication. The number, the shape and the size of replicates depend upon the nature of the experimental material.

3) Local Control: The process of dividing the whole experimental material into a group of homogeneous plots (called Blocks) in such a manner that the plots within the block is homogeneous and plots between the blocks is heterogeneous. The blocking is done perpendicular to the direction of the fertility gradient.

Basic Designs: There are three basic designs.

- CRD – Completely Randomized Design
- RBD – Randomized Block Design
- LSD – Latin Square Design

The name 'basic design' is due to the fact that these were the first designs to be discovered by Prof.R.A.Fisher (Father of Design of Experiments).

Analysis of Variance (ANOVA): According to Prof. R.A. Fisher, Analysis of variance is the separation of variance ascribable to one group of causes from the variance ascribable to other group of causes:

i. Assignable causes
ii. Chance causes.

chance causes (factors) is known as experimental error or simply error.

Treatments: The objects of comparison in an experiment are defined as treatments.
For example:(i) suppose an Agronomist wishes to know the effect of different spacings on the yield of a crop, different spacings will be treatments. Each spacing will be called a treatment.
(ii) If different doses of fertilizer are tried in an experiment to test the responses of a crop to the fertilizer doses, the different doses will be treatments and each dose will be a treatment.
Experimental unit: Experimental unit is the object to which treatment is applied to record the observations.
For example (i) In laboratory, insects may be kept in groups of five or six. To each group, different insecticides will be applied to know the efficacy of the insecticides. In this study different groups of insects will be the experimental unit.
(ii) If treatments are different varieties, then the objects to which treatments are applied to make observations will be different plot of land. The plots will be called experimental units.

Blocks: In agricultural experiments, most of the times we divide the whole experimental unit (field) into relatively homogeneous sub-groups or strata. These strata, which are more uniform amongst themselves than the field as a whole are known as blocks.

COMPLETELY RANDOMIZED DESIGN (CRD):

When the treatments are arranged randomly over the predetermined homogeneous set of experimental units, design is known as Completely Randomized Design. Incidentally, CRD is the only design where relaxation of not applying each treatment equal no. of times is allowed. However, this should not be used indiscriminately.

Applicability:

When the experimental material is homogeneous, CRD is adopted. Normally this condition is not achieved in the field experiments. Thus, CRD is applied in Laboratory experiments or Pot experiments or in the Greenhouse**.**

Mathematical Model

$Y_{ij} = \mu + T_i + e_{ij}$

Where μ = General Effect

T_i = Effect due to applying ith treatment in the jth plot

e_{ij} = Error due to applying ith treatment in the jth plot

Y_{ij} = Yield due to applying ith treatment in the jth plot

LAYOUT

T_1	T_3	T_4	T_1
T_5	T_2	T_5	T_3
T_2	T_4	T_1	T_2
T_3	T_3	T_5	T_4
T_4	T_1	T_2	T_3

ANOVA

Sources of Variation	D.F.	S.S.	M.S.S. = S.S./D.F.	F
Treatment	t-1	S1	S1/(t-1) =VT	VT/VE
Error	(N-1)-(t-1)	S2	S2/(N-1)-(t-1) =VE	
Total	N-1	S		

Correction Factor (C.F.) = G^2/N

Total Sum of Squares (T.S.S.) = $\sum Y^2_{ij}$ - C.F. = S

Treatment Sum of Squares (Tr.S.S.) = $\sum T_i^2/r$ - C.F. = S1

Error Sum of Squares (E.S.S.) = T.S.S.- Tr.S.S.= S2

S.E./Plot = $\sqrt{VE}$

S.E.diff. mean = $\sqrt{2x\ VE/r}$

C.D. = t 0.05 (for error d.f.) x S.E.(d)

C.V. = S.E./Plot/G.M. X 100

RANDOMIZED BLOCK DESIGN (RBD):

It is an arrangement of 'v' treatments in 'b' blocks in such a way that each treatment occurs once and only once in a block.

Applicability

When the fertility gradient in the field is in one known direction, RBD is applied. In agricultural field experiments RBD is mostly used.

Mathematical Model

$Y_{ij} = \mu + T_i + b_j + e_{ij}$

Where μ = General Effect

T_i = Effect due to applying i^{th} treatment in the j^{th} plot

e_{ij} = Error due to applying i^{th} treatment in the j^{th} plot

b_j = Effect due to applying j^{th} block

Y_{ij}= Yield due to applying i^{th} treatment in the j^{th} plot

e_{ij} = Error due to applying i^{th} treatment in the j^{th} block

Analysis

	R_1	R_2	-	Rr	Total	Mean
T_1	Y_{11}	Y_{12}	-	y_{1r}	T_1	t_1
T_2	Y_{21}	Y_{22}	-	Y_{2r}	T_2	t_2
-	-	-	-	-	-	-
-	-	-	-	-	-	-
T_v	Y_{v1}	Y_{v2}	-	Y_{vr}	T_v	t_v
Total	R_1	R_2	-	R_3	G	G.M.

Correction Factor (C.F.) = G^2/N

Total Sum of Squares (T.S.S.) = $\sum Y^2_{ij}$ - C.F. = S

Replication Sum of Squares (R.S.S.) = $\sum R_j^2/t$ –C.F. = S1

Treatment Sum of Squares (Tr.S.S.) = $\sum T_i^2/r$ - C.F. = S2

Error Sum of Squares (E.S.S.) = T.S.S.-R.S.S.- Tr.S.S.= S3

S.E./Plot = $\sqrt{VE}$

S.E.diff. mean = $\sqrt{2x\ VE/r}$

C.D. = t 0.05 (for error d.f.) x S.E.(d)

C.V. = S.E./Plot/G.M. X 100

ANOVA

Sources of Variation	D.F.	S.S.	M.S.S. = S.S./D.F.	F
Replication	r-1	S1	S1/(r-1) =VR	VR/VE
Treatment	t-1	S2	S2/(t-1) =VT	VT/VE
Error	(r-1) (t-1)	S3	S3/(r-1)(t-1) =VE	
Total	N-1	S		

Advantages:

- Increased Precision is obtained due to using Local Control
- Any no. of treatments can be included. If large no. of homogeneous units are available, Large no. of treatments can be included
- The analysis is simple. It remains simple even if some plots are missing
- The amount of information in RBD is more than that of CRD. Thus RBD is more efficient than CRD

Disadvantages

- RBD is not suitable for large no. of treatments, because it increases the block size and heterogeneity of the blocks which increases the experimental error.
- For this disadvantage, RBD is a Versatile design which is most frequently used in agricultural experiments.