Statistical Methods

(Test of statistical significance , Analysis of variance- Random Block Design and its application in plant breeding and genetics; Correlation and Regression)

Course Code –BOTY 4204

Course Title- Techniques in plant sciences , biostatistics and bioinformatics

By – Dr. Alok Kumar Shrivastava
Department of Botany
Mahatma Gandhi Central University, Motihari

Unit 4- Statistical Methods

Central tendency, dispersion, standard error, coefficient of variation; Probability distributions (normal, binomial of Poission) and Confidence limits. Test of statistical significance (t-test, Chi-square): Analysis of variance-Random Block Design and its application in plant breeding and genetics; Correlation and Regression.

T-test

- A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. In other words we can sat t-test is used to compare the mean of two given sample. T-statistic was introduced by William Sealy Gosset (pet name ‘Student’)in 1908.
- The t-test is one of many tests used for the purpose of hypothesis testing in statistics which allows testing of an assumption applicable to a population.
- A t-test is used when the population parameters (mean, and standard deviation) are not known.
- Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.
- There are several different types of t-test that can be performed depending on the data and type of analysis required.

- ✓ The t-test tells you how significant the differences between groups are; in other words it lets you know if those differences (measured in mean) could have happened by chance.
- ✓ Outcome of the t-test produces t-value which is then compared against a value obtained from a critical value table (called the T-distribution table).
- ✓ T -value measures the size of the difference relative to the variation in your sample data. In other words t- value is simply the calculated difference represented in units of standard error.
- ✓ Find the t-value by dividing the difference between group means by the standard error of difference between the groups.
- ✓ the greater the magnitude of t-value the greater the evidence against null hypothesis.

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

X_1 = Mean of sample1

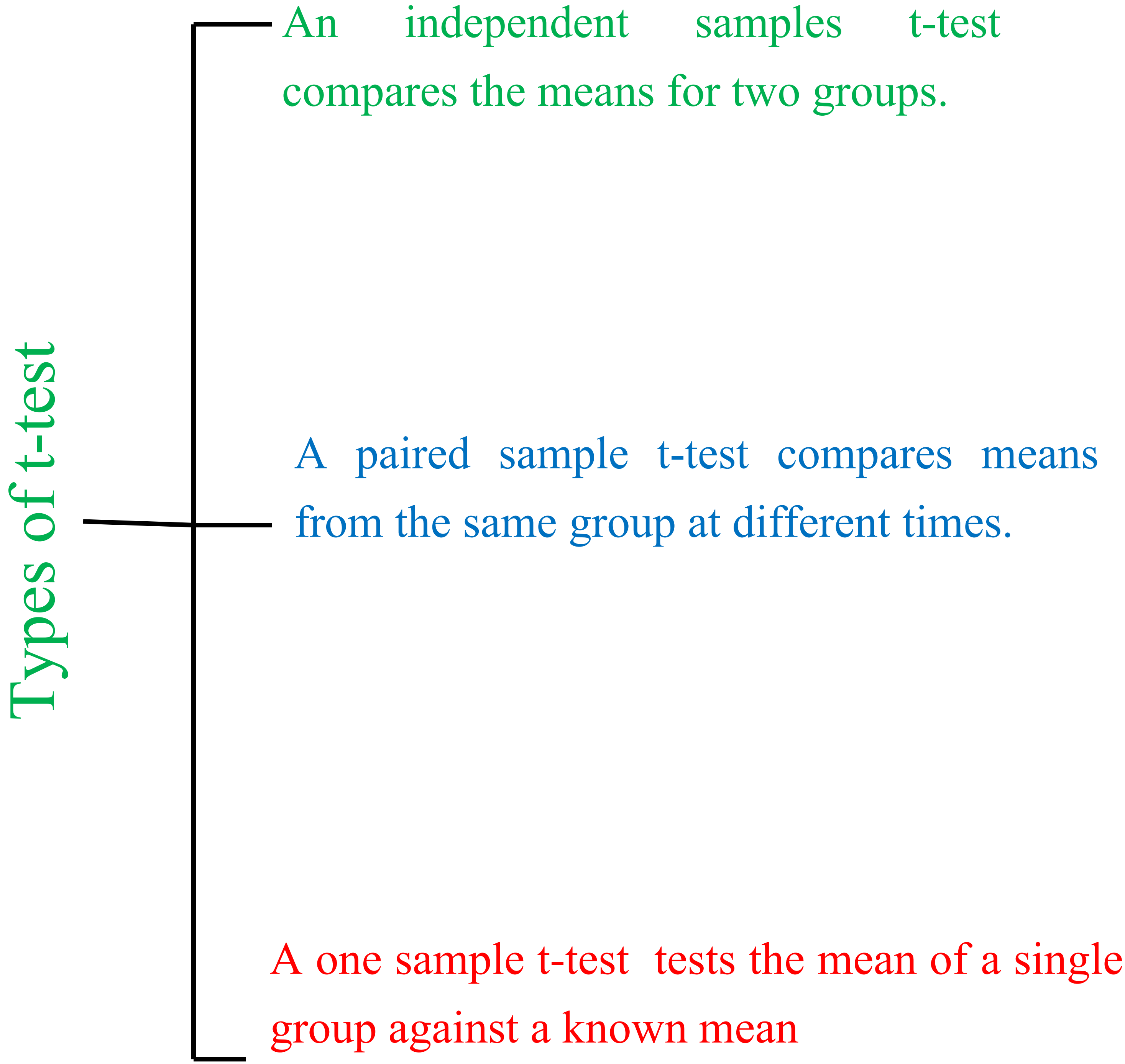
X_2= Mean of sample 2

n_1 = Size of sample 1

n_2= Size of sample 2

$(S_1)^2$ =Variance of sample1

$(S_2)^2$ = Variance of sample1



Types of t-test

An independent samples t-test compares the means for two groups.

A paired sample t-test compares means from the same group at different times.

A one sample t-test tests the mean of a single group against a known mean

ANOVA

- ANOVA also known as analysis of variance, is used to compare multiple sample with a single test.
- There are two types of ANOVA: one way and two way. One way and two way refers to the number of independent variables (factors) in your analysis of variance test.
- One way ANOVA or unidirectional ANOVA is used to compare the difference between the three or more samples/ groups of a single independent variable.
- The two way ANOVA compares the mean differences between groups that have been split on two independent variables. The primary purpose of a two way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

ANCOVA

- Analysis of covariance is used to test the main and interaction effects of categorical variable on a continuous dependent variable, controlling for the effects of selected other continuous variables, which co-vary with the dependent.
- ANCOVA is used to compare one variable in two or more populations while considering other variables while ANOVA is used to compare and contrast the means of two or more populations.

covariates

- A covariates represents a source of variation that has not been controlled in the experiment and is believed to affect the dependent variable. The aim of ANCOVA is to remove the effects of such uncontrolled variation, in order to increase statistical power and to ensure an accurate measurement of the true relationship between independent and dependent variables.

MANOVA

- MANOVA (Multiple analysis of variance) allows us to test the effect of one or more independent variable on two or more dependent variables. In addition , MANOVA can also detect the difference in co-relation between dependent variables given the groups of independent variables.

MANCOVA

- Multivariate analysis of covariance is an extension of ANCOVA methods to cover cases where there is more than one dependent variable and where the control of concomitant continuous independent variables covariates is required. The main benefit of the MANCOVA over MANOVA is the factoring out of noise or error that has been introduced by covariant.
- Like all tests in the ANOVA family, the primary aim of this is to test for significant differences between group means.

Chi – square test

✓A chi-square statistic is a test that measures how expectations compare to actual observed data. Te data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables and drawn from a large enough sample.

✓This test is commonly used for testing relationships between categorical variables

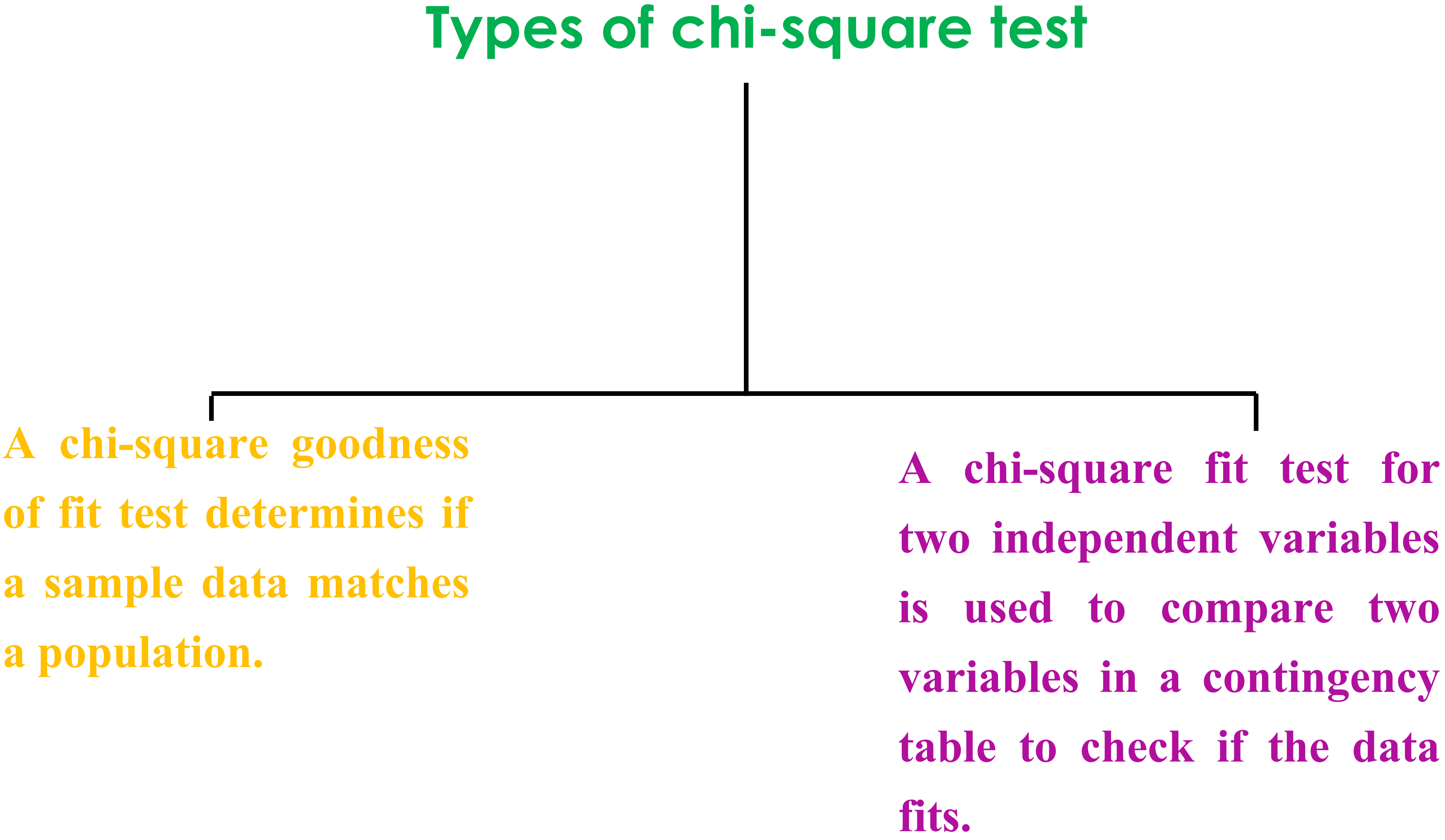
✓A chi-square test tests a null hypothesis about the relationship between two variables.

✓To calculate chi square, we take the square of the difference between the observed and expected values and divide it by the expected value.

✓A very small chi-square value means that your observed data fits your expected data extremely well in other words , there is a relationship. A very large chi-square value means that your observed data does not fits very well in other words , there is not a relationship.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O= Observed frequency
E=Expected frequency



Types of chi-square test

A chi-square goodness of fit test determines if a sample data matches a population.

A chi-square fit test for two independent variables is used to compare two variables in a contingency table to check if the data fits.

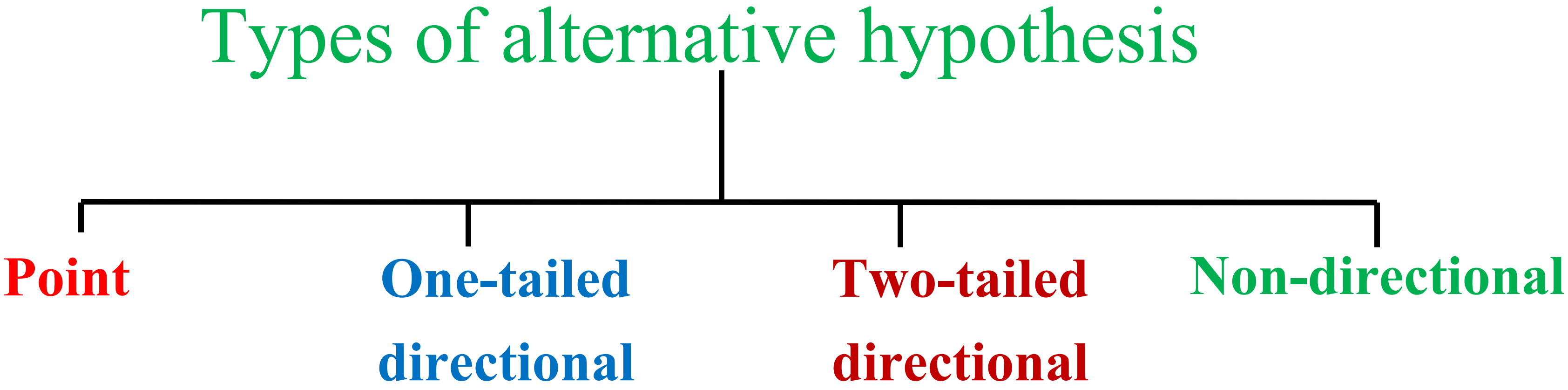
Null hypothesis

✓ According to null hypothesis there is no significant difference between specified populations, any observed difference being due sampling or experimental error. It is denoted by H0.

✓ A null hypothesis is a hypothesis that says there is no statistical significance between the two variables.

Alternative hypothesis

An alternative hypothesis is a hypothesis that states there is statistically significant relationship between the two variables. It is just alternative to the null and denoted by H1 or Ha. The concept of alternative hypothesis given by **Jerzy Neyman and Egon Pearson.**



Types of alternative hypothesis

Point

One-tailed directional

Two-tailed directional

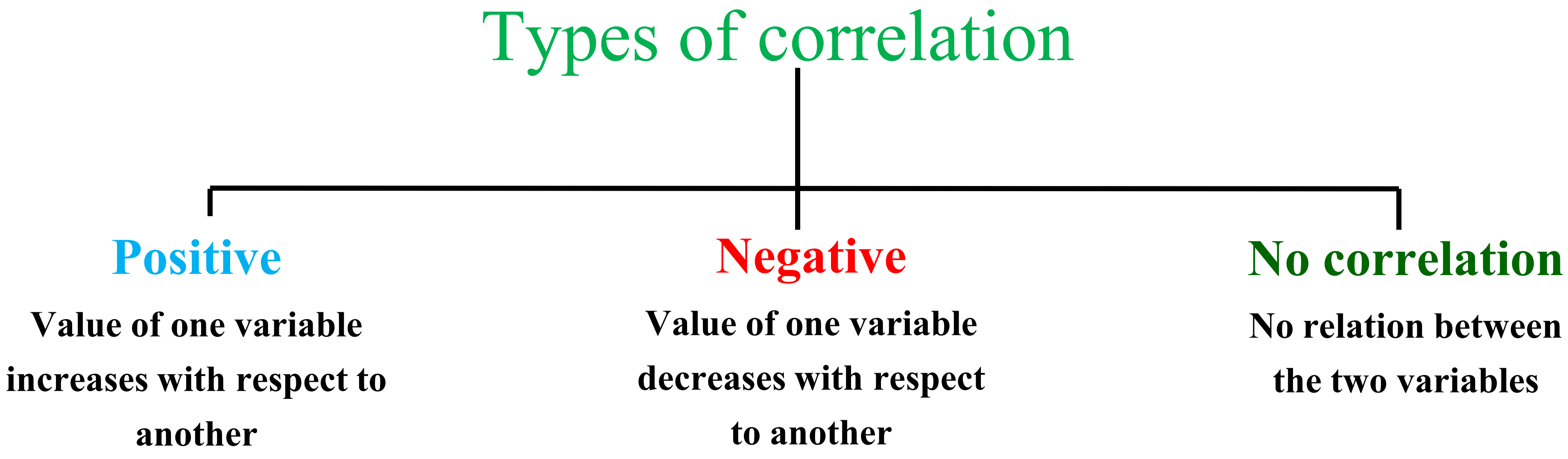
Non-directional

Correlation

✓ It is a statistical measurement of the relationship between two variables. It is a bivariate analysis. It is used to quantify the degree to which two variables are related. Through the correlation analysis one can correlation coefficient which is a statistical measure of the strength of the relationship between the relative movements of two variables. It is represented by r –value. According to Cauchy Schwarz inequality it has a value between +1 to -1. when the r value is closer to +1 or -1 , it indicates that there is a stronger linear relationship between the two variables. A correlation of – 0.97 is a strong negative correlation while a correlation of 0.10 would be a weak positive correlation. if calculated value greater than +1 or less than -1 it means that there is an error in the correlation measurement.

✓ -1 correlation indicates a perfect negative correlation, meaning that as one variable goes up and other goes down.

✓ +1 correlation indicates a perfect positive correlation, meaning that as both variable goes up or down in parallel.



Types of correlation

Positive

Value of one variable increases with respect to another

Negative

Value of one variable decreases with respect to another

No correlation

No relation between the two variables

Pearson correlation coefficient

- Pearson correlation coefficient also known as r, R or Pearson's r.
- It is a measure of the strength and direction of the linear relationship between two variables that is defined as the covariance of the variables divided by the product of their standard deviations.
- This is the best known and most commonly used type of correlation coefficient.

Correlation Coefficient Formula

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2}\sqrt{(Y - \overline{Y})^2}}$$

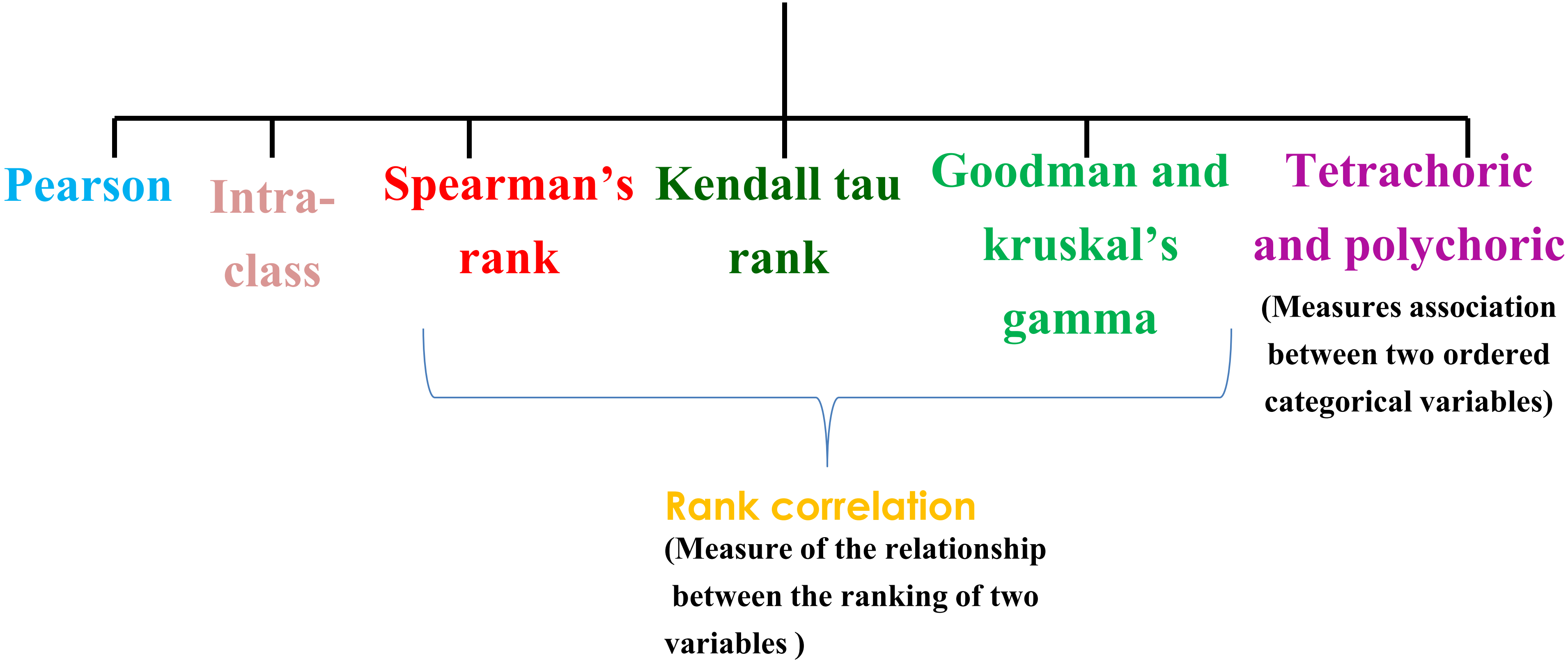
OR

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

OR

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

Types of correlation coefficient



Pearson

Intra-class

Spearman's rank

Kendall tau rank

Goodman and kruskal's gamma

Tetrachoric and polychoric

(Measures association between two ordered categorical variables)

Rank correlation

(Measure of the relationship between the ranking of two variables)

Regression

Regression analysis is a powerful statistical method that allow you to examine the relationship between two or more variables of interest . There are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable. Regression analysis is used in stats to find trends in data. The term regression is coined by Francis Galton.

Linear relationship

Graph of linear equation forms a straight line. A linear relationship have curve whose slope is constant.

Linear regression equation-

$$Y=a+bX$$

Y= dependent variable

X= independent variable

b= slope of line

a= Y intercept

Non-linear relationship

A non linear relationship between two entities in which change in one entity does not correspond with constant change in other entity. The graph for a non linear relationship is curved. Slope of curve changes in non linear relationship.

Correlation vs. regression

- Correlation is a statistical measure that defines co-relationship or association of two variables while regression describes how an independent variable associated with the dependent variable..
- In correlation there is no difference in dependent and independent variables but in regression both variables are different.
- The objective of correlation is to find a value expressing the relationship between variables while regression is used to estimate value of a random variable based on the values of a fixed variable.
- Correlation describe a linear relationship between two variables while Regression fit the best line and estimate one variable based on another variable.

R vs. R Square $(R)^2$

- ✓ ‘R’ called the linear correlation coefficient which measures the strength and direction of linear relationship of two variables. It is sometimes referred to as the Pearson product moment correlation coefficient in the honour of its developer Karl Pearson.
- ✓ The coefficient of determination also called R-square value. It is useful because it gives the proportion of the variance of one variable that is predictable from the other variable. It is a measure that allow us to determine how certain one can be in making predictions from certain model/graph. The coefficient of determination is the ratio of the explained variation to the total variation
- ✓ R–square value is the square of the correlation. It measures the proportion of variation in the dependent variable that can be attributed to the independent variable. The R-square value is always between 0 and 1 inclusive. Perfect positive linear association.

Experimental design

- Experimental design are various types of plot arrangement which are used to test a set of treatments to draw a valid conclusion about a particular problem.
- In other words experimental design is the process of planning and study to meet a specific objectives.
- Experimental design are broadly classified into two; single- factor experiments (a single factor varies while all others are kept constant). and multi-factor experiments or factorial experiments (effects of more than one factor are to be determined).
- It is developed by R.A. Fisher in 1920.

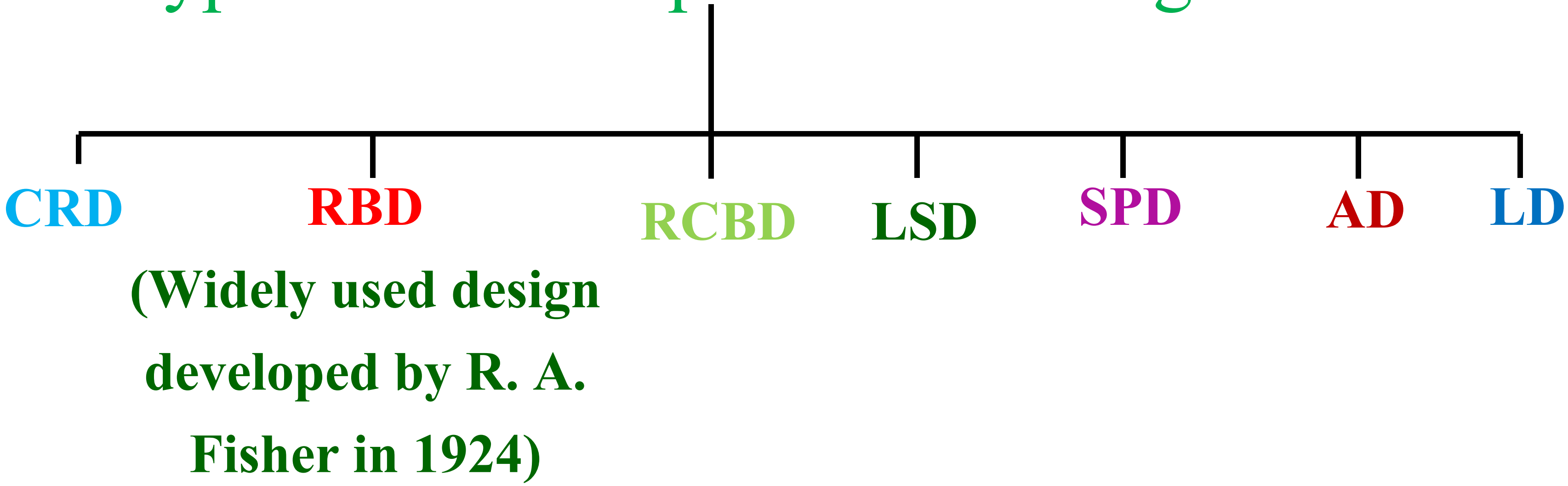
Objectives of experimental design

- Reduce the experimental error.
- In reducing the soil heterogeneity.
- Increase precision of experiments.
- In screening off various treatments.
- In assessment of variance and covariance.
- Used in proper interpretation of scientific results and drawing valid conclusion.
- Show the direction of better results.
- In partitioning of variation into different components.

Principle of experimental design

- Replication-replication of the treatment under investigation or to provide an estimate of experimental error.
- Randomization-the allocation of the treatment to the different experimental units by a random process is known as randomization.
- Local control-the principal of making use of greater homogeneity in groups of experimental units for reducing experimental error.

Types of basic experimental designs



CRD

RBD

(Widely used design developed by R. A. Fisher in 1924)

RCBD

LSD

SPD

AD

LD

CRD-Completely randomized designs

RBD-Randomized block designs

RCBD-Randomized complete block designs

LSD- Latin square design

SPD-Split plot designs

AD- Augmented designs

LD- lattice designs (alpha lattice design)

Need of experimental design in breeding

- To develop new varieties.
- Problem with regards to the estimation of differences between candidates and testing their significance are not considered.
- Getting the answer to question which the experimenter wants to know.
- The average value of all observations in a population.

Completely randomized design (CRD)

- CRD is a statistical experimental design where the treatments are assigned completely at random so that each treatment unit has the same chance of receiving any one treatment.
- In CRD ,any difference among experimental units receiving the same treatment is considered as an experimental error
- CRD is applicable only when the experimental material is homogenous.
- CRD is generally applicable to the lab experimental conditions it is not used in field experiments.

Advantages of CRD

- Easy to understand and calculate the variance.
- The number of replication can vary from treatment to treatment.
- It has high flexibility and thus any number of treatments can be used.
- Simple statistical analysis is required in the analysis of CRD.
- CRD provides maximum number of degree of freedom.

Disadvantages of CRD

- CRD can be applied only to homogenous experiments.
- The principle of local control is not used in CRD

Randomized block design (RBD)

- It is used when the experimental materials is not homogenous and fertility gradient is moving one direction.
- RBD is the most commonly used experimental design in agriculture.
- The number of equal plot in each block equal to the number of treatment.
- In case of field experiment materials is divided into a number of equal blocks
- The design is based on of three principles(Replication, Randomization and Local control) of experimental designs.

Random block design

- This method was introduced by S. Bernstein.
- Blocking is the arranging of experimental units in groups(block) that are similar to one another.
- Typically blocking factor is a source of variability that is not of primary interest to the experimenter.
- With a random block design , the experimenter divides subjects into subgroups called blocks, such that the variability within blocks is less tan the variability between blocks.
- Blocking is used to remove the effects of a few of the most important nuisance variables. Randomization is then used to reduce the contamination effects of the remaining nuisance variables. For important nuisance variables , blocking will yield higher significance in the variables of interest than randomizing.

Laying out of RBD

1- The experimental material is divided into blocks consisting of homogenous experimental units and each block is divided into number of treatments equal to the total number of treatments.

2- Randomization should be taken within each block and treatments are applied following the random number table.

3-collection and analysis of data: after the collection of data from the individual experimental unit(treatments) ANOVA(analysis of variance) table is formed.

4- computation of critical differences –critical differences is the differences between the treatment means, which places the treatments statistically as well as significantly apart. Otherwise if the differences of two treatments mean is less than CD it can be concluded both the treatments are on par.

Advantages of RBD

- It is more efficient and accurate to CRD.
- Chance of error in RBD is comparatively less.
- Flexibility is also very high in RBD thus any number of treatments and any number of replications can be used.
- When material is heterogeneous and number of treatment more than twenty.
- Analysis of sample relatively simple and easy.
- Errors of any treatment can be isolated.

Disadvantages of RBD

- Not applicable for very large number of treatment.
- It can not be applied if heterogeneity of the plot is very high.
- With large number of treatments, the possibility of experimental errors will be high.

CRD vs. RBD

- Randomization is done treatment wise in CRD while replication or block wise in RBD.
- Total variation is divided into two components (treatments and error)in CRD while into three components (blocks, treatments, and error) in RBD.

Latin square design(LSD)

- LSD is a design where the experimental material is divided into ‘m’ rows, ‘m’ columns and ‘m’ treatments-assigned by randomization method to rows and column.
- The randomization is in such a way that each treatment occurs only once in each row and in each column.

Advantages of LSD

- Analysis is relatively simple but complicated than CRD and RBD.
- Statistical analysis is simple if one value is missing.
- Most efficient design when compared to CRD and RBD.

Disadvantages of LSD

- Not suitable for agricultural experiments.
- When two or more values are missing analysis is complicated.
- Difficult when treatments are more than ten.

Randomized complete block design(RCBD)

- RCBD is a standard design for agricultural experiments in which similar experimental units are grouped into blocks or replicates. It is used to control variation in an experiments.

Thank you