STA301 - Statistics and Probability Glossary By www.virtualians.pk

Pie Chart : Pie Chart consists of a circle which is divided into two or more mars in accordance with the number of distinct classes that we have in our data.

Statistical Inference : Statistical Inference is an estimate or prediction or some other generalization about a Population based on information contained in a sample.

Statistics : Statistics is that science which enables to draw conclustions about various phenomena one the basis of real data collected on sample basis.

Sample : Sample is that part of the Population from which information is collected.

What is meant by order?. : Order: Arrangement of objects in ascending or descending way is known as order.

Population : The collection of all individuals,items or data under consideration in statistical study is called Population.

Nominal Scale : The classification or grouping of observatoins into mutually exclusive qualitative categories is said to constitute a nominal scale e.g students are classified as male and female.

Ordinal Scale : It includes the characteristic of a nominal scale and in addition has the property of ordering or ranking of measurments e.g the performance of students can be rated as excellent,good or poor.

Interval Scale : A measurment scale possessing a constant interval size but not true zero point is called an Interval Scale.

Ratio Scale : It is a special kind of an interval scale in which the scale of measurment has a true zero point as its origin.

Median : Median of a set of values arranged in ascending or descending order of magnitude is defined as middle value if the number of values is odd and mean of two middle values if the number of values is even.Median is a value at or below which 50% of data lie.

Average : A single value which intended to represent a distribution or a set of data as a whole is called an average. It is more or less a central value around which the observations tend to cluster so it is called measure of central tendency. Since measure of central tendency indicate the location of the distribution on X axis so it is also called measure of location.

Mean Deviation : The mean deviation is defined as the arithmetic mean of the deviations measured either from the mean or from the median, all deviations being counted as positive.

Chebshev's Theorm : Chebshev's Theorem states that "For any number K greater than one at least 1-1/k2 of the data values fall with in K standard deviations of the mean i.e. Within the interval.

Moments : Moments are the arithmetic means of the powers to which the deviations are raised.

Kurtosis : kurtosis is the degree of peakness of a distribution usually taken relative to a normal distribution.

Correlation : Correlation is a measure of the strength or the degree of relationship between two random variables. OR Interdependence of two variables is called correlation.

Venn Diagram : A diagram that is understood to represent sets by circular regions, parts of circular regions or their complements with respect to a rectangle representing the space S is called a Venn diagram. The Venn diagrams are used to represent sets and subsets in a pictorial way and to verify the relationship among sets and subsets.

Mutually Exclusive Event : Two events are said to be mutually exclusive events if and only if they can not both occur together at the same time. OR Two events are said to be mutually exclusive events if the occurrence of one event discard the occurrence of other event.

Independent events : Two events A and B in the same sample space S, are defined to be independent (or statistically independent) if the probability that one event occurs, is not affected by whether the other event has or has not occured.

Random variable : A numerical quantity whose value is determined by the outcome of a random experiment is called a random variable.

Distribution Function : The function which gives the probability of the event that X takes a value less than or equal TO a specified value x is called a distribution function and is also called the cumulative distribution function.

Cumulative Distribution Function : The function which gives the probability of the event that X takes a value less than or equal TO a specified value x is called a cumulative distribution function and is also called the distribution function.

Sampling Frame : A sampling frame is a complete list of all the elements in the population.

Sampling Error : The sampling error is the difference between the the sample statistic and the population parameter.

Probability Samples : Probability samples are those in which following the sampling plan each unit in the poplation has a known probability of being included in the sample.

Non probability samples : Non probability samples are those in which the sample elements are the arbitrarily selected by the sampler because in this judgment the elements thus chosen will most effectively represent the Population.

Frequency Polygon : A frequency polygon is obtained by plotting the class frequencies against the mid-points of the classes, and connecting the points so obtained by straight line segments.

Variable : A measurable quantity which can vary from one individual or object to another is called a variable.

Constant : A quantity which can assume only one value is called a constant

Event. : the possible outcomes of an experiment is known as event.

Data. : A well defined collection of objects is known as data.

Mode : The mode is a value which occures most frequently in a set of data i.e. It indicates the most common result

Box and Whisker plot : A Box and Whisker plot provides a graphical representation of data through its five number summary.

The five number summary : A five number summary consists of X0, Q1, median, Q3, and Xm. It enables us to find the shape of the distribution without drawing a graph.

EXHAUSTIVE EVENTS : Two or more than two mutually exclusive events are said to be exhaustive events when their union constitute the entire sample space

Equally likely events : Two events A and B are said to be equally likely when one event is as likely to occur as other

Probability : Probability is defined as the ratio of favorable cases over equally likely cases.

Table : Table is a systematic arrangement of data into vertical columns and horizontal rows.

Tabulation : The process of arranging data into rows and columns is called tabulation.

Classification : The process of arranging data in classes or categories according to some common characteristics present in the data is called classification.

Class Mark or Mid Point : The class mark or mid point is that value which divides a class into two equall parts.

Mid Poin or Class Mark : The mid point or class mark is that value which divides a class into two equal parts.

Measure of location : A single value which intended to represent a distribution or a set of data as a whole is called an average. It is more or less a central value around which the observations tend to cluster so it is called measure of central tendency. Since measure of central tendency indicate the location of the distribution on X axis so it is also called measure of location.

The Semi-interquartile Range : The quartile deviation or the Semi-interquartile Range is defined as half of the difference between the first and third quartiles.

The coefficient of variation : The coefficient of variation expresses the standard deviation as the percentage of the arithmetic mean.

Disjoint Set : Two sets A and B are said to be disjoint Sets if they have no elements in common.

DISTRIBUTION FUNCTION: : The distribution function of a random variable X, denoted by F(x), is defined by F(x) = P(X < x). The function F(x) gives the probability of the event that X takes a value LESS THAN OR EQUAL TO a specified value x. The distribution function is abbreviated to d.f. And is also called the cumulative distribution function (cdf) as it is the cumulative probability function of the random variable X from the smallest value up to a specific value x.

Experimental design: : An experimental design is a set of rules or a plan to collect the data relevant to the problem under investigation in such a way as to provide the basis for valid and objective inferences about the stated problem. The plan usually consists of collection of the treatments, specification of experimental layout, allocation of treatments.

Experimental Unit: : An experimental unit is the basic unit to which the experiment is performed. It is the basic unit to which the treatment is applied and in which the variable under investigation is measured and analyzed.

Randomized Deigns; : These designs are those in which treatments are applied to experimental units randomly and conclusions are supported by the statistical results.

Basic Randomized Designs: : Randomization Replication Local Control

Randomization: : It is a random process of assigning treatments to the experimental unit. The random process implies that every possible allocation of treatments has the same probability.

Replication: : The second principle of an experimental design is replication which is the repetition of the basic experiment. It is a complete run of all the treatments to be tested in the experiment.

Local Control: : It is used to bring all extraneous sources of variations under control. For this purpose we use Local Control, a term referring to the amount of balancing, blocking and grouping of the experimental units.

Complete Randomized Designs : In this design treatments are applied to the experimental units completely at random, that is randomization is done without any restrictions. Design is completely flexible, any number of treatments and any number of units per treatments can be applied.

ANOVA : Analysis of variance is defined as the procedure by means of which the total variability of the set of data measured by total sum of square is partitioned into components that measure different sources of variations. The procedure thus permits the decomposition of the total SS into to the component SS which are corresponding to the real and suspected sources of variations.

Randomized complete block Design (RCB): : Randomized complete block Design (RCB) is a design in which • Experimental material is divided into groups or blocks in such a manner that experimental units within a particular block are relatively homogeneous. • Each block contains complete set of treatments i.e. It constitutes a replication of treatments. • Treatments are assigned at random to the experimental units with in each block which means the randomization is restricted with blocks.

Latin Square Design: : LS design is an arrangement of k treatments in a k*k square, where the treatments are grouped in blocks in two directions, the direction being orthogonal to each other and to the treatments, and where the treatments appear once and only once in each in each direction. It should be noted that in Latin square design, the number of rows, the number of columns and number of treatments must be equal

Critical Value : The value that separates the critical region from the acceptance region, is called the critical value(s).

Level of significance : Level of significance of a test is the probability used as a standard for rejecting null hypothesis Ho when Ho is assumed to be true. The level of significance acts as a basis for determining the critical region of the test.

Statistics 2 : Statistics is a science of facts and figures.

Deciles : Deciles are those nine quantities that divide the distribution into ten equall parts.

Percentiles : Percentiles are those ninety nine quantities that divide the distribution into hundred equall parts

Arithmetic Mean : Arithmetic Mean is a value obtained by dividing the sum of the observations by their numbers.

Geometric Mean : The Geometric Mean G, of a set of n positive values is defined as the positive nth root of their product.

Absolute Measure of Dispersion : An absolute measure of dispersion is one that measures the dispersion in terms of the same units, or in the square of units as the units of the data.

Dispersion : The variability that exists between data set.

Relative Measure of Dispersion : A Relative Measure of Dispersion is one that measures the dispersion in terms of a ratio, coefficient or percentage and is independent of the units of measurement.

Range : The range is defined as the difference between the maximum and minimum values of a data set.

Quartile Deviation : The quartile deviation is defined as half of the difference between the first and third quartiles.

Set : A set is any well defined collection or list of distinct objects.

Standard error of estimate : The degree of scatter of the observed values about the regression line measured by what is called standard deviation of regression or standard error of estimate.

Class of Sets : A set of sets is called a class.

Primary Data : The data published or used by an organization which originally collected them are called primary data thus the primary data are thr first hand information collected, complied, and published by an organization for a certain purpose.

Secondary Data : The data published or used by an organization other than the one which origninally collected them are known as secondary data.

Harmonic Mean : Harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of the values.

Quartiles : Quartiles are those three quantities that divide the distribution into four equal parts.

Quantiles : Collectively the quartiles, the deciles,percentiles and other values obtained by equall sub-division of the data are called quantiles.

Index Number : An Index Number is a statistical measure which shows changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession etc.

Standard Deviation : Standard Deviation is defined as the positive square root of the mean of the squared deviations of the values from their mean.

Variance : Variance is defined as the square of the standard deviation.

Regression : Dependence of one variable on the other variable is called regression. OR Estimation or prediction of one variable on the basis of other variable is called regression.

Random Experiment : An experiment which produces different results even though it is repeated a large number of times under essentially similar conditions is called a random experiment.

Sub Set : A set that consists of some elements of an other set is called a subset of that set.

Non-Sampling Error : Such errors which are not attributable to sampling but arise in the process of data collection even if a complete count is carried out.

Skewness : Skewness is the lack of symmetry in a distribution around some central value (mean,median or mode).It is thus the degree of a symmetry.

Permutation : an arrangement of all or some of a set of objects in a definite order is called permutation.

Universal Set : All sets are subsets of one particular set called universal set.

Sample Space : The set or collection of all possible outcomes of an experiment is called the sample space.

Conditional Probability : The probability of the occurence of an event A when it is known that some other event B has already occured is called the conditional probability.

Degrees of freedom : Degrees of freedom can be defined as the number of observations in the sample minus the number of population parameters that are estimated from the sample data (from those observations

P value : The p-value is a property of the data, and it indicates “how improbable” the obtained result really is.

Test Statistic : A statistic (i.e. A function of sample data not containing any parameter), which provides a basis for testing a null hypothesis, is called a test statistics.

Addition law : A probability law used to compute the probability of a union of two events, denoted A and B. It is P(AÈB)=P(A)+P(B)-P(AÇB). For mutually exclusive events, because P(AÇB)=0, it reduces to P(AÈB)=P(A)+P(B).

Alternative hypothesis : The hypothesis concluded to be true if the null hypothesis is rejected.

ANOVA table : A table used to summarize the analysis of variance computations and results. It contains columns showing the source of variation, the sum of squares, the degrees of freedom, the mean square, and the F values.

Bayes' theorem : A method used to compute posterior probabilities.

Binomial probability distribution : A probability distribution showing the probability of x successes in n trials of a binomial experiment.

Binomial probability function : The function used to compute probabilities in a binomial experiment.

Blocking : The process of using the same or similar experimental units for all treatments. The purpose of blocking is to remove a source of variation from the error term and hence provide a more powerful test for a difference in population or treatment means.

Box plot : A graphical summary of data. A box, drawn from the first to the third quartiles, shows the location of the middle 50% of the data. Dashed lines, called whiskers, extending from the ends of the box show the location of data values greater than the third quartile and data values less than the first quartile. The locations of any outliers are also noted.

Central limit theorem : A theorem that enables one to use the normal probability distribution to approximate the sampling distribution of the sample mean and sample proportion whenever the sample size is large.

Consistency : A property of a point estimator that is present whenever larger sample sizes tend to provide point estimates closer to the population parameter

Histogram : A graphical presentation of a frequency distribution, relative frequency distribution, or percent frequency distribution of quantitative data constructed by placing the class intervals on the horizontal axis and the frequencies on the vertical axis.

Null hypothesis : The hypothesis tentatively assumed true in the hypothesis testing procedure. Or A null hypothesis, generally denoted by the symbol H0, is any hypothesis which is to be tested for possible rejection or nullification under the assumption that it is true.

Normal probability distribution : A continuous probability distribution. Its probability density function is bell shaped and determined by its mean m and standard deviation s.

Observation : The set of measurements obtained for a single element.

Ogive : A graph of a cumulative distribution.

One-tailed test : A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in one tail of the sampling distribution. Or The entire rejection region lies in only one of the two tails, either in the right tail or in the left tai, of the sampling distribution of the test-statistic, is called a one-tailed test or one-sided test.

Parameter : numerical characteristic of a population, such as a population mean, a population standard deviation, a population proportion, and so on.

Point estimate : A single numerical value used as an estimate of a population parameter.

Point estimator : The sample statistic that provides the point estimate of the population parameter.

Poisson probability distribution : A probability distribution showing the probability of x occurrences of an event over a specified interval of time or space.

Poisson probability function : The function used to compute Poisson probabilities.

Population parameter : A numerical value used as a summary measure for a population of data (e.g., the population mean, the population variance, and the population standard deviation).

Posterior probabilities : Revised probabilities of events based on additional information.

Power curve : A graph of the probability of rejecting H0 for all possible values of the population parameter not satisfying the null hypothesis. The power curve provides the probability of correctly rejecting the null hypothesis.

Power : The probability of correctly rejecting H0 when it is false.

Probability density function : A function used to compute probabilities for a continuous random variable. The area under the graph of a probability density function over an interval represents probability.

Probability function : A function, denoted by f(x), that provides the probability that x assumes a particular value for a discrete random variable.

Qualitative data : Data that are labels or names used to identify an attribute of each element. Qualitative data may be nonnumeric or numeric.

Qualitative variable : A variable with qualitative data.

Quantitative data : Data that indicate how much or how many of something. Quantitative data are always numeric.

T Distribution : A family of probability distributions that can be used to develop interval estimates of a population mean whenever the population standard deviation is unknown and the population has a normal or near-normal probability distribution.

Target population : The population about which inferences are made.

Treatment : Different levels of a factor.

Tree diagram : A graphical representation helpful in identifying the sample points of an experiment involving multiple steps.

Two-tailed test : A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in either tail of the sampling distribution.

Type I error : The error of rejecting H0 when it is true.

Type II error - : The error of accepting H0 when it is false.

Unbiasedness : A property of a point estimator when the expected value of the point estimator is equal to the population parameter it estimates.

Union of events A and B : The event containing all sample points that are in A, in B, or in both. The union is denoted AÈB.

Types of Experimental Designs: : Systematic Designs Randomized design

Systematic Designs: : These designs are those in which treatments are applied to the experimental units by some systematic manner that is choice of the experimenter

Acceptance and rejection region: : All possible values which a test-statistic may assume can be divided into two mutually exclusive groups: One group consisting of values which appear to be consistent with the null hypothesis (i.e. Values which appear to support the null hypothesis), and the other having values which lead to the rejection of the null hypothesis. The first group is called the acceptance region and the second set of values is known as the rejection region for a test

Type I error: : When we perform a hypothesis test, we derive evidence from the sample in the form of a test statistics. There is a possibility that sample may lead us to make a wrong decision. We may reject the hypothesis when it is in fact true. This type of error is called an error of first kind or type I-error. The probability of committing a type I error is denoted by α. Thus α is the probability of rejecting null hypothesis Ho when Ho true.

Type II error: : When we perform a hypothesis test, we derive evidence from the sample in the form of a test statistics. There is a possibility that sample may lead us to make a wrong decision. We may accept the hypothesis when it is in fact false. This type of error is called an error of second kind or a Type II error. The probability of committing a type II error is denoted by β. Thus β is the probability of accepting null hypothesis Ho when Ho false.

Class midpoint : The point in each class that is halfway between the lower and upper class limits.

Complement of event A : The event consisting of all sample points that are not in A.

Dependent variable : The variable that is being predicted or explained. It is denoted by y.

Descriptive statistics : Tabular, graphical, and numerical methods used to summarize data.

Dot plot : A simple graphical summary of data with each observation represented by a dot placed above a horizontal axis that shows the range of values for the observations.

Discrete random variable : A random variable that may assume either a finite number of values or an infinite sequence of values.

Empirical rule : A rule that states the percentages of items that are within one, two, and three standard deviations from the mean for mound-shaped, or bell-shaped, distributions.

Experiment : A process that generates well-defined outcomes.

Binomial experiment : A probability experiment having the following four properties: consists of n identical trials, two outcomes (success and failure) are possible on each trial, probability of success does not change from trial to trail, and the trials are independent.

Factorial experiment : An experimental design that allows statistical conclusions about two or more factors.

Five-number summary : An exploratory data analysis technique that uses the following five numbers to summarize the data set: smallest value, first quartile, median, third quartile, and largest value.

Frame : A list of the sampling units for a study. The sample is drawn by selecting units from the frame.

Frequency distribution : A tabular summary of data showing the number (or frequency) of items in each of several non-overlapping classes.

Grouped data : Data available in class intervals as summarized by a frequency distribution. Individual values of the original data are not available.

Independent variable : The variable that is doing the predicting or explaining. It is denoted by x.

Intersection of A and B : The event containing all sample points that are in both A and B. The intersection is denoted AÇB.

Joint probability : The probability of two events both occurring; that is, the probability of the intersection of two events.

Judgment sampling : A nonprobabilistic method of sampling whereby element selection is based on the judgment of the person doing the study.

Interquartile range (IQR) : A measure of variability, defined to be the difference between the third and first quartiles.

Least squares method : The method used to develop the estimated regression equation. It minimizes the sum of squared residuals (the deviations between the observed values of the dependent variable, yi, and the estimated values of the dependent variable, yi)

Regression equation : The equation that describes how the mean or expected value of the dependent variable is related to the independent variable.

Rejection region : The range of values that will lead to the rejection of a null hypothesis.

Replication : The number of times each experimental condition is repeated in an experiment.

Residual : The difference between the observed value of the dependent variable and the value predicted using the estimated regression equation.

Sample point : An element of the sample space. A sample point represents an experimental outcome.

Sample statistic : A numerical value used as a summary measure for a sample (e.g., the sample mean, the sample variance, and the sample standard deviation). The value of the sample statistic is used to estimate the value of the population parameter.

Sampled population : The population from which the sample is taken.

Sampling unit : The units selected for sampling. A sampling unit may include several elements.

Sampling with replacement : Once an element has been included in the sample, it is returned to the population. A previously selected element can be selected again and therefore may appear in the sample more than once.

Sampling without replacement : Once an element has been included in the sample, it is removed from the population and cannot be selected a second time.

Scatter diagram : A graph of bivariate data in which the independent variable is on the horizontal axis and the dependent variable is on the vertical axis.

Simple linear regression : Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line.

Simple random sampling : Finite population: a sample selected such that each possible sample of size n has the same probability of being selected. Infinite population: a sample selected such that each element comes from the same population and the elements are selected independently.

Standard error : The standard deviation of a point estimator.

Stem-and-leaf display : An exploratory data analysis technique that simultaneously rank orders quantitative data and provides insight about the shape of the distribution.

Stratified random sampling : A probability sampling method in which the population is first divided into strata and a simple random sample is then taken from each stratum.

Hypergeometric probability function : The function used to compute the probability of x successes in n trials when the trials are dependent.

Multiplication law : A probability law used to compute the probability of an intersection of two events, denoted A and B. It is P(AÇB)=P(A)P(B|A) or P(AÇB)= P(B)P(A|B). For independent events it reduces to P(AÇB)=P(A)P(B).

Goodness of fit test : A statistical test conducted to determine whether to reject a hypothesized probability distribution for a population.

Sampling distribution : A probability distribution consisting of all possible values of a sample statistic.