STA301 Formulas All formulas from handouts

Statistics and Probability (Virtual University of Pakistan)

Formulas

FROM LECTURE 01 TO 22

Mean:

$$\frac{\sum fX}{\sum f}$$

Weighted Mean:

$$\overline{X}_w = \frac{\sum W_i X_i}{\sum W_i}$$

Mean Deviation:

Ungroup Data

$$M.D = \frac{\sum |d|}{n}$$

Group Data

$$M.D = \frac{\sum f_i |d_i|}{\sum f}$$

$$|d| = (X - \overline{X})$$

Coefficient of Mean Deviation:

$$Co\text{-efficient of } M.D(\text{for mean}) = \frac{M.D}{Mean}$$

$$Co\text{-efficient of } M.D(\text{for median}) = \frac{M.D}{Median}$$

Standard Deviation:

Ungroup Data

$$S.D = \sqrt{\frac{\sum (X - \overline{X})^2}{n}}$$

Group Data

$$S.D = \sqrt{\frac{\sum f(X - \overline{X})^2}{\sum f}}$$

Shortcut Formula for Ungroup data

$$S.D = \sqrt{\left\{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2\right\}}$$

Shortcut Formula for Group Data

$$S.D = \sqrt{\left\{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2\right\}}$$

Co-Efficient of Standard Deviation:

$$= \frac{S.D}{\overline{X}}$$

Variance:

Ungroup Data

$$\text{Variance} = \frac{\sum\left(X - \overline{X}\right)^2}{n}$$

Group Data

$$\text{Variance} = \frac{\sum f\left(X - \overline{X}\right)^2}{\sum f}$$

Co-Efficient of Variation:

$$C.V = \frac{S.D}{\overline{X}} \times 100$$

Median:

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

$$Median = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

Harmonic Mean:

Ungroup Data

$$\overline{X} = \frac{n}{\sum\left(\frac{1}{x}\right)}$$

Group Data

$$\overline{X} = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$$

Average Speed:

$$\text{Average Speed} = \frac{\text{Total Distance Travelled}}{\text{Total Time Taken}}$$

$$\text{Time} = \frac{\text{Distance}}{\text{Speed}}$$

$$\text{Speed} = \frac{\text{Distance}}{\text{Time}}$$

$$\text{Distance} = \text{Speed} \times \text{Time}$$

Mode:

$$\widehat{X} = l + \frac{(fm - f1)}{(fm - f1) + (fm - f2)} \times h$$

Empirical Relation between Mean, Median and Mode:

$$Mode = 3Median - 2Mean$$

Range:

$$Range = X^m - X^0$$

Mid-Range:

$$\text{Mid Range} = \frac{X_0 + X_m}{2}$$

Mid Quatile Range:

$$\text{Mid Range} = \frac{Q_1 + Q_3}{2}$$

Coefficient of Dispersion:

$$\text{Coefficient of Dispersion} = \frac{X_m - X_0}{X_m + X_0}$$

Geometric Mean:

Ungroup Data

$$G.M = anti\log\left[\frac{\sum \log X}{n}\right]$$

Group Data

$$G.M = anti\log\left[\frac{\sum f \log X}{\sum f}\right]$$

Quartile Deviation:

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

$$Co-efficient \text{ of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

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

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

Skewness:

$$SK = \frac{Mean - Mode}{S.D}$$

Pearson's Coefficient of Skewness

$$SK = \frac{3(Mean - Median)}{S.D}$$

Bowley's coefficient of Skewness

$$SK = \frac{Q_1 + Q_3 - 2(Median)}{Q_3 - Q_1}$$

Kurtosis:

$$K = \frac{Q.D}{P_{90} - P_{10}}$$

Moments:

Ungroup Data

$$m_1 = \frac{\sum (X_i - \bar{X})}{n} = 0$$

$$m_2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

$$m_3 = \frac{\sum (X_i - \bar{X})^3}{n}$$

Group Data

$$m_3 = \frac{\sum f(X_i - \bar{X})^3}{\sum f}$$

Moment Ratios:

$$b_1 = \frac{(m_3)^2}{(m_2)^3} \text{ and } b_2 = \frac{m_4}{(m_2)^2}$$

Equation of a Straight Line

$$Y = a + bX$$

Standard Error of Estimate:

$$s_{yx} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

Normal Equations:

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Correlation:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})\sum (Y - \bar{Y})}}$$

PEARSON'S COEFFICIENT OF CORRELATION

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

Linear Correlation of Coefficient:

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

Regression Line of Y on X:

$$byx = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

Division of Circle:

$$= \frac{\text{Cell Frequency}}{\text{Total Frequency}} \times 360$$

Binomial Coefficient:

$$({}^nC_r)$$

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^{n-r}\ b^r$$

1. Cumulative Law:

$A \cup B = B \cup A$

$A \cap B = B \cap A$

2. Associative Law:

$(A \cup B) \cup C = A \cup (B \cup C)$

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

3. Distributive Law:

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

4. Idempotent Law:

$A \cup A = A$

$A \cap A = A$

5. Identity Law:

$A \cup S = S$

$A \cap S = A$

$A \cup \phi = A$

$A \cap \phi = \phi 0$

6. Complementation Law:

$A \cup \overline{A} = S$

$A \overline{\cap} A = \phi$

$(\overline{A}) = A$

$\overline{S} = \phi$

$\phi = \overline{S}$

7. De'Morgan's Law:

$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$

$\overline{(A \cap B)} = \overline{A} \cup \overline{B}$

Rule of Permutation:

$$= \frac{n!}{(n-r)!}$$

Rule of Combination:

$$= \frac{n!}{r!(n-r)!}$$

Probability Formula:

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

$$= P(A) = \frac{\text{No. of Favourable Outcomes}}{\text{Total No. of Possible Outcomes}}$$

Mutually Exclusive Events:

If A and B mutually exclusive events,

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

Complementation Law of Probability:

$$P(\overline{A}) = 1 - P(A)$$

Addition Law:

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

Conditional Probability:

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

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

Multiplication Law:

$$P(A \cap B) = P(A)P(B/A)$$

Baye's Theorem:

$$P(A_1 | B) = \frac{P(A_1).P(B|A_1)}{P(A_1).P(B|A_1) + P(A_2).P(B|A_2)}$$

Computation of Standard Deviation Probability Distribution:

$$S.D.(X) = \sqrt{\frac{\sum X^2 P(X)}{\sum P(X)} - \left[\frac{\sum XP(X)}{\sum P(X)}\right]^2}$$

$$S.D.(X) = \sqrt{\sum X^2 P(X) - \left[\sum XP(X)\right]^2}$$

As we know,

$$\sum P(X) = 1$$

Coefficient of Variation:

$$= \frac{\sigma}{\mu} \times 100$$

Total Cards = 52

Diamond = 13
Heart = 13
Club = 13
Spades = 13

Face Cards = 12

Queen = 4
Jack = 4
King = 4
Jocker = 4

FROM LECTURE 23 TO 45

Mathematical Expectations:

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

E(X) is also called the mean of X and is usually denoted by the letter μ .
If the values are equally likely:

$$E(X) = \frac{1}{n} \sum X_i$$

If X is a discrete random variable and if a and b are constants, then:

$$E = (aX + b) = aE(X) + b$$

If μ is a Population Mean then:

$$E = (X - \mu)2 = E(X2) - [E(X)]2$$

Shortcut Formula for the Variance in Mathematical Expectations:

$$\sigma 2 = E(X) - [E(X)]2$$

Skewness of Probability:

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

Kurtosis of Probability:

$$\beta_1 = \frac{\mu_2^4}{\mu_2}$$

Mean on Probability:

$$E(X) = \sum XP(X)$$

If the Parameters in distribution are a and b then

$$\mu = \frac{a+b}{2}$$

Mean of the Population Distribution

$$\mu = E(X) = \sum xf(x)$$

Mean of the SAMPLIN DISTRIBUTION of $\overline{X}$

$$\mu_{\bar{x}} = E(\overline{X}) = \sum \bar{x} f(\bar{x})$$

$$\mu_{\bar{x}} = \mu$$

Variance of Probability:

$$Var(X) = \sum X2\ P(X) - [\sum XP(X)]2$$

Expected value of Variance

$$Var(X)\ or\ \sigma 2 = E(X-\mu)2 = \sum (x_i - \mu)2\ f(x)$$

If the Parameters in distribution are a and b then

$$\sigma^2 = \frac{(b-a)^2}{12}$$

Variance of the Population Distribution

$$\sigma^2 = \sum x^2 f(x) - \left[\sum xf(x)\right]^2$$

Variance of the SAMPLIN DISTRIBUTION of $\overline{X}$

Ungroup Data

$$\sigma^2 \overline{X} = \sum \overline{X}^2 - \left[\sum (\overline{X})\right]^2$$

Group Data

$$\sigma^2 \overline{X} = \sum \overline{X}^2 f(\overline{X}) - \left[\sum \overline{X} f(\overline{X})\right]^2$$

In case of sampling with replacement:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In case of sampling without replacement from a finite population:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Finite Population Correction (fpc) = $\sqrt{\frac{N-n}{N-1}}$

Binomial Distribution:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

IN CASE OF NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

Joint Probability Function:

$$f(x, y)$$

kth Moment:

$$m'_r = \frac{\sum X_i^r}{n}$$

Origin of Random Variable

μk

$$E(Xk) = \sum X_i k \, f(x)$$

Mean of Random Variable

μk

$$E(X - \mu)k = \sum (X_i - \mu)k \, f(x)$$

Chebychev's Inequality:

In the case of a Discrete Probability Distribution

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

$$\sigma = \sqrt{E[(X - \mu)^2]}$$

$$\mu = E(X)$$

MARGINAL PROBABILITY:

Marginal Probability function of X

$$g(x_i) = \sum_{j=1}^{n} f(x_i, y_j)$$

Marginal Probability function of Y

$$h(y_i) = \sum_{i=1}^{m} f(x_i, y_j) = P(Y = y_j)$$

Conditional Probability Function:

Conditional Probability function for X

$$\mathrm{f}(x_i \,|\, y_j) = P(X = x_i \,|\, Y = y_j)$$

Conditional Probability function for Y

$$\mathrm{f}(y_j \,|\, x_i) = P(Y = y_j \,|\, X = x_i)$$

HYPER GEOMETRIC PROBABILITY DISTRIBUTION:

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

Where
N = number of units in the population,
n = number of units in the sample, and
k = number of successes in the population.

HYPER GEOMETRIC DISTRIBUTION:

Mean of Hyper Geometric Distributions

$$\mu = n\frac{k}{N}$$

Variance of Hyper Geometric Distribution

$$\sigma^2 = n\frac{k}{N}\frac{N-k}{N}\frac{N-n}{N-1}$$

Binomial Probability Distribution:

$$\lim_{\substack{n\to\infty \\ \mathrm{p}\to 0}} \mathrm{b(x;n,p)} = \frac{e^{-\mu}\mu^x}{x!}$$

$$x = 0, 1, 2, ..., \infty$$

$$e = 2.17828$$

$$\mu = np$$

UNIFORM DISTRIBUTION:

A random variable X is said to be uniformly distributed if its density function is defined as

$$f(x)=\frac{1}{b-a},\quad a\le x\le b$$

Poisson Process Formula:

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

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

Where
λ = average number of occurrences of the outcome of interest per unit of time,
t = number of time-units under consideration, and
x = number of occurrences of the outcome of interest in t units of time.

NORMAL DISTRIBUTION:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$$

Where:

$$\pi=3.1416\ or\ \frac{22}{7}$$

$$e=2.71828$$

Moment Ratios in Probability:

The moment ratios of the normal distribution come out to be 0 and 3 respectively:

$$\beta_1=\frac{\mu_3^2}{\mu_2^3}=\frac{0^2}{(\sigma^2)^3}=0$$

$$\beta_2=\frac{\mu^4}{\mu_2^2}=\frac{3\sigma^4}{(\sigma^2)^2}=3$$

Standardization Formula:

$$Z=\frac{X-\mu}{\sigma}$$

In case of $\bar{X}$

$$Z=\frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Proportion of Successes in the Population:

$$p = \frac{X}{N} \text{ or } \hat{p} = \frac{X}{n}$$

p = Proportion of Success
X = Number of successes in the population
n = Sampling distribution of p

SAMPLING DISTRIBUTION OF DIFFERENCES BETWEEN MEANS:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

and

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}$$

$$\mu_{\bar{x}_1 - \bar{x}_2} = \sum (\bar{x}_1 - \bar{x}_2)\ f(\bar{x}_1 - \bar{x}_2)$$

and

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sum d^2 f(d) - \left[\sum df(d)\right]^2$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}$$

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}}$$

SAMPLING DISTRIBUTION OF $\hat{p}$

$$\mu_{\hat{p}} = \sum \hat{p} f(\hat{p})$$

Standard Deviation of the Sampling Distribution:

$$\sigma_{\hat{p}1 - \hat{p}2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Modified Formula for the Sample Variance:

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

Relative Efficiency:

$$E_f = \frac{Var(T_2)}{Var(T_1)}$$

Difference between the Means of Two Populations:

$$(\overline{x_1} - \overline{x_2}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

CONFIDENCE INTERVAL FOR A POPULATION PROPORTION (P):

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where

$\hat{p}$ = Proportion of "success" in hte sample

n = Sample size

$z_{\alpha/2}$ = 1.96 for 95% confidence

2.58 for 99% confidence

CONFIDENCE INTERVAL FOR P1-P2:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p})}{n_2}}$$

SAMPLE SIZE FOR ESTIMATING POPULATION PROPORTION:

$$\hat{p} = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

t-DISTRIBUTION:

$$f(x) = \frac{1}{\sqrt{v}\beta\left(\frac{1}{2}, \frac{v}{2}\right)}\left(1 + \frac{x^2}{v}\right)^{-(v+1)/2}$$

Variance of the t-distribution

$$\sigma^2 = \frac{v}{v-2} \quad for\ v > 2$$

F-Distribution:

$$f(x)=\frac{\Gamma\left[(v_1+v_2)/2\right](v_1/v_2)^{v_1/2}x^{(v_1/2)-1}}{\Gamma(v_1/2)\Gamma(v_2/2)\left[1+v_1x/v_2\right]^{(v_1+v_2)/2}},\qquad 0<x<\infty$$

Mean of the F-distribution:

$$For\ v2>2$$

$$=\frac{v_2}{v_2-2}$$

Variance of the F-distribution:

$$For\ v2>4$$

$$\sigma^2=\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$$

Mean Square:

$$Mean\ Square=\frac{Sum\ of\ Squares}{Degrees\ of\ Freedom}$$

Least Significant Difference:

$$LSD=t_{\alpha/2,(v)}\sqrt{\frac{2(MSE)}{r}}$$

Definitions

FROM LECTURE 01 TO 22

Statistics:
Statistics is that science which enables to draw conclusions about various phenomena one the basis of real data collected on sample basis.
OR
Statistics is a science of facts and figures.

INFERENTIAL STATISTICS:
That branch of Statistics which enables us to draw conclusions or inferences about various phenomena on the basis of real data collected on sample basis.

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

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.

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

Nominal Scale:
The classification or grouping of observations 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 measurements e.g. the performance of students can be rated as excellent, good or poor.

Interval Scale:
A measurement 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 measurement has a true zero point as its origin.

Biased Errors:
An error is said to be biased when the observed value is consistently and constantly higher or lower than the true value.

Unbiased Errors or Random Errors:
An error, on the other hand, is said to be unbiased when the deviations, i.e. the excesses and defects, from the true value tend to occur equally often.

Primary Data:
The data published or used by an organization which originally collected them are called primary data thus the primary data are the 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 originally collected them are known as secondary data.

DIRECT PERSONAL INVESTIGATION:
In this method, an investigator collects the information personally from the individuals concerned. Since he interviews the informants himself, the information collected is generally considered quite accurate and complete.

INDIRECT INVESTIGATION:
Sometimes the direct sources do not exist or the informants hesitate to respond for some reason or other. In such a case, third parties or witnesses having information are interviewed.

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

Sample:
A sample is a group of units selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions about the larger group.
OR
Sample is that part of the Population from which information is collected.

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 sample statistic and the population parameter.

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.

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.

Standard error:
The standard deviation of a point estimator.
AND
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.

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

NON-RANDOM SAMPLING:
Nonrandom sampling' implies that kind of sampling in which the population units are drawn into the sample by using one's personal judgment. This type of sampling is also known as purposive sampling.

Quota Sampling:
Quota sampling is a method of sampling widely used in opinion polling and market research. Interviewers are each given a quota of subjects of specified type to attempt to recruit for example, an interviewer might be told to go out and select 20 adult men and 20 adult women, 10 teenage girls and 10 teenage boys so that they could interview them about their television viewing.

RANDOM SAMPLING:
The theory of statistical sampling rests on the assumption that the selection of the sample units has been carried out in a random manner.
By random sampling we mean sampling that has been done by adopting the lottery method.

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.

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.

SIMPLE BAR CHART:
A simple bar chart consists of horizontal or vertical bars of equal width and lengths proportional to values they represent.

MULTIPLE BAR CHARTS:
This kind of a chart consists of a set of grouped bars, the lengths of which are proportionate to the values of our variables, and each of which is shaded or colored differently in order to aid identification.

CLASS BOUNDARIES:
The true class limits of a class are known as its class boundaries.

HISTOGRAM:
A histogram consists of a set of adjacent rectangles whose bases are marked off by class boundaries along the X-axis, and whose heights are proportional to the frequencies associated with the respective classes.

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.

FREQUENCY CURVE:
When the frequency polygon is **smoothed**, we obtain what may be called the frequency curve.

CUMULATIVE FREQUENCY DISTRIBUTION:
As in the case of the frequency distribution of a discrete variable, if we start adding the frequencies of our frequency table column-wise, we obtain the column of cumulative frequencies.

AVERAGES (I.E. MEASURES OF CENTRAL TENDENCY):
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 Arithmetic, Geometric and Harmonic means
Are averages that are mathematical in character, and give an indication of the magnitude of the observed values.

The Median
Indicates the middle position while the mode provides information about the most frequent value in the distribution or the set of data.

THE MODE:
The Mode is defined as that value which occurs most frequently in a set of data i.e. it indicates the most common result.

DOT PLOT:
The horizontal axis of a dot plot contains a scale for the quantitative variable that we want to represent. The numerical value of each measurement in the data set is located on the horizontal scale by a dot.

GROUPING ERROR:
"Grouping error" refers to the error that is introduced by the assumption that all the values falling in a class are equal to the mid-point of the class interval.

Ogive:
A graph of a cumulative distribution.

Dispersion:
The variability that exists between data set.

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

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

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

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

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

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

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.

Relative measure of dispersion:
Relative measure of dispersion is one that is expressed in the form of a ratio, co-efficient of percentage and is independent of the units of measurement.

COEFFICIENT OF QUARTILE DEVIATION:
The Coefficient of Quartile Deviation is a pure number and is used for *COMPARING* the variation in two or more sets of data.

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.

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

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

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.

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

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. OR Correlation is a technique which measures the strength of association between two variables.

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.

SAMPLE SPACE:
A set consisting of all possible outcomes that can result from a random experiment (real or conceptual), can be defined as the sample space for the experiment and is denoted by the letter S. Each possible outcome is a member of the sample space, and is called a sample point in that space.

SIMPLE & COMPOUND EVENTS:
An event that contains exactly one sample point is defined as a simple event. A compound event contains more than one sample point, and is produced by the union of simple events.

MUTUALLY EXCLUSIVE EVENTS:
Two events A and B of a single experiment are said to be mutually exclusive or disjoint if and only if they cannot both occur at the same time i.e. they have no points in common.

EXHAUSTIVE EVENTS:
Events are said to be collectively exhaustive, when the union of mutually exclusive events is equal to the entire sample space S.

EQUALLY LIKELY EVENTS:
Two events A and B are said to be equally likely, when one event is as likely to occur as the other. In other words, each event should occur in equal number in repeated trials.

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.

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

MOMENTS:
A moment designates the power to which deviations are raised before averaging them.

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

SUBJECTIVE OR PERSONALISTIC PROBABILITY:
Probability in this sense is purely subjective, and is based on whatever evidence is available to the individual. It has a disadvantage that two or more persons faced with the same evidence may arrive at different probabilities.

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

MULTIPLICATION LAW:
"The probability that two events A and B will both occur is equal to the probability that one of the events will occur multiplied by the conditional probability that the other event will occur given that the first event has already occurred."

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 occurred,

Baye's theorem:
A method used to compute posterior probabilities.

RANDOM VARIABLE:
Such a numerical quantity whose value is determined by the outcome of a random experiment is called a random variable.

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

Set:
A set is any well-defined collection or list of distinct objects, e.g. a group of students, the books in a library, the integers between 1 and 100, all human beings on the earth, etc.

SUBSET:
A set that consists of some elements of another set, is called a subset of that set. For example, if B is a subset of A, then every member of set B is also a member of set A. If B is a subset of A, we write: $(B \subset A)$

PROPER SUBSET:
If a set B contains some but not all of the elements of another set A, while A contains each element of B, then the set B is defined to be a proper subset of A.

Universal Set:
The original set of which all the sets we talk about, are subsets, is called the universal set (or the space) and is generally denoted by S or Ω.

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, named after the English logician John Venn (1834-1923).
The Venn diagrams are used to represent sets and subsets in a pictorial way and to verify the relationship among sets and subsets.

UNION OF SETS:
The union or sum of two sets A and B, denoted by $(A \cup B)$, and read as "A union B", means the set of all elements that belong to at least one of the sets A and B, that is $(A \cup B)$

INTERSECTION OF SETS:
The intersection of two sets A and B, denoted by A ∩B, and read as "A intersection B", means that the set of all elements that belong to both A and B.

OPERATIONS ON SETS:
The basic operations are union, intersection, difference and complementation.

SET DIFFERENCE:
The difference of two sets A and B, denoted by A – B or by A – (A ∩ B), is the set of all elements of A which do not belong to B.

DISJOINT SETS:
Two sets A and B are defined to be disjoint or mutually exclusive or non-overlapping when they have no elements in common, i.e. when their intersection is an empty set i.e. $A \cap B = \phi$

Conjoint Sets:
Two sets A and B are said to be conjoint when they have at least one element in common.

COMPLEMENTATION:
The particular difference S – A, that is, the set of all those elements of S which do not belong to A, is called the complement of A and is denoted by□A or by Ac.

POWER SET:
The class of ALL subsets of a set A is called the Power Set of A and is denoted by P(A). For example, if A = {H, T}, then P(A) = {φ, {H}, {T}, {H, T}}.

CARTESIAN PRODUCT OF SETS:
The Cartesian product of sets A and B, denoted by A × B, (read as "A cross B"), is a set that contains all ordered pairs (x, y), where x belongs to A and y belongs to B.
This set is also called the Cartesian set of A and B set of A and B, named after the French mathematician Rene' Descartes (1596-1605).

TREE DIAGRAM:
The "tree" is constructed from the left to the right. A "tree diagram" is a useful device for enumerating all the possible outcomes of two or more sequential events.

PERMUTATION:
A permutation is any ordered subset from a set of n distinct objects.
OR
An arrangement of all or some of a set of objects in a definite order is called permutation.

COMBINATION:
A combination is any subset of r objects, selected without regard to their order, from a set of n distinct objects.

OCCURRENCE OF AN EVENT:
An event A is said to occur if and only if the outcome of the experiment corresponds to some element of A.

PARTITION OF SETS:
A partition of a set S is a sub-division of the set into non-empty subsets that are disjoint and exhaustive, i.e. their union is the set S itself.

CLASS OF SETS:
A set of sets is called a class. For example, in a set of lines, each line is a set of points.

DISTRIBUTION FUNCTION:
The distribution function of a random variable X, denoted by F(x), is defined by F(x) = P(X < x).

FROM LECTURE 23 TO 45

Chebychev's Inequality:
If X is a random variable having mean μ and variance σ2 > 0, and k is any positive constant, then the probability that a value of X falls within k standard deviations of the mean is at least.

CONTINUOUS RANDOM VARIABLE:
A random variable X is defined to be continuous if it can assume every possible value in an interval [a, b], a < b, where a and b may be –∞ and +∞ respectively.

JOINT DISTRIBUTIONS:
The distribution of two or more random variables which are observed simultaneously when an experiment is performed is called their JOINT distribution.

BIVARIATE PROBABILITY FUNCTION:
The joint or bivariate probability distribution consisting of all pairs of values (xi, yj).

MARGINAL PROBABILITY FUNCTIONS:
The point to be understood here is that, from the joint probability function for (X, Y), we can obtain the INDIVIDUAL probability function of X and Y. Such individual probability functions are called MARGINAL probability functions.

CONDITIONAL PROBABILITY FUNCTION:
Let X and Y be two discrete r.v.'s with joint probability function f(x, y). Then the conditional probability function for X given Y = y, denoted as f(x|y).

INDEPENDENCE:
Two discrete r.v.'s X and Y are said to be statistically independent, if and only if, for all possible pairs of values (xi, yj) the joint probability function f(x, y) can be expressed as the *product* of the two marginal probability functions.

COVARIANCE OF TWO RANDOM VARIABLES:
The covariance of two r.v.'s X and Y is a numerical measure of the extent to which their values tend to increase or decrease *together*. It is denoted by σ XY or Cov (X, Y), and is defined as the expected value of the product.

BINOMIAL DISTRIBUTION:
The binomial distribution is a very important discrete probability distribution. It was discovered by James Bernoulli about the year 1700.

PROPERTIES OF A BINOMIAL EXPERIMENT:

- Every trial results in a success or a failure.
- The successive trials are independent.
- The probability of success, p, remains constant from trial to trial.
- The number of trials, n, is fixed in advanced.

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.

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.

PROPERTIES OF A BINOMIAL EXPERIMENT:

- Every item selected will either be defective (i.e. success) or not defective (i.e. failure)
- Every item drawn is independent of every other item
- The probability of obtaining a defective item i.e. 7% is the same (constant) for all items. (This probability figure is according to relative frequency definition of probability.

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

PROPERTIES OF HYPERGEOMETRIC EXPERIMENT:

- The outcomes of each trial may be classified into one of two categories, success and failure.
- The probability of success changes on each trial.
- The successive trials are not independent.
- The experiment is repeated a fixed number of times.

Hyper Geometric Distribution:

There are many experiments in which the condition of independence is violated and the probability of success does not remain constant for all trials. Such experiments are called hyper geometric experiments.

PROPERTIES OF THE HYPERGEOMETRIC DISTRIBUTION:

- If N becomes indefinitely large, the hyper geometric probability distribution tends to the BINOMIAL probability distribution.
- The above property will be best understood with reference to the following important points:
- There are two ways of drawing a sample from a population, sampling with replacement, and sampling without replacement.
- Also, a sample can be drawn from either a finite population or an infinite population.

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.

PROPERTIES OF POISSON DISTRIBUTION:

- It is a limiting approximation to the binomial distribution, when p, the probability of success is very small but n, the number of trials is so large that the product np = μ is of a moderate size;
- a distribution in its own right by considering a POISSON PROCESS where events occur randomly over a specified interval of time or space or length.

POISSON PROCESS:

It may be defined as a physical process governed at least in part by some random mechanism.

NORMAL DISTRIBUTION:

A continuous random variable is said to be normally distributed with mean μ and standard deviation σ if its probability density function is given by (Formula of Normal Distribution)

THE STANDARD NORMAL DISTRIBUTION:

A normal distribution whose mean is zero and whose standard deviation is 1 is known as the standard normal distribution.

THE PROCESS OF STANDARDIZATION:

In other words, the standardization formula converts our normal distribution to the one whose mean is 0 and whose standard deviation is equal to 1.

SAMPLING DISTRIBUTION:

The probability distribution of any statistic (such as the mean, the standard deviation, the proportion of successes in a sample, etc.) is known as its sampling distribution.

CENTRAL LIMIT THEOREM:
The theorem states that:
"If a variable X from a population has mean μ and finite variance $\sigma 2$, then the sampling distribution of the sample mean–X approaches a normal distribution with mean μ and variance $\sigma 2/n$ as the sample size n approaches infinity.**"**

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

SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION:
In this regard, the first point to be noted is that, whenever the elements of a population can be classified into two categories, technically called "success" and "failure", we may be interested in the proportion of "successes" in the population.

POINT ESTIMATION:
Point estimation of a population parameter provides as an estimate a single value calculated from the sample that is likely to be close in magnitude to the unknown parameter.

UNBIASEDNESS:
An estimator is defined to be unbiased if the statistic used as an estimator has its expected value equal to the true value of the population parameter being estimated.

CONSISTENCY:
An estimator $\hat{\theta}$ is said to be a consistent estimator of the parameter θ if, for any arbitrarily small positive quantity e.

EFFICIENCY:
An unbiased estimator is defined to be efficient if the variance of its sampling distribution is smaller than that of the sampling distribution of any other unbiased estimator of the same parameter.

METHODS OF POINTESTIMATION:
- The Method of Moments
- The Method of Least Squares
- The Method of Maximum Likelihood

These methods give estimates which may differ as the methods are based on different theories of estimation.

THE METHOD OF LEAST SQUARES:
The method of Least Squares, which is due to Gauss (1777-1855) and Markov (1856-1922), is based on the theory of linear estimation. It is regarded as one of the important methods of point estimation.

METHOD OF MAXIMUM LIKELIHOOD:
This method was introduced in 1922 by Sir Ronald A. Fisher (1890-1962).The mathematical technique of finding Maximum Likelihood Estimators is a bit *advanced*, and involves the concept of the Likelihood Function.

HYPOTHESIS-TESTING:
It is a procedure which enables us to decide on the basis of information obtained from sample data whether to accept or reject a statement or an assumption about the value of a population parameter.
NULL HYPOTHESIS:
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.

ALTERNATIVE HYPOTHESIS:
An alternative hypothesis is any other hypothesis which we are willing to accept when the null hypothesis H0 is rejected. It is customarily denoted by H1 or HA.

TYPE-I AND TYPE-II ERRORS:
On the basis of sample information, we may reject a null hypothesis H0, when it is, in fact, true or we may accept a null hypothesis H0, when it is actually false. The probability of making a Type I error is conventionally denoted by α and that of committing a Type II error is indicated by β.

TEST-STATISTIC:
A statistic (i.e. a function of the sample data not containing any parameters), which provides a basis for testing a null hypothesis, is called a test-statistic.

PROPERTIES OF STUDENT'S t-DISTRIBUTION:
i) The t-distribution is bell-shaped and symmetric about the value t = 0, ranging from $-\infty$ to ∞.
ii) The number of degrees of freedom determines the shape of the t-distribution.

PROPERTIES OF F-DISTRIBUTION:
1. The F-distribution is a continuous distribution ranging from zero to plus infinity.
2. The curve of the F-distribution is positively skewed.

ANALYSIS OF VARIANCE (ANOVA):
It is a procedure which enables us to test the hypothesis of equality of several population means.

EXPERIMENTAL DESIGN:
By an experimental design, we mean a plan used to collect the data relevant to the problem under study in such a way as to provide a basis for valid and objective inference about the stated problem. The plan usually includes:

- The selection of treatments, whose effects are to be studied,
- The specification of the experimental layout, and
- The assignment of treatments to the experimental units.

SYSTEMATIC AND RANDOMIZED DESIGNS:
In this course, we will be discussing only the randomized designs, and, in this regard, it should be noted that for the randomized designs, the analysis of the collected data is carried out through the technique known as Analysis of Variance.

THE COMPLETELY RANDOMIZED DESIGN (CR DESIGN):
A completely randomized (CR) design, which is the simplest type of the basic designs, may be defined as a design in which the treatments are assigned to experimental units completely at random, i.e. the randomization is done without any restrictions.

THE RANDOMIZED COMPLETE BLOCK DESIGN (RCB DESIGN):
A randomized complete block (RCB) design is the one in which

- The experimental material (which is not homogeneous overall) is divided into groups or blocks in such a manner that the experimental units within a particular block are relatively homogeneous.
- Each block contains a complete set of treatments, i.e., it constitutes a replication of treatments.
- The treatments are allocated at random to the experimental units within each block, which means the randomization is restricted. (A new randomization is made for every block.)The object of this type of arrangement is to bring the variability of the experimental material under control.

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)

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.

THE LEAST SIGNIFICANT DIFFERENCE (LSD) TEST:
According to this procedure, we compute the smallest difference that would be judged significant, and compare the absolute values of all differences of means with it. This smallest difference is called the least significant difference or LSD.

PROPERTIES OF THE CHI-SQUARE DISTRIBUTION:
The Chi-Square ($\chi 2$) distribution has the following properties:
1. It is a continuous distribution ranging from 0 to $+\infty$ .The number of degrees of freedom determines the shape of the chi-square distribution. (Thus, there is a different chi-square distribution for each number of degrees of freedom.
As such, it is a whole family of distributions.)
2. The curve of a chi-square distribution is positively skewed. The skewness decreases as ν increases.

ASSUMPTIONS OF THE CHI-SQUARE TEST OF GOODNESS OF FIT:
While applying the chi-square test of goodness of fit, certain requirements must be satisfied, three of which are as follows:
1. The total number of observations (i.e. the sample size) should be at least 50.
2. The expected number ei in any of the categories should not be less than 5. (So, when the expected frequency ei in any category is less than 5, we may combine this category with one or more of the other categories to get $ei \geq 5$.)
3. The observations in the sample or the frequencies of the categories should be independent.

DEGREES OF FREEDOM:
As you will recall, when discussing the t-distribution, the chi-square distribution, and the F-distribution, it was conveyed to you that the parameters that exists in the equations of those distributions are known as degrees of freedom.

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

LATEST STATISTICAL DEFINITION:
Statistics is a science of decision making for governing the state affairs. It collects analyzes, manages, monitors, interprets, evaluates and validates information. Statistics is Information Science and Information Science is Statistics. It is an applicable science as its tools are applied to all sciences including humanities and social sciences.