Probability

Introduction: Probability is a measure of how likely it is for an event to happen

Outcome: If the experiment is conducted getting result is called outcome.

Trial: the Experiment is known trial

Experiment is two types:

I. Deterministic Experiment
II. Probabilistic Experiment

Deterministic Experiment:

The outcome is unique or certain is called deterministic experiment

Probabilistic Experiment:

The result or outcome is not unique, but may be one of the several possible outcomes. This type of experiment is called probabilistic experiment.

Exhaustive Events:

The total no. of possible outcomes in any trial are known as exhaustive events or cases.

Example 1: Tossing of a coin is a trial and getting head or tail. There are two exhaustive cases.

Example 2: In throwing of a die there are six exhaustive events or cases (1, 2, 3, 4, 5, and 6)

Mutually Exclusive Events:

Events are said to be mutually exclusive. If two or more events cannot be occur simultaneously in the same trial.

Example 1: In throwing of a die all the six faces are mutually exclusive events

Example 2: In tossing of a coin the events head and tail are mutually exclusive events

Independent Events:

Several events are said to be independent if the happening of an event is not affected by supplementary knowledge concerning the occurrence of any no. of the remaining events.

Example: In tossing an unbiased coin the event of getting head in the first task is independent of getting head in the first task is independent of getting head in the second and third are independent units.

Favorable Events:

The no of cases favorable to an event in a trial is the no. of outcomes which entire the happening of the event is favorable event.

Example: In throwing of two dies the no. of cases favorable to getting the sum is 5 is (1, 4), (4, 1), (2, 3), (3, 2).

Probability:

If a trial result "n" exhaustive and mutually exclusive events and "m" of them are favorable to the happening to an event "E" then probability of happening of event is given by

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

Where m = No. of favorable cases

n = Total no. of cases.

Additional theorem or Law:

If A and B are two events then probability then probability of

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

If A, B, C is any three events then the probability of

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

Multiplication theorem or Law:

If A and B are two independent events then $\begin{bmatrix}\cup='Or'\,operator\\ \cap='And'\,operator\end{bmatrix}$

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

If A and B are two events are mutually exclusive events

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

Where $P(A \cap B) = 0$

$P(A \cup B) = P(A) + P(B) - 0$

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

Conditional Probability:

If A and B are two events, the conditional probability of B, when the event A has already happened is denoted by $\frac{B}{A}$.

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

$$P(A \cap B) = P(A) \times P\left(\frac{B}{A}\right) \text{----------} (I)$$

Similarly conditional probability occurrence of A assuming that the event B has already happened is denoted as $\frac{A}{B}$

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

$$P(A \cap B) = P(B) \times P\left(\frac{A}{B}\right) \text{----------} (II)$$

∪ - or
∩ - And

From the equations I and II

$$P(A \cap B) = P(A) \times P\left(\frac{B}{A}\right) = P(B) \times P\left(\frac{A}{B}\right)$$

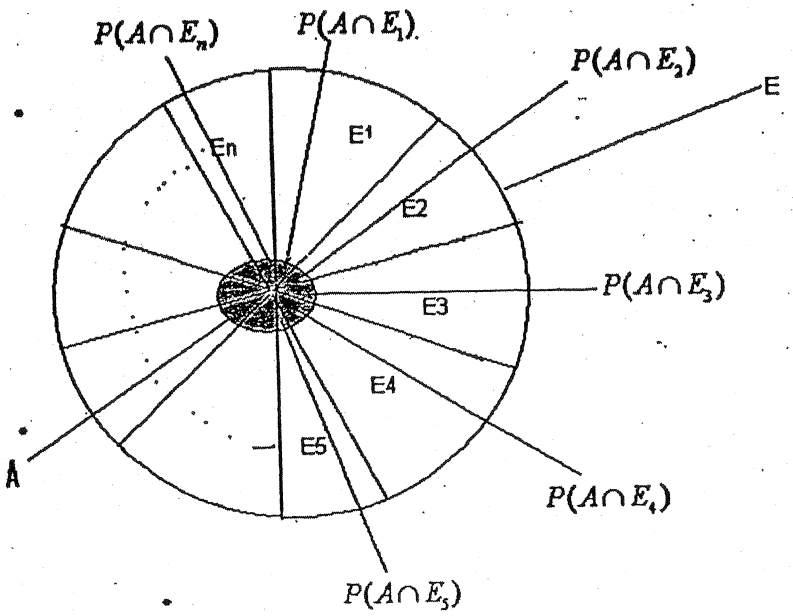
Bay's theorem or Inverse Probability theorem:

Statement:

Let $E_1, E_2, E_3, \ldots\ldots E_n$ are mutually exclusive events and $P(E) \neq 0$. Therefore any arbitrary event A which is subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$ then

$$P\left(\frac{E_i}{A}\right) = \frac{P(E_i)P\left(\frac{A}{E_i}\right)}{\sum_{i=1}^{n} P(E_i)P\left(\frac{A}{E_i}\right)}$$

Proof:



$$A = (A\cap E_1)\cup(A\cap E_2)\cup(A\cap E_3)\text{------}\cup(A\cap E_n)$$

The inner circle represents the event "A". A can occur along with the $E_1, E_2, E_3, \ldots\ldots E_n$ are mutually exclusive and exhaustive events. That is, $A\cap E_1, A\cap E_2, A\cap E_3, \ldots\ldots A\cap E_n$ are also mutually exclusive events.

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

If events are mutually exclusive $P(A\cap B)=0$

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

$$A=(A\cap E_1)\cup(A\cap E_2)\cup(A\cap E_3)-------\cup(A\cap E_n)$$

$$P(A)=P[(A\cap E_1)\cup(A\cap E_2)\cup(A\cap E_3)-------\cup(A\cap E_n)]$$

$$=P(A\cap E_1)+P(A\cap E_2)+P(A\cap E_3)-------+P(A\cap E_n)$$

$$P(A)=\sum_{i=1}^{n}P(A\cap E_i)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(I)$$

By using conditional probability

$$P\left(\frac{A}{E_i}\right)=\frac{P(A\cap E_i)}{P(E_i)}$$

$$P(A\cap E_i)=P(E_i)P\left(\frac{A}{E_i}\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(II)$$

$P(A\cap E_i)$ in second equation value substituting equation (I)

$$P(A)=\sum_{i=1}^{n}P(E_i)P\left(\frac{A}{E_i}\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(III)$$

$$P\left(\frac{E_i}{A}\right)=\frac{P(A\cap E_i)}{P(A)}$$

$$P(A\cap E_i)=P(A)P\left(\frac{E_i}{A}\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(IV)$$

From equations II and IV

$$P(A\cap B)=P(E_i)P\left(\frac{A}{E_i}\right)=P(A)P\left(\frac{E_i}{A}\right)$$

$$P(E_i)P\left(\frac{A}{E_i}\right)=P(A)P\left(\frac{E_i}{A}\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(V)$$

$$P(E_i)P\left(\frac{A}{E_i}\right)=\left[\sum_{i=1}^{n}P(E_i)P\left(\frac{A}{E_i}\right)\right]P\left(\frac{E_i}{A}\right)$$

$$P\left(\frac{E_i}{A}\right)=\frac{P(E_i)P\left(\frac{A}{E_i}\right)}{\sum_{i=1}^{n}P(E_i)P\left(\frac{A}{E_i}\right)}$$

Discrete Distributions

Binomial Distribution:

Definition: Let X be a discrete random variable it assume only Non – negative values. Then the probability mass function is defined by

$$P(X=x)={}^{n}C_{r}p^{r}q^{n-r} \qquad X=1,2,3.....n$$

Where n = No. of trials

x = No. of success

p = Probability of Success

q = Probability of failure

Conditions/Assumptions/ Rules:

1. The no. of trials must be fixed that "n".
2. Probability of success (p) value must be constant.
3. The no. of trials must be independent to each other

Properties:

1. Binomial distribution mean is 'np'
2. Binomial distribution variance is 'npq'
3. Binomial distribution standard deviation is '

Biased – coin or coin is unknown

If biased coin is tossed then the probability of getting head is not equal to probability of getting tail.

Unbiased – coin or coin is known:

If an unbiased coin is tossed then the probability of getting head is equal to probability of getting tail.

Poisson distribution:

If X is a discrete random variable and is set to follow Poisson distribution it assume only positive values or non – negative values than the probability mass function is given by

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

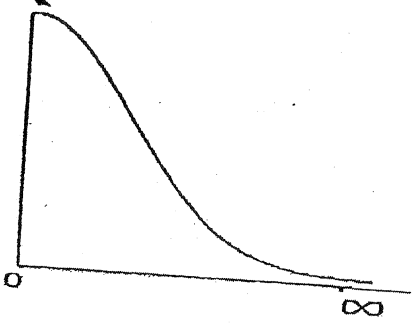
Assumptions:

1. The no. Of trials is infinite or large. i.e.,
2. Constant probability of success is small i.e.,
3. np (Binomial mean or parameter or average) = is passion distribution mean or parameter.

Properties:

1. Poisson distribution Mean is
2. Poisson distribution variance is
3. Mean and variance of Poisson distribution is same then
4. Standard deviation of Poisson distribution is

Poisson distribution curve is positively skewed



Normal Distribution:

If X is continuous random variable and is said to follow normal distribution then the probability density function is given by

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

$-\infty < x < \infty$

$-\infty < \mu < \infty$

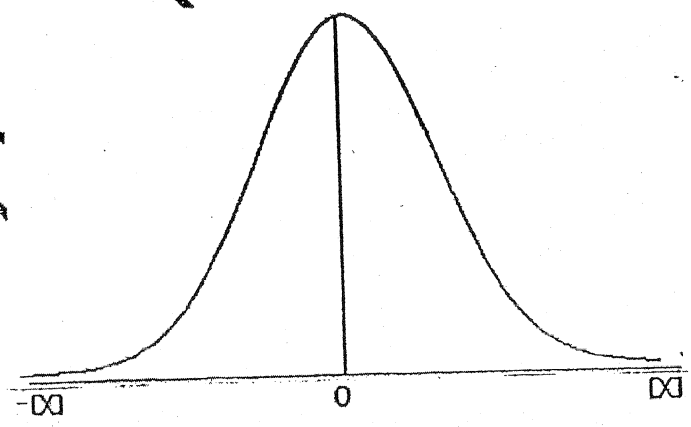
$0 < \sigma < \infty$

Properties/Assumptions:

1. The random variable is must be continuous
2. Normal distribution mean is μ
3. Normal distribution variance
4. Normal distribution standard deviation
5. Mean and variance of the normal distribution are called parameters.
6. If X follows standard normal distribution then probability dencity function is given by Where 'Z' is the standard normal variation

Then $z=\frac{x-\mu}{\sigma}$

Normal curve is symmetrical curve.



Importance:

Normal distribution plays a very important role in statistical theory.

1. Most of the distributions (Binomial, Poisson etc) can be approximated by normal distribution.
2. Many of the sampling distributions (chi-square, t, F distributions) tends to normal distributions for large sample theory.
3. Normal distribution finds the large applications in statistical quality control in industry for setting control limits.

Exponential Distribution:

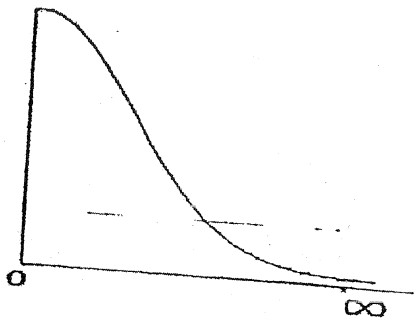
A continuous distribution variable 'X' is said to follow exponential distribution then the probability distribution then the probability density function is given by

$$f(x) = \lambda e^{-\lambda x}$$

$$0 \le x \le \infty$$

Assumptions:

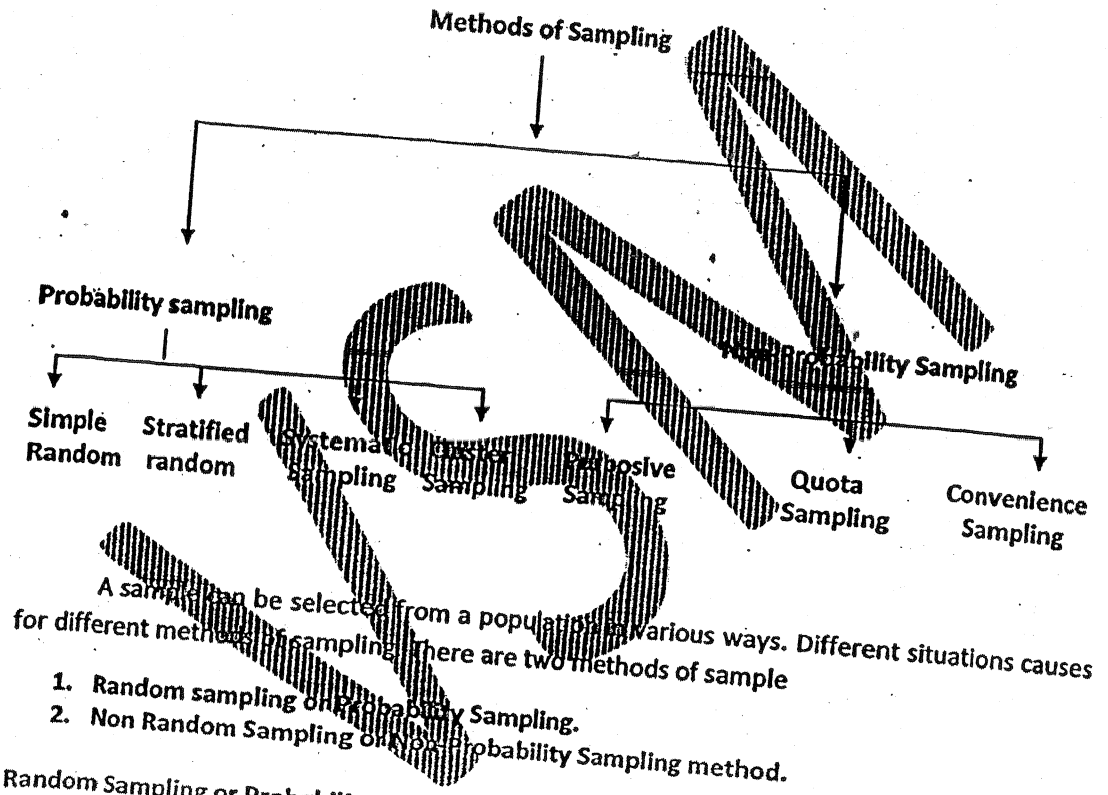
1. Exponential distribution Mean is $\frac{1}{\lambda}$ then exponential distribution variance is $\frac{1}{\lambda^2}$. Then exponential distribution standard deviation is $\frac{1}{\lambda}$
2. Mean and standard deviation of the exponential distribution are the same.
3. It assume only non – negative or positive values
4. Exponential distribution curve is right side curve or positively skewed.



Sampling

Sample: Any part of the population is called sample. or set of data drawn from the population is called sample.

Ex: if we want to purchase a bag of rice. The entire bag of rice is called population and the handful rice which we will test is called sample.



A sample can be selected from a population in various ways. Different situations causes for different methods of sampling. There are two methods of sample

1. Random sampling or Probability Sampling.
2. Non Random Sampling or Non Probability Sampling method.

Random Sampling or Probability sampling method:

Random or probability sampling is the scientific technique of drawing samples from the population according to some laws of chance in which each unit in the universe or population has some definite pre-assigned probability of being selected in the sample. It is of two types.

i) Un-Restricted Sampling
ii) Restricted Sampling

Un-Restricted Sampling:

Simple Random sampling:

It is the method of selection of a sample in such a way that each and every member of population or universe has an equal chance or probability of being included in the sample.

It is two types of Methods:

1. Lottery Method
2. Random numbers method

Lottery Method

It is the simplest, most common and important method of obtaining a random sample. Under this method all the members of population or universe are serially numbered on small slips of a paper which are put in drum and thoroughly mixed by vibrating the drawn. After mixing, the numbered slips are drawn out of the drum one by one according to the sample size. The number of slips so drawn constitutes a random sample.

Merits:

- All the units in the universe have an equal chance of being selected
- It is simple and easy to apply
- It is representative
- It is free from bias
- It does not require prior knowledge of the true composition of the universe sampling errors are easily assessed.

Restricted Sampling methods

It is three types:

1. Stratified sampling.
2. Systematic sampling
3. Cluster sampling.

Stratified sampling

In stratified random sampling the population is divided into strata (groups) before the sample is drawn. Under stratified sampling the population is divided into several sub-populations that are individually more homogeneous than the total population (the different sub-populations are called 'strata') and then we select items from each stratum to constitute a sample. Stratified sampling is suitable in those cases where the population is heterogeneous but there is homogeneity within each of the groups of the strata.

Advantages:

I. It is a representative sample of the heterogeneous population.
II. It gives higher statistical efficiency
III. It is a self-weighting sample
IV. Population mean can be estimated by calculating sample mean

Disadvantages:

I. It may be difficult to divide the population into heterogeneous groups
II. There may be over lapping of different strata of the population which will provide an unrepresentative.

Systematic Sampling:

In this method every elementary unit of the population it is arranged in order and the sample units are distributed equally at regular intervals.

There are 3 steps:

1. Sampling interval K is determined

$$K = \frac{\text{No. of units in the population}}{\text{No. of units desired in the sample}}$$

2. One unit between the first and K^{th} in the population list is randomly chosen.
3. Add K^{th} unit to the randomly chosen number.

Example: Consider 1000 households, from which we want to select 50 units. To select the first unit, we randomly pick one number between 1 to 20 say 17. So our sample is starting with 17, 37, 57..... Please note that only first item was randomly selected. The rest are systematically selected. This is a very popular method because; we need only one random number.

Advantages:

- It is most suitable where the population units are serially numbered or serially arranged.
- It requires less time
- It is cheaper than simple random sampling
- It is easy to check whether every k^{th} unit is included in the sample
- It is statistically more efficient then simple random sampling

Disadvantage: It may not provide a desirable result due to large variation in the items selected.

Cluster Sampling:

Cluster means group. In Cluster sampling, homogeneous population is divided into heterogeneous groups. Those groups are called clusters and then some these clusters are randomly selected for inclusion in overall sample.

Suppose, a researcher wants to select a random sample of 1200 households out of 60000 estimated households in city for a survey. A direct sample of individual households would be difficult to select because a list of household does not exist and would be too costly to prepare. Instead, he can select a random sample of areas, wards. The number of wards to be selected depends on the average number of estimated households per ward. Suppose, the average number of households per ward is 200. Then six wards comprise the sample size of 1200. The application of cluster sampling in social science research, demographic studies, large scale surveys of political and social behavior, attitude survey.

Merits

- Cluster sampling is much easier and much easy to apply. It is widely applied when population is large.
- Its cost is much less compared to other sampling methods.
- It promotes convenience of field word as it could be done in compact places.

Non-Random Sampling Or Non-Probability Sampling method

In this type of sampling, items for the sample are selected deliberately by the researcher; his choice concerning the items remains supreme.

I. Purposive sampling
II. Quota sampling
III. Convenience sampling

Restricted Sampling

Purposive sampling:

Purposive sampling is the method of sampling by which a sample is drawn from population based entirely on the personal judgment of the investigator. It is also known as judgment sampling or deliberate sampling. Randomness finds no place in it and so the sample drawn under this method cannot be subjected to mathematical concepts used in computing sampling error.

Quota sampling

In quota sampling method, quotas are fixed according to the basic parameters of the population determined earlier and each field investigator is assigned with quotas of no. of elementary units to interview.

Suppose 2, 00,000 students are appeared for a competitive examination and we need to Select 1% of them based on quota sampling. The classification of quota may be as follows

Example: Classification of samples

Category	Quota
General	1000
Sport	600
NRI	100
SC/ST	300
Total	2000

Un-Restricted Sampling

Convenience or Accidental Sampling:

The convenience sampling is a non-probability sampling and selecting sample units based on just 'hit and miss' fashion i.e., interviewing people whatever sampling units that are conveniently available. This method is also called accidentally are included in the sample. If a person is to submit a project report on labor-management relations in textile industry and he takes a textile mill close to his office and interviews some people over there, he is following the convenience sampling method.

Advantages of Random or Probability Sampling

1. **Random sampling's objective is unbiased. As a result it is defensible before the superiors or even before the court of law.**
2. **The size of sample depends on demonstrable statistical method and therefore, it has justification for the expenditure.**
3. **It provides a more accurate method of drawing conclusions about characteristics of the population as the parameters.**
4. **It is used to draw the statistical inference.**
5. **The samples may be combined and evaluated even though accomplished by different individuals.**

ESTIMATION

Definition: When the data are collected by sampling from a population, the most important objective of statistical analysis is to draw inferences or generalization about that population from the information embodied in the sample. Statistical estimation, or briefly estimation is concerned with the methods by which population characteristics are estimated form sample information.

With respect to estimating a parameter, the following two types of estimates are possible:

- Point estimation
- Interval estimation

Point estimation

The point estimation in a single number which is used as an estimate of the unknown population parameter. The procedure in point estimation is to select a random sample of 'n' observations $X_1, X_2, X_3 \ldots\ldots X_n$ from a population f (X, θ) and then to use some preconceived method to arrive from these observations at a number say $\hat{\theta}$ (read theta hat)which we accept

as an estimator of θ. the estimator $\hat{\theta}$ is a single point on the real number scale on thus the name point estimation, $\hat{\theta}$ depends on the random variables that generate the sample and hence, it too is a random variable with its own sampling distribution.

{Notes: The symbol θ is generally used to denote a parameter that could be a mean, median or some measure of variability, etc.}

Interval estimation or confidence level:

As distinguished form a point estimate which provides one single value of the parameter. An interval estimate of a population parameter is a statement of two values between which is estimated that the parameter lies. An interval estimate would always be specified by two values, i.e., the lower one and the upper one. In more technical terms, interval estimation refers to the estimations of a parameter by a random interval called the confidence interval, whose end points L and U with L < U, are functions of the observed random variables such that the probability that the inequality $L<\theta<U$ is satisfied in terms of predetermined number. L and U are called the confidence limits and are the random end points of interval estimate.

If we estimate the average income of the people living in a village as Rs.875 it will be a point estimate the average income of the could lie between Rs.800 and Rs.950, it will be an interval estimate.

On comparing these two methods of estimation we find that point estimation has an advantage as much as it provides and exact value for the parameter under investigation.

Properties of a good estimator:

A distinction is made between an estimate and an estimator. The numerical value of the sample mean is said to be an estimate of the population mean figure, for example, the sample mean $\bar{x}$ is an estimator of the population mean.

A good estimator, as common sense dictates, is close to the parameter being estimated. Its quality is to evaluated in terms of the following properties.

1. Unbiasedness:

An estimator is said to be unbiased if its expected value is identified with the population parameter being estimated. That is if $\hat{\theta}$ is an unbiased estimate of θ, then we must have

$E(\hat{\theta})=\theta$ many estimators are "Asymptomatically Unbiased" in the sense of the biases reduce to practically insignificant values zero when 'n' becomes sufficiently large. The estimator s^2 is an example.

2. Consistency:

If an estimator, say $\hat{\theta}$, approaches the parameter θ closer and closer as the sample size 'n' increases, $\hat{\theta}$ is said to be a consistent estimator of θ. Stating somewhat more rigorously. The estimator $\hat{\theta}$ is said to be a consistent estimator of θ if as 'n' approaches infinity, the probability approaches 1 that $\hat{\theta}$ will differ from the parameter θ by not more than an arbitrary small constant.

3. Efficiency:

The concept of efficiency refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with the smaller variance (for, a given sample size) is said to be relatively more efficient. Stated in a somewhat different language, estimator $\hat{\theta}_1$ is said to be more efficient than another estimator $\hat{\theta}_2$ for if the variance of the estimator, the more concentrated is the distribution of the estimator around the parameter being estimated.

4. Sufficiency:

An estimator is said to be sufficient if it contains as much information as is possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that if a sufficient estimator exits, it is absolutely unnecessary to consider any other estimator: a sufficient estimator ensures that all information a sample can furnish with respect to the estimation of a parameter is utilized.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties.

Matrices

Definition: A matrix is defined as a rectangular array of numbers of symbols arranged by brackets. Now days, it plays major role in modern mathematics, having wide applications. Matrix is more useful for practical business problems and therefore, they form an important part of business mathematics. The matrix form suits very well for game theory, linear programming, budgeting for by – products etc.

Types of Matrix:

Addition of two Matrixes:

If A and B are two matrixes are some order in then the matrices the matrices addition can be defined as adding the corresponding element. For the given example is

$$A=\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix};B=\begin{bmatrix}2 & 3\\ 3 & 1\end{bmatrix}$$

$$A+B=\begin{bmatrix}1+2 & 2+3\\ 2+3 & 1+1\end{bmatrix}$$

$$A+B=\begin{bmatrix}3 & 5\\ 5 & 2\end{bmatrix}$$

Properties of Matrix Addition:

1. Matrix addition is commutative, if A+B = B+A
2. Matrix addition is associative, if (A+B)+C = A +(B+C)
3. For a matrix of A of dimension m×n, if exists another matrix B of the same dimension, such A+B = B+A = 0. Then B is known as additive inverse or negative of A and is denoted by -A

Matrix Multiplication:

If A and B are any two matrixes then the matrixes multiplication can be defined as no. of columns in first matrix equal to no. of rows in second matrix. For the given example is

$$A=\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix};B=\begin{bmatrix}2 & 3\\ 3 & 1\end{bmatrix}$$

$$AB=\begin{bmatrix}1\times2+2\times3 & 1\times3+2\times1\\ 2\times2+1\times3 & 2\times3+1\times1\end{bmatrix}$$

$$=\begin{bmatrix}8 & 5\\ 7 & 7\end{bmatrix}$$

Properties of Matrix Multiplication:

- All the matrices cannot be multiplied by each other. Matrix multiplication is not Commutative, i.e., $AB \neq BA$.
- Matrix multiplication is associative, i.e., $A(BC) = (AB)C$.
- Matrix Multiplication is distributive i.e., $A(B+C) = AB+AC$

Row Matrix:

A matrix which has exactly one row is called row matrix.

Eg: $[2 \quad 3 \quad 4]$

Columns Matrix:

A matrix which has exactly one column is called column matrix.

Eg: $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$

Square Matrix:

A matrix in which the no. of rows is equal to no. of columns is called square matrix.

Eg: $A = \begin{bmatrix} 1 & 3 \\ 2 & 3 \end{bmatrix}_{2\times 2}$

Null or Zero Matrix:

A matrix which all the elements is zero called null or zero matrix.

Eg: $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}_{2\times 2}$

Diagonal Matrix:

A square matrix in which all non-diagonal elements are zero is called a diagonal matrix.

Eg: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

Scalar Matrix:

A diagonal matrix whose diagonal elements are equal called a scalar matrix.

Eg: $\begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$

Identity Matrix:

A diagonal matrix whose diagonal elements are equal to '1' is called identity Matrix.

$$I=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}_{2\times 2} \text{ Or } \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}_{3\times 3}$$

Triangular Matrix:

A square matrices $A = (a_{ij})_{m\times n}$ whose elements $a_{ij} = 0$ for $I > j$ is called upper triangular matrix.

Eg: $\begin{bmatrix}5 & 2 & 3\\ 0 & 4 & 2\\ 0 & 0 & 3\end{bmatrix}$

A square matrices $A = (a_{ij})_{m\times n}$ whose elements $a_{ij} = 0$ for $I < j$ is called lower triangular matrix.

Eg: $\begin{bmatrix}5 & 0 & 0\\ 6 & 4 & 0\\ 2 & 1 & 3\end{bmatrix}$.

Transpose Matrix:

A matrix obtained by interchanging rows and columns of a matrix 'A' is called the transpose of 'A' and is denoted by [illegible] (read as transpose).

Inverse Matrix:

The concept of inverse matrix is very useful in solving simultaneous equations, input-output analysis and regression analysis. There are methods of finding the inverse of matrix.

- Using ad joint matrix (cofactor method)
- Gauss elimination method

Cramer's rule:

We can find the solution of a system of linear equations where the number of equations is equal to the number of variables with a simple rule using determinants or Cramer's rule.

Set:

A set is defined as a collection of objective or elements. No attempt is made to define an element. It is like a point in a plane. But one can tell whether or not an element belongs to a given set. A set can be illustrated from the following example.

Examples:

- The set of all possible outcomes in throwing a dice.
- The set of all integers from 1 to 10.
- The set of all commerce students in Andhra University.
- The set of all vowels in English alphabets.

Progressions

Progressions: If we write down a succession of numbers which follow a certain pattern, then such a succession is called a progression.

Arithmetic Progressions: Let us consider the following succession of numbers

3,8, 13,18,23,28,33,38, ...

From this succession of numbers we can guess the pattern from which these numbers are written. It is clear that the difference between any two consecutive numbers is the same.

For example: 8-3 = 5, 13-8=5, 18-13=5 and so on. Such a succession is called an arithmetical progression.

Geometric Progressions: Now let us consider the following succession of numbers

3, 6, 12, 24, 48, 96, 192 ...

From this succession of numbers we can guess the pattern from which they are written. It is clear that the ratio of any two consecutive numbers is the same. For example

$\frac{6}{3}=2, \frac{12}{6}=2, \frac{24}{12}=2$ and so on. Such a succession is called a geometric progression.

Harmonic Progressions: Let us consider the following succession of numbers.

$\frac{1}{2}, \frac{1}{5}, \frac{1}{8}, \frac{1}{11}, \ldots\ldots\ldots$

From this succession also we can guess the pattern from which these numbers are written. It is clear that the numerator is fixed and that denominators of all those numbers follow a certain pattern. In fact the difference between any two consecutive denominators is always the same.

For example: 5-2 = 3, 8-5=3, 11-8 = 3 and so on . Such successions is called a Harmonic progressions

Applications of Baye's theorem:

The Baye's theorem is useful in revising the original probability estimates of known outcomes as we gain additional information about these outcomes. The prior probabilities when changed in the light of new information are called revised or posterior probabilities.

Suppose A_1, A_2, A_3,.....A_n represent 'n' mutually exclusive and exhaustive events with prior marginal probabilities $P(A_1), P(A_2), P(A_3), \ldots\ldots\ldots P(A_n)$. Let 'B' be an arbitrary event with P (B) ≠ 0 for which conditional probabilities $P\left(\frac{B}{A_1}\right), P\left(\frac{B}{A_2}\right), P\left(\frac{B}{A_3}\right), \ldots\ldots\ldots P\left(\frac{B}{A_n}\right)$ are also known. Given the information that outcome B has occurred the revised (or posterior) probabilities with help of Baye's theorem using the formula

$$P\left(\frac{A_i}{B}\right) = \frac{P\left(\frac{B}{A_i}\right)P(A_i)}{\sum_{i=1}^{n} P\left(\frac{B}{A_i}\right)P(A_i)}$$

FUNCTIONS

In economics it is quite often to find related variables where the changes in one variable depend on the change in the other variable.

For example, supply and demand; price and demand; and advertisement and sales are all related pairs of variables. The way in which one variable depends on other variable is described by means of 'functions' which provide all mathematics and its applications.

The demand for a commodity may depend on its price and the supply. General if there are n independent variables $x_1, x_2, x_3, x_4, \ldots\ldots$ X_n and y is the dependent variable, then we write.

$$y = f(x_1, x_2, x_3, x_4, \ldots\ldots, x_n).$$

If x is the production of a commodity and R, K, and L a the raw materials, capital and labors respectively then the production function is written as

$$x = f(R, K, L).$$

Types of functions:

Polynomial function:

$$f(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots\ldots\ldots + a_{n-1} x + a_n.$$

where n is positive integer and $a_0, a_1, a_2, \ldots\ldots\ldots\ldots a_{n-1}, a_n$ are all constant is defined as polynomial function

Constant function:

A zero degree polynomial function is defined as a constant function

For example $f(x) = 5$

$f(x) = k$ where k is constant are all constant functions.

Linear function:

A polynomial function of degree one is a linear function. Other words, $y = f(x) = a x + b$ is linear function where a and b are constants.

Quadratic functions:

Second degree polynomial function is called quadratic function. In other words, $y = f(x) = ax + bx + c$ is a quadratic function where a, b, c are all constants.

The graph of quadratic function is called parabola.

Logarithmic and Exponential functions:

In the equation

$$e^y = x,$$

y is called the logarithm of x to base a.

we write this equation in an alternative form

$$y = \log^e x,\ x > 0,\ e > 0,\ e \neq 1.$$

This is called the logarithmic functions

Exponential functions:

An exponential function is of the form,

$x = e^y$, $e > 0$ where e is the base

it should be noted that an exponential function is different from linear function like $y = ex$ or $x = Ky$ and power functions like $y = x^e$ or

$x = y^k$

24

QT

Testing of Hypothesis

Population: The aggregate of all units pertaining to a study is called population or universe.

Sample: Set of data drawn from the population is called sample. The process of selection a sample from the population is called sampling.

Example: Suppose there are 3000 students in a college and 250 students are selected in order to estimate the average height of students.
This number of 250 students constitutes a sample and the total number of 3000 students is population.

Population size (N) is finite or sometimes infinite and sample size (n) is always finite

Parameter: Population constants [Population Mean μ and population variance "σ^2"] are called parameters.

Statistic: Sample constant [sample Mean $\bar{x}$ and sample variance s^2] are called statistic

In practice parameter values are not known and the estimates based on the sample values are generally used. Thus, statistic which may be recorded as an estimate of parameter obtained from the sample.

Sampling distribution of a statistics

If we draw sample of size 'n' from a given finite population of size 'N' then the total no. of possible samples is called sampling distribution.

Example:

$$^{N}c_{n} = \frac{N!}{(N-n)!n!} = k$$

$$^{4}c_{2} = \frac{4!}{(4-2)!2!}$$

$$= \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1}$$

$$^{4}c_{2} = 6$$

$$(1,2,3,4) = (1,2),(1,3),(1,4),(2,3),(2,4),(3,4)$$

Test of Hypothesis or Test of significance: A very important aspect of the sampling theory is the study of test of significance which enables us to decide on the basis of the sample results. The deviation between the observed sample statistic and hypothetical parameter value.

A test of statistical hypothesis is a two action decision problem after the experimental sample values have been obtained the two actions being acceptance (or) rejection of hypothesis under consideration.

Testing of hypothesis are two types:

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)

Null hypothesis (H_0)

It is usually a hypothesis of no difference is called null hypothesis. It is usually denoted by "H_0". It should be completely impartial and should have no brief for any party or company nor should be allow his personal views to utilize the decision.

Example: Let us consider the light bulbs problem. Suppose that the bulbs manufactured under some standard manufacturing process have an average life of " hours and is proposed to test a new procedure "" for manufacturing light bulbs. Thus, we have two populations of bulbs those manufacture by standard process and those manufacture by new process.

In this problem the following three hypotheses may be set up

1. Standard process is greater than new process.
2. Standard process is less than to new process
3. There is no difference between standard process and new process.

Null hypothesis (H_0): There is no difference between new process and standard process.

Alternative hypothesis (H_1):

Any hypothesis which is complimentary to the null hypothesis is called alternative hypothesis, which is denoted by H_1.

Example: Above light bulbs alternative hypothesis (H_1) is: New process is better than standard process (or) new process is inferior to standard process.

Let us, suppose that the bulbs manufactured under some standard manufacturing process have an average life of 'μ_1' hours. If 'μ_2' is the mean life of the bulbs manufactured by the new process.

One Tailed and Two Tailed test:

One Tailed test: A test of any statistical hypothesis where the alternative hypothesis is one tailed [left tailed or right tailed] is called as one tailed test.

Two Tailed test: A test of any statistical hypothesis where the alternative hypothesis two tailed test such as null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative hypothesis is $H_0 : \mu_1 \neq \mu_2$ is known as Two Tailed test.

Type – I Error (α): Reject Null hypothesis (H_0) when it is true. Probability of Type –I error is

Probability of type –I error $P(Type - I) = \alpha$

Type –II Error (β): Accept Null hypothesis (H_1) when it is wrong.
Probability of type –II error $P(Type - II) = \beta$

Procedure for Testing of Hypothesis:

We know summarize below the various steps in testing of statistical hypothesis in a systematic manner.

1. Null Hypothesis: Set up the Null Hypothesis (H_0)
2. Alternative Hypothesis: Set up the Alternative Hypothesis (H_1). This will enable us to decide whether we have to use a single – tailed (right or left tailed) test or two – test .
3. Level of Significance: Choose the appropriate levee of Significance (α) depending on reliability of the estimates and permissible risk. This is to be decide before sample is drawn, i.e., α is fixed in advance.
4. Test Statistic (or Test criterion): Compute the test statistic

Under the null hypothesis

$$Z = \frac{t - E(t)}{S.E(t)}$$

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of Z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Small Sample Tests

The sample size is less than 30 (n<30) this type of sample are called small samples.

In this section, we will study the following test.

- t –test for single mean
- t-test for difference of two mean (unpaired t -test)
- Paired t –test
- t –test for single correlation
- F-test for single variance
- χ^2 - test for single variance
- χ^2 -test for goodness of fit.

t –test

Let X_i (I =1, 2, 3, 4 ...n) be a random sample of size 'n' drawn from the normal population with mean μ and variance σ^2 the t –test is defined as

$$t = \frac{|\bar{x} - \mu|}{\frac{S}{\sqrt{n}}}$$

Where $\bar{x}$ = Sample mean
μ = Population mean
S = Standard deviation

$$S = \sqrt{\frac{\sum (x - \bar{x})}{n - 1}}$$

n = number of samples or observations

Applications of t –test:

The t – test was wide no .of applications

- To test the significance difference between sample and population mean
- To test the significance of the difference between two means.
- To test the significance of an observed sample correlation co – efficient and population correlation co-efficient.

Test for Single Mean:

If a random sample of size 'n' has been drawn from the normal population with specified mean μ_0

Now under the null hypothesis H_0, the test statistic t is given by

$$t = \frac{|\bar{x} - \mu|}{\frac{S}{\sqrt{n}}} \text{ follows (n-1) degrees of freedom}$$

Where $\bar{x}$ = Sample mean
μ = Population mean
S = Standard deviation

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

Conclusion: Now we compare the calculated value and tabulated value.
If calculated value of t is less than tabulated value of t. then we accept null hypothesis (H_0) at certain level of significance.
If calculated value of t is greater than tabulated value of t. then we reject null hypothesis (H_0) at certain level of significance.

Confidence interval for population mean μ formulae is $\bar{x} \pm (t_{tab}) \frac{S}{\sqrt{n}}$

95% Confidence limits for the population mean μ are $\bar{x} \pm (t_{0.05}) \frac{S}{\sqrt{n}}$

99% Confidence limits for the population mean μ are $\bar{x} \pm (t_{0.01}) \frac{S}{\sqrt{n}}$

90% Confidence limits for the population mean μ are $\bar{x} \pm (t_{0.10}) \frac{S}{\sqrt{n}}$

Assumptions:

- Sample observations have drawn from the normal populations
- Sample observations are independent
- The population standard deviation is unknown.

Test for Difference of Means (unpaired t –test)

Suppose we want to test if the two independent samples x and y of sizes n_1 and n_2 have been drawn from the same population.

Null hypothesis:

There is no significance difference between two means. Now under the null hypothesis H_0 the test statistic t is given by.

$$t = \frac{|\bar{x} - \bar{y}|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ follows } (n_1+n_2-1) \text{ degrees of freedom}$$

Where n_1 = no. of observations of first sample.

n_2 = no. of observations of second sample.

$\bar{x}$ = First sample mean

$\bar{y}$ = Second sample mean

S = Standard deviations of two samples.

$$S = \sqrt{\frac{\sum (x-\bar{x})^2 - \sum (y-\bar{y})^2}{n_1 + n_2 - 2}}$$

Conclusion: Now we compare the calculated value and tabulated value.

If calculated value of t is less than tabulated value of t. then we accept null hypothesis (H_0) at certain level of significance.

If calculated value of t is greater than tabulated value of t. then we reject null hypothesis (H_0) at certain level of significance.

Confidence interval for $|\mu_1 - \mu_2|$ i.e., for the difference in the two means of independent populations formulae is $|\bar{x} - \bar{y}| \pm (t_{tab}) S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

95% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x} - \bar{y}| \pm (t_{0.05}) S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

99% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x} - \bar{y}| \pm (t_{0.01}) S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

90% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x} - \bar{y}| \pm (t_{0.10}) S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Paired t – test for difference of two means

Suppose two samples are not independent but sample observations are paired ($n_1 = n_2$) together. The pair of observations (x_i, y_i) i = 1, 2, 3 ....n corresponds to the i^{th} sample unit. To test if the sample means differs significantly or not.

For example: if we want to test the efficiency of a particular drug say for inducing sleep. Let (x_i, y_i) be the readings in hours of sleep of the i^{th} individual before and after the drug is given respectively. Here instead of applying t –test for difference of two means, we apply the paired t – test

Null hypothesis: There is no significance difference between two means

Now under the null hypothesis (H_0), the test statistic t is given by

$$t = \frac{|\bar{d}|}{\frac{S}{\sqrt{n}}} \text{ follows (n-1) degrees of freedom}$$

Where $d = y - x$

n = no. of observations ($n = n_1 = n_2$).

$$S = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}}$$

Conclusion: Now we compare the calculated value and tabulated value.
If calculated value of t is less than tabulated value of t. then we accept null hypothesis (H_0) at certain level of significance.
If calculated value of t is greater than tabulated value of t. then we reject null hypothesis (H_0) at certain level of significance.

Confidence interval for difference between the two means formulae is $|\bar{d}| \pm (t_{\alpha h}) \frac{S}{\sqrt{n}}$

95% Confidence limits for difference between the two means are $|\bar{d}| \pm (t_{0.05}) \frac{S}{\sqrt{n}}$

99% Confidence limits for difference between the two means are $|\bar{d}| \pm (t_{0.01}) \frac{S}{\sqrt{n}}$

90% Confidence limits for difference between the two means are $|\bar{d}| \pm (t_{0.10}) \frac{S}{\sqrt{n}}$

Degrees of Freedom:

Degree of freedom is no .of observations – no of independent constraints. It is used to find tabulated value for Small sample tests (χ^2 - test, t – test and F – test).

Chi – Square (χ^2) – test

A very powerful test for testing the significance between the theory and experiment. It was first discovered by Karl Pearson.

If O_i (i = 1, 2, 3, ... n) is a set of observed frequencies and if E_i (1 = 1, 2, 3, ...n) is a set of expected frequencies then chi – square (χ^2) is given by

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right] \text{ follows (n-1) degrees of freedom}$$

Conditions or Properties:

- The sample observations should be independent
- Sum of observed frequencies = sum of expected frequency ($\sum O_i = \sum E_i$)
- The total frequency (N) should be greater than 50.

- Theoretical cell frequency should be greater than 5

  If any theoretical cell frequency is less than 5 then for the application of chi – square test it is pooled with preceding or succeeding frequency so that the pooled frequency more than 5 and exist for the degrees of freedom last in pooling.

Analysis of Variance (ANOVA)

Analysis of variance is a powerful statistical tool for significantly the test based on t – test is an adequate procedure only for testing the significance between the sample means.

In a situation when we have two or more samples to consider at a time an alternative procedure is called analysis of variance.

Eg: Suppose five fertilizers are applied at random to four plots each in a field consists of 20 plots of the same shape and same size and the yield of wheat on each to these plots is given we may be interested to finding out whether the affect of these fertilizers the yields is significantly different or in other words.

The answer of this problem is providing by the technique of analysis of variance (ANOVA) is to test the homogeneity of the several means (more than two means).

Assumptions for ANOVA test:

ANOVA test is based on the test statistics F for the validity of the F test in ANOVA the following assumptions are.

- The sample observations are independent
- Various treatments and environment effects or additive in nature.
- The sample have been drawn from the normal population

In the following selections we will discuss the analysis of variance

- One way classification
- Two way classification.

Large Samples

If the sample size in greater than or equal to 30 ($n \geq 30$). Then it is called a "Large Sample".

In this section, we will study the following tests which are based upon on a large sample

- Test for single mean
- Test for two means
- Test for single proportion
- Test for two proportion
- Test for two standard deviations
- Test for single correlation
- Test for two correlation

Test for Single Mean

Let Xi (i = 1, 2, 3... n) be a random sample size 'n' drawn form a normal population with mean μ

Null hypothesis: There is no significance difference between the sample mean and the population mean.

Now under null hypothesis (H_0), the test statistic is

$$Z = \frac{|\bar{x} - \mu|}{\frac{\sigma}{\sqrt{n}}}$$ (If standard deviation is known)

Where $\bar{x}$ = Sample mean
μ = Population mean
σ = Standard deviation

$$Z = \frac{|\bar{x} - \mu|}{\frac{s}{\sqrt{n}}}$$ (If standard deviation is unknown)

Where s = Sample.Standard deviation.

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for population mean μ formulae is $\bar{x} \pm (Z_{tab}) \frac{\sigma}{\sqrt{n}}$

95% Confidence limits for the population mean μ are $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

99% Confidence limits for the population mean μ are $\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$

90% Confidence limits for the population mean μ are $\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$

Test for two Means or Test of significance for difference of two Means

Let $\bar{x}_1$ be the mean of a sample of size n_1 drawn from a population with mean μ_1 and variance σ_1^2 and $\bar{x}_2$ be the mean of another independent sample of size n_2 drawn from another population with mean μ_2 and variance σ_2^2.

Null hypothesis (H_0): There is no significance difference between two population means.

Now under null hypothesis H_0, the test statistic is

$$Z = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where n_1 = no. of observations of first sample.

n_2 = no. of observations of second sample.

$\bar{x}_1$ = First sample mean

$\bar{x}_2$ = Second sample mean

σ_1 = Standard deviations of first sample

σ_2 = Standard deviation of second sample

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for $|\mu_1 - \mu_2|$ i.e., for the difference in the two means of populations formulae is $|\bar{x}_1 - \bar{x}_2| \pm (Z_{\alpha\delta}) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

95% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x}_1 - \bar{x}_2| \pm (1.96) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

99% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x}_1 - \bar{x}_2| \pm (2.58) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

90% Confidence limits for $|\mu_1 - \mu_2|$ are $|\bar{x}_1 - \bar{x}_2| \pm (1.645) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Test for Single Proportion

If a random sample size of n. let X is a no. of persons possessing the given attribute. Then the sample proportion of success is $p = \frac{X}{n}$, we have proved that $E(p) = P$.

Null hypothesis (H_0): There is no significance difference between sample proportion and population proportion.

Now under null hypothesis H_0, the test statistic is

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

Where 'p' = Sample proportion

P = Population proportion

Q = 1-P

n = no. of observations or samples

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for population proportion P is $|p-P| \pm (Z_{tab})\sqrt{\frac{PQ}{n}}$

95% Confidence interval for population proportion P are $|p-P| \pm (1.96)\sqrt{\frac{PQ}{n}}$

99% Confidence interval for population proportion P are $|p-P| \pm (2.58)\sqrt{\frac{PQ}{n}}$

90% Confidence interval for population proportion P are $|p-P| \pm (1.645)\sqrt{\frac{PQ}{n}}$

Test for two proportions

Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 form the two populations respectively. Then sample proportions are

$p_1 = \frac{X_1}{n_1}$ And $p_2 = \frac{X_2}{n_2}$

Null hypothesis (H_0): There is no significance difference between two population Proportions.

Now under null hypothesis H_0, the test statistic is

$$Z = \frac{|p_1 - p_2|}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Where n_1 = no. of observations of first sample.

n_2 = no. of observations of second sample.

p_1 = First sample proportion

p_2 = Second sample proportion

P = Population proportion

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

$$Q = 1 - P$$

Conclusion: Now we compare the calculated value of Z with table value of (Z_α)

If calculated value of Z is less than tabulated value of z then we accept null hypothesis at certain level of significance.

If calculated value of Z is greater than tabulated value of Z then we accept alternative hypothesis at certain level of significance.

Confidence interval for $|P_1 - P_2|$ i.e., for the difference in the two proportions of populations formulae is $|p_1 - p_2| \pm (Z_{\alpha\alpha}) \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

95% Confidence limits for $|P_1 - P_2|$ are $|p_1 - p_2| \pm (1.96) \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

99% Confidence limits for $|P_1 - P_2|$ are $|p_1 - p_2| \pm (2.58) \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

90% Confidence limits for $|P_1 - P_2|$ are $|p_1 - p_2| \pm (1.645) \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Varalu

Measure of Central Tendency

Introduction: Consideration of data in to single value mostly it is at centre and it carries important properties of data. Also known as Measure of Location or centering the constant. In statistics measure central tendency is called averages.

There are five types of measure of central tendency or average which are commonly used these are

1. Arithmetic mean
2. Median
3. Mode
4. Geometric Mean
5. Harmonic mean

Rule or Requisites a good measure of Central tendency

The following are the characteristic to be satisfied by a good measure of central tendency.

1. It should be rigidly defined
2. It should be easy to understand and easy to calculate
3. It should be based on all observation.
4. It should be suitable for further mathematical treatment
5. It should not be effected much by a extreme values

Simple Arithmetic Mean:

Mean is obtained by adding together all the items and by dividing this total by the no. of items.

Mean for ungrouped data

Let 'X' takes a values $x_1, x_2, \ldots, x_n$ be "n" observations then arithmetic mean is defined as

$$\overline{X} = \frac{\sum x}{n}$$

Where $\sum$ is the capital sigma of the Greek alphabet and it is used in mathematics to denote sum of values. And 'n' is no. of observations.

Arithmetic Mean for Grouped data

Discrete Series:

Let 'X' denote $x_1, x_2, x_3, \ldots\ldots x_n$, and their corresponding frequencies are $f_1, f_2, f_3, \ldots\ldots f_n$, then arithmetic mean is defined as

$$\overline{X} = \frac{\sum fx}{N}$$

Where $N = \sum f$

Continuous Series:

Let "m" denote $m_1, m_2, m_3, \ldots\ldots m_n$, and their corresponding frequencies are $f_1, f_2, f_3, \ldots$ F_n then arithmetic Mean is defined as

$$\overline{X} = \frac{\sum fm}{N}$$

Where $N = \sum f$

Merits:

1. It is rigidly defined
2. It is easy to calculate and easy to understand
3. It is based on all observations
4. It is suitable for further mathematical treatment

Demerits:

1. It may not be represented in actual data so it is theoretical
2. The extreme values have greater effect on mean
3. It cannot be calculated if all the values are not known
4. Mean may lead to fallacious conditions in the absence of original observations

Median:

Median of the distribution is the value of the variable which divides it into two equal parts

Median for ungrouped data:

If a number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude

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

In case of even number of observations there are '2' meddle terms and median is obtain by taking the arithmetic mean of the middle terms.

$$Median = \frac{n}{2} and \frac{n}{2}+1^{th} terms$$

Median for Grouped data: when the date is grouped. Then Geometric mean is

Discrete Series:

When the data is the discrete series then Median is :

The steps for calculating Median are given below:

1. Find $\frac{N}{2}$. Where $N = \sum f$
2. Cumulative frequencies (c.f) is just greater than (>) $\frac{N}{2}$
3. The corresponding value of "X" is Median

Continuous Series:

When the data is continuous series then Median is

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

Where L = lower limit of the median class

f = frequency of the median class

c = cumulative frequency of the class preceding the median class.

N = total frequency ($\sum f$)

Merits:

1. It rigidly defined
2. It easy to understand and easy to calculate
3. It can be calculated for distributions with "open –end" class.
4. It is not affected by extreme values
5. It deals with the qualitative characteristics.

Demerits:

1. It is not subject to algebraic treatment
2. It cannot represent the irregular distribution series.
3. It is positional average and is based on the middle item
4. It does not have sampling stability
5. It is an estimate in case of a series containing even number of items

Mode:

Mode of the distribution can be defined as most frequently occurring values.

Examples

1. The average height of an India (Male) is 5^{1}-6^{11}.
2. The average size of the shoes sold in a shop is "7"

Mode for ungrouped data:

X : 5,9,6,5,7,5,10,35,78.

Here Mode is '5'. Because '5' is most repeated value

Mode for grouped data:

Discrete Series:

When the data is discrete the mode of the given distribution as the following steps:

1. Value of the variable noted in the first column.
2. Frequencies are noted in second column.
3. Sum of 2-2 frequencies are noted in third column.
4. Ignoring the first frequency and sum of 2-2 frequencies in fourth column.
5. Sum of 3-3 frequencies noted in the fifth column.
6. Ignoring the first frequency and sum of 3-3 frequencies are noted in the sixth column.

7. Ignoring the two frequencies and sum of 3-3 frequencies are noted in seventh column and
8. Repeat this procedure as our requirement.
9. After completing the table and identify the maximum frequency in each column and write down the corresponding value of the variable in analysis table and take which value is most frequently occurred.

Continuous Series:

When the data is continuous then mode is

$$Mode = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$$

Where L = Lower limit of the model class

f_1 = Frequency of the model class

f_0 = Frequency of the preceding model class

f_2 = Frequency of the succeeding model class

h or c = width of the class interval.

Merits:

1. Mode is readily comprehensible and easy to calculate
2. Mode is not all affected by extreme values.
3. Mode can be conveniently located even if the frequency distribution has class intervals of unequal magnitude.
4. Provided the model class and the classes proceeding and succeeding if all the same magnitude.
5. Mode is the average to width used to used to find ideal size

Example: Business forecasting in the manufacture of readymade garments, shoes etc.

6. It can be located in some cases by inspection.

Demerits:

1. There are different formulae for its calculations which ordinarily give different answers.
2. Mode is determinate some series have two or more than two modes.
3. It cannot be subjected to algebraic treatments.

For example: The combined mode cannot be calculated for the modes of two series.

4. It is an unstable measure as it affected much by sampling fluctuations.
5. Mode for the series with unequal class intervals cannot calculate.

Geometric Mean:

Geometric mean of a set of 'n' observations is the n^{th} of their product. Thus, geometric mean is denoted by "G".

G.M for Ungrouped Data: when the date is ungrouped. Then Geometric mean is

$$G = Anti\log\left(\frac{\sum \log x}{n}\right)$$

G.M for groped Data: when the date is grouped. Then Geometric mean is

Discrete Series:

$$G = Anti\log\left(\frac{\sum f \log x}{N}\right)$$

$$where\ 'N' = \sum f$$

Continuous Series:

$$G = Anti\log\left(\frac{\sum f \log m}{N}\right)$$

Harmonic Mean:

Harmonic mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocal of the human values

Harmonic Mean for Ungrouped data: when the date is ungrouped. Then Harmonic mean is

$$H.M = \frac{n}{\sum\left(\frac{1}{x}\right)}$$

Harmonic Mean for grouped data: when the date is grouped. Then Harmonic mean is

Discrete Series:

$$H.M = \frac{N}{\sum\left(\frac{f}{x}\right)}$$

Continuous Series:

$$H.M = \frac{N}{\sum\left(\frac{f}{m}\right)}$$

Measure of Dispersion

Literal meaning of dispersion is "scatteredness". We study dispersion to have an idea of spread about the central values of given distribution is called as 'Measure of Dispersion'.

Characteristic for good measure of Dispersion:

The dispersion for a good measure of dispersion is the same as those for an good measure of central tendency,

1. It should be rigidly defined
2. It should be easy to calculate and easy to understand
3. It should be based on all observations.
4. It should be amenable to further mathematical treatment.
5. It should be affected as little as possible by fluctuations of sampling.

Measure of Dispersion as follows:

- Range
- Quartile Deviation

- Mean Deviation
- Standard deviation

Range:

Range is the difference between the maximum value and minimum value of the given data.

$$\text{Range} = \text{Maximum value} - \text{minimum value}$$

$$\text{Co-efficient of Range} = \frac{\text{Maximum value} - \text{Minimum value}}{\text{Maximum value} + \text{Minimum value}}$$

Quartile Deviation:

It is the difference between third quartile and first quartile of the data divided by two.

$$\text{Quartile Deviation } (Q.D) = \frac{Q_3 - Q_1}{2}$$

$$\text{Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation:

Frequencies distribution, then mean deviation from the average "A" is given by

Represents the modules are the absolute value of the deviation then the sign is ignored.

Ungrouped series:

Grouped series ():

Standard Deviation:

So usually denoted by the Greek letters "σ" is the square root of A.M of the square of the deviation of the given values from their A.M.

Standard deviation for Ungrouped Data:

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

Standard deviation for grouped Data:

Discrete series: when the data is discrete series then standard deviation is

$$\sigma = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

Continuous Series: when the data is continuous series then standard deviation is

$$\sigma = \sqrt{\frac{\sum fm^2}{N} - \left(\frac{\sum fm}{N}\right)^2}$$

Co – efficient of Variation:

100 times the co – efficient of dispersion based upon standard deviation is called co – efficient of variation.

$$\text{Co-efficient Dispersion} = \frac{\sigma}{\bar{x}}$$

$$\text{Co -effiecient of Variation} = \frac{\sigma}{\bar{x}} \times 100$$

It is very useful to find the variation between two series. A Series which has less co-efficient of variation that series has more homogeneous. A series which has large co-efficient of variation that series has more heterogeneous.

Skewness

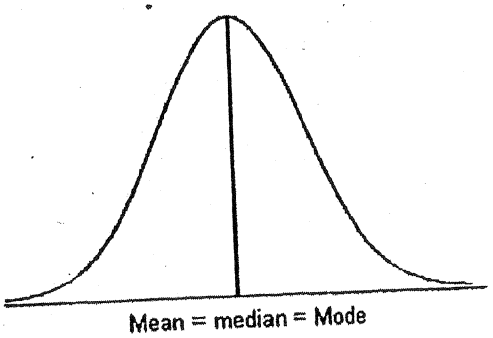
Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution. When distribution is not symmetrical (or asymmetrical) it is called a skewed distribution. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution.

Types of Skewness:

Skewness may be three types

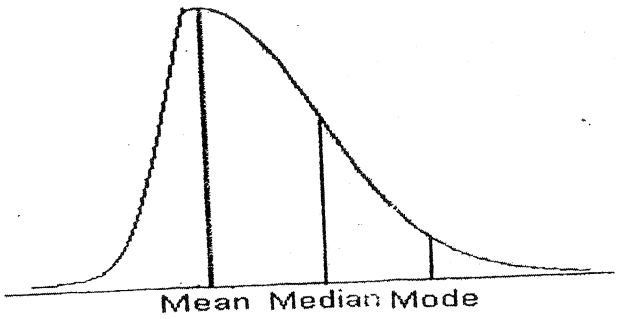
1. Symmetrical distribution
2. Positively skewed distribution
3. Negatively skewed distribution

Symmetrical Distribution



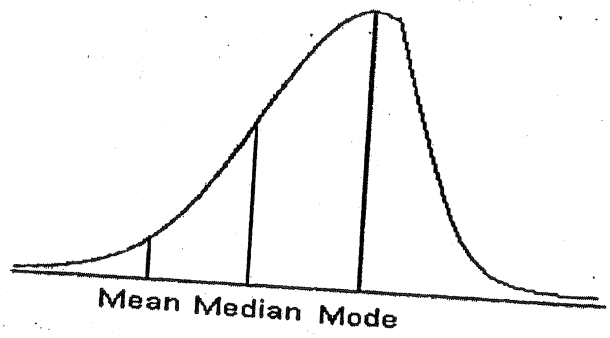
It is clear from the above diagram that in symmetrical distribution the value of mean, median and mode coincide (mean = median = mode). The spread of the frequencies is the same on both sides of the centre point of the curve.

Positively skewed distribution



The above diagram that in positively skewed distribution the value of mean is maximum and that of mode least – the median lies in between the two (mean > median > mode). In positively skewed distribution the frequencies are spread out over a greater range of values on the high values end of the curve (right hand side) than they are on the low value end.

Negatively Skewed distribution



The above diagram that in negatively skewed distribution the value of mode is maximum and that of mean least – the median lies in between the two (mean < median < mode). In negatively skewed distribution the frequencies are spread out over a greater range of values on the low values end of the curve (left hand side) than they are on the low value end.

Karl Pearson's Coefficient of Skewness

It is based upon the difference between mean and median. The difference is divided by standard deviation to give a relative measure. The formula thus becomes.

$$S_k = \frac{Mean - Mode}{S.D(\sigma)} \text{ Or } \frac{3(\text{Mean-Median})}{\text{S.D}(\sigma)}$$

Bowley's coefficient of skewness : It is based on Quartiles

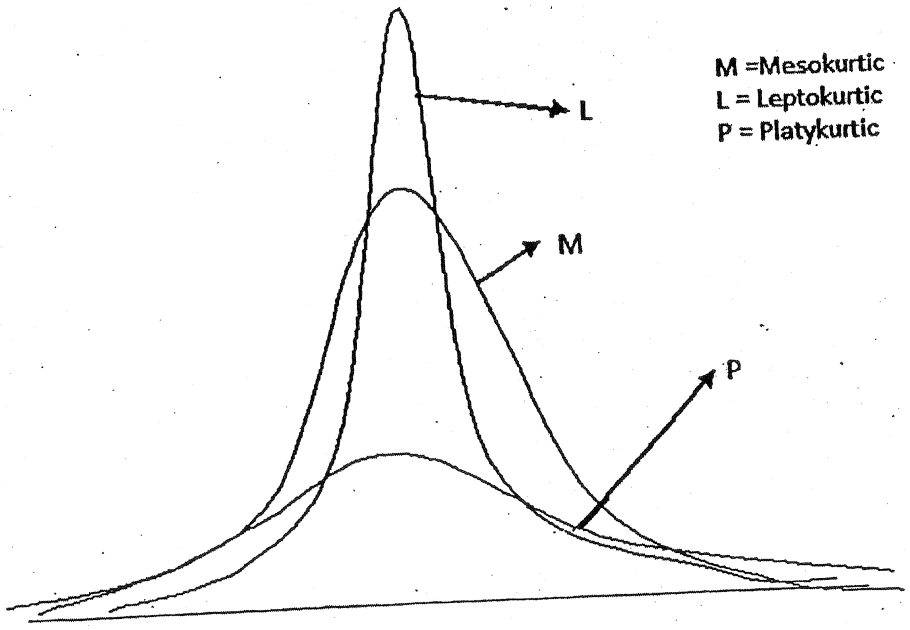
$$S_k = \frac{Q_3 + Q_1 - 2(Median)}{Q_3 - Q_1}$$

Kurtosis

Kurtosis is the degree of peakedness of a distribution usually taken relative to a normal distribution. Kurtosis in Greek means "bulginess". In a statistics kurtosis refers to the degree of flatness or peekedness in the region about the mode of a frequency curve the degree of kurtosis of a distribution is measured relative to the peakedness of normal curve.

In other words measures kurtosis tell us the extent to which a distribution is more peaked or flat – topped than the normal curve. If a curve is more peaked than the normal curve is called 'leptokurtic', if a curve is more flat-topped than the normal curve, it is called 'platykurtic', the normal curve itself is known as kurtosis of excess.

The following diagram illustrates the shape of three different curves mentioned above.



The above diagram clearly shows that these curves differ widely with regard to convexity, an attribute which Karl Pearson referred to "kurtosis". Curve 'M' is a normal one and is called 'mesokurtic'. Curve is more peaked than 'M' and is called Leptokurtic (L). A leptokurtic curve is narrower central portion and higher tails than does the normal curve 'P' is less peaked and it is called 'Platy kurtic'

Correlation

First we know some basic terms:

Uni-variate Distribution: the distribution which involves one variable is called univariate distribution

Bi-variate Distribution: The distribution which involves two or more variables is called Bi-variate distribution.

Correlation:

The relation between two variables is called correlation. It is used to measure the relationship between two variables.

If the change one variable affects a change in other variable, the variables, are correlated. Correlation broadly classified into three ways.

Positive Correlation

If two variables deviated in the same direction. If increase in one variable in a corresponding increase in other variable. (Or)

If decrease in one variable a corresponding decrease in the other variable. This is the same direction. This type of correlation is called positive correlation.

Example:

1. Height and weight of certain group of persons
2. Income and expenditure.
3. Rainfall and agricultural production.

Negative Correlation:

If tow variables deviated in the opposite direction. If the increase in one variable in a corresponding decrease in other variable. (Or)

If the decrease in one variable in a corresponding increase in the other variable. This is the opposite direction this type of correlation is called Negative correlation.

Example: Price and demand of commodity.

Zero Correlation or Independent Correlation:

There is no relation between two variables. This is also known as zero correlaton

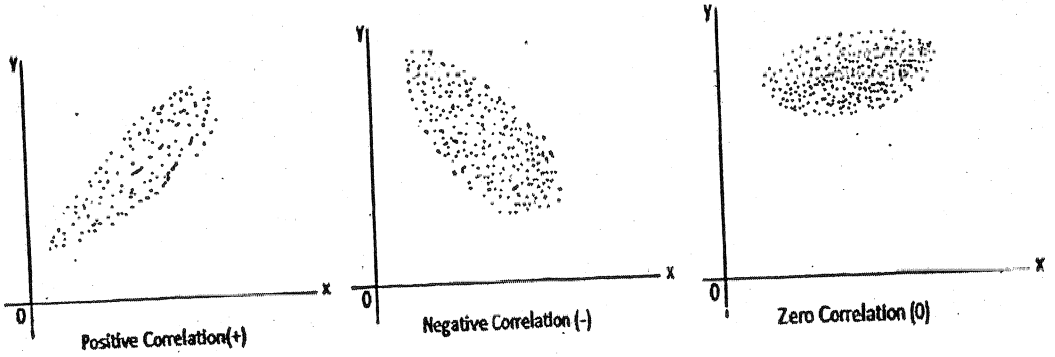
Example: Beautiful and Intelligence.

Scatter Diagram:

It is the simplest way of the diagrammatic representation of the Bi – variate data. For Bi-variate distribution if the values of the variables (x,y) I = 1,2,3....n are plotted along the x-axis and y- axis respectively in the xy-plane.

The diagram of dots obtained is known as scattered diagram. From this scattered diagram we can form a fairly good, though vague, idea whether the variables are correlated or not.

Example: if the dots are very dense, that is very close to each other. We should expect a fairly good amount of correlation between the variables. If the dots are widely scattered, we should expect a bad correlation.



Karl Pearson's Correlation Coefficient:

This is used to measure the degree of linear relationship between two variables. If x and y are two variables then "Karl Pearson's correlation coefficient (r)" is

$$r_{xy} \text{ or } r(x,y) = \frac{\frac{1}{n}\sum xy - \overline{xy}}{\sqrt{\left[\frac{1}{n}\sum x^2 - (\bar{x})^2\right]\left[\frac{1}{n}\sum y^2 - (\bar{y})^2\right]}}$$

Properties of the Karl Pearson's correlation co-efficient:

The two variables x and y are linearly related. In other words, this scatter diagram of the data will give a straight line curve.

1. Correlation co-efficient is always lies between -1 and +1. That is, $-1 \leq r \leq +1$.
   If r = +1, the correlation is perfect and positive.
   If r = 0, the correlation is zero, there is no relation between two variables.
   If r = -1, the correlation is perfect and negative.
2. Correlation coefficient is independent of change of origin and scale that is r(x, y) = r(u, v)
3. Independent variables are un-correlated
4. Karl Pearson's correlation co-efficient deals with the quantitative characteristics only.

Probable error of Correlation Coefficient:

If r(x, y) is correlation coefficient in a sample of "n" pairs of observations then standard error is given by

$$\text{Standard Error (S.E)} = \frac{1-r^2}{\sqrt{n}}$$

Probable error of correlation coefficient is defined as

$$P.E\,(r) = 0.675(S.E)$$

$$= 0.675\left(\frac{1-r^2}{\sqrt{n}}\right)$$

Where P.E = probable error of correlation coefficient.

Spearmen's Rank Correlation Coefficient:

X_i, y_j be the ranks of two characteristics A and B respectively the spearmen's correlation coefficient is denoted by

$$R(x,y) = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where 'n' = number of observations

d = the rank of x – rank of y (R(x)-R(y))

Spearmen's Tide Rank correlation co-efficient: Let (x, y) be the same repeated ranks of the two characteristics A and B then spearmen's tide rank correlation coefficient is denoted by

$$R(x,y)=1-\frac{6\left(\sum d^2+T_x+T_y\right)}{n(n^2-1)}$$

Where n = number of observations

T_x = Tide rank in x-series

$$T_x=\frac{\sum_{i=1}^{n} m_i(m_i^2-1)}{n(n^2-1)}$$

Where m_i = number of repeated times in i^{th} highest value.

T_y= Tide rank in y-series

$$T_y=\frac{\sum_{j=1}^{n} m_j(m_j^2-1)}{n(n^2-1)}$$

Where m_j= number of repeated times in j^{th} highest value.

Regression

The literal meaning of regression analysis is "stepping back towards the average". The regression analysis was first derived by "Sir. Francis Galton". The regression analysis is the mathematical measure of average relation between two or more variables in terms of the origin unit of the data.

The main aim of regression analysis is to estimate or predict unknown values from the given known values.

Example: $y=a+bx$

Simple Regression or Linear Regression:

The average relation between one dependent variable and one independent variable is called regression.

Example: $y = a + bx$

Multiple Regressions:

The average relationship between one dependent variable and two or more independent variables is called multiple regression

Example: $y = a + b_1x_1 + b_2x_2 + b_3x_3 \ldots\ldots + b_nx_n$

Dependent and Independent Variables

In the regression analysis, there are two types of variables. One is dependent variable and other one is independent variable.

The variable whose value influenced or predicted is known as dependent variable or explained variables .Which values influences or prediction by other variables in known as independent variable or explanatory variable.

Regression Lines:

Let (x_i, y_j) I =1, 2, 3, ...n be the bi-variate data, 'y' is dependent variable and 'x' is independent variable then regression equation of 'y' on 'x' is defined as

$$y = a + bx$$

Where a is a constant

b is the regression coefficient

The regression equation of 'x' on 'y' is defined as

$$x = a + by$$

Any line passes through the points $\bar{x}$ and $\bar{y}$ respectively then the regression equation of 'y' on 'x' is defined as

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Where b_{yx} is the regression coefficient of 'y' on 'x'.

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

or

$$b_{yx} = \frac{\frac{1}{n}\sum xy - \overline{xy}}{\frac{1}{n}\sum x^2 - (\bar{x})^2}$$

The regression equation of 'x' on 'y' is defined as

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where b_{xy} is regression co efficient of 'x' on 'y'.

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

or

$$b_{xy} = \frac{\frac{1}{n}\sum xy - \overline{xy}}{\frac{1}{n}\sum y^2 - (\bar{y})^2}$$

Properties of Regression Co efficient:

- Correlation coefficient is geometric mean of two regression coefficients

$$r = \sqrt{b_{yx} \times b_{xy}}$$

- The arithmetic mean of two regression coefficient is greater than the correlation coefficient.

$$\frac{b_{yx} + b_{xy}}{2} > r$$

- Regression coefficient is independent of change of origin and scale.