**SYLLABUS**

**Subject – Statistics**

|  |  |
| --- | --- |
| UNIT – I | Meaning, definition, significance, scope and limitations of statistical investigation process of data collection, primary and secondary data. Method of sampling, preparation of questionnaire, classification and tabulation of data preparation of statistical series and its type. |
| UNIT – II | Measurement of central tendency – mean, median, quartile, mode, geometric mean and harmonic mean. |
| UNIT – III | Dispersion and skewness, analysis of time series – meaning, importance, components, decomposition of time series measurement of long term trends, measurements of cyclical and irregular fluctuation. |
| UNIT – IV | Correlation meaning, definition, type and degree of correlation, methods of correlation, regression analysis meaning uses difference between correlation and regression, linear regression equation, calculation of coefficient of regression. |
| UNIT – V | Probability, Permutation and Development of Probability, Permutation and Combination, Calculation of Probability, Addition and Multiplication. |

**UNIT — I**

**STATISTICS**

The word “Statistics” of English language has either been derived from the Latin word status or Italian word statistics and meaning of this term is “An organised political state.

**Meaning:** The science of collecting, analysing and interpreting such data or Numerical data relating to an aggregate of individuals.

E.g:- Statistics of National Income, Statistics of Automobile Accidents, Production Statistics, etc.

**Definition:** - “The classified facts relating the condition of the people in a state specially those facts which can be stated in members or in tables of members or in any tabular or classified arrangements.”

**-Webster**

“Statistics may be regarded as (i) the study of population (ii) The study of variation (iii) The study of method of reduction of data”

**-R.A. Fisher.**

**Nature /Features /Characteristics of statistics**

* It is an aggregate of facts.
* Analysis of multiplicity of causes.
* It is numerically expressed.
* It is estimated according to reasonable standard of accuracy.
* It is collected for pre-determined purpose.
* It is collected in a systematic manner.

**Division of Statistics**

Theoretical Statistical Methods Applied

**Theoretical:** Mathematical theory which is the basis of the science of statistics is called theoretical statistics.

**Statistical Methods:** By this method we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes.

**Few Methods are:-**

1. Collection of Data (2) Classification (3) Tabulation (4) Presentation (5) Analysis (6) Interpretation (7) Forecasting.

**Applied: -** It deals with the application of rules and principles developed for specific problem in different disciplines.

**Eg: -** Time series, Sampling, Statistical Quality control, design of experiments.

**Functions of Statistics:**-

* It presents facts in a definite form.
* It simplifies mass of figures
* It facilitates comparison
* It helps in prediction
* It helps in formulating suitable & policies.

**Scope of Statistics:-**

1. Statistics and state or govt.
2. Statistics and business or management.

* Marketing
* Production
* Finance
* Banking
* Control
* Research and Development
* Purchases

1. Statistics and Economics
   * Measures National Income
   * Money Market analysis
   * Analysis of competition, monopoly, oligopoly,
   * Analysis of Population etc.
2. Statistics and science
3. Statistics and Research

**Limitations:-**

1. It is not deal with items but deals with aggregates.
2. Only on expert can use it
3. It is not the only method to analyze the problem.
4. It can be misused etc.

**Statistical Investigation**

**Meaning:** In general it means as a statistical survey.

In brief. Scientific and systematic collection of data and their analysis with the help of various statistical method and their interpretation.

**Stages of Statistical Investigation:-**

* Planning of Investigation
* Collection of Data
* Editing of Data
* Presentation of Data

1. Classification
2. Tabulation
3. Diagrams
4. Graphs
   * Analysis of Data
   * Interrelation of Data or Report Preparation

**Types of Statistical Investigation:-**

1. Experiment or survey investigation
2. Complete or sample investigation
3. Official, semi-official, Non official investigation
4. Confidential or open investigation
5. General purpose and specific purpose investigation
6. Original or repetitive investigation.

**METHODS OF SAMPLING**

**Meaning: -** The process of obtaining a sample and its subsequent analysis and interpretation is known as sampling and the process of obtaining the sample if the first stage of sampling.

The various methods of sampling can broadly be divided into:

1. Random sampling method
2. Non Random sampling method

**Random Sampling Method**

**I Simple Random Sampling:** - In this method each and every item of the population is given an equal chance of being included in the sample.

(a) Lottery Method (b) Table of Random Numbers

**Merits:**

Equal opportunity to each item.

Better way of judgment

Easy analysis and accuracy

**Limitations:**

Different in investigation

Expensive and time consuming

For filed survey it is not good

**II Stratified Sampling:-** In this it is important to divided the population into homogeneous group called strata. Then a sample may be taken from each group by simple random method.

**Merit:-** More representative sample is used.

Grater accuracy

Geographically Concentrated

**Limitations:** Utmost care must be exercised due to homogeneous group deviation. In the absence of skilled supervisor sample selection will be difficult.

**III Systematic Sampling:-** This method is popularly used in those cases where a complete list of the population from which sampling is to be drawn is available. The method is to be select k th item from the list where k refers to the sampling interval.

**Merits: -** It can be more convenient.

**Limitation: -** Can be Baised.

**IV Multi- Stage Sampling:** - This method refers to a sampling procedure which is carried out in several stages.

**Merit:** - It gives flexibility in Sampling

**Limitation: -** It is difficult and less accurate

**Non Random Sampling Method:-**

* + 1. **Judgment Sampling: -** The choice of sample items depends exclusively on the judgment of the investigator or the investigator exercises his judgement in the choice of sample items. This is an simple method of sampling.
    2. **Quota Sampling: -** Quotas are set up according to given criteria, but, within the quotas the selection of sample items depends on personal judgment.
    3. **Convenience Sampling: -** It is also known as chunk. A chunk is a fraction of one population taken for investigation because of its convenient availability. That is why a chunk is selected neither by probability nor by judgment but by convenience.

**Size of Sample**:- It depends upon the following things:-

Cost aspects. The degree of accuracy desired. Time, etc. Normally it is 5% or 10% of the total population.

**Limitation of overall sampling Method:-**

Some time result may be inaccurate and misleading due to wrong sampling.

Its always needs superiors and experts to analyze the sample.

It may not give information about the overall defects. In production or any study.

It Becomes Biased due to following reason:-

1. Faulty process of selection
2. Faulty work during the collection of information
3. Faulty methods of analysis etc.

Statistics in Business and Industry

In general a typical organization has three levels: Strategic, Managerial and Operational. This classification is somewhat general and arbitrary. However, this can help identify and emphasize different tools to be directed at the different levels.

We envision Statistics to play important roles at all these levels.

Strategic Level (Top of an Organization)

At the strategic level the most emphasis should be on Statistical Thinking which has the following components:

i) Notion of a Process,

ii) Notion of measurement and data based decisions,

iii) Understanding and dealing with variation,

iv) Statistical tools, and

v) Systematic approach.

The notion of process thinking is fundamental to any organizational change. Every action has some inputs and outputs which need to be identified in every context. Decisions at the strategic level should be based on facts supported by appropriate data and this requires an understanding of variation (Deming 1986).

Business and Industry have seen the arrival and demise of many programs such as Total Quality Management. Embracing any program that comes along without firm commitment and understanding is doomed to failure.

Managerial Level (Middle Level)

This is the level at which systems are devised for implementation of the directions taken by upper management. In particular, systems for robust product and process design, process control and improvement, and training are the responsibility of middle management. Understanding of some statistical tools and statistical thinking are prerequisites for those who are designing these systems.

Operational Level

At this level the methods are implemented through the system built at the managerial level. Understanding of statistical tools such as control charting, capability analysis, design of experiments, measurement system analysis, regression analysis etc are essential. Appropriate statistical tools need to be used by operational people as part of their daily work. People in some areas need to know the details only of certain statistical tools. For instance, an operator responsible for maintaining stability of a process by charting should know the workings of a control chart but need not know a lot about design of experiments. On the other hand an engineer responsible for process improvement should be knowledgeable in several aspects of statistical process control and design of experiments.

Implementation: General Issues

Commitment of Top Management

For the success of any program that affects an organization as a whole, full and highly visible commitment of senior management is essential. Thus the vision and values of the top level management is highly important. Employees must perceive active leadership and involvement of senior people in implementation.

Top management has to asses the situation early and to decide to allocate the needed resources. They have to decide in advance what role they can and will play. For example the success of the Six Sigma program at General Electric and Motorola is due to the commitment of its senior management.

Recognition of the potential benefits of implementation in the beginning can help focus on what is needed. Top management must recognize that, in addition to help solving problems and improving processes, statistical tools can help increase customer satisfaction and help measure the performance of the organization. Implementation of statistical tools is an ongoing process and it helps the organization to be a learning organization and a knowledge based enterprise. Knowledge based organizations will be the successful ones in the long run.

Role of Statisticians

The traditional role of a statistician in business or industry has been to act as a consultant to projects, or to train some workers in certain tools such as statistical process control and design of experiments. This role has to be broadened. Statisticians need to teach at all levels of an organization (senior managers, scientists

and engineers, middle managers and operational people). Statisticians have to be leaders, facilitators, aide to management etc. They have to identify the role of statistics in various business functions and also have to interact with the outside world. These require broadening statistician’s skills set. Statisticians need to ac-

quire communication skills and have to be good communicators. We have to keep the statistical tools appropriate and simple. We also need to make sure that the implementation adds value to the organization.

**UNIT-II**

**PROCESS OF DATA COLLECTION**

**Data:** - A bundle of Information or bunch of information.

**Data Collection:** Collecting Information for some relevant purpose & placed in relation to each other.

**Types of Data:-**

1. **Raw Data:-** When we collect data through schedules and questionnaires or some other method eg:- Classification, tabulation etc.
2. **Processed Data:-** When we use the above raw data for application of different methods of analysing of data. Like using correlation, Z-test, T-test on data. That will be known as processed data.

**Sources of Data Collection:-**

1. **Internal Data: -** When data is collected by problem the internal source for any specific

It purpose.

1. **External Data: -** This type of data collected by the external source.
2. **Primary Data: -** It is original and collected first time. it is like raw material and it is required large sum of money, energy and time.
3. **Secondary Data: -** Secondary data are those already in existence and which have been collected for some other purpose than answering of the question at hand.
4. **Qualitative Data: -** Which can not be measurable but only there presence and absence in a group of individual can be noted are called qualitative data.
5. **Quantitative Data: -** The characteristics which can be measured directly are known as quantitative data.

**Collection of Data: -** It means the methods that are to be employed for obtaining the required information from the units under investigations.

**Methods of Data Collection:- (Primary Data)**

* Direct Personal Interviews
* By observation
* By Survey
* By questionnaires

**Difference between Primary and secondary data:-**

|  |  |  |
| --- | --- | --- |
| **Points** | **Primary Data** | **Secondary Data** |
| 1. **Originality** | Primary data are original i.e., collected first time. | Secondary data are not original, i.e.., they are already in existence and are used by the investigator. |
| 1. **Organisation** | Primary data are like raw material. | Secondary data are in the from of finished product. They have passed through statistical methods. |
| 1. **Purpose** | Primary data are according to the object of investigation and are used without correction. | Secondary data are collected for some other purpose and are corrected before use. |
| 1. **Expenditure** | The collection of primary data require large sum, energy and time. | Secondary data are easily available from secondary sources (published or unpublished). |
| 1. **Precautions** | Precautions are not necessary in the use of primary data. | Precautions are necessary in the use of secondary data. |

**Preparation of Questionnaires:-**

This method of data collection is quit popular, particularly in case of big enquires, it is adopted by individuals, research workers. Private and public organization and even by government also.

A questionnaires consists of number of question printed or type in a definite order on a form or set of forms. The respondents have to answer the question on their own.

**Importance:-**

1. Low cost and universal
2. Free from biases.
3. Respondents have adequate time to respond
4. Fairly approachable

**Demerits:-**

(i) Low rate of return

(ii) Fill on educated respondents

(iii) Slowest method of Response

**Preparation of Questionnaires: -** It is considered as the heart of a survey operation. Hence it should be very carefully constructed. If it is not properly set up and carefully constructed.

|  |  |  |
| --- | --- | --- |
| Step I | :- | Prepare it in a general form. |
| Step II | :- | Prepare sequence of question. |
| Step III | :- | Emphasize on question formulation and wordings |
| Step IV | :- | Ask Logical and not misleading questions. |
| Step V | :- | Personal questions should be left to the end. |
| Step VI | :- | Technical terms and vague expressions should be availed classification and Tabulation of Data |

**Classification & Tabulation of Data**

After collecting and editing of data an important step towards processing that classification. It is grouping of related facts into different classes.

**Types of classification:-**

1. **Geographical:-** On the basis of location difference between the various items. E.g. Sugar Cave, wheat, rice, for various states.
2. **Chronological:-** On the basis of time

e.g.-

|  |  |
| --- | --- |
| **Year** | **Sales** |
| 1997 | 1,84,408 |
| 1998 | 1,84,400 |
| 1999 | 1,05,000 |

1. **Qualitative classification: -** Data classified on the basis of some attribute or quality such as, colour of hair, literacy, religion etc.

**Population**

1. **Quantitative Classification: -** When data is quantify on some units like height, weight, income, sales etc.

**Tabulation of Data**

A table is a systematic arrangement of statistical data in columns and Rows.

**Part of Table:-**

1. Table number
2. Title of the Table
3. Caption
4. Stub
5. Body of the table
6. Head note
7. Foot Note

**Types of Table:-**

(i) Simple and Complex Table:-

**(a) Simple or one-way table:-**

|  |  |
| --- | --- |
| **Age** | **No. of Employees** |
| 25 | 10 |
| 30 | 7 |
| 35 | 12 |
| 40 | 9 |
| 45 | 6 |

**(b) Two way Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Age** | **Males** | **Females** | **Total** |
| 25 | 25 | 15 | 40 |
| 30 | 20 | 25 | 45 |
| 35 | 24 | 20 | 44 |
| 40 | 18 | 10 | 28 |
| 45 | 10 | 8 | 18 |
| **Total** | **97** | **78** | **175** |

**2) General Purpose and Specific Purpose Table:-** General purpose table, also known as the reference table or repository tables, which provides information for general use or reference.

Special purpose are also known as summary or analytical tables which provides information for one particular discussion or specific purpose.

**UNIT-III**

**Measure of Central tendency**

The point around which the observations concentrate in general in the central part of the data is called central value of the data and the tendency of the observations to concentrate around a central point is known as Central Tendency.

**Objects of Statistical Average:**

* To get a single value that describes the characteristics of the entire group
* To facilitate comparison

**Functions of Statistical Average:**

* Gives information about the whole group
* Becomes the basis of future planning and actions
* Provides a basis for analysis
* Traces mathematical relationships
* Helps in decision making

**Requisites of an Ideal Average:**

* Simple and rigid definition
* Easy to understand
* Simple and easy to compute
* Based on all observations
* Least affected by extreme values
* Least affected by fluctuations of sampling
* Capable of further algebric treatment

***ARITHMETIC MEAN (****clip_image006****)***

Arithmetic Mean of a group of observations is the quotient obtained by dividing the sum of all observations by their number. It is the most commonly used average or measure of the central tendency applicable only in case of quantitative data. Arithmetic mean is also simply called “mean”. Arithmetic mean is denoted by clip_image006.

**Merits of Arithmetic Mean:**

* It is rigidly defined.
* It is easy to calculate and simple to follow.
* It is based on all the observations.
* It is readily put to algebraic treatment.
* It is least affected by fluctuations of sampling.
* It is not necessary to arrange the data in ascending or descending order.

**Demerits of Arithmetic Mean:**

* The arithmetic mean is highly affected by extreme values.
* It cannot average the ratios and percentages properly.
* It cannot be computed accurately if any item is missing.
* The mean sometimes does not coincide with any of the observed value.
* It cannot be determined by inspection.
* It cannot be calculated in case of open ended classes.

**Methods of Calculating Arithmetic Mean:**

* Direct Method
* Short cut method
* Step deviation method

**Use of Arithmetic Mean:**

Arithmetic Mean is recommended in following situation:

* When the frequency distribution is symmetrical.
* When we need a stable average.
* When other measures such as standard deviation, coefficient of correlation are to be computed later.

***MEDIAN (M)***

The median is that value of the variable which divides the group into two equal parts, one part comprising of all values greater and other of all values less than the median. For calculation of median the data has to be arranged in either ascending or descending order. Median is denoted by **M**.

**Merits of Median:**

* It is easily understood and easy to calculate.
* It is rigidly defined.
* It can sometimes be located by simple inspection and can also be computed graphically.
* It is positional average therefore not affected at all by extreme observations.
* It is only average to be used while dealing with qualitative data like intelligence, honesty etc.
* It is especially useful in case of open end classes since only the position and not the value of items must be known.
* It is not affected by extreme values.

**Demerits of Median:**

* For calculation, it is necessary to arrange data in ascending or descending order.
* Since it is a positional average, its value is not determined by each and every observation.
* It is not suitable for further algebric treatment.
* It is not accurate for large data.
* The value of median is more affected by sampling fluctuations than the value of the arithmetic mean.

**Uses of Median:**

The use of median is recommended in the following situations:

* When there are open-ended classes provided it does not fall in those classes.
* When exceptionally large or small values occur at the ends of the frequency distribution.
* When the observation cannot be measured numerically but can be ranked in order.
* To determine the typical value in the problems concerning distribution of wealth etc.

***MODE (Z)***

Mode is the value which occurs the greatest number of times in the data. The word mode has been derived from the French word **‘La Mode’** which implies fashion. The Mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values. Mode is denoted by **Z.**

**Merits of Mode:**

* It is easy to understand and simple to calculate.
* It is not affected by extreme large or small values.
* It can be located only by inspection in ungrouped data and discrete frequency distribution.
* It can be useful for qualitative data.
* It can be computed in open-end frequency table.
* It can be located graphically.

**Demerits of Mode:**

* It is not well defined.
* It is not based on all the values.
* It is suitable for large values and it will not be well defined if the data consists of small number of values.
* It is not capable of further mathematical treatment.
* Sometimes, the data has one or more than one mode and sometimes the data has no mode at all.

**Uses of Mode:**

The use of mode is recommended in the following situations:

* When a quick approximate measure of central tendency is desired.
* When the measure of central tendency should be the most typical value.

***GEOMETRIC MEAN (G.M)***

The geometric mean also called geometric average is the nth root of the product of n non-negative quantities. Geometric Mean is denoted by **G.M**.

**Properties of Geometric Mean:**

* The geometric mean is less than arithmetic mean, G.M<A.M
* The product of the items remains unchanged if each item is replaced by the geometric mean.
* The geometric mean of the ratio of corresponding observations in two series is equal to the ratios their geometric means.
* The geometric mean of the products of corresponding items in two series.

**Merits of Geometric Mean:**

* It is rigidly defined and its value is a precise figure.
* It is based on all observations.
* It is capable of further algebraic treatment.
* It is not much affected by fluctuation of sampling.
* It is not affected by extreme values.

**Demerits of Geometric Mean:**

* It cannot be calculated if any of the observation is zero or negative.
* Its calculation is rather difficult.
* It is not easy to understand.
* It may not coincide with any of the observations.

**Uses of Geometric Mean:**

* Geometric Mean is appropriate when:
  + Large observations are to be given less weight.
  + We find the relative changes such as the average rate of population growth, the average rate of intrest etc.
  + Where some of the observations are too small and/or too large.
* Also used for construction of Index Numbers.

***HARMONIC MEAN (H.M)***

Harmonic mean is another measure of central tendency. Harmonic mean is also useful for quantitative data. Harmonic mean is quotient of “number of the given values” and “sum of the reciprocals of the given values”. It is denoted by **H.M.**

**Merits of Harmonic Mean:**

* It is based on all observations.
* It not much affected by the fluctuation of sampling.
* It is capable of algebraic treatment.
* It is an appropriate average for averaging ratios and rates.
* It does not give much weight to the large items and gives greater importance to small items.

**Demerits of Harmonic Mean:**

* Its calculation is difficult.
* It gives high weight-age to the small items.
* It cannot be calculated if any one of the items is zero.
* It is usually a value which does not exist in the given data.

**Uses of Harmonic Mean:**

* Harmonic mean is better in computation of average speed, average price etc. under certain conditions.

**Dispersion**

The Dispersion (Known as Scatter, spread or variations) measures the extent to which the items vary from some central value. The measures of dispersion is also called the average of second order (Central tendency is called average of first order).

The two distributions of statistical data may be symmetrical and have common means, median or mode, yet they may differ widely in the scatter or their values about the measures of central tendency.

**Significance/ objectives of Dispersion-**

* To judge the reliability of average
* To compare the two an more series
* To facilitate control
* To facilitate the use of other statistical measures.

**Properties of good Measure of Dispersion**

* Simple to understand
* Easy to calculate
* Rigidly defined
* Based on all items
* Sampling stability
* Not unduly affected by extreme items.
* Good for further algebraic treatment

Dispersion

1. Mean Deviation (coefficient of M.D)
2. Standard Deviation

Based on all items

Graphic Method

Based on selected Items

1. Range (coefficient of Range)
2. Inter-quartile, coefficient of Range (IQR), (IQR)

Lorenz Curve

1. **Range: -** Range (R) is defined as the difference between the value of largest item and value of smallest item included in the distributions. Only two extreme of values are taken into considerations. It also does not consider the frequency at all series.
2. **Quartile Deviation: -** Quartile Deviation is half of the difference between upper quartile (Q3) and lower quartile (Q1). It is very much affected by sampling distribution.
3. **Mean Deviation: -** Mean Deviation or Average Deviation (δAlpha) is arithmetic average of deviation of all the values taken from a statistical average (Mean, Median, and Mode) of the series. In taking deviation of values, algebraic sign + and – are also treated as positive deviations. This is also known as first absolute moment.
4. **Standard Deviation:-** The standard deviation is the positive root of the arithmetic mean of the squared deviation of various values from their arithmetic mean. The S.D. is denoted as σ Sigma.

**Method of calculating standard Deviation-**

1. Direct Method 2. Short-cut-Method 3. Step deviations Method

**Properties**

Fixed Relationship among measures of dispersion in a normal distribution there is a fixed relationship between quartile Deviation, Mean Deviation and Standard Deviation Q.D = 2/3 σ, Mean Deviation = 4/5σ.

**Distinction between mean deviation and standard deviation**

|  |  |  |
| --- | --- | --- |
| **Base** | **Mean Deviation** | **Standard Deviation** |
| 1. Algebric Sign | Actual +, - Signs are ignored and all deviation are taken as positive | Actual signs +, - are not ignored whereas they are squared logically to be ignored. |
| 1. Use of Measure | Mean deviation can be computed from mean, median, mode | Standard deviation is computed through mean only |
| 1. Formula | M.D or δ = | S.D or σ = |
| 1. Further algebraic Treatment | It is not capable of further algebraic treatment. | It is capable of further algebraic treatment |
| 1. Simplicity | M.D is simple to understand and easy to calculate | S.D is somewhat complex than mean deviation. |
| 1. Based | It is based on simple average of sum of absolute deviation | It is based on square root of the average of the squared deviation |

**Variance**

The square of the standard deviation is called variance. In other words the arithmetic mean of the squares of the deviation from arithmetic mean of various values is called variance and is denoted as σ2. Variance is also known as second movement from mean. In other way, the positive root of the variance is called S.D.

Coefficient of Variations- To compare the dispersion between two and more series we define coefficient of S.D. The expression is x 100 = known as coefficient of variations.

σ

X

**Interpretation of Coefficient of Variance-**

|  |  |
| --- | --- |
| **Value of variance** | **Interpretation** |
| Smaller the value of σ2 | Lesser the variability or greater the uniformity/ stable/ homogenous of population |
| Larger the value of σ2 | Greater the variability or lesser the uniformity/ consistency of the population |

**Dispersion**

**Range = R**

|  |  |  |
| --- | --- | --- |
| **Individual Series** | **Discrete Series** | **Continuous Series** |
| Range = L-S  Where L=Largest,  S=Smallest Observation |  |  |
| Coefficient of Range |  |  |

**Quartile Deviation - Q.D.**

|  |  |  |
| --- | --- | --- |
| **Individual Series** | **Discrete Series** | **Continuous Series** |
|  |  |  |
| Coefficient of Q.D. |  |  |

**Mean Deviation - M.D.** **(“Through actual mean, mode, median)**

|  |  |  |
| --- | --- | --- |
| **Individual Series** | **Discrete Series** | **Continuous Series** |
|  |  |  |
| Coefficient of |  |  |
| Mean |  |  |
| Coefficient of |  |  |
| (Mode) |  |  |
| Coefficient of |  |  |

**Standard Deviation =  can be calculated through mean only**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Individual Series** | **Discrete Series** | **Continuous Series** |
| Direct  (Through actual mean) |  |  |  |
| Indirect (Through assumed mean) |  |  |  |

“A Time Series” is a series of statistical data recorded in accordance with their time of occurrence. Here it is noted that it is a set of observation taken at specified times usually (but not always) at equal intervals. Thus a set of data depending on the time (which may be year, quarter, month, day etc.) is called a “Time Series”.

Today the use of time series analysis is not merely confined to economists and businessmen, but it extensively used by scientists, sociologist, biologists, astronomist, geologists, research workers etc.

**Some example of time series are**

1. The population of a country in different years.
2. The annual production of coal in India over the last ten years.
3. Deposits received by bank in a year.
4. The daily closing price of a share in the Bombay Stock Exchange.
5. The monthly sales of departmental store for the last six months.
6. Hourly temperature recorded by the store for the last six months.

**According to Patterson** “A timeseries consists of statistical data which are collected. Recorded or observed over successive increments.

**Utility or importance of Time Series**

The very important use of time series analysis is its use in forecasting future information and behaviour.

1. It enables us to predict or forecast the behavior of the phenomenon in future. Which is very essential for business planning. On the basis of past information, the trend can be estimated and projections can also be made for the uncertain future. It assists in reducing, the risk and uncertainties of business and industry.
2. It helps in the evaluation of current achievement by review and evaluation of progress made through a plan can be done on the basis of time series.
3. It helps in the analysis of past behavior of the phenomenon under consideration. What changes had taken place in the past, what factor were responsible for these changes, under that conditions these changes took place, etc. are certain issues which could be studied and analysed by time series.
4. It helps in making comparative studies in the values of different phenomenon at different times or place. It provides a scientific basis for making comparison by studying and isolating the effects of various components of a time series.
5. The segregation and study of the various components of time series is of paramount importance to a businessman in the planning of future operations and the formulation of executive and policy decisions.
6. On the basis of the past performance of the various sectors of economy, we can determine future requirements and a suitable policy can be formulated to get desired and predetermined objectives.

**Causes of variation in time series**

If the values of a phenomenon are observed at different periods of time, the values so obtained will show appreciable variations.

The following factors are generally affect any time series

1. Changing of tastes, habits and fashions of the people.
2. Changing of customs, conventions of the people.
3. Rituals and festivals.
4. Political movements, government policies.
5. War, Famines, Drought, Flood, Earthquakes and Epidemic etc.
6. Unusual weather or seasons.

**Components of Time Series**

“A time series may be defined as a collection of readings belonging to different time periods of some economic variable or composite of variable.

Eg. The retail price of a particular commodity are influenced by a number of factors namely the crop yield which further depends on weather conditions, irrigation facilities, fertilizers used, transportation facilities, consumer demand etc.

The various forces affecting the values of a phenomenon in a time series may be broadly classified into the following four categories, commonly known as the components of a time series.

1. Secular Trend (i.e. long-term smooth, regular movement)
2. Seasonal variation (periodic movement, the period being not greater than one year)
3. Cyclical Variation (periodic movement with period greater than one year)
4. Irregular or Random Variation.

Components of Time Series

Long-Term

Short-Term

Secular Trend (T)

Cyclical Variations (C)

Seasonal Variation (S)

Irregular or Random (I)

1. **Secular Trend** :- It is the matter of common sense that there might be violent variations in a time series during a short span of time, however in a long run, it has a tendency either to rise or fall. This tendency or trend of variation may be either upward or down set on over a long time period. This is known as ‘Secular trend’ or ‘Simple trend. It is but natural that population growth, technological progress medical facilities production, prices etc. are not judge over a day, month or year they shores. The movement of upward, downward or constant over a fairly long period.

**Broadly the trends are divided under two heads:**

1. Linear Trends, and
2. Non- Linear Trends
3. **Linear Trends: -** If we plot the values of time series on graph it shows the straight line i.e. growth rate is constant. Although in practice linear trend is commonly used but it is rearely found in economics and business data.
4. **Non-Linear Trends:** In business or economics generally growth is slow in the begging and them it is rapid for some time period after which it becomes stable for some time period and finally retards gradually. It is not linear it forms a curve known as non linear trends.
5. **Seasonal Variation:** As we Heard season the first things comes in our mind is spring, summer, autumn and winter. Generally seasonal variations are occur due to changes in weather condition, customer, tradition fashion etc.

Seasonal variations represent a periodic movement where the period is not longer than one year. The factors, which mainly cause this type of variation in time series, are the climatic changes of the different seasons. For example

1. Sale of woolens go up in winter.
2. Sale of raincoat and umbrella go up in rainy season.
3. Prices of food grains decrease with the arrival of new crop.
4. Sale of cooler, refrigerator etc. rise during the summer season.

Another variation occurs due to man-made convention and customs. Which people follow at different times like Durga Pooja, Dashehra, Deepawali, Ide. X-max etc. The seasonal variations may take place per day per week or per month. For example:

1. Sale of departmental stores go up in festivals.
2. Sale of cloths and Jewelry pick up in marriages.
3. Sale of Paint, furniture and electronics goes up during festivals like, Deepawali, Ide, X-max etc.
4. Sale of vehicles increase considerably during Durga Pooja and Dasherhra.
5. **Cyclical Variations:** Most of the business activities are often characterized by recurrence of periods of prosperity and slump constituting a business cycle. Cyclical variations are another type of periodic movement, with a period more than one year. Such movements are fairly regular and oscillatory in nature. One complete period is called a ‘cycle’ cyclical variations are not as regular as seasonal variation, but the sequence of changes, marked by prosperity, decline, depression and recovery, remains more of less regular.

**Phases of Business Cycle**

**Prosperity (Boom)**

**Normal**

**Depression**

1. **Irregular or Random Variation:** Irregular or random variation are such variation which are completely unpredictable in character. These are caused by factors which are either wholly unaccountable or caused by such unforeseen events like Earthquakes, flood, drought famines, epidemic etc, and some man-made situations like strikes lock-outs wart etc.

**Mathematical Models for Analysis of time Series**

Though there are many models by which a time series can be analyzed, two models commonly used for decomposition of a time series into various components are

1. **Additive Model :-** According to the additive model, the decomposition of time series is done on the assumption that the effect of various components are additives in nature, i.e. U = T+S+C+R

Where, U, is the time series value and T, S, C, and R stand for trend seasonal, cyclical and random variation.

In this model ‘S, C and R are absolute quantities and can have positive or negative values. The model assumes that the four components of the time series are independent of each other and non-has any effect whatsoever on the remaining three components.

1. **Multiplication Model :** According to the multiplication model, the decomposition of a time series on the assumption that the effects of the four components of a time series (T, S, C and R) are not necessarily independent of each other. In fact, the model presumes that their effects are interdependent According to this model.

U = T × S × C × R

**Measurement of Trend or Secular Trend**

The different methods of determining the trend component of a time series are

Measurement of Trend or Secular Trend

1. Freehand Method or Graphic Method
2. Semi-average Method
3. Moving Average Method
4. Least Square method

Straight Line Trend Equation

Quadratic Trend Equation

Exponential Trend Equation

1. **Moving Average Method:** Moving average method is very commonly used for the isolation of trend and in smoothing out fluctuations in time series. In this method, a series of arithmetic means of successive observation, known as moving averages, as calculated from the given data, and these moving average are used as trend values.
2. **Yearly moving average is given by**

**Illustration1 Calculate 3 yearly moving averages:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Years | : | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 |
| Earning(Lakhs) | : | 80 | 90 | 70 | 60 | 110 | 50 | 40 | 30 |

**Working Rule**

1. Add the values of the first3 years (namely 1979, 1981 i.e., 80+90+70=240) and place the total against the middle year1980.
2. Leave the first year’s value and add up the values of the next 3 years (i.e., 1980, 1981, 1982, viz., 90+70+70+60 = 220) and place the total against the middle year i.e., year 1981.

**Illustration2** Calculate 5 yearly moving averages and seven year moving average for the following data:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | : | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 |
| Sales (‘000 Rs.) | : | 123 | 140 | 110 | 98 | 104 | 133 | 95 | 105 | 150 | 135 |

**Calculation of Moving Averages when the period is Even:**

If the period of the moving average is even, centre point of the group will lie between two years. It is, therefore, necessary to adjust or shift (technically known as centre) these average so that they coincide with the years. For example

4-yearly moving average is calculated as:

**Step 1 :** Add the values of first four year, and place the total between the 2nd and 3rd year.

**Step 2 :** Leave the first year value and then add the for values of the next four years and place the total in between the 3rd and 4th year **Continue this process until the last year is taken into account.**

**Step 3 : Divide 4** yearly moving totals 4. It will give **4 yearly moving average.**

**Step 4 :** Add first two moving averages and divide it by 2 to get the **moving average centered.** Place it against 3rd year. Leave the first moving average and then add next two moving average and divide by 2 to get the next moving average centered. Place it against the 4th Year. Continue this process till the last moving average is included.

**Alternative Procedure:** In this procedure step 1 and 2 are same as above.

**Step 3:** Addfirst two 4 yearly moving total place it against 3rd year. Leave the first moving total and then add nexttwo moving total to get the next moving total centred. Place it against the 4th year. Continue this process till the last moving total is included.

**Step 4 :**  Diving these centered moving totals by 8. It will give 8 yearly moving average. This procedure will more clear by following illustration.

**Illustration** Construction a four-yearly centered moving average from the following data :

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | : | 1970 | 1975 | 1980 | 1985 | 1990 | 1995 | 2000 |
| Imported Cotton (in ‘000) | : | 129 | 131 | 106 | 91 | 95 | 84 | 93 |

**Method of Least Squares**

It is an appropriate mathematical technique to determined an equation which best fits on a given observation relating to two variables. In this procedure for fitting a live to a set of observation the sum of the squared deviations between the calculated and observed values in minimised. Therefore the technique is named as “Least-Squares method.” And the line so obtained is known as ‘Best fit line”.

We know that sum of the deviations from the arithmetic mean is zero. Therefore the sum of the deviations from the line of the best fit is zero.

1. , i.e., the sum of the deviations of the actual values of y and computed values of y is zero.
2. is least, i.e., the sum of the squares of deviations from the actual and the computed value of y is least.

That is why it is called the method of least squares and the line obtained by this method is called the ‘line of best fit’

This method may be used either to fit a straight line trend or parabolic trend straight line trend is represented by the equation y= a + bx where y represents the estimated values of the trend x represents the deviations in the time period. A and b are constants.

‘a’ represents intercept of the line of the y no is and ‘b’ represent the slope of the line i.e. it gives the changes in the value of y for per unit change in the value of x if b>0 it show and growth rate and if b<0 it shows decline rate.

**Merits:**

1. This is the only method of measuring trend which provides the future values authentically very convincing and reliable.
2. This method is used for forecasting the series for example.
3. If other factors are not so effective no share market, this method can provide very reliable information about the movement of the share of a company.
4. This method has no scope for personal bias of the investigator.
5. It is only method which gives the rate of growth per annum.

**Demerits:-**

1. The method required mathematical ability. Some items it involves tedious and complicated calculations.
2. The method has no flexibility i.e. if even a single term is added to series it makes necessary to do all the calculations again.
3. Estimations and predictions by this method are based only on long term variations and the impact of cyclical, seasonal and irregular variations are completely ignored.

**Computation of Trend Values by the Least Squares Method**

We know straight lines trend is given by y= a+bx in order to determine the values of the constants and b the following two normal equations are to be solved.

Where n represents number of years (months or any other period) for which data are given:

sum of actual values of y variable.

represents sum of deviations from the origin.

x2 represents sum of deviations from the origin.

represents sum of the deviations from the origin and actual values.

**Remarks :-** The variable x can be measured from any point of time as origin. But if middle time period is taken as origin and deviations are taken from the middle time period it provides the above normal equation would be reduced to the

🡪=na+0=na 🡪 Thus a =

b +0+= Thus b =

**UNIT-IV**

**Correlation**

**Introduction**

1. Correlation is a statistical tool & it enables us to measure and analyse the degree or extent to which two or more variable fluctuate/vary/change w.e.t. to each other.
2. For example – Demand is affected by price and price in turn is also affected by demand. Therefore we can say that demand and price are affected by each other & hence are correlated. the other example of correlated variable are –
3. While studying correlation between 2 variables use should make clear that there must be cause and effect relationship between these variables. for e.g. – when price of a certain commodity is changed (↑ or ↑) its demand also changed (↑ or ↑) so there is case & effect relationship between demand and price thus correlation exists between them. Take another eg. where height of students; as well as height of tree increases, then one cannot call it a case of correlation because neither height of students is affected by height of three nor height of tree is affected by height of students, so there is no cause & effect relationship between these 2 so no correlation exists between these 2 variables.
4. In correlation both the variables may be mutually influencing each other so neither can be designated as cause and the other effect for e.g. –

Price ↑→ Demand ↓

Demand ↓→ Price ↑

So, both price & demand are affected by each other therefore use cannot tell in real sense which one is cause and which one is cause and which one is effect.

**definitions of Correlation**

1. “If 2 or more quantities vary is sympathy, so that movements is one tend to be accompanied by corresponding movements in the other(s), then they are said to be correlated”. **Connor.**
2. “Correlation means that between 2 series or groups of data there exists some casual correction”. **WI King**
3. “Analysis of Correlation between 2 or more variables is usually called correlation.” **A.M. Turtle**
4. “Correlation analysis attempts to determine the degree of relationship between variables.

**Ya Lun chou**

**types of correlation**

Correlation

Positive Negative Correlation

Simple & Multiple Correlation

Partial & Total Correlation

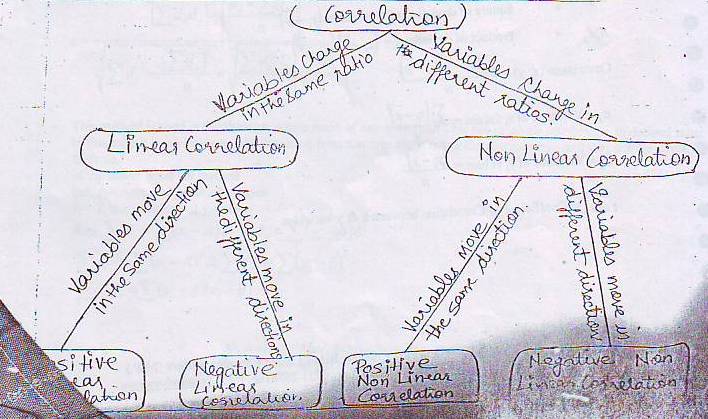
Liner & Non Linear Correlation

|  |  |  |
| --- | --- | --- |
|  | **Positive Correlation** | **Negative Correlation** |
| 1 | Value of 2 variables move in the same direction i.e. when increase/decrease in value of one variable will cause increase or decrease in value of other variable. | Value of 2 variables move in opposite direction i.e. when one variable increased, other variable decreases when one variable is decreased, other variable increase. |
| 2 | E.g. Supply & Price  So, supply and price are …….correlated  P = Price/Unit  Q = quantity Supplied | E.g. Demand & Price  So, Demand & Price vely correlated  P = Price/Unit  Q = quantity Supplied |

|  |  |  |
| --- | --- | --- |
|  | **simple Correlation** | **multiple Correlation** |
| 1 | In simple correlation, the relationship is confined to 2 variables only, i.e. the effect of only one variable is studied | The relationship between more than 2 variables is studied. |
| 2 | E.g. Demand & Price  Demand depends on → Price  This is case of simple correlation because relationship is confined to only one factor (that affects demand) i.e. price so we have to find correlation between demand & price.  If, demand = Y  If, demand – X  Then, Correlation between Y & X | E.g. Demand & Price  Demand depends on → Price  Demand on → income  This is case of multiple correlations because 2 factors (Price & Income) that affects demand are taken. We have to find correlation between demand & price.  Demand & Price  If, demand = Y  Price = X1  Price = X2  Then  Correlation between Y & X1  Correlation between Y & X2 |

|  |  |
| --- | --- |
| **simple Correlation** | **multiple Correlation** |
| In partial correlation though more than 2 factors are involved but correlation is studies only between to be constant.  E.g.  X1 Y = Demand  Y X1 = Price  X2 X2 = Income | In total correlation relationship between all the variables is studied i.e., none of item is assumed to be constant  E.g.  X1 Y = Demand  Y X1 = Price  X2 X2 = Income |
| If we study correlation between Y & X1 & assume X2 to be constant it is a case of partial correlation. this is what we do in law of demand – assume factors other than price as constant (Ceteris paribus – Keeping other things constant) | If we assume that income is not constant i.e. we study the effect of both price & income on demand, it is a case of total correlation.  In other words, cataris paribus assumption is relaxed in this case. |

|  |  |  |
| --- | --- | --- |
|  | **linear Correlation** | **non-linear Correlation** |
| 1 | In linear correlation, due to unit, change value of one variable there is constant change in the value of other variable. The graph for such a relationship is straight line. E.G. – If in a factory no of workers are doubled, the production output is also doubled, and correlation would be linear. | In non linear or curvilinear correlation, due to unit, change value of one variable, the change in the value of other variable is not constant. the graph for such a relationship is a curve. E.G. – The amount spent on advertisement will not bring the change in the amount of sales in the same ratio, it means the variation. |
| 2 | If the changed in 2 variables are in the same direction and in the constant ratio, it is linear positive correlation  Y  X   |  |  | | --- | --- | | X | Y | | 2 | 3 | | 4 | 6 | | 6 | 9 | | 8 | 12 | | If the change in 2 variables is in the same direction but not in constant ratio, the correlation is non linear positive.   |  |  | | --- | --- | | X | Y  Y  X | | 50 | 10 | | 55 | 12 | | 60 | 15 | | 90 | 30 | | 100 | 45 | |
| 3 | If changes in 2 variables are in the opposite direction but in constant ratio, the correlation is linear negative. For eg. every 5% ↑ is price of a good is associated with 10% decrease in demand the correlation between price and demand would be linear negative.  Y  X   |  |  | | --- | --- | | X | Y | | 2 | 21 | | 4 | 18 | | 6 | 15 | | 8 | 12 | | 10 | 9 | | If changes in 2 variables are in opposite direction and not in constant ratio, the correlation is non linear negative. For eg: - every 5% ↑in price of good is associated with 20% to 10%↓in demand, the correlation between price & demand would be non linear negative.  Y  X   |  |  | | --- | --- | | X | Y | | 80 | 50 | | 55 | 60 | | 50 | 75 | | 90 | 130 | |

****

**Type – 1 [Based on karl Pearson’s cofficient of correlation]**

Before use move to numerical, use understand the basic notions & concepts –

dx = Deviations of xi value from mean = (xi - )

x = Mean of x value [Average of X values] =

n = No. of observations

dy = Deviation of y value from mean = (y - )

= Mean of y values =

d2x = Square of deviation of x values = (xi - )2

d2y = Square of deviation of x values = (yi - )2

dxdy = Product of deviations = (xi - ) (yi - )

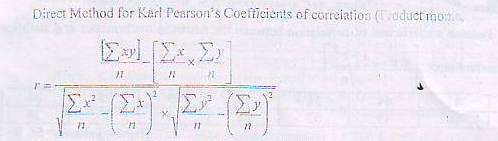
Covariance (x,y) =

σx = Variance of xi values =

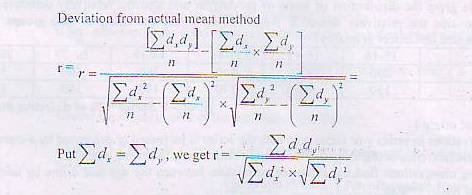
σy = Variance of yi values =

r or rxy = coefficient of correlation between x 7 y variables.

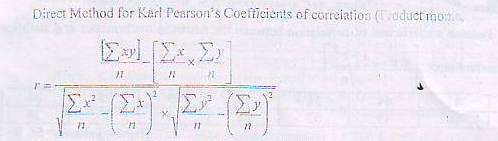
Direct Method for Karl Pearson’s Coefficient of correlation



Deviation from actual mean method



Deviation from assumed mean method (Short Cut Method)



This method is used in the situation where mean of any series (x or y) is not in whole number, i.e. in decimal value. in this case it is advisable to take deviation from assumed mean rather than actual mean and then use the above formula.

In the above short cut method

Let, A = Assumed mean of X series

B = Assumed mean of y series

then Σdx = Σ(xi – A) & Σdy = Σ(yi – B) &

Σdx 2= Σ(xi – A)2 & Σdy2= Σ(yi – B)2

Σdxdy= Σ(xi – A)(xi – B)

**Regression Analysis**

The dictionary meaning of regression is “Stepping Back”. The term was first used by a British Biometrician” Sir Francis Galton 1822 – 1911) is 1877. He found in his study the relationship between the heights of father & sons. In this study he described “That son deviated less on the average from the mean height of the race than their fathers, whether the father’s were above or below the average, son tended to go back or regress between two or more variables in terms of the original unit of the data.

**Meaning**

Regression Analysis is a statistical tool to study the nature extent of functional relationship between two or more variable and to estimate the unknown values of dependent variable from the known values of independent variable.

**Dependent Variables –** The variable which is predicted on the basis of another variable is called dependent or explained variable (usually devoted as y)

**Independent variable –** The variable which is used to predict another variable called independent variable (denoted usually as X)

**Definition**

Statistical techniques which attempts to establish the nature of the relationship between variable and thereby provide a mechanism for prediction and forecasting is known as regression Analysis.

**– Ya-lun-Chon”**

**Importance/uses of Regression Analysis**

* Forecasting
* Utility in Economic and business area
* Indispensible for goods planning
* Useful for statistical estimates.
* Study between more than two variable possible
* Determination of the rate of change in variable
* Measurement of degree and direction of correlation
* Applicable in the problems having cause and effect relationship
* Regression Analysis is to estimate errors
* Regression Coefficient (bxy & byx) facilitates to calculate of determination ® & coefficient or correlation (r)

**Regression Lines**

The lines of best fit expressing mutual average relationship between two variables are known as regression lines – there are two lines of regression

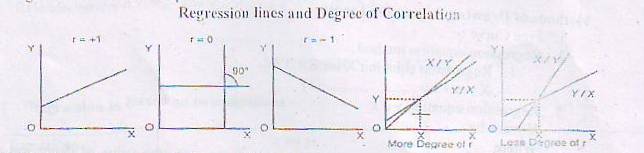
**Why are two Regression lines –**

1. While constructing the lines of regression of x on y is treated as independent variables where as ‘x’ is treated as treated as dependent variable. This gives most probable values of ‘X’ for gives values of y. the same will be there for y on x.

**Relationship between correlation & REgression**

1. When there is perfect correlation between two series (r = 1) the regression with coincide and there will be only one regression line.
2. When there is no correction (r = o)> Both the lines will cut each other at point.
3. Where there is more degree of correction, say (r = 70 or more the two regression line with be next to each other whereas when less degree of correction. Say (r= 10 on less) the two regression line will be a parted from each other.

**regression lines and degree of correlation**



**difference between Correlation and REgression Analysis**

the correlation and regression analysis, both, help us in studying the relationship between two variables yet they differ in their approach and objectives. The choice between the two depends on the purpose of analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Base** | **Correlation** | **Regression** |
| 1 | meaning | Correlation means relationship between two or more variables in which movement in one have corresponding movements in other | regression means step ping back or returning to the average value, i.e., it express average relationship between two or more variables. |
| 2 | relationship | correlation need not imply cause and effect relationship between the variables under study | regression analysis clearly indicates the cause and effect relationship. the variable(s) constituting causes(s) is taken as independent variables(s) and the variable constituting the variable consenting the effect is taken as dependent variable. |
| 3 | object | correlation is meant for co-variation of the two variables. the degree of their co-variation is also reflected in correlation. but correlation does not study the nature of relationship. | regression tells use about the relative movement in the variable. We can predict the value of one variable by taking into account the value of the other variable. |
| 4 | nature | there may be nonsense correlation of the variable has no practical relevance | There is nothing like nonsense regression. |
| 5 | measure | correlation coefficient is a relative measure of the linear relationship between X and Y. It is a pure number lying between 1 and +1 | the regression coefficient is absolute measure representing the change in the value of variable. We can obtain the value of the dependent variable. |
| 6 | application | correlation analysis has limited application as it is confined only to the study of linear relationship between the variables. | Regression analysis studies linear as well as non linear relationship between variables and therefore, has much wider application. |

**Why least square is the Best?**

When data are plotted on the diagram there is no limit to the number of straight lines that could be drawn on any scatter diagram. Obviously many lines would not fit the data and disregarded. If all the points on the diagram fall on a line, that line certainly would the best fitting line but such a situation is rare and ideal. Since points are usually scatters, we need a criterion by which the best fitting line can be determined.

**Methods of Drawing Regression Lines –**

1. Free curve –
2. regression equation x on y,

X = a + by …………………………….(1)

1. regression equation y on x

y = a + bx

where

‘a’ is that point where regression lines touches y axis (the value of dependent variable value when value or independent variable is zero)

‘b’ is the slop of the said line (The amount of change in the value of the dependent variable per unit change)

Change in independent variable)

A and b constants can be calculated through –

Σ(x = a + by) (by multiplying ‘Σ’)

Σx = Na + bΣy (1)

Σx (y = a + bx) (by multiplying Σx)

Σxy = Σxa + bΣx2 (2)

**kinsds of regression analysis**

1. linear and Non- Linear Regression
2. Simple and Multiple Regression

**functions of regression lines –**

1. to make the best estimate –
2. to indicate the nature and extent of correlation

**regression equations –**

the regression equation’s express the regression lines, as there are two regression lines there are two regression equations –

Explanation is given in formulae –

**regression lines**

1. regression equation of x on y

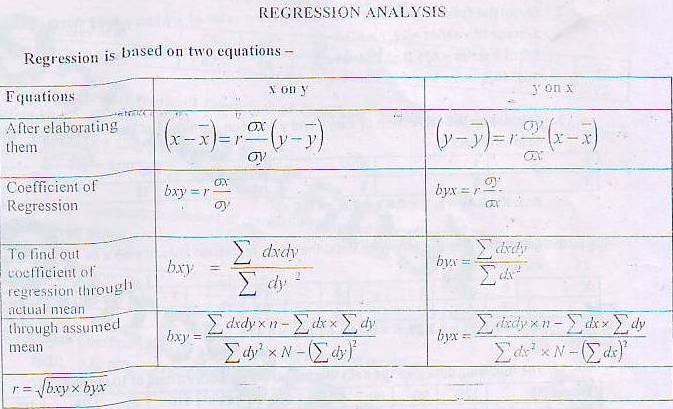
x – X = bxy (y – y)

where bxy = regression coefficient of X on Y

1. Regression euation of y on x

y – Y = bxy (x – x) where bxy = regression coefficient of Y on X



****

**regression coefficient –** There are two regression coefficient like regression equation, they are (bxy and byx)

Properties of regression coefficients –

* Same sign – Both coefficient have the same either positive on negative
* Both cannot by greater than one – If one Regression is greater than “One” or unity. Other must be less than one.
* Independent of origin – Regression coefficient are independent of origin but not of scale.
* A.M.> ‘r’ – mean of regression coefficient is greater than ‘r’
* R is G.M. – Correlation coefficient is geometric mean between the regression coefficient
* R, bxy and bxy – They all have same sign

**UNIT-V**

**ProBABILITY**

## Introduction

Probability has a very old history, it was originated in the games of chance related to gambling. For instance, throwing of dice or coin and drawing cards from a pack. Jerome Cardan (1501~1576), an Italian mathematician was the first man to write a book on the subject “Book on Games of chance” which was published in 1663 after his death. The probability formulae and techniques were developed by Jacob Bernoulli(1654-1705), De Moivre (1667-1754), Thomas Bayes(1702-1761) and Joseph Lagrange(1736- 1813). Pierre Simon, Laplace in the nineteenth century unified all these early ideas and compiled the first general theory of probability.

In the beginning, the probability theory was successfully applied at the gambling tables. But after some time, it was applied in the solution of social, political, economic and business problems. In fact, it has become a part of our everyday lives. We face uncertainty in personal and management decisions and use probability theory. Probability constitutes the foundation of statistical theory.

## Approaches

There are mainly three approaches to probability

1. Classical approach
2. Empirical approach
3. Axiomatic approach

Few terms can be defined / explained with reference to simple experiments relating to tossing of coins, throwing of a diee or drawing cards from a pack of cards.

Random Experiment

An experiment can be considered as a random experiment if all the possible outcomes are known in advance and none of the outcomes can be predicted with certainty. e.g. throwing a dice, tossing a coin etc.

Trial & Event

When a random experiment is performed, it is called a trial and outcome or combinations of outcomes are termed as events. For example

1. When a coin is tossed repeatedly, the result is not unique. We may get any of the two faces; head or tail. Thus, throwing a coin is a random experiment and getting of a head or tail is an event.
2. In the similar manner, when a dice is thrown, it is called a random experiment. Getting anyof the faces 1, 2, 3, 4, 5 or 6 is an event. Getting an odd no. or an even no., getting no. greater than 3 or lower than five, these are called events.
3. Similarly, drawing of two balls from an urn containing ‘a’ red balls and ‘b’ white balls is a trial and getting of both red balls, or both white balls, or one red and one white ball are events.

Exhaustive Cases

When a random experiment is done, there are some outcomes; the total numbers of possible outcomes are called exhaustive cases for the experiment. For e.g. when a coin is tossed, we can get head (H) or tail (T). Hence exhaustive no. of cases is 2 (i.e. H,T) If two coins are tossed, the various possibilities areHH, HT, TH, TT (number of exhaustive cases are four) where HT means, Head on first coin and Tail on second coin, and TH means, Tail on first coin and Head on second coin and so on.

In case of toss of three coins, number of outcomes is

= (H,T) x (H,T) x(H,T)

= (HH,HT,TH,TT) x (H,T)

= (HHH, HHT, HTH, HTT, THH, THT, TTH, TTT)

No. of possible outcomes is 8 = 23. In general, in a throw of n coins, the exhaustive no. of cases is 2n.

In a throw of a die, exhaustive number of cases is 6, since we can get any one of the six faces marked 1, 2, 3, 4, 5 and 6. If two dice are thrown, the possible outcomes are

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (1,1) | (1,2) | (1,3) | (1,4) | (1,5) | (1,6) |
| (2,1) | (2,2) | (2,3) | (2,4) | (2,5) | (2,6) |
| (3,1) | (3,2) | (3,3) | (3,4) | (3,5) | (3,6) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (4,1) | (4,2) | (4,3) | (4,4) | (4,5) | (4,6) |
| (5,1) | (4,2) | (4,3) | (4,4) | (4,5) | (4,6) |
| (6,1) | (4,2) | (4,3) | (4,4) | (4,5) | (4,6) |

* 1. total no. is 36, where (i,j) means number i on the first die and j on the second die, i and j both taking the values from 1 to 6. In the case of throw of two dice, no. of possible outcome = 62 = 36 and in the case of throw of three dice, no. of possible outcome = 63 = 216; in the case of throw of n dice, no. of possible outcome = 6n

Favourable cases or events

The number of outcomes of a random experiment which result in the happening of anevent are named as the cases favourable to the event.

* + 1. When a toss of two coins takes place, the no. of cases favourable to the event ‘exactly one head is two (i.e. TH or HT) and for getting ‘two heads’ is one (i.e. HH)
    2. When a card is drawn from a pack of cards, the no. of cases favourable to the event ‘getting a diamond’ are 13 and getting ‘an ace of spade’ is one.

Mutually Exclusive events or cases

Two or more events are considered as mutually exclusive if the happening of any one of them excludes the happening of all others in the same experiment. For example, in toss of a coin, the events ‘head’ and ‘tail’ are mutually exclusive because if head comes, we can’t get tail and if tail comes we can’t get head. Similarly, in the throw of a die, the six faces numbered 1, 2, 3, 4, 5 and 6 are mutually exclusive. Thus, events are said to be mutually exclusive of no two or more of them can happen simultaneously.

Classical/priori Probability

It is the oldest and simplest approach. Under this approach, there is no need to physically perform the experiment. The basic assumption is that the outcomes of a random experiment are equally likely. e.g. in a throw of a dice, occurrence of 1,2,3,4,5,6 are equally likely event.

If a random experiment results in N exhaustive, mutually exclusive and equally likely outcomes out of which mare favourable to the happening of an event X then the probability of occurrence of X i.e. P(X) is given by

P(X) =

*Favourable Exhaustive*

*cases cases*

*m*

= *N*

**Example 1.** Abag containing 10 green and 20 red balls. Aball is drawn at random. What is the probability that it is green.

**Sol.** Total number of balls in the bag = 10+20 = 30 Number of green balls = 10

Probabilityof getting a green ball =

10 1

= 30 = 3

Empirical Probability

The classical definition is difficult to apply as soon as we move from the field of coins, cards, dice and other games of chance. It may not explain the actual results in certain cases e.g if a coin is tossed 20 times , we may get 14 heads and 6 tails. The probability of head is thus 0.7 and tail is 0.3. However, if experiment is carried out large number of times, we should expect approximately equal number of heads and tails.

If an experiment is performed **repeatedly** under essentially homogeneous and identical conditions then the limiting value of the ratio of the number of times the event occurs to the number of trials, as the number of trials become indefinitely large is known as the probability of happening of the event.

*m*

P(X) = lim *N*

N  

AxiomaticApproach

The axiomatic Probability theory is an attempt at constructing a theory of probability which is free from inadequacies of both the classical and empirical approaches. It plays an important role in rendering a reasonable amount of comprehensibility and tractability to the understanding of chance phenomenon at least in the initial stages of any scientific inquiry into their structure and composition where other approaches are less comprehensible and tractable.

## Addition Law

The probability of occurrence of either event A or event B of two mutually exclusive events is equal to the sum of their individual probability.

Mathematically, we can represent as P(A  B) = P(A) + P (B)

**Proof :-** If an event Acan happen in a ways and B in a ways then

1 2

The number of ways in which either event can happen in a + a ways.

1 2

Total number of possibilities is n.

Then by definition , the probability of either the first or second event happening is

a1 + a 2 =

n

a1 + a 2

n n

a1

But

n

= P(A)

And a 2

n

= P(B)

Hence P(A  B) = P(A) + P (B)

The theorem can be extended to three or more mutually exclusive events, thus P(A B C) = P(A) + P (B) + P(C)

**Example 2.** A deck of 52 cards, one card is drawn. What is the probability that it is either a king or a queen?

**Sol.** There is four kings and four queens in a pack of 52 cards.

The probability of drawing a card that is king = 4

52

The probability of drawing a card that is queen = 4

52

Since the events are mutually exclusive, the probability that the card drawn is either a king or

a queen = 4 +

52

4 = 2

52 13

If two events A & B are not **mutually exclusive** (joint events) then the addition law can be stated as follows

The probability of the occurrence of either event A or event B or both is equal to the probability that event Aoccurs, plus the probability that event B occurs minus the probability that both events occur. I can be shown as

P(A  B) = P(A) + P (B) – P (A  B)

**Example 3.** The managing committee of Residents Welfare Association formed a sub-committee of 5 persons to look into electricity problem. Profiles of the 5 persons are

|  |  |
| --- | --- |
| Male age | 40 |
| Male age | 43 |
| Female age | 38 |
| Female age | 27 |
| Male age | 65 |

If a chairperson has to be selected from this, what is the probability that he would be either female or over 32 years.

**Sol.** P (female or over 32) = P(female) + P(over 32) – P(female and over 32)

2 4 1

= + - = 1 5 5 5

**Example 4.** What is the probability of picking a card that was a heart or a spade.

**Sol.** Using the addition rule,

P(heart or spade) = P(heart) + P(spade) – P(heart and spade)

= 13 + 13 - 1 = 1

52 52 2 2

## Multiplication Law

It states that if two events A and B are independent , the probability that they both will occur is equal to the product of their individual probability.

P(A and B) = P(A) X P(B)

It can be extended to three or more independent events. P(A,B and C) = P(A) X P(B) X P(C)

**Proof :-** If an event A can happen in n1ways of which a1 are successful and B in n2ways of which a2 are successful then

The number of successful happening in both cases is a1x a2 . Total number of possibilities is n1 x n2.

Then by definition , the probability of the occurrence of both events is

a1 x a2 =

n1 x n2

a1 x a2 n n

But

and

a1 = P(A)/n

a2 = P(B)/n

Hence P(A and B) = P(A) x P (B)

**Example 5.** In order to marry with a girl, a man wants these qualities White complexion – the probability of getting this is one in fifty.

Etiquettes – the probability is one in hundred.

Dowry – the probability of getting this is one in Twenty.

Calculate the probability of his getting married to such a girl when the possession of these three attributes is independent.

**Sol.** Probability of a girl with white complexion =

Probability of a girl with handsome dowry =

1 = 0.05

20

1 = 0.02

50

Probability of a girl with etiquettes =

1 = 0.01

100

The probability of simultaneously occurrence of all these qualities = 0.05 x 0.02 x 0.01

= 0.00001

## Conditional Probability

The multiplication theorem described above is not applicable in case of dependent events. Two events A and B are said to be dependent when B can occur only when A is known to have occurred. The probability attached to such an event is called conditional probability and is denoted by P(A/B) i.e. probability of A given that B has occurred.

*Similary*,

*P*( *A* / *B*)  *P*( *A*  *B*) / *P*(*B*);

*P*(*B* / *A*)  *P*( *A*  *B*) / *P*( *A*);

*P*(*B*)  0

*P*( *A*)  0

*Symbolically*, *we write*

*P*( *A*  *B*)  *P*( *A* / *B*)  *P*(*B*)

*or P*(*B*  *A*)  *P*(*B* / *A*)  *P*( *A*)