**Study Material (Lecture Notes)**

**Business Statistics**

**107 QUAN**

**Instructor’s Name**

*Md. Izhar Alam, PhD*

*Assistant Professor*

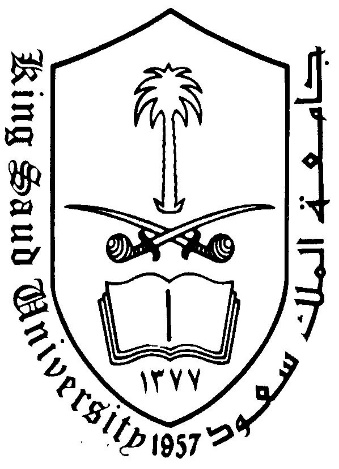
*Department of Finance*

*College of Business Administration*

*King Saud University, Muzahimiyah*

*E- mail:* [*mialam@ksu.edu.sa*](mailto:mialam@ksu.edu.sa)

*Mob No: + 966 536108067*



KING SAUD UNIVERSITY

(MUZAHIMIYAH)

**Main Objectives of the Course Specification**

*Business statistics teaches students to extract the best possible information from data in order to aid decision making, particularly in terms of sales forecasting, quality control and market research. You are also taught to determine the type of data which is needed, the way it should be collected and how it should be analyzed. After this course, you should be able to express a generally question as a statistical one, to use statistical tools for relevant calculations, and to apply graphical techniques for displaying data. The course will focus on descriptive statistics. Indeed, the main objective of Business Statistics is to describe data and make evidence based decisions using inferential statistics. This course should lead you to perform statistical analyses, interpret their results, and make inferences about the population from sample data.*

|  |  |  |
| --- | --- | --- |
| List of Topics | No. of  Weeks | Contact Hours |
| **Data and Variables:** Collection of Data; Sampling and Sample Designs;Classification and Tabulation of Data;Diagrammatic and Graphic Presentation; | 1 | 3 |
| **Descriptive measures:**  Central Tendency- Mean, Median, Mode, Variation, Shape, Covariance,  Mean Deviation and Standard Deviation, Coefficient of Correlation | 4 | 12 |
| **Discrete probability distributions:**  probability distribution for a discrete random variable, binomial distribution, Poisson distribution  **Continuous probability distribution:** Normal distribution | 3 | 9 |
| **Confidence interval estimation** | 1 | 3 |
| **Chi-square tests:**  Chi-square test for the difference between two proportions, Chi-square test for differences among more than two proportions, Chi-square test of independence | 2 | 6 |
| **Simple Linear Regression** | 2 | 6 |
| **Multiple Regression** | 2 | 6 |

**Recommended Textbooks:**

1. David M. Levine, Timothy C. Krehbiel, & Mark L Berenson, Business Statistics: A First Course plus MyStatLab with Pearson eText -- Access Card Package, Pearson.
2. Anderson, D. R., Sweeney, D. J., & Williams, T. A. Essentials of Modern Business Statistics with Microsoft Office Excel, South-Western: Mason, OH.
3. Berenson, ML, Levine, D, Krehbiel, TC, Watson, J, Jayne, N & Turner, LW. Business Statistics: Concepts and Applications, Pearson Education, Frenchs Forest, New South Wales.
4. Groebner, DF, Shannon, PW, Fry, PC & Smith, KD. Business Statistics: A Decision-making Approach, Prentice Hall, Harlow, England.
5. Keller, G. Statistics for Management and Economics, South-Western Cengage Learning, Belmont, California.

**Chapter- 1**

# **Statistics: Introduction**

A set of numbers collected to study particular situations is known as **data**. These data are presented in systematic form in order to draw some direct inferences from the same. Also some other terms and quantities are calculated from the data to make better interpretations.

The study associated with all of the above is called **statistics**. Therefore, statistics contains collection and presentation of data, analyzing the data on the basis of the measures of central value, dimension etc.

The purpose to study business statistics in this course is to understand the basic statistical methods that are useful in decision making.

**Basic Definitions**

* **Statistics:** The collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.
* **Data:** A set of numbers collected to study particular situations is known as data. It refers to any group of measurements that happen to interest us. These measurements provide information the decision maker uses.
* **Primary Data:** Primary data are measurements observed and recorded as part of original study. These are data not available elsewhere.
* **Secondary Data:** Data which are not originally collected but rather obtained from published or unpublished sources are called secondary data.
* **Variable:** Characteristic or attribute that can assume different values at different times, places or situations.
* **Random Variable:** A variable whose values are determined by chance.
* **Population:** All subjects possessing a common characteristic that is being studied.
* **Sample:** A sub- group or sub- set of the population.
* **Parameter:** Characteristic or measure obtained from a population.
* **Statistic** (*not to be confused with Statistics*): Characteristic or measure obtained from a sample.
* **Descriptive Statistics:** Collection, organization, summarization, and presentation of data.
* **Inferential Statistics**: Generalizing from samples to populations using probabilities. Performing hypothesis testing, determining relationships between variables, and making predictions.
* **Qualitative Variables:** Variables which assume non-numerical values.
* **Quantitative Variables:** Variables which assume numerical values.
* **Discrete Variables:** Variables which assume a finite or countable number of possible values. Usually obtained by counting.
* **Continuous Variables:** Variables which assume an infinite number of possible values. Usually obtained by measurement.
* **Nominal Level:** Level of measurement which classifies data into mutually exclusive, all-inclusive categories in which no order or ranking can be imposed on the data.
* **Ordinal Level:** Level of measurement which classifies data into categories that can be ranked. Differences between the ranks do not exist.
* **Interval Level:** Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.
* **Ratio Level:** Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.

**Collection of Data**

Data may be obtained either from the primary source or the secondary source. A primary source is one that itself collects the data whereas a secondary source is one that makes available data which were collected by some other agency.

**Choice between Primary and Secondary Data:** the investigator must decide at the outset whether he will use primary data or secondary data in an investigation. The choice between the two depends mainly on the following considerations:

* Nature and scope of the enquiry;
* Availability of time;
* Degree of accuracy desired; and
* The collecting agency, i.e., whether an individual, an institute or a Government body.

It may be pointed out that most statistical analysis rests upon secondary data. Primary data are generally used in those cases where the secondary data do not provide an adequate basis for analysis.

**Methods of Collecting Primary Data:**

* Direct personal interviews;
* Indirect oral interviews;
* Information from correspondents;
* Mailed questionnaire method; and
* Schedules sent through enumerators.

**Sources of Secondary Data:**

* Published sources; and
* Unpublished sources

**Editing Primary and Secondary Data:**

Once the data have been obtained either from primary or secondary source, the next step in a statistical investigation is to edit the data, i.e., to scrutinize the data. While editing primary data the following considerations need attention:

* The data should be complete;
* The data should be consistent;
* The data should be accurate; and
* The data should be homogeneous.

**Precautions in the Use of Secondary Data:**

* Whether the data are suitable for the purpose of investigation;
* Whether the data are adequate for investigation; and
* Whether the data are reliable or not.

**Sampling and Sample Designs**

When secondary data are not available for the problem under study, a decision may be taken to collect primary data. The required information may be obtained by following either the census method or the sample method.

**Census Method:**

Information on population can be collected in two ways – census method and sample method. In census method every element of the population is included in the investigation. For example, if we study the average annual income of the families of a particular village or area, and if there are 1000 families in that area, we must study the income of all 1000 families. In this method no family is left out, as each family is a unit.

**Merits and limitations of Census method:**

**Mertis:**

* + - 1. The data are collected from each and every item of the population
      2. The results are more accurate and reliable, because every item of the universe is required.
      3. Intensive study is possible.
      4. The data collected may be used for various surveys, analyses etc.

**Limitations:**

1. It requires a large number of enumerators and it is a costly method
2. It requires more money, labour, time energy etc.
3. It is not possible in some circumstances where the universe is infinite.

## **Sample:**

Statisticians use the word **sample** to describe a portion chosen from the population. A finite subset of statistical individuals defined in a population is called a sample. The number of units in a sample is called the **sample size.**

## **Sampling frame:**

For adopting any sampling procedure it is essential to have a list identifying each sampling unit by a number. Such a list or map is called sampling frame. A list of voters, a list of house holders, a list of villages in a district, a list of farmers etc. are a few examples of sampling frame.

## **Principles of Sampling:**

Samples have to provide good estimates. The following principle tell us that the sample methods provide such good estimates

## **Principle of statistical regularity:**

A moderately large number of units chosen at random from a large group are almost sure on the average to possess the characteristics of the large group.

## **Principle of Inertia of large numbers:**

Other things being equal, as the sample size increases, the results tend to be more accurate and reliable.

## **Principle of Validity:**

This states that the sampling methods provide valid estimates about the population units (parameters).

## **Principle of Optimization:**

This principle takes into account the desirability of obtaining a sampling design which gives optimum results. This minimizes the risk or loss of the sampling design.

The foremost purpose of sampling is to gather maximum information about the population under consideration at minimum cost, time and human power.

## **Types of Sampling:**

The technique of selecting a sample is of fundamental importance in sampling theory and it depends upon the nature of investigation. The sampling procedures which are commonly used may be classified as

1. Probability sampling.
2. Non-probability sampling.
3. Mixed sampling.

## **Probability sampling (Random sampling):**

A probability sample is one where the selection of units from the population is made according to known probabilities. (eg.) Simple random sample, probability proportional to sample size etc.

## **Non-Probability sampling:**

It is the one where discretion is used to select ‘representative’ units from the population (or) to infer that a sample is ‘representative’ of the population. This method is called **judgement or purposive** sampling. This method is mainly used for opinion surveys; A common type of judgement sample used in surveys is quota sample. This method is not used in general because of prejudice and bias of the enumerator. However if the enumerator is experienced and expert, this method may yield valuable results. For example, in the market research survey of the performance of their new car, the sample was all new car purchasers.

## **Mixed Sampling:**

Here samples are selected partly according to some probability and partly according to a fixed sampling rule; they are termed as mixed samples and the technique of selecting such samples is known as **mixed sampling**.

## **Methods of selection of samples:**

Here we shall consider the following three methods:

1. Simple random sampling.
2. Stratified random sampling.
3. Systematic random sampling.

## **Simple random sampling:**

A simple random sample from finite population is a sample selected such that each possible sample combination has equal probability of being chosen. It is also called unrestricted random sampling.

## **Simple random sampling without replacement:**

In this method the population elements can enter the sample only once (ie) the units once selected is not returned to the population before the next draw.

## **Simple random sampling with replacement:**

In this method the population units may enter the sample more than once. Simple random sampling may be with or without replacement.

**Frequency Distribution**

# **Introduction:**

Frequency distribution is a series when a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order of magnitude in a series. It is simply a table in which the data are grouped into classes and the number of cases which fall in each class are recorded. It shows the frequency of occurrence of different values of a single Phenomenon.

# **A frequency distribution is constructed for three main reasons:**

* + 1. To facilitate the analysis of data.
    2. To estimate frequencies of the unknown population distribution from the distribution of sample data and
    3. To facilitate the computation of various statistical measures

# **Raw data:**

The statistical data collected are generally raw data or ungrouped data. Let us consider the daily wages (in SR) of 30 laborers in a factory.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 80 | 70 | 55 | 50 | 60 | 65 | 40 | 30 | 80 | 90 |
| 75 | 45 | 35 | 65 | 70 | 80 | 82 | 55 | 65 | 80 |
| 60 | 55 | 38 | 65 | 75 | 85 | 90 | 65 | 45 | 75 |

The above figures are nothing but raw or ungrouped data and they are recorded as they occur without any pre consideration. This representation of data does not furnish any useful information and is rather confusing to mind. A better way to express the figures in an ascending or descending order of magnitude and is commonly known as array. But this does not reduce the bulk of the data. The above data when formed into an array is in the following form:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 30 | 35 | 38 | 40 | 45 | 45 | 50 | 55 | 55 | 55 |
| 60 | 60 | 65 | 65 | 65 | 65 | 65 | 65 | 70 | 70 |
| 75 | 75 | 75 | 80 | 80 | 80 | 80 | 85 | 90 | 90 |

The array helps us to see at once the maximum and minimum values. It also gives a rough idea of the distribution of the items over the range. When we have a large number of items, the formation of an array is very difficult, tedious and cumbersome. The Condensation should be directed for better understanding and may be done in two ways, depending on the nature of the data.

**Example:**

***In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0 | 3 | 2 | 1 | 5 | 6 | 2 |
| 2 | 1 | 0 | 3 | 4 | 2 | 1 | 6 |
| 3 | 2 | 1 | 5 | 3 | 3 | 2 | 4 |
| 2 | 2 | 3 | 0 | 2 | 1 | 4 | 5 |
| 3 | 3 | 4 | 4 | 1 | 2 | 4 | 5 |

***Represent the data in the form of a discrete frequency distribution.***

**Solution:**

Frequency distribution of the number of children

|  |  |  |
| --- | --- | --- |
| **Number of** | **Tally** | **Frequency** |
| **Children** | **Marks** |  |
| 0 |  | 3 |
| 1 |  | 7 |
| 2 |  | 10 |
| 3 |  | 8 |
| 4 |  | 6 |
| 5 |  | 4 |
| 6 |  | 2 |
|  | Total | 40 |

**b) Continuous frequency distribution:**

In this form of distribution refers to groups of values. This becomes necessary in the case of some variables which can take any fractional value and in which case an exact measurement is not possible. Hence a discrete variable can be presented in the form of a continuous frequency distribution.

Wage distribution of 100 employees

|  |  |
| --- | --- |
| **Weekly wages** | **Number of** |
| **(SR)** | **employees** |
| 50-100 | 4 |
| 100-150 | 12 |
| 150-200 | 22 |
| 200-250 | 33 |
| 250-300 | 16 |
| 300-350 | 8 |
| 350-400 | 5 |
| Total | 100 |

**Nature of class:**

The following are some basic technical terms when a continuous frequency distribution is formed or data are classified according to class intervals.

**a)** **Class limits:**

The class limits are the lowest and the highest values that can be included in the class. For example, take the class 30-40. The lowest value of the class is 30 and highest class is 40. In statistical calculations, lower class limit is denoted by L and upper class limit by U.

**b) Class Interval:**

The class interval may be defined as the size of each grouping of data. For example, 50-75, 75-100, 100-125…are class intervals. Each grouping begins with the lower limit of a class interval and ends at the lower limit of the next succeeding class interval

**c) Width or size of the class interval:**

The difference between the lower and upper class limits is called Width or size of class interval and is denoted by ‘C’.

**d) Range:**

The difference between largest and smallest value of the observation is called The Range and is denoted by ‘R’ ie

R = Largest value – Smallest value

* 1. =L - S

1. **Mid-value or mid-point:**

The central point of a class interval is called the mid value or mid-point. It is found out by adding the upper and lower limits of a class and dividing the sum by 2.

i.e., Mid- Value =

For example, if the class interval is 20-30 then the mid-value is = = 25

**f) Frequency:**

Number of observations falling within a particular class interval is called frequency of that class.

Let us consider the frequency distribution of weights if persons working in a company.

|  |  |
| --- | --- |
| **Weight** | **Number of** |
| **(in kgs)** | **persons** |
| 30-40 | 25 |
| 40-50 | 53 |
| 50-60 | 77 |
| 60-70 | 95 |
| 70-80 | 80 |
| 80-90 | 60 |
| 90-100 | 30 |
| Total | 420 |

In the above example, the class frequencies are 25,53,77,95,80,60,30. The total frequency is equal to 420. The total frequencies indicate the total number of observations considered in a frequency distribution.

In the above example, the class frequencies are 25,53,77,95,80,60,30. The total frequency is equal to 420. The total frequencies indicate the total number of observations considered in a frequency distribution.

**g) Number of class intervals:**

The number of class interval in a frequency is matter of importance. The number of class interval should not be too many. For an ideal frequency distribution, the number of class intervals can vary from 5 to 15. To decide the number of class intervals for the frequency distribution in the whole data, we choose the lowest and the highest of the values. The difference between them will enable us to decide the class intervals.

Thus the number of class intervals can be fixed arbitrarily keeping in view the nature of problem under study or it can be decided with the help of Sturges’ Rule. According to him, the number of classes can be determined by the formula

K = 1 + 3. 322 log10 N

Where N = Total number of observations; log = logarithm of the number

K = Number of class intervals.

Thus if the number of observation is 10, then the number of class intervals is

K = 1 + 3. 322 log 10 = 4.322 @ 4

If 100 observations are being studied, the number of class interval is

K = 1 + 3. 322 log 100 = 7.644 @ 8 and so on.

**h) Size of the class interval:**

Since the size of the class interval is inversely proportional to the number of class interval in a given distribution. The approximate value of the size (or width or magnitude) of the class interval ‘C’ is obtained by using Sturges’ rule as

Size of Class- Interval, C =

Range

=

1+3.322 log10 N

Where Range = Largest Value – Smallest Value in the distribution.

**Types of class intervals:**

There are three methods of classifying the data according to class intervals namely

* 1. Exclusive method
  2. Inclusive method
  3. Open-end classes

1. **Exclusive method:**

When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class; it is known as the exclusive method of classification.

**Example:**

|  |  |
| --- | --- |
| **Expenditure** | **No. of families** |
| **(SR)** |  |
| 0 -5000 | 60 |
| 5000-10000 | 95 |
| 10000-15000 | 122 |
| 15000-20000 | 83 |
| 20000-25000 | 40 |
| Total | 400 |

It is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class. In the above example, there are so families whose expenditure is between SR.0 and SR.4999.99. A family whose expenditure is SR.5000 would be included in the class interval 5000-10000. This method is widely used in practice.

**b) Inclusive method:**

In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval.

**Example:**

|  |  |
| --- | --- |
| **Class interval** | **Frequency** |
| 5-9 | 7 |
| 10-14 | 12 |
| 15-19 | 15 |
| 20-29 | 21 |
| 30-34 | 10 |
| 35-39 | 5 |
| Total | 70 |

Thus to decide whether to use the inclusive method or the exclusive method, it is important to determine whether the variable under observation in a continuous or discrete one.

In case of continuous variables, the exclusive method must be used. The inclusive method should be used in case of discrete variable.

**c) Open end classes:**

A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. The necessity of open end classes arises in a number of practical situations, particularly relating to economic and medical data when there are few very high values or few very low values which are far apart from the majority of observations.

**Example:**

|  |  |
| --- | --- |
| **Salary Range** | **No of** |
|  | **workers** |
| Below 2000 | 7 |
| 2000 – 4000 | 5 |
| 4000 – 6000 | 6 |
| 6000 – 8000 | 4 |
| 8000 and | 3 |
| above |  |

**Construction of frequency table:**

Constructing a frequency distribution depends on the nature of the given data. Hence, the following general consideration may be borne in mind for ensuring meaningful classification of data.

1. The number of classes should preferably be between 5 and 20. However there is no rigidity about it.
2. As far as possible one should avoid values of class intervals as 3,7,11, 26….etc. preferably one should have class-intervals of either five or multiples of 5 like 10, 20, 25, 100 etc.
3. The starting point i.e. the lower limit of the first class, should either be zero or 5 or multiple of 5.
4. To ensure continuity and to get correct class interval we should adopt “exclusive” method.
5. Wherever possible, it is desirable to use class interval of equal sizes.

**Preparation of frequency table:**

**Example 1:**

Let us consider the weights in kg of 50 college students.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 42 | 62 | 46 | 54 | 41 | 37 | 54 | 44 | 32 | 45 |
| 47 | 50 | 58 | 49 | 51 | 42 | 46 | 37 | 42 | 39 |
| 54 | 39 | 51 | 58 | 47 | 64 | 43 | 48 | 49 | 48 |
| 49 | 61 | 41 | 40 | 58 | 49 | 59 | 57 | 57 | 34 |
| 56 | 38 | 45 | 52 | 46 | 40 | 63 | 41 | 51 | 41 |

Here the size of the class interval as per Sturges’ rule is obtained as follows

Size of Class Interval, C =

=

= = 5

Thus the number of class interval is 7 and size of each class is 5. The required size of each class is 5. The required frequency distribution is prepared using tally marks as given below:

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **Tally marks** | **Frequency** |
| 30-35 |  | 2 |
| 35-40 |  | 6 |
| 40-45 |  | 12 |
| 45-50 |  | 14 |
| 50-55 |  | 6 |
| 55-60 |  | 6 |
| 60-65 |  | 4 |
| Total |  | 50 |

**Example 2:**

***Given below are the number of tools produced by workers in a factory.***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 43 | 18 | 25 | 18 | 39 | 44 | 19 | 20 | 20 | 26 |
| 40 | 45 | 38 | 25 | 13 | 14 | 27 | 41 | 42 | 17 |
| 34 | 31 | 32 | 27 | 33 | 37 | 25 | 26 | 32 | 25 |
| 33 | 34 | 35 | 46 | 29 | 34 | 31 | 34 | 35 | 24 |
| 28 | 30 | 41 | 32 | 29 | 28 | 30 | 31 | 30 | 34 |
| 31 | 35 | 36 | 29 | 26 | 32 | 36 | 35 | 36 | 37 |
| 32 | 23 | 22 | 29 | 33 | 37 | 33 | 27 | 24 | 36 |
| 23 | 42 | 29 | 37 | 29 | 23 | 44 | 41 | 45 | 39 |
| 21 | 21 | 42 | 22 | 28 | 22 | 15 | 16 | 17 | 28 |
| 22 | 29 | 35 | 31 | 27 | 40 | 23 | 32 | 40 | 37 |

***Construct frequency distribution with inclusive type of class interval. Also find.***

1. ***How many workers produced more than 38 tools?***
2. ***How many workers produced less than 23 tools?***

**Solution:**

Using Sturges’ formula for determining the number of class intervals, we have

Number of class intervals = 1+ 3.322 log10N

* 1+ 3.322 log10100
* 7.6

Sizes of class interval =

= = 5

Hence taking the magnitude of class intervals as 5, we have 7 classes 13-17, 18-22… 43-47 are the classes by inclusive type. Using tally marks, the required frequency distribution is obtain in the following table-

|  |  |  |
| --- | --- | --- |
| **Class** | **Tally Marks** | **Number of** |
| **Interval** |  | **tools produced** |
|  |  | **(Frequency)** |
| 13-17 |  | 6 |
| 18-22 |  | 11 |
| 23-27 |  | 18 |
| 28-32 |  | 25 |
| 33-37 |  | 22 |
| 38-42 |  | 11 |
| 43-47 |  | 7 |
| **Total** |  | **100** |

**Cumulative frequency table:**

**Example:**

|  |  |  |  |
| --- | --- | --- | --- |
| Age group (in yrs) | No of Women | Less than Cumulative frequency | More than Cumulative frequency |
| 15-20 | 3 | 3 | 64 |
| 20-25 | 7 | 10 | 61 |
| 25-30 | 15 | 25 | 54 |
| 30-35 | 21 | 46 | 39 |
| 35-40 | 12 | 58 | 18 |
| 40-45 | 6 | 64 | 6 |

**Less than cumulative frequency distribution table**

|  |  |
| --- | --- |
| End values upper | Less than Cumulative |
| limit | frequency |
| Less than 20 | 3 |
| Less than 25 | 10 |
| Less than 30 | 25 |
| Less than 35 | 46 |
| Less than 40 | 58 |
| Less than 45 | 64 |

**(b) More than** **cumulative frequency distribution table**

|  |  |
| --- | --- |
| End values lower | Cumulative frequency |
| limit | more than |
| 15 and above | 64 |
| 20 and above | 61 |
| 25 and above | 54 |
| 30 and above | 39 |
| 35 and above | 18 |
| 40 and above | 6 |

**Conversion of cumulative frequency to simple Frequency:**

If we have only cumulative frequency ‘either less than or more than’, we can convert it into simple frequencies. For example if we have ‘less than Cumulative frequency, we can convert this to simple frequency by the method given below:

|  |  |  |
| --- | --- | --- |
| **Class interval** | **‘ less than’** | **Simple frequency** |
|  | **Cumulative frequency** |  |
| 15-20 | 3 | 3 |
| 20-25 | 10 | 10 - 3=7 |
| 25-30 | 25 | 25 - 10 = 15 |
| 30-35 | 46 | 46 - 25 = 21 |
| 35-40 | 58 | 58 - 46 = 12 |
| 40-45 | 64 | 64 - 58 =6 |

Method of converting ‘more than’ cumulative frequency to simple frequency is given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class interval** | **‘ more than’** | **Simple frequency** | | |  |
|  | **Cumulative frequency** |  |  |  |  |
| 15-20 | 64 | 64 | - 61 = | | 3 |
| 20-25 | 61 | 61 | - 54 = | | 7 |
| 25-30 | 54 | 54 | -39 = 15 | | |
| 30-35 | 39 | 39 | - 18 = | | 21 |
| 35-40 | 18 | 18 | - 6 | = 12 | |
| 40-45 | 6 | 6 - 0 | | = | 6 |

**Diagrammatic and Graphical Representation**

# **Introduction:**

In the previous chapter, we have discussed the techniques of classification and tabulation that help in summarizing the collected data and presenting them in a systematic manner.

# **Diagrams:**

A diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationship. If we draw diagrams on the basis of the data collected they will easily be understood.

# **Significance of Diagrams and Graphs:**

Diagrams and graphs are extremely useful because of the following reasons.

* + 1. They are attractive and impressive.
    2. They make data simple and intelligible.
    3. They make comparison possible
    4. They save time and labour.
    5. They have universal utility.
    6. They give more information.
    7. They have a great memorizing effect.

# **Types of diagrams:**

In practice, a very large variety of diagrams are in use and new ones are constantly being added. For the sake of convenience and simplicity, they may be divided under the following heads:

* + 1. One-dimensional diagrams
    2. Two-dimensional diagrams
    3. Three-dimensional diagrams
    4. Pictograms and Cartograms

# **One-dimensional diagrams:**

In such diagrams, only one-dimensional measurement, i.e height is used and the width is not considered. These diagrams are in the form of bar or line charts and can be classified as

1. Line Diagram
2. Simple Diagram
3. Multiple Bar Diagram
4. Sub-divided Bar Diagram
5. Percentage Bar Diagram

# **Line Diagram:**

# **Example:**

***Show the following data by a line chart:***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No. of children | 0 | 1 | 2 | 3 | 4 | 5 |
| Frequency | 10 | 14 | 9 | 6 | 4 | 2 |

**Line Diagram**

**No. of Children**

# **Simple Bar Diagram:**

# **Example:**

***Represent the following data by a bar diagram.***

|  |  |
| --- | --- |
| Year | Production (in tones) |
| 1991 | 45 |
| 1992 | 40 |
| 1993 | 42 |
| 1994 | 55 |
| 1995 | 50 |

# **Solution:**

**Simple Bar Diagram**

1991 1992 1993 1994 1995

**Year**

# **Multiple Bar Diagram:**

# **Example:**

**Draw a multiple bar diagram for the following data.**

|  |  |  |
| --- | --- | --- |
| Year | Profit before tax  ( in lakhs of rupees ) | Profit after tax  ( in lakhs of rupees ) |
| 1998 | 195 | 80 |
| 1999 | 200 | 87 |
| 2000 | 165 | 45 |
| 2001 | 140 | 32 |

# **Solution:**

**Multiple bar Diagram**



**1998 1999 2000 2001**

**Year**

**Pie Diagram or Circular Diagram:**

# **Example:**

**Draw a Pie diagram for the following data of production of sugar in quintals of various countries.**

|  |  |
| --- | --- |
| **Country** | **Production of Sugar (in quintals)** |
| Cuba | 62 |
| Australia | 47 |
| India | 35 |
| Japan | 16 |
| Egypt | 6 |

# **Solution:**

The values are expressed in terms of degree as follows.

|  |  |  |
| --- | --- | --- |
| **Country** | **Production of Sugar** | |
| **In Quintals** | **In Degrees** |
| Cuba | 62 | 134 |
| Australia | 47 | 102 |
| India | 35 | 76 |
| Japan | 16 | 35 |
| Egypt | 6 | 13 |
| Total | 166 | 360 |

# **Graphs:**

A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. Even a common man can understand the message of data from the graph. Comparisons can be made between two or more phenomena very easily with the help of a graph.

However here we shall discuss only some important types of graphs which are more popular and they are-

* + 1. Histogram
    2. Frequency Polygon
    3. Frequency Curve
    4. Ogive
    5. Lorenz Curve

# **Histogram:**

# **Example: *Draw a histogram for the following data.***

|  |  |
| --- | --- |
| Daily Wages | Number of Workers |
| 0-50 | 8 |
| 50-100 | 16 |
| 100-150 | 27 |
| 150-200 | 19 |
| 200-250 | 10 |
| 250-300 | 6 |

**Example:**

***For the following data, draw a histogram.***

|  |  |
| --- | --- |
| Marks | Number of Students |
| 21-30 | 6 |
| 31-40 | 15 |
| 41-50 | 22 |
| 51-60 | 31 |
| 61-70 | 17 |
| 71-80 | 9 |

# **Solution:**

For drawing a histogram, the frequency distribution should be continuous. If it is not continuous, then first make it continuous as follows.

|  |  |
| --- | --- |
| Marks | Number of Students |
| 20.5-30.5 | 6 |
| 30.5-40.5 | 15 |
| 40.5-50.5 | 22 |
| 50.5-60.5 | 31 |
| 60.5-70.5 | 17 |
| 70.5-80.5 | 9 |

# **Frequency Polygon:**

# **Example:**

**Draw a frequency polygon for the following data.**

|  |  |
| --- | --- |
| Weight (in kg) | Number of Students |
| 30-35 | 4 |
| 35-40 | 7 |
| 40-45 | 10 |
| 45-50 | 18 |
| 50-55 | 14 |
| 55-60 | 8 |
| 60-65 | 3 |

# **Frequency Curve:**

# **Example:**

**Draw a frequency curve for the following data.**

|  |  |
| --- | --- |
| Monthly Income (in SR) | No. of family |
| 0-1000 | 21 |
| 1000-2000 | 35 |
| 2000-3000 | 56 |
| 3000-4000 | 74 |
| 4000-5000 | 63 |
| 5000-6000 | 40 |
| 6000-7000 | 29 |
| 7000-8000 | 14 |

# **Ogives:**

There are two methods of constructing ogive namely:

* + - 1. The ‘ less than ogive’ method
      2. The ‘more than ogive’ method.

In less than ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve. In more than ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

# **Example: Draw the Ogives for the following data.**

|  |  |
| --- | --- |
| Class interval | Frequency |
| 20-30 | 4 |
| 30-40 | 6 |
| 40-50 | 13 |
| 50-60 | 25 |
| 60-70 | 32 |
| 70-80 | 19 |
| 80-90 | 8 |
| 90-100 | 3 |

# **Solution:**

|  |  |  |
| --- | --- | --- |
| Class limit | Less than ogive | More than ogive |
| 20 | 0 | 110 |
| 30 | 4 | 106 |
| 40 | 10 | 100 |
| 50 | 23 | 87 |
| 60 | 48 | 62 |
| 70 | 80 | 30 |
| 80 | 99 | 11 |
| 90 | 107 | 3 |
| 100 | 110 | 0 |

**Cumulative frequency**

**Chapter- 2**

**Measures of Central Tendency**

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations.

**Types of Averages:**

There are five averages. Among them *mean, median and mode* are called ***simple averages*** and the other two averages *geometric mean and harmonic mean* are called ***special averages***.

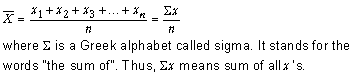
# **Characteristics for a good or an ideal average:**

The following properties should possess for an ideal average.

* 1. It should be rigidly defined.
  2. It should be easy to understand and compute.
  3. It should be based on all items in the data.
  4. Its definition shall be in the form of a mathematical formula.
  5. It should be capable of further algebraic treatment.
  6. It should have sampling stability.
  7. It should be capable of being used in further statistical computations or processing.

**Arithmetic mean**

The arithmetic mean (or, simply average or mean) of a set of numbers is obtained by dividing the sum of numbers of the set by the number of numbers. If the variable x assumes n values x1, x2 …xn then the mean, is given by



# **Example*:* Calculate the mean for 2, 4, 6, 8, and 10.**

**Solution:** Mean = = = 6

**(i) Direct method :** If the observations  x1,x2,x3........xn have frequencies  f1,f2,f3........fn  respectively, then the mean is given by :

http://xamidea.in/images/learning/Statistics/image005.gif

This method of finding the mean is called the direct method.

# **Example:**

Given the following frequency distribution, calculate the arithmetic mean

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Marks (x)** | 50 | 55 | 60 | 65 | 70 | 75 |
| **No of Students (f)** | 2 | 5 | 4 | 4 | 5 | 5 |

**Solution:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Marks (x)** | 50 | 55 | 60 | 65 | 70 | 75 | Total |
| **No of Students (f)** | 2 | 5 | 4 | 4 | 5 | 5 | 25 |
| **fx** | 100 | 275 | 240 | 260 | 350 | 375 | 1600 |

http://xamidea.in/images/learning/Statistics/image005.gif

= = 64

**(ii) Short cut method:** In some problems, where the number of variables is large or the values of xiorfiare larger, then the calculations become tedious. To overcome this difficulty, we use short cut or deviation method in which an approximate mean, called assumed mean is taken. This assumed mean is taken preferably near the middle, say A, and the deviation di =xi − Ais calculated for each variable http://xamidea.in/images/learning/Statistics/image008.gifThen the mean is given by the formula:

http://xamidea.in/images/learning/Statistics/image009.gif

**Mean for a grouped frequency distribution**

# **Example: *Given the following frequency distribution, calculate the arithmetic mean***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Marks (x) | 50 | 55 | 60 | 65 | 70 | 75 |
| No of Students (f) | 2 | 5 | 4 | 4 | 5 | 5 |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | f | fx | d=x-A | fd |
| 50 | 2 | 100 | -10 | -20 |
| 55 | 5 | 275 | -5 | -25 |
| **60** | 4 | 240 | 0 | 00 |
| 65 | 4 | 260 |  | 20 |
| 70 | 5 | 350 |  | 50 |
| 75 | 5 | 375 |  | 75 |
|  | 25 | 1600 |  | 100 |

**By Direct method:**

http://xamidea.in/images/learning/Statistics/image005.gif

= = 64

**By Short-cut method:**

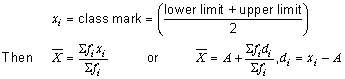
*x* *A* *fd*

*N*

= 60 + = 60 + 4 = 64

**Mean for a grouped frequency distribution**

Find the class mark or mid-value x, of each class, as



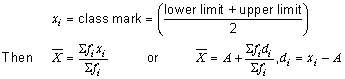
# **Example:**

***Following is the distribution of persons according to different income groups. Calculate arithmetic mean.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Income SR (100) | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
| Number of persons | 6 | 8 | 10 | 12 | 7 | 4 | 3 |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Income C.I | Number of Persons (f) | Mid X | d = x A  c | fd |
| 0-10 | 6 | 5 | -3 | -18 |
| 10-20 | 8 | 15 | -2 | -16 |
| 20-30 | 10 | 25 | -1 | -10 |
| 30-40 | 12 | A =35 | 0 | 0 |
| 40-50 | 7 | 45 | 1 | 7 |
| 50-60 | 4 | 55 | 2 | 8 |
| 60-70 | 3 | 65 | 3 | 9 |
| Total | 50 |  |  | -20 |



= 35 + × 10

= 35 – 4 = 31

# **Merits and demerits of Arithmetic mean:**

# **Merits:**

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

# **Demerits:**

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

# **Harmonic mean (H.M.):**

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x1,x2…..xn are n observations,

For a frequency distribution

**Example:**

***From the given data calculate H. M. 5, 10, 17, 24, and 30.***

**Solution:**

|  |  |
| --- | --- |
| X | 1  *x* |
| 5 | 0.2000 |
| 10 | 0.1000 |
| 17 | 0.0588 |
| 24 | 0.0417 |
| 30 | 0.0333 |
| Total | 0.4338 |

Hence,

= = 11.52

**Example:**

***The marks secured by some students of a class are given below. Calculate the harmonic mean.***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Marks** | 20 | 21 | 22 | 23 | 24 | 25 |
| **Number of Students** | 4 | 2 | 7 | 1 | 3 | 1 |

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| Marks  *X* | No of students f | 1  *x* | (1/*x*) |
| 20 | 4 | 0.0500 | 0.2000 |
| 21 | 2 | 0.0476 | 0.0952 |
| 22 | 7 | 0.0454 | 0.3178 |
| 23 | 1 | 0.0435 | 0.0435 |
| 24 | 3 | 0.0417 | 0.1251 |
| 25 | 1 | 0.0400 | 0.0400 |
|  | 18 |  | 0.8216 |

Hence,

**=**  = 21.91

**Geometric Mean (G.M.):**

The geometric mean of a series containing n observations is the nth root of the product of the values. If x1, x2…, xn are observations then

G.M. =

= (*x1.x2.x3……xn*)(1/n)

Log G.M. = (log*x1 + logx2 + logx3 +……+ logxn*)

Log G.M. **=**

G.M. = Antilog

**Example:**

***Calculate the geometric mean (G.M.) of the following series of monthly income of a batch of families 180, 250, 490, 1400, 1050.***

**Solution:**

|  |  |
| --- | --- |
| **x** | **Log x** |
| 180 | 2.2553 |
| 250 | 2.3979 |
| 490 | 2.6902 |
| 1400 | 3.1461 |
| 1050 | 3.0212 |
|  | 13.5107 |

G.M. = Antilog = Antilog = Antilog 2.70 = 503.6

**Example:**

**Calculate the average income per head from the data given below .Use geometric mean.**

|  |  |  |
| --- | --- | --- |
| Class of people | Number of families | Monthly income per head (SR) |
| Landlords | 2 | 5000 |
| Cultivators | 100 | 400 |
| Landless – labours | 50 | 200 |
| Money – lenders | 4 | 3750 |
| Office Assistants | 6 | 3000 |
| Shop keepers | 8 | 750 |
| Carpenters | 6 | 600 |
| Weavers | 10 | 300 |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class of people** | **Annual income ( SR) X** | **Number of families (f)** | **Log x** | **f logx** |
| Landlords | 5000 | 2 | 3.6990 | 7.398 |
| Cultivators | 400 | 100 | 2.6021 | 260.210 |
| Landless – labours | 200 | 50 | 2.3010 | 115.050 |
| Money – lenders | 3750 | 4 | 3.5740 | 14.296 |
| Office Assistants | 3000 | 6 | 3.4771 | 20.863 |
| Shop keepers | 750 | 8 | 2.8751 | 23.2008 |
| Carpenters | 600 | 6 | 2.7782 | 16.669 |
| Weavers | 300 | 10 | 2.4771 | 24.771 |
|  |  | 186 |  | 482.257 |

G.M. = Antilog

= Antilog

= Antilog (2.5928)

= SR 391.50

**Combined Mean:**

If the arithmetic averages and the number of items in two or more related groups are known, the combined or the composite mean of the entire group can be obtained by

Combined Mean, X =

**Example:**

Find the combined mean for the data given below:

n1 = 20; mean (x1) = 4; n2 = 30 and mean (x2) = 3

**Solution:**

Combined Mean, X = = = = = 3.4

**Positional Averages (Median and Mode):**

These averages are based on the position of the given observation in a series, arranged in an ascending or descending order. The magnitude or the size of the values does matter as was in the case of arithmetic mean. It is because of the basic difference that the median and mode are called the positional measures of an average.

**Median:**

The median is the middle value of a distribution i.e., median of a distribution is the value of the variable which divides it into two equal parts. It is the value of the variable such that the number of observations above it is equal to the number of observations below it.

**Ungrouped or Raw data:**

Arrange the given values in the increasing or decreasing order. If the numbers of values are odd, median is the middle value. If the numbers of values are even, median is the mean of middle two values.

By formula,

Median, Md = th item

# **When odd numbers of values are given:-**

# **Example:**

***Find median for the following data***

***25, 18, 27, 10, 8, 30, 42, 20, 53***

# **Solution:**

Arranging the data in the increasing order 8, 10, 18, 20, 25, 27, 30, 42, 53

Here, numbers of observations are odd (N= 9)

Hence, Median, Md = th item = th item= th item

The middle value is the 5th item i.e., 25 is the median value.

**When even numbers of values are given:-**

# **Example:**

***Find median for the following data***

***5, 8, 12, 30, 18, 10, 2, 22***

# **Solution:**

Arranging the data in the increasing order 2, 5, 8, 10, 12, 18, 22, 30

Here median is the mean of the middle two items (ie) mean of (10,12) ie

= 11

**Example:**

***The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Accountancy.***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Marks (Statistics) | 53 | 55 | 52 | 32 | 30 | 60 | 47 | 46 | 35 | 28 |
| Marks (Accountancy) | 57 | 45 | 24 | 31 | 25 | 84 | 43 | 80 | 32 | 72 |

**Solution:**

For such question, median is the most suitable measure of central tendency. The marks in the two subjects are first arranged in increasing order as follows:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Marks in Statistics | 28 | 30 | 32 | 35 | 46 | 47 | 52 | 53 | 55 | 60 |
| Marks in Accountancy | 24 | 25 | 31 | 32 | 43 | 45 | 57 | 72 | 80 | 84 |

Median value for Statistics = (Mean of 5th and 6th items) = = 46.5

Median value for Accountancy = (Mean of 5th and 6th items) = = 44

Therefore, the level of knowledge in Statistics is higher than that in Accountancy.

**Grouped Data:**

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

**Discrete Series:**

*Step1:* Find cumulative frequencies.

*Step 2:* Find

*Step 3:* See in the cumulative frequencies the value just greater than

*Step4:* Then the corresponding value of x will be median.

**Example:**

***The following data are pertaining to the number of members in a family. Find median size of the family.***

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number of members **x** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Frequency F | 1 | 3 | 5 | 6 | 10 | 13 | 9 | 5 | 3 | 2 | 2 | 1 |

**Solution:**

|  |  |  |
| --- | --- | --- |
| **X** | **f** | **cf** |
| 1 | 1 | 1 |
| 2 | 3 | 4 |
| 3 | 5 | 9 |
| 4 | 6 | 15 |
| 5 | 10 | 25 |
| 6 | 13 | 38 |
| 7 | 9 | 47 |
| 8 | 5 | 52 |
| 9 | 3 | 55 |
| 10 | 2 | 57 |
| 11 | 2 | 59 |
| 12 | 1 | 60 |
| N= | 60 |  |

Median = Size of item = Size of item = 30.5th item.

The cumulative frequency just greater than 30.5 is 38 and the value of x corresponding to 38 is 6. Hence the median size is 6 members per family.

***Note:***

*It is an appropriate method because a fractional value given by mean does not indicate the average number of members in a family.*

**Continuous Series:**

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step 2: Find

Step3: See in the cumulative frequency the value first greater than Then the corresponding class interval is called the Median Class. Then apply the formula for Median,

Md = *l+ ×h*

Where,

*l* = lower limit of the median class

 Σfi = n = number of Observations

*f* = frequency of the median class

*h* = size of the median class (assuming class size to be equal)

*cf* = cumulative frequency of the class preceding the median class.

N = Total frequency.

**Note:**

*If the class intervals are given in inclusive type convert them into exclusive type and call it as true class interval and consider lower limit in this.*

**Example:**

***The following table gives the frequency distribution of 325 workers of a factory, according to their average monthly income in a certain year.***

|  |  |
| --- | --- |
| Income group (in Rs) | Number of workers |
| Below 100 | 1 |
| 100-150 | 20 |
| 150-200 | 42 |
| 200-250 | 55 |
| 250-300 | 62 |
| 300-350 | 45 |
| 350-400 | 30 |
| 400-450 | 25 |
| 450-500 | 15 |
| 500-550 | 18 |
| 550-600 | 10 |
| 600 and above | 2 |
| N= | 325 |

***Calculate median income.***

**Solution:**

|  |  |  |
| --- | --- | --- |
| **Income group (Class-interval)** | **Number of workers (Frequency)** | **Cumulative frequency c.f** |
| Below 100 | 1 | 1 |
| 100-150 | 20 | 21 |
| 150-200 | 42 | 63 |
| 200-250 | 55 | 118 |
| 250-300 | 62 | 180 |
| 300-350 | 45 | 225 |
| 350-400 | 30 | 255 |
| 400-450 | 25 | 280 |
| 450-500 | 15 | 295 |
| 500-550 | 18 | 313 |
| 550-600 | 10 | 323 |
| 600 and above | 2 | 325 |
|  | 325 |  |

Here, = = 162.5

So, l = 250; n/2 = 162.5; cf = 118; f = 62 and h = 50

Median, Md = *l+ ×h*

= 250+ ×50

= 250 + 35.89 = 285.89

**Example:**

***Calculate median from the following data:***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Class Interval** | 1. 4 | 5- 9 | 10- 14 | 15- 19 | 20- 24 | 25- 29 | 30- 34 | 35- 39 |
| **Frequency** | 5 | 8 | 10 | 12 | 7 | 6 | 3 | 2 |

**Solution:**

Here, class intervals are in inclusive type so first we should convert it into exclusive type as done below:

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **Frequency** | **Cumulative Frequency (cf)** |
| 0.5- 4.5 | 5 | 5 |
| 4.5- 9.5 | 8 | 13 |
| 9.5- 14.5 | 10 | 23 |
| 14.5- 19.5 | 12 | 35 |
| 19.5- 24.5 | 7 | 42 |
| 24.5- 29.5 | 6 | 48 |
| 29.5- 34.5 | 3 | 51 |
| 34.5- 39.5 | 2 | 53 |
|  | N= 53 |  |

Here, = = 26.5

So, l = 14.5; n/2 = 26.5; cf = 23; f = 12 and h = 5

Median, Md = *l+ ×h*

= 14.5+ ×5

= 14.5 + 1.46

= 15.96

**Example:**

***Following are the daily wages of workers in a textile. Find the median.***

|  |  |
| --- | --- |
| Wages (in SR.) | Number of workers |
| less than 100 | 5 |
| less than 200 | 12 |
| less than 300 | 20 |
| less than 400 | 32 |
| less than 500 | 40 |
| less than 600 | 45 |
| less than 700 | 52 |
| less than 800 | 60 |
| less than 900 | 68 |
| less than 1000 | 75 |

**Solution:**

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 100, hence the width of the class interval equal to 100.

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **f** | **c.f** |
| 0-100 | 5 | 5 |
| 100-200 | 7 | 12 |
| 200-300 | 8 | 20 |
| 300- 400 | 12 | 32 |
| 400-500 | 8 | 40 |
| 500-600 | 5 | 45 |
| 600-700 | 7 | 52 |
| 700-800 | 8 | 60 |
| 800-900 | 8 | 68 |
| 900-1000 | 7 | 75 |
|  | N= 75 |  |

Here, = = 37.5

So, l = 400; n/2 = 37.5; cf = 32; f = 8 and h = 100

Median, Md = *l+ ×h*

= 400+ ×100

= 400 + 68.75

= 468.75

**Example: *Find median for the data given below.***

|  |  |
| --- | --- |
| **Marks** | **Number of students** |
| Greater than 10 | 70 |
| Greater than 20 | 62 |
| Greater than 30 | 50 |
| Greater than 40 | 38 |
| Greater than 50 | 30 |
| Greater than 60 | 24 |
| Greater than 70 | 17 |
| Greater than 80 | 9 |
| Greater than 90 | 4 |

**Solution:**

Here we are given lower limit and more than cumulative frequencies.

|  |  |  |  |
| --- | --- | --- | --- |
| Class interval | f | More than c.f | Less than c.f |
| 10-20 | 8 | 70 | 8 |
| 20-30 | 12 | 62 | 20 |
| 30-40 | 12 | 50 | 32 |
| 40-50 | 8 | 38 | 40 |
| 50-60 | 6 | 30 | 46 |
| 60-70 | 7 | 24 | 53 |
| 70-80 | 8 | 17 | 61 |
| 80-90 | 5 | 9 | 66 |
| 90-100 | 4 | 4 | 70 |
|  | 70 |  |  |

Here, = = 35

So, l = 40; n/2 = 35; cf = 32; f = 38 and h = 10

Median, Md = *l+ ×h*

= 40+ ×10 = 40 + 3.75 = 43.75

**Example:**

***Compute median for the following data.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mid-Value | 5 | 15 | 25 | 35 | 45 | 55 | 65 | 75 |
| Frequency | 7 | 10 | 15 | 17 | 8 | 4 | 6 | 7 |

**Solution :**

Here values in multiples of 10, so width of the class interval is 10.

|  |  |  |  |
| --- | --- | --- | --- |
| Mid x | C.I | f | c.f |
| 5 | 0-10 | 7 | 7 |
| 15 | 10-20 | 10 | 17 |
| 25 | 20-30 | 15 | 32 |
| 35 | 30-40 | 17 | 49 |
| 45 | 40-50 | 8 | 57 |
| 55 | 50-60 | 4 | 61 |
| 65 | 60-70 | 6 | 67 |
| 75 | 70-80 | 7 | 74 |
|  |  | N= 74 |  |

Here, = = 37

So, l = 30; n/2 = 37; cf = 32; f = 17 and h = 10

Median, Md = *l+ ×h*

= 30+ ×10

= 30 + 2.94

= 32.94

**Quartiles:**

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower) quartile (Q1) marks off the first one-fourth, the third (upper) quartile (Q3) marks off the three-fourth.

**Raw or ungrouped data:**

First arrange the given data in the increasing order and use the formula for Q1 and Q3 then quartile deviation, Q.D. is given by

Q.D. =

Where, Q1 = )th item and Q3 = )th item

**Example:**

Compute quartiles for the data given below 25,18, 30, 8, 15, 5, 10, 35, 40, 45

**Solution:**

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

Q1 = )th item

= )th item

= (2.75) th item.

= 2nd item + ) (3rd item- 2nd item)

= 8 + ) (10 - 8)

= 8 + )×2

= 9.5

Q3 = )th item

= 3(2.75) th item.

= 8.25th item

= 8th item + )(9th item- 8th item)

= 35 +) (40 - 35)

= 35 + 1.25

= 36.25

**Discrete Series:**

*Step1:* Find cumulative frequencies

*Step2:* Find )

*Step3:* See in the cumulative frequencies, the value just greater than )then the corresponding value of *x* is Q1

*Step4:* Find 3)

*Step5:* See in the cumulative frequencies, the value just greater than 3) then the corresponding value of x is Q3.

**Example:**

***Compute quartiles for the data given below.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 5 | 8 | 12 | 15 | 19 | 24 | 30 |
| f | 4 | 3 | 2 | 4 | 5 | 2 | 4 |

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| **x** | **f** | **c.f** |  |
|  |
| 5 | 4 | 4 | |
| 8 | 3 | 7 | |
| 12 | 2 | 9 | |
| 15 | 4 | 13 | |
| 19 | 5 | 18 | |
| 24 | 2 | 20 | |
| 30 | 4 | 24 | |
| Total | 24 |  | |

Q1 = )th item

= )th item

= )th item

= 6.25th = 8

Q3 = )th item

=  th item

= 18.75th item = 24

**Continuous Series:**

Step1: Find cumulative frequencies;

Step2: Find)

Step3: See in the cumulative frequencies, the value just greater), then the corresponding class interval is called first quartile class.

Step4: Find 3), See in the cumulative frequencies the value just greater than 3) then the corresponding class interval is called 3rd quartile class. Then apply the respective formulae

Q1 = *l1+ ×h1*

Q3 = *l3+ ×h3*

Where *l*1 = lower limit of the first quartile class

*f*1 = frequency of the first quartile class

*h*1 = width of the first quartile class

*cf*1 = c.f. preceding the first quartile class

*l*3 = 1ower limit of the 3rd quartile class

*f*3 = frequency of the 3rd quartile class

h3 = width of the 3rd quartile class

*cf*3 = c.f. preceding the 3rd quartile class

**Example:**

***The following series relates to the marks secured by students in an examination.***

|  |  |
| --- | --- |
| **Marks** | **No. of students** |
| 0-10 | 11 |
| 10-20 | 18 |
| 20-30 | 25 |
| 30-40 | 28 |
| 40-50 | 30 |
| 50-60 | 33 |
| 60-70 | 22 |
| 70-80 | 15 |
| 80-90 | 12 |
| 90-100 | 10 |

***Find the quartiles.***

**Solution:**

|  |  |  |
| --- | --- | --- |
| **C.I.** | **f** | **cf** |
| 0-10 | 11 | 11 |
| 10-20 | 18 | 29 |
| 20-30 | 25 | 54 |
| 30-40 | 28 | 82 |
| 40-50 | 30 | 112 |
| 50-60 | 33 | 145 |
| 60-70 | 22 | 167 |
| 70-80 | 15 | 182 |
| 80-90 | 12 | 194 |
| 90-100 | 10 | 204 |
|  | 204 |  |

Here, ) = ) = 51 and 3) = 3× 51 = 153

Q1 = *l1+ ×h1*

= 20+ ×10 = 28.8

Q3 = *l3+ ×h3*

= 60+ ×10 = 63.63

**Deciles:**

These are the values, which divide the total number of observation into 10 equal parts. These are 9 deciles D1, D2…D9. These are all called first decile, second decile…etc.

**Deciles for Raw data or ungrouped data**

**Example: *Compute D5 for the data given below 5, 24, 36, 12, 20, 8***.

**Solution:**

Arranging the given values in the increasing order 5, 8, 12, 20, 24, 36

D5 = 5 observation

= 5observation

= observation

= 3rd item +( [ 4th item – 3rd item]

= 12 + ( [ 20 – 12]

= 16.

**Deciles for Grouped data:**

Same as quartile.

# **Percentiles:**

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The percentile (Pk) is that value of the variable up to which lie exactly k% of the total number of observations.

# **Relationship:**

P25 = Q1; P50 = D5 = Q2 = Median and P75 = Q3

**Percentile for Raw Data or Ungrouped Data:**

**Example: *Calculate P15 for the data given below: 5, 24, 36 , 12 , 20 , 8.***

**Solution:**

Arranging the given values in the increasing order. 5, 8, 12, 20, 24, 36

P15 = 15 item

= 15 item

= (1.05)th item

= 1st item + 0.5(2nd item – 1st item)

= 5 + 0.5(8 - 5)

= 5.15

**Percentile for Grouped Data:**

# **Example: *Find P53 for the following frequency distribution.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class Interval | 0-5 | 5-10 | 10-15 | 15-20 | 20-25 | 25-30 | 30-35 | 35-40 |
| Frequency | 5 | 8 | 12 | 16 | 20 | 10 | 4 | 3 |

# **Solution:**

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **Frequency** | **cf** |
| 0-5 | 5 | 5 |
| 5-10 | 8 | 13 |
| 10-15 | 12 | 25 |
| 15-20 | 16 | 41 |
| 20-25 | 20 | 61 |
| 25-30 | 10 | 71 |
| 30-35 | 4 | 75 |
| 35-40 | 3 | 78 |
| Total | 78 |  |

P53 = *l+ ×h* = *20+ ×5 =* 20.085

**Mode:**

The mode or modal value of a distribution is that value of the variable for which the frequency is the maximum. It refers to that value in a distribution which occurs most frequently. It shows the center of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order for shoes or ready-made garments the modal size helps because these sizes and other sizes around in common demand.

# **Computation of the mode:**

# **Ungrouped or Raw Data:**

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

# **Example:**

2, 7, 10, 15, 10, 17, 8, 10, 2

Mode = M0 = 10

In some cases the mode may be absent while in some cases there may be more than one mode

# **Example:** (1) 12, 10, 15, 24, 30 (no mode)

(2) 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

the modes are 7 and 10

# **Grouped Data:**

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

**Continuous distribution:**

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the following formula:

Mode, Mo = *l+ ×h*

Where, *l* = lower limit of the modal class

*f* = frequency of the modal class

f1 = frequency of the class preceding the modal class

f2 = frequency of the class following the modal class.

*h* = size of the modal class

***Remarks:***

1. *If (2f1-f0-f2) comes out to be zero, then mode is obtained by the following formula taking absolute differences within vertical lines;*
2. *Mode, Mo = l+ ×h*
3. *If mode lies in the first class interval, then f is taken as zero.*
4. *The computation of mode poses no problem in distributions with open-end classes, unless the modal value lies in the open-end class.*

# **Example: *Calculate mode for the following:***

|  |  |
| --- | --- |
| **C- I** | **f** |
| 0-50 | 5 |
| 50-100 | 14 |
| 100-150 | 40 |
| 150-200 | 91 |
| 200-250 | 150 |
| 250-300 | 87 |
| 300-350 | 60 |
| 350-400 | 38 |
| 400 and above | 15 |

# **Solution:**

The highest frequency is 150 and corresponding class interval is 200 – 250, which is the modal class.

Here, *l* = 200; *f* = 150; *f1*= 91; *f2* = 87 and *h* = 50

Mode, Mo = *l+ ×h*

*=* 200+ ×50

*=* 200+ 24.18

*= 224.18*

**Determination of Modal class:**

For a frequency distribution modal class corresponds to the maximum frequency. But in any one (or more) of the following cases-

1. If the maximum frequency is repeated
2. If the maximum frequency occurs in the beginning or at the end of the distribution
3. If there are irregularities in the distribution, the modal class is determined by the method of grouping.

**Steps for Calculation:**

We prepare a grouping table with 6 columns

1. In column I, we write down the given frequencies;
2. Column II is obtained by combining the frequencies two by two;
3. Leave the 1st frequency and combine the remaining frequencies two by two and write in column III;
4. Column IV is obtained by combining the frequencies three by three;
5. Leave the 1st frequency and combine the remaining frequencies three by three and write in column V;
6. Leave the 1st and 2nd frequencies and combine the remaining frequencies three by three and write in column VI.

Mark the highest frequency in each column. Then form an analysis table to find the modal class. After finding the modal class, use the formula to calculate the modal value.

# **Example:**

***Calculate mode for the following frequency distribution.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class interval | 1. 5 | 5- 10 | 10- 15 | 15- 20 | 20- 25 | 25- 30 | 30- 35 | 35- 40 |
| Frequency | 9 | 12 | 15 | 16 | 17 | 15 | 10 | 13 |

# **Grouping Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| C I | f | 2 | 3 | 4 | 5 | 6 |
| 0- 5  5-10  10-15  15-20  20-25  25-30  30-35  35-40 | 9  12  15  16  **17**  15  10  13 | 21  31  **32**  23 | 27  **33**  25 | 36  **48** | **43**  42 | **48**  38 |

**Analysis Table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Columns** | **0-5** | **5-10** | **10-15** | **15-20** | **20-25** | **25-30** | **30-35** | **35-40** |
| 1 |  | 1 | 1 | 1 | 1 | 1 |  |  |
| 2 | 1 |
| 3 | 1 |  |
| 4 | 1 | 1 | 1 |
| 5 | 1 |  |
| 6 | 1 | 1 | 1 |
| Total |  | 1 | 2 | 4 | 5 | 2 |  |  |

The maximum occurred corresponding to 20-25, and hence it is the modal class.

Here, *l* = 20; *f* = 16; *f1*= 15; *f2* = 17 and *h* = 5

Mode, Mo = *l+ ×h*

*=* 20+ ×5

*=* 20+ ×5

*So, Mode, Mo = l+ ×h*

= 20+ ×5

*=* 20*+ =* 20+ 1.67 *=* 21.67

# **Empirical Relationship between Averages**

In a symmetrical distribution the three simple averages mean = median = mode. For a moderately asymmetrical distribution, the relationship between them are brought by Prof. Karl Pearson as

*Mode = 3 Median - 2 Mean*

# **Example:**

***If the mean and median of a moderately asymmetrical series are 26.8 and 27.9 respectively, what would be its most probable mode?***

# **Solution:**

Using the empirical formula Mode = 3 median 2 mean

= 3 27.9 2 26.8

= 30.1

**Example:**

***In a moderately asymmetrical distribution the values of mode and mean are 32.1 and 35.4 respectively. Find the median value.***

# **Solution:**

Using empirical Formula

Median =

= = 34.3

**Measures of Dispersion – Skewness and Kurtosis**

# **Introduction:**

The measures of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. For example consider the following marks of two students.

|  |  |
| --- | --- |
| Student I | Student II |
| 68 | 85 |
| 75 | 90 |
| 65 | 80 |
| 67 | 25 |
| 70 | 65 |

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

# **Characteristics of a good measure of dispersion:**

An ideal measure of dispersion is expected to possess the following properties

* + 1. It should be rigidly defined
    2. It should be based on all the items.
    3. It should not be unduly affected by extreme items.
    4. It should lend itself for algebraic manipulation.
    5. It should be simple to understand and easy to calculate

# **Absolute and Relative Measures:**

There are two kinds of measures of dispersion, namely (1).Absolute measure of dispersion and (2).Relative measure of dispersion.

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations. They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations.

The various absolute and relative measures of dispersion are listed below.

|  |  |
| --- | --- |
| Absolute measure | Relative measure |
| Range | Co-efficient of Range |
| Quartile deviation | Co-efficient of Quartile deviation |
| Mean deviation | Co-efficient of Mean deviation |
| Standard deviation | Co-efficient of variation |

# **Range and coefficient of Range:**

**Range:**

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S.

Where L = Largest value. S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed,

**Method 1:**

L = Upper boundary of the highest class

S = Lower boundary of the lowest class

**Method 2:**

L = Mid value of the highest class.

S = Mid value of the lowest class.

**Co-efficient of Range:**

Co-efficient of Range =

**Example:**

Find the value of range and its co-efficient for the following data.

7, 9, 6, 8, 11, 10, 4

**Solution:**

L=11, S = 4.

Range = L – S

= 11- 4 = 7

Co-efficient of Range = = = = 0.4667

**Example: *Calculate range and its co efficient from the following distribution.***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Size:** | 60- 63 | 63- 66 | 66- 69 | 69- 72 | 72- 75 |
| **Number:** | 5 | 18 | 42 | 27 | 8 |

**Solution:**

L = Upper boundary of the highest class = 75

S = Lower boundary of the lowest class = 60

Range = L – S = 75 – 60 = 15

Co-efficient of Range = = = = 0.1111

**Quartile Deviation and Co efficient of Quartile Deviation:**

**Quartile Deviation (Q.D.):**

**Definition:** Quartile Deviation is half of the difference between the first and third quartiles. Hence, it is called Semi Inter Quartile Range.

In symbol, Q.D. =

Among the quartiles Q1, Q2 and Q3, the range Q3- Q1 is called inter quartile range and, semi inter quartile range.

**Co-efficient of Quartile Deviation:**

Co-efficient of Quartile Deviation =

**Example:** Find the Quartile Deviation for the following data:

391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

# **Solution:**

Arrange the given values in ascending order.

384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488

Position of Q1 is = = 2.75th item

Q1 = 2nd value + 0.75 (3rd value – 2nd value)

= 391 + 0.75 (407 – 391)

= 391 + 0.75 16

= 391 + 12

= 403

Position of Q3 is = 3×2.75 = 8.25th item

Q3 = 8th value + 0.25 (9th value – 8th value)

= 777 + 0.25 (1490 – 777)

= 777 + 0.25 (713)

= 777 + 178.25 = 955.25

Q.D. = = = = 276.125

**Example:**

***Weekly wages of labors are given below. Calculate Q.D. and Coefficient of Q.D.***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Weekly Wage (Rs.) | :100 | 200 | 400 | 500 | 600 |
| No. of Weeks | : 5 | 8 | 21 | 12 | 6 |

**Solution:**

|  |  |  |
| --- | --- | --- |
| Weekly Wage (Rs.) | No. of Weeks | Cum. No. of Weeks |
| 100 | 5 | 5 |
| 200 | 8 | 13 |
| 400 | 21 | 34 |
| 500 | 12 | 46 |
| 600 | 6 | 52 |
| Total | N=52 |  |

Position of Q1 is = = 13.25th item

Q1 = 13th value + 0.25 (14th Value – 13th value)

= 13th value + 0.25 (400 – 200)

= 200 + 0.25 (400 – 200)

= 200 + 0.25 (200)

= 200 + 50 = 250

Position of Q3 is = 3×13.25 = 39.25th item

Q3 = 39th value + 0.75 (40th value – 39th value)

= 500 + 0.75 (500 – 500)

= 500 + 0.75 0

= 500

Q.D. = = = = 125

Co-efficient of Quartile Deviation = = = = 0.33

**Example:**

***For the date given below, give the quartile deviation and coefficient of quartile deviation.***

|  |  |  |  |
| --- | --- | --- | --- |
| X : 351 – 500 501 – 650 | 651 – 800 | 801–950 | 951–1100 |
| f : 48 189 | 88 | 4 | 28 |

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| x | f | True class Intervals | Cumulative frequency |
| 351- 500 | 48 | 350.5- 500.5 | 48 |
| 501- 650 | 189 | 500.5- 650.5 | 237 |
| 651- 800 | 88 | 650.5- 800.5 | 325 |
| 801- 950 | 47 | 800.5- 950.5 | 372 |
| 951- 1100 | 28 | 950.5- 1100.5 | 400 |
| Total | N = 400 |  |  |

Since, N/4 = 100

Therefore, Q1 Class is 500.5- 650.5

Hence, *l1= 500.5; n/4 = 100; cf1 =48; f1= 189; h1= 150*

Q1 = *l1+ ×h1*

= 500.5+ ×150 = 541.77

Now, for Q3

= 3*×*100= 300

Hence, Q3 Class is 650.5- 800.5

*l3= 650.5; = 300; cf3= 237; f3= 88; h3= 150*

Q3 = *l3+ ×h3*

= 650.5+ ×150 = 757.89

Q.D. = = = = 108.06

Co-efficient of Quartile Deviation = = = = 0.1663

**Mean Deviation and Coefficient of Mean Deviation:**

**Mean Deviation:** The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation is measure of dispersion based on all items in a distribution.

**Definition:**

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored.

We usually compute mean deviation about any one of the three averages mean, median or mode. Sometimes mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

**Coefficient of mean deviation:**

Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

Coefficient of Mean Deviation =

*If the result is desired in percentage*,

The coefficient of mean deviation = ×100

**Computation of mean deviation – Individual Series:**

* + - 1. Calculate the average mean, median or mode of the series.
      2. Take the deviations of items from average ignoring signs and denote these deviations by |D|.
      3. Compute the total of these deviations, i.e., |D|
      4. Divide this total obtained by the number of items.

Symbolically,

M.D. =

**Example:**

***Calculate mean deviation from mean and median for the following data:***

***100, 150, 200, 250, 360, 490, 500, 600, and 671.***

***Also calculate co- efficient of M.D.***

**Solution:**

Mean = = = = 369

Now arrange the data in ascending order

100, 150, 200, 250, 360, 490, 500, 600, 671

Median, Md = Value of ()th item = ()th item = 5th item = 360

|  |  |  |
| --- | --- | --- |
| **X** | **D ⃒X Mean⃒** | **⃒D⃒ ⃒x Md⃒** |
| 100 | 269 | 260 |
| 150 | 219 | 210 |
| 200 | 169 | 160 |
| 250 | 119 | 110 |
| 360 | 9 | 0 |
| 490 | 121 | 130 |
| 500 | 131 | 140 |
| 600 | 231 | 240 |
| 671 | 302 | 311 |
| 3321 | 1570 | 1561 |

M.D. from mean = = = 174.44

Co-efficient of M.D. = = = 0.47

M.D. from median = = = 173.44

Co-efficient of M.D. = = = 0.48

**Mean Deviation- Discrete Series:**

M.D. =

**Example:**

***Compute Mean deviation from mean and median from the following data:***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Height in cms | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 |
| No. of persons | 15 | 20 | 32 | 35 | 33 | 22 | 20 | 10 | 8 |

***Also compute coefficient of mean deviation.***

**Solution:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Height X | No. of persons f | d= x- A A =162 | fd | |D| =  |X- mean| | f|D| |
| 158 | 15 | - 4 | - 60 | 3.51 | 52.65 |
| 159 | 20 | - 3 | - 60 | 2.51 | 50.20 |
| 160 | 32 | - 2 | - 64 | 1.51 | 48.32 |
| 161 | 35 | - 1 | - 35 | 0.51 | 17.85 |
| **162** | 33 | 0 | 0 | 0.49 | 16.17 |
| 163 | 22 | 1 | 22 | 1.49 | 32.78 |
| 164 | 20 | 2 | 40 | 2.49 | 49.80 |
| 165 | 10 | 3 | 30 | 3.49 | 34.90 |
| 166 | 8 | 4 | 32 | 4.49 | 35.92 |
|  | 195 |  | - 95 |  | 338.59 |

Mean = = = 162 – 0.49 = 161.51

M.D. = = = 1.74

Coefficient of M.D. = = = 0.0108

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Height x | No. of persons f | c.f. | D =  ⃒X Median⃒ | f ⃒D⃒ |
| 158 | 15 | 15 | 3 | 45 |
| 159 | 20 | 35 | 2 | 40 |
| 160 | 32 | 67 | 1 | 32 |
| 161 | 35 | 102 | 0 | 0 |
| 162 | 33 | 135 | 1 | 33 |
| 163 | 22 | 157 | 2 | 44 |
| 164 | 20 | 177 | 3 | 60 |
| 165 | 10 | 187 | 4 | 40 |
| 166 | 8 | 195 | 5 | 40 |
|  | 195 |  |  | 334 |

Median = Size of ()th item = ()th item = (96)th item = 161

M.D. = = = 1.71

Coefficient of M.D. = = = 0.0106

**Mean Deviation-Continuous Series:**

The method of calculating mean deviation in a continuous series same as the discrete series. In continuous series we have to find out the mid points of the various classes and take deviation of these points from the average selected. Thus

M.D. =

**Example:**

***Find out the mean deviation from mean and median from the following series.***

|  |  |
| --- | --- |
| **Age in years** | **No of persons** |
| 0-10 | 20 |
| 10-20 | 25 |
| 20-30 | 32 |
| 30-40 | 40 |
| 40-50 | 42 |
| 50-60 | 35 |
| 60-70 | 10 |
| 70-80 | 8 |

***Also compute co-efficient of mean deviation.***

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | m | f | d = m A  c  (A=35,C=10) | fd | D   ⃒m  x⃒ | f ⃒D⃒ |
| 0-10 | 5 | 20 | -3 | -60 | 31.5 | 630.0 |
| 10-20 | 15 | 25 | -2 | -50 | 21.5 | 537.5 |
| 20-30 | 25 | 32 | -1 | -32 | 11.5 | 368.0 |
| 30-40 | **35** | 40 | 0 | 0 | 1.5 | 60.0 |
| 40-50 | 45 | 42 | 1 | 42 | 8.5 | 357.0 |
| 50-60 | 55 | 35 | 2 | 70 | 18.5 | 647.5 |
| 60-70 | 65 | 10 | 3 | 30 | 28.5 | 285.0 |
| 70-80 | 75 | 8 | 4 | 32 | 38.5 | 308.0 |
|  |  | 212 |  | 32 |  | 3193.0 |

Mean =

= = 36.5

M.D. = = = 15.06

**Calculation of median and M.D. from median:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | m | f | c.f | |D| = |m-Md| | f |D| |
| 0-10 | 5 | 20 | 20 | 32.25 | 645.00 |
| 10-20 | 15 | 25 | 45 | 22.25 | 556.25 |
| 20-30 | 25 | 32 | 77 | 12.25 | 392.00 |
| 30-40 | 35 | 40 | 117 | 2.25 | 90.00 |
| 40-50 | 45 | 42 | 159 | 7.75 | 325.50 |
| 50-60 | 55 | 35 | 194 | 17.75 | 621.25 |
| 60-70 | 65 | 10 | 204 | 27.75 | 277.50 |
| 70-80 | 75 | 8 | 212 | 37.75 | 302.00 |
| N= 212 | | | | Total | 3209.50 |

= = 106

*l= 30; cf= 77; f= 40; h= 10*

Median, Md = *l+ ×h=* 30+ ×10= 37.25

M.D. = = = 15.14

Coefficient of M.D. = = = 0.41

# **Merits and Demerits of M.D:**

# **Merits:**

1. It is simple to understand and easy to compute.
2. It is rigidly defined.
3. It is based on all items of the series.
4. It is not much affected by the fluctuations of sampling.
5. It is less affected by the extreme items.
6. It is flexible, because it can be calculated from any average.
7. It is better measure of comparison.

# **Demerits:**

1. It is not a very accurate measure of dispersion.
2. It is not suitable for further mathematical calculation.
3. It is rarely used. It is not as popular as standard deviation.
4. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

# **Standard Deviation and Coefficient of variation:**

# **Standard Deviation:**

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

# **Definition:**

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by the Greek letter  (sigma).

# **Calculation of Standard deviation-Individual Series:**

There are two methods of calculating Standard deviation in an individual series.

* + - 1. Deviations taken from Actual mean; and
      2. Deviation taken from Assumed mean

1. **Deviation taken from Actual mean:**

This method is adopted when the mean is a whole number.

Steps:

1. Find out the actual mean of the series ( *x* )
2. Find out the deviation of each value from the mean (*x = X- X*)
3. Square the deviations and take the total of squared deviations x2
4. Divide the total ( x2 ) by the number of observation, ()
5. The square root of () is standard deviation.

Thus, Standard Deviation (SD or ) = =

1. **Deviations taken from assumed mean:**

This method is adopted when the arithmetic mean is fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, we apply short– cut method; deviations are taken from an assumed mean. The formula is:

S.D. or 

Where d-stands for the deviation from assumed mean = (X-A)

# **Steps:**

* 1. Assume any one of the item in the series as an average (A)
  2. Find out the deviations from the assumed mean; i.e., X-A denoted by d and also the total of the deviations d
  3. Square the deviations; i.e., d2 and add up the squares of deviations, i.e, d2
  4. Then substitute the values in the following formula:

S.D. or 

**Example:**

Calculate the standard deviation from the following data. 14, 22, 9, 15, 20, 17, 12, 11

**Solution:**

*Deviations from actual mean.*

|  |  |  |
| --- | --- | --- |
| Values (X) |  |  |
| 14 | -1 | 1 |
| 22 | 7 | 49 |
| 9 | -6 | 36 |
| 15 | 0 | 0 |
| 20 | 5 | 25 |
| 17 | 2 | 4 |
| 12 | -3 | 9 |
| 11 | -4 | 16 |
| 120 |  | 140 |

Mean, (*X bar*) = = = 15

Thus, Standard Deviation (SD or ) = = = = = 4.18

**Example:**

The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Student Nos : 1 2 3 | 4 | 5 6 | 7 | 8 | 9 | 10 |
| Marks : 43 48 65 | 57 | 31 60 | 37 | 48 | 78 | 59 |

**Solution:** (Deviations from assumed mean)

|  |  |  |  |
| --- | --- | --- | --- |
| Nos. | Marks (x) | d=X-A (A=57) | d2 |
| 1 | 43 | -14 | 196 |
| 2 | 48 | -9 | 81 |
| 3 | 65 | 8 | 64 |
| 4 | 57 | 0 | 0 |
| 5 | 31 | -26 | 676 |
| 6 | 60 | 3 | 9 |
| 7 | 37 | -20 | 400 |
| 8 | 48 | -9 | 81 |
| 9 | 78 | 21 | 441 |
| 10 | 59 | 2 | 4 |
| n = 10 |  | d=-44 | d2 =1952 |

S.D. or   = = = 13.26

# **Calculation of standard deviation:**

# **Discrete Series:**

There are three methods for calculating standard deviation in discrete series:

1. Actual mean methods: If the actual mean in fractions, the calculation takes lot of time and labour; and as such this method is rarely used in practice.
2. Assumed mean method: Here deviations are taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.
3. Step-deviation method: If the variable values are in equal intervals, then we adopt this method.

**Example:**

***Calculate Standard deviation from the following data.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X : | 20 | 22 | 25 | 31 | 35 | 40 | 42 | 45 |
| f : | 5 | 12 | 15 | 20 | 25 | 14 | 10 | 6 |

**Solution:**

Deviations from assumed mean

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | f | d = x –A (A = 31) | d2 | fd | fd2 |
| 20 | 5 | -11 | 121 | -55 | 605 |
| 22 | 12 | -9 | 81 | -108 | 972 |
| 25 | 15 | -6 | 36 | -90 | 540 |
| 31 | 20 | 0 | 0 | 0 | 0 |
| 35 | 25 | 4 | 16 | 100 | 400 |
| 40 | 14 | 9 | 81 | 126 | 1134 |
| 42 | 10 | 11 | 121 | 110 | 1210 |
| 45 | 6 | 14 | 196 | 84 | 1176 |
|  | N=107 |  |  | fd=167 | fd2= 6037 |

S.D. or   = = = 7.35

**Calculation of Standard Deviation –Continuous Series:**

In the continuous series the method of calculating standard deviation is almost the same as in a discrete series. But in a continuous series, mid-values of the class intervals are to be found out. The step- deviation method is widely used.

**Coefficient of Variation:**

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

The coefficient of variation is obtained by dividing the standard deviation by the mean and multiplies it by 100. Symbolically,

Coefficient of Variation (CV) = =

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

**Example:**

***In two factories A and B located in the same industrial area, the average weekly wages (in SR) and the standard deviations are as follows:***

|  |  |  |  |
| --- | --- | --- | --- |
| **Factory** | **Average (x)** | **Standard Deviation (σ)** | **No. of workers** |
| **A B** | **34.5**  **28.5** | **5**  **4.5** | **476**  **524** |

1. ***Which factory A or B pays out a larger amount as weekly wages?***
2. ***Which factory A or B has greater variability in individual wages?***

**Solution:**

Given N1= 476; X1= 34.5 and σ1= 5

N2 = 524, X2 = 28.5, 2 = 4.5

1. Total wages paid by factory A

= 34.5 476

= SR16.422

Total wages paid by factory B

= 28.5 524

= SR.14,934.

Therefore factory A pays out larger amount as weekly wages.

1. C.V. of distribution of weekly wages of factory A and B are

CV (A) = = = 14.49

CV (B) = = = 15.79

Factory B has greater variability in individual wages, since C.V. of factory B is greater than C.V of factory A.

# **Example:**

***Prices of a particular commodity in five years in two cities are given below:***

|  |  |
| --- | --- |
| ***Price in city A*** | ***Price in city B*** |
| ***20*** | ***10*** |
| ***22*** | ***20*** |
| ***19*** | ***18*** |
| ***23*** | ***12*** |
| ***16*** | ***15*** |

***Which city has more stable prices?***

# **Solution:** *Actual mean method*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City A | | | City B | | |
| Prices (X) | Deviations from X=20 dx | dx2 | Prices (Y) | Deviations from Y =15 dy | dy2 |
| 20 | 0 | 0 | 10 | -5 | 25 |
| 22 | 2 | 4 | 20 | 5 | 25 |
| 19 | -1 | 1 | 18 | 3 | 9 |
| 23 | 3 | 9 | 12 | -3 | 9 |
| 16 | -4 | 16 | 15 | 0 | 0 |
| x=100 | dx=0 | dx2=30 | y=75 | dy=0 | dy2 =68 |

**City A:**

Mean = = = 20;

SD ( = = = = 2.45

CV (City A) = = = = 12.25%

**City B:**

Mean = = = 15;

SD ( = = = = 3.69

CV (City B) = = = = 24.6%

Therefore, City A had more stable prices than City B, because the coefficient of variation is less in City A.

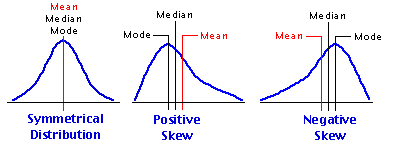
**Skewness**

**Meaning:**

Skewness means ‘lack of symmetry’. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution. If in a distribution mean median mode, then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

**Symmetrical distribution:**

It is clear from the diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the center point of the curve



**Positively skewed distribution:**

It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

**Negatively skewed distribution:**

It is clear from the above diagram, in a negatively skewed distribution, the value of the mode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

**Measures of skewness:**

The important measures of skewness are

1. Karl – Pearason’ s coefficient of skewness
2. Bowley’ s coefficient of skewness

**Karl – Pearson’s Coefficient of skewness:**

According to Karl – Pearson, the absolute measure of skewness = mean – mode. This measure is not suitable for making valid comparison of the skewness in two or more distributions because the unit of measurement may be different in different series. To avoid this difficulty use relative measure of skewness called Karl – Pearson’ s coefficient of skewness given by

Karl – Pearson’s Coefficient Skewness =

In case of mode is ill – defined, the coefficient can be determined by the formula:

Karl – Pearson’s Coefficient Skewness =

# **Bowley’ s Coefficient of skewness:**

In Karl – Pearson’ s method of measuring skewness the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the median; ie.,

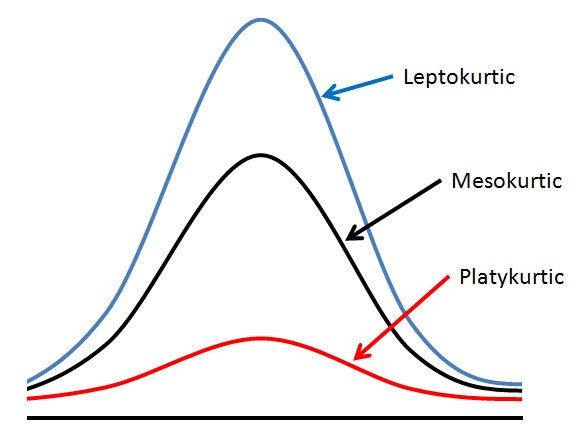
Median – Q1 = Q3 – Median. But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula:

Bowley’ s Coefficient of skewness (sk) =

**Kurtosis**

The expression ‘ Kurtosis’ is used to describe the peakedness of a curve. The three measures – central tendency, dispersion and skewness describe the characteristics of frequency distributions. But these studies will not give us a clear picture of the characteristics of a distribution.

As far as the measurement of shape is concerned, we have two characteristics – skewness which refers to asymmetry of a series and kurtosis which measures the peakedness of a normal curve. All the frequency curves expose different degrees of flatness or peakedness. This characteristic of frequency curve is termed as kurtosis. Measure of kurtosis denote the shape of top of a frequency curve. Measure of kurtosis tell us the extent to which a distribution is more peaked or more flat topped than the normal curve, which is symmetrical and bell-shaped, is designated as Mesokurtic. If a curve is relatively more narrow and peaked at the top, it is designated as Leptokurtic. If the frequency curve is more flat than normal curve, it is designated as platykurtic.



# **Measure of Kurtosis:**

The measure of kurtosis of a frequency distribution based moments is denoted by 2 and is given by



If 2 =3, the distribution is said to be normal and the curve is mesokurtic.

If 2 >3, the distribution is said to be more peaked and the curve is leptokurtic.

If 2< 3, the distribution is said to be flat topped and the curve is platykurtic.

**Correlation**

# **Introduction:**

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only variable such as height, weight, ages, marks, wages, etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-Variate Analysis. Sometimes the variables may be inter-related. In health sciences we study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure, etc. The nature and strength of relationship may be examined by correlation and Regression analysis.

Thus Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

# **Uses of correlation:**

1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

# **Types of Correlation:**

Correlation is classified into various types. The most important ones are

1. Positive and negative.
2. Linear and non-linear.
3. Partial and total.
4. Simple and Multiple.

# **Positive and Negative Correlation:**

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (ie) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.

If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.

# **Linear and Non-linear correlation:**

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | 2 | 4 | 6 | 8 | 10 | 12 |
| Y | 3 | 6 | 9 | 12 | 15 | 18 |

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

# **Simple and Multiple correlation:**

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlation.

# **Partial and total correlation:**

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

# **Computation of correlation:**

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by ‘r’.

**Co-variation:**

The covariation between the variables x and y is defined as-

Cov (XY) =

=

Where, X bar is the mean of X and Y bar is the mean of Y. x and y are deviations from its mean.

**Karl Pearson’s Coefficient of Correlation:**

It is most widely used method in practice and it is known as Pearsonian Coefficient of Correlation. It is denoted by ‘r’. The formula for calculating ‘r’ is-

1. r = ; Where =
2. r =
3. r =

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series respectively.

**Properties of Correlation Coefficient:**

*Property 1:* Correlation coefficient lies between –1 and +1.

*Property 2:* ‘r’ is independent of change of origin and scale.

*Property 3:* It is a pure number independent of units of measurement.

*Property 4:* Independent variables are uncorrelated but the converse is not true.

*Property 5:* Correlation coefficient is the geometric mean of two regression coefficients.

*Property 6:* The correlation coefficient of x and y is symmetric. rxy = ryx.

# **Limitations:**

* 1. Correlation coefficient assumes linear relationship regardless of the assumption is correct or not.
  2. Extreme items of variables are being unduly operated on correlation coefficient.
  3. Existence of correlation does not necessarily indicate cause- effect relation.

# **Interpretation:**

The following rules helps in interpreting the value of ‘ r’ .

1. When r = 1, there is perfect +ve relationship between the variables.
2. When r = -1, there is perfect –ve relationship between the variables.
3. When r = 0, there is no relationship between the variables.
4. If the correlation is +1 or –1, it signifies that there is a high degree of correlation. (+ve or –ve) between the two variables.

If r is near to zero (ie) 0.1,-0.1, (or) 0.2 there is less correlation.

# **Example:**

Find Karl Pearson’s coefficient of correlation from the following data between height of father (x) and son (y).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 64 | 65 | 66 | 67 | 68 | 69 | 70 |
| Y | 66 | 67 | 65 | 68 | 70 | 68 | 72 |

Comment on the result.

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y | *x* = *X* *X*  *x* = x – 67 | *x2* | *y* = *Y* *Y*  *y* = Y - 68 | *y2* | *xy* |
| 64 | 66 | -3 | 9 | -2 | 4 | 6 |
| 65 | 67 | -2 | 4 | -1 | 1 | 2 |
| 66 | 65 | -1 | 1 | -3 | 9 | 3 |
| 67 | 68 | 0 | 0 | 0 | 0 | 0 |
| 68 | 70 | 1 | 1 | 2 | 4 | 2 |
| 69 | 68 | 2 | 4 | 0 | 0 | 0 |
| 70 | 72 | 3 | 9 | 4 | 16 | 12 |
| 469 | 476 | 0 | 28 | 0 | 34 | 25 |

Mean of X = = = 67;

Mean of Y = = = 68.

Hence, Karl Pearson’s Coefficient of Correlation, r = = = = = 0.81. Since r = + 0.81, the variables are highly positively correlated i. e., tall fathers have tall sons.

**Example:**

***Calculate coefficient of correlation from the following data.***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Y | 9 | 8 | 10 | 12 | 11 | 13 | 14 | 16 | 15 |

**Example:**

***Calculate Pearson’s Coefficient of correlation.***

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 45 | 55 | 56 | 58 | 60 | 65 | 68 | 70 | 75 | 80 | 85 |
| Y | 56 | 50 | 48 | 60 | 62 | 64 | 65 | 70 | 74 | 82 | 90 |

**Rank Correlation**

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc. The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. It is defined as-

ρ =

Where, ρ (rho) = rank correlation coefficient;

= sum of squares of differences between the pairs of ranks; and

N = number of pairs of observations.

The value of ρ lies between –1 and +1. If ρ = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If ρ = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5th rank, the common rank to be assigned to each item is = = 5.5 which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is (m3- m). A slight formula is used when there is more than one item having the same value. The formula is-

ρ =

Where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

**Example:**

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Price of tea | 88 | 90 | 95 | 70 | 60 | 75 | 50 |
| Price of coffee | 120 | 134 | 150 | 115 | 110 | 140 | 100 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Price of tea** | **Rank** | **Price of coffee** | **Rank** | **D** | **D2** |
| 88 | 3 | 120 | 4 | 1 | 1 |
| 90 | 2 | 134 | 3 | 1 | 1 |
| 95 | 1 | 150 | 1 | 0 | 0 |
| 70 | 5 | 115 | 5 | 0 | 0 |
| 60 | 6 | 110 | 6 | 0 | 0 |
| 75 | 4 | 140 | 2 | 2 | 4 |
| 50 | 7 | 100 | 7 | 0 | 0 |
|  |  |  |  |  | ∑D2 = 6 |

ρ = = = = 1- 0.1071 = 0.8929

The relation between price of tea and coffee is positive at 0.89. Based on quality the association between price of tea and price of coffee is highly positive.

**Example:**

***In an evaluation of answer script the following marks are awarded by the examiners.***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***1st*** | ***88*** | ***95*** | ***70*** | ***960*** | ***50*** | ***80*** | ***75*** | ***85*** |
| ***2nd*** | ***84*** | ***90*** | ***88*** | ***55*** | ***48*** | ***85*** | ***82*** | ***72*** |

***Do you agree the evaluation by the two examiners is fair?***

**Solution:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **x** | **R1** | **y** | **R2** | **D** | **D2** |
| 88 | 2 | 84 | 4 | 2 | 4 |
| 95 | 1 | 90 | 1 | 0 | 0 |
| 70 | 6 | 88 | 2 | 4 | 16 |
| 60 | 7 | 55 | 7 | 0 | 0 |
| 50 | 8 | 48 | 8 | 0 | 0 |
| 75 | 5 | 82 | 5 | 0 | 0 |
| 80 | 4 | 85 | 3 | 1 | 1 |
| 85 | 3 | 75 | 6 | 3 | 9 |
|  |  |  |  |  | 30 |

ρ = = = = 1- 0.357 = 0.643

ρ = 0.643 shows fair in awarding marks in the sense that uniformity has arisen in evaluating the answer scripts between the two examiners.

**Example:**

***Rank Correlation for tied observations. Following are the marks obtained by 10 students in a class in two tests.***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Students*** | ***A*** | ***B*** | ***C*** | ***D*** | ***E*** | ***F*** | ***G*** | ***H*** | ***I*** | ***J*** |
| ***Test 1*** | ***70*** | ***68*** | ***67*** | ***55*** | ***60*** | ***60*** | ***75*** | ***63*** | ***60*** | ***72*** |
| ***Test 2*** | ***65*** | ***65*** | ***80*** | ***60*** | ***68*** | ***58*** | ***75*** | ***63*** | ***60*** | ***70*** |

***Calculate the rank correlation coefficient between the marks of two tests.***

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Student | Test 1 | R1 | Test 2 | R2 | D | D2 |
| A | 70 | 3 | 65 | 5.5 | -2.5 | 6.25 |
| B | 68 | 4 | 65 | 5.5 | -1.5 | 2.25 |
| C | 67 | 5 | 80 | 1.0 | 4.0 | 16.00 |
| D | 55 | 10 | 60 | 8.5 | 1.5 | 2.25 |
| E | 60 | 8 | 68 | 4.0 | 4.0 | 16.00 |
| F | 60 | 8 | 58 | 10.0 | -2.0 | 4.00 |
| G | 75 | 1 | 75 | 2.0 | -1.0 | 1.00 |
| H | 63 | 6 | 62 | 7.0 | -1.0 | 1.00 |
| I | 60 | 8 | 60 | 8.5 | 0.5 | 0.25 |
| J | 72 | 2 | 70 | 3.0 | -1.0 | 1.00 |
|  |  |  |  |  |  | ∑D2 = 50.00 |

60 is repeated 3 times in test 1. 60,65 is repeated twice in test 2. m = 3; m = 2; m = 2.

ρ =

ρ =

ρ = = = 0.68

**Interpretation:** There is uniformity in the performance of students in the two tests.

**Chapter- 3**

**Basic Concepts of Theory of Probability**

**Objectives:**

* To examine the use of probability theory in decision making;
* To explain the different ways probabilities arise; and
* To develop rules for calculating different types of probabilities.

**Chapter Contents:**

* Basic Terminology;
* Three Types of Probability;
* Probability Rules;
* Probabilities under Conditions of Statistical Independence; and
* Probabilities under Conditions of Statistical Dependence.

**Basic Terminology**

***Probability:***

Probability is the chance something will happen. It is expressed as fractions or as decimals between zero and one.

Assigning a probability of zero means that something can never happen; a probability of 1 indicates that something will be always happen.

***Event:***

In probability theory, an event is one or more of the possible outcomes of doing something. If we toss a coin, getting a tail would be an event, and getting a head would be another event.

***Experiment:***

The activity that produces such an event is called an experiment in probability theory. In a coin- toss experiment, what is the probability of the event head? The answer is ½ or 0.5.

***Sample space:***

The set of all possible outcomes of an experiment is called the sample space for the experiment. . In a coin- toss experiment, the sample space is S = {head, tail}

***Mutually exclusive events:***

Events are said to be mutually exclusive if one and only one of them can take place at a time. In a coin- toss experiment, for example, we have two possible outcomes, heads and tails. On any toss, either heads or tails may turn up, but not both. As a result, the events heads and tails on a single toss are said to be mutually exclusive.

**Types of Probability**

There are three basic ways of classifying probability which are as follow:

1. Classical approach of probability (Classical Probability);
2. Relative frequency approach;
3. Subjective approach.

**Classical probability**

It defines the probability that an event will occur as-

Probability of an event =

***Example:***

What is the probability of getting a head on one toss?

**Solution:** P (head) = = = 0.5

**Note:** *Classical probability is often called a priori probability because if we keep using orderly examples such as fair coins, unbiased dice, and standard decks of cards, we can state the answer in advance (a priori) without tossing a coin, rolling a die, or drawing a card.*

**Shortcomings of the classical probability:**

This approach is useful when we deal with card games, dice games, coin tosses, and the like, but has serious problems when we try to apply it to the less orderly decision problems we encounter in management.

**Relative frequency approach**

In the 1800s, British statisticians, interested in a theoretical foundation for calculating risk of losses in life insurance and commercial insurance, began defining probabilities from statistical data collected on births and deaths.

This method uses the relative frequencies of past occurrences as probabilities. We determine how often something has happened in the past and use that figure to predict the probability that it will happen again in the future. For example, suppose an insurance company knows from past actuarial data that of all males 40 years old, about 60 out of every 100,000 will die within a year period. Using this method, the company estimates the probability of death for that age group as

= 0.0006

Note: when we use the relative frequency approach to establish probabilities, our probability figure will gain accuracy as we increase the number of observations.

**Subjective Probabilities**

Subjective probabilities are based on the beliefs of the person making the probability assessment. In fact, subjective probability can be defined as the probability assigned to an event by an individual, based on whatever evidence is available. This evidence may be in the form of relative frequency of past occurrences, or it may be just an educated guess.

Subjective probability assignments are often found when events occur only once or at most a very few times.

**Probability Rules**

1. **Addition Rules; and**
2. **Multiplication Rules.**

Probability of Event A Happening = P(A)

**Marginal or Unconditional Probability:**

A single probability means that only one event can take place. It is called a marginal or unconditional probability. For example, let us suppose that 50 members of a college class drew tickets to see which student would get a free trip to the National Museum. Any one of the students could calculate his or her chances of winning as:

P (Winning) = = 0.02

In this case, a student’s chance is 1 in 50 because we are certain that the possible events are mutually exclusive, that is, only one student can win at a time.

**Addition Rule for Mutually Exclusive Events:**

Probability of Either A or B Happening **=** P (A or B) = P(A) + P(B)

***Example: Five equally capable students (A, B, C, D and E) are waiting for a summer internship in a company that has announced that it will hire only one of the five by random drawing. What is the probability that A will get internship in the company?***

**Solution:** P (A) = = 0.2

However, if we ask, “what is the probability that either A or B will get internship in the company?”

Probability of Either A or B will get Internship **=** P (A or B) = P(A) + P(B)

= = = 0.4

***Example: Following table gives data on the sizes of families in a certain town. What is the probability that a family chosen at random from this town will have four or more children?***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***No. of Children*** | ***0*** | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6 or more*** |
| ***Proportion of Families having this many children*** | ***0.05*** | ***0.10*** | ***0.30*** | ***0.25*** | ***0.15*** | ***0.10*** | ***0.05*** |

**Solution:** P(4,5, 6 or more) = P(4) +P(5) + P(6 or more)

= 0.15 + 0.10 + 0.05

= 0.30

**Note:** For any event A, either A happens or it doesn’t. So the events A and not A are exclusive and exhaustive.

P(A) + P(not A) = 1

Or, P(A) = 1- P(not A)

**Example: On the basis of the above table, what is the probability of a family having 5 or fewer children?**

**Solution:** P(0, 1, 2, 3, 4, 5) = P(0) +P(1) + P(2) + P(3) +P(4) + P(5)

= 0.05 + 0.10 + 0.30 + 0.25 + 0.15 + 0.10

= 0.95

**OR,**

P(0, 1, 2, 3, 4, 5) = 1- P(6 or more)

= 1- 0.05 = 0.95

**Addition Rule for Events that are Not Mutually Exclusive:**

P (A or B) = P (A) + P (B) – P (AB)

[Probability of A or B happening when A and B are not mutually exclusive = Probability of A happening + Probability of B happening – Probability of A and B happening together]

***Example: Determine the probability of drawing either an ace or a heart.***

**Solution: P(Ace or Heart) = P(Ace) + P(Heart) – P(Ace and Heart)**

**=**  - = =

***Example:*** The employees of a certain company have elected 5 of their number to represent them on the employee- management productivity council. Profiles of the 5 are as follows:

1. Male – age 30 years;
2. Male – age 32 years;
3. Female – age 45 years;
4. Female – age 20 years; and
5. Male – age 40 years.

This group decides to elect a spokesperson by drawing a name from a hat. *What is the probability the spokesperson will be either female or over 35?*

**Solution:** P(Female or over 35 years) = P(Female) + P(Over 35 years) – P(Female and Over 35)

= - =

**Probabilities under Conditions of Statistical Independence**

When two events happen, the outcome of the first even may or may not have an effect on the outcome of the second event. That is, the events may be either dependent or independent. Here, we first examine events that are statistically independent. The occurrence of one event has no effect on the probability of the occurrence of any other event is called statistically independent probability.

There are three types of probabilities under statistical independence:

1. Marginal Probability;
2. Joint Probability; and
3. Conditional Probability.

***Marginal Probability under Statistical Independence:***

A marginal or unconditional probability is the simple probability of the occurrence of an event. In a fair coin toss, P(H) = 0.5, and P(T)= 0.5. This is true for every toss, no matter how many tosses have been made or what their outcomes have been. Every toss stands alone and in no way connected with any other toss. Thus, the outcomes of each toss of a fair coin is an event that is statistically independent of the outcomes of every other toss of the coin.

***Joint Probability under Statistical Independence:***

The probability of two or more independent events occurring together or in succession is the product of their marginal probabilities. Mathematically,

P(AB) = P(A) × P(B)

Where

P(AB) = probability of events A and B occurring together or in succession (this is known as a *joint probability*);

P(A) = marginal probability of event A occurring; and

P(B) = marginal probability of event B occurring.

In terms of the fair coin example, the probability of heads on two successive tosses is the probability of heads on first toss (H1) times the probability of heads on the second toss (H2). That is, *P(H1H2) = P(H1) × P(H2).* We have shown that the events are statistically independent, because the probability of any outcome is not affected by any preceding outcome. Therefore, the probability of heads on any toss is 0.5, and *P(H1H2) = 0.5 × 0.5 = 0.25.*

Likewise, the probability of getting heads on three successive tosses is *P*(*H1H2H3*) = *0.5×0.5×0.5=* 0.125.

We can make the probabilities of events even more explicit using a probability tree.

|  |  |  |
| --- | --- | --- |
| Probability tree of one toss | Probability tree of two tosses | Probability tree of third toss |
| http://flylib.com/books/3/287/1/html/2/images/11fig03.jpg | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **1 Toss** | | **2 Tosses** | | **3 Tosses** | |
| Possible Outcomes | Probability | Possible Outcomes | Probability | Possible Outcomes | Probability |
| H1 | 0.5 | H1H2 | 0.25 | H1H2H3 | 0.125 |
| T1 | 0.5 | H1T2 | 0.25 | H1H2T3 | 0.125 |
|  | **1.0** | T1H2 | 0.25 | H1T2H3 | 0.125 |
|  |  | T1T2 | 0.25 | H1T2T3 | 0.125 |
|  |  |  | **1.00** | T1H2H3 | 0.125 |
|  |  |  |  | T1H2T3 | 0.125 |
|  |  |  |  | T1T2H3 | 0.125 |
|  |  |  |  | T1T2T3 | 0.125 |
| **The sum of the probabilities of all the possible outcomes must always equal 1.** | | | | | **1.000** |

**Example 1:** ***What is the probability of getting tails, heads, tails in that order on three successive tosses of a fair coin?***

**Solution:** *P(T1H2T3) = P(T1) × P(H2) × P(T3)= 0.5×0.5×0.5= 0.125*

**Example 2:** ***What is the probability of getting tails, tails, heads in that order on three successive tosses of a fair coin?***

**Solution:** *P(T1T2H3) = P(T1) × P(T2) × P(H3)= 0.5×0.5×0.5= 0.125*

**Example 3:** ***What is the probability of at least two heads on three tosses?***

**Solution:** Recalling that the probabilities of mutually exclusive events are additive, we can note the possible ways that at least two heads on three tosses can occur, and we can sum their individual probabilities. The outcomes satisfying the requirement are H1H2H3, H1H2T3, H1T2H3, and T1H2H3. Because each of these has an individual probability of 0.125, the sum is 0.5. Thus, the probability of at least two heads on three tosses is 0.5.

**Example 4:** ***What is the probability of at least one tail on three tosses?***

**Solution:** there is only one case in which no tails occur, namely H1H2H3. Therefore, we can simply subtract for the answer:

1. P(H1H2H3) = 1- 0.125 = 0.875

The probability of at least one tail occurring in three successive tosses is 0.875.

Example 5: What is the probability of at least one head on two tosses?

Solution: The possible ways at least one head may occur are H1H2, H1T2, T1H2. Each of these has a probability of 0.25. Therefore, the probability of at least one head on two tosses is 0.75. Alternatively, we could consider the case in which no head occurs- namely, T1T2- and subtract its probability from 1; that is,

1. P(T1T2) = 1- 0.25 = 0.75

***Conditional Probability under Statistical Independence***

Thus far, we have discussed two types of probabilities, marginal (or unconditional) probability and joint probability. Symbolically, marginal probability is P (A) and joint probability is P (AB). Besides these two, there is one other type of probability, known as conditional probability. Symbolically, conditional probability is written as- P (B│A) and is read as “the probability of *event B* given that *event A* has occurred.”

For statistically independent events, the conditional probability of even B given that event A has occurred is simply the probability of event B. That is-

P (B│A) = P(B)

Independent events are those whose probabilities are in no way affected by the occurrence of each other. Symbolically, P (B│A) = P(B).

***Summary of three types of probabilities under statistical independence:***

|  |  |  |
| --- | --- | --- |
| **Types of Probability** | **Symbol** | **Formula** |
| Marginal | P(A) | (PA) |
| Joint | P(AB) | P(A)×P(B) |
| Conditional | P (B│A) | P(B) |

**Probabilities under Conditions of Statistical Dependence**

Statistical dependence exists when the probability of some event is dependent on or affected by the occurrence of some other event. Just as with independent events, there are three types of probabilities under statistical dependence:

1. Conditional;
2. Joint; and
3. Marginal.

***Conditional Probabilities under Statistical Dependence:***

The formula for conditional probability under statistical dependence is given as:

P (B│A) =

Suppose we have a box containing 10 balls distributed as follows:

* 3 are coloured and dotted;
* 1 is coloured and striped;
* 2 are gray and dotted; and
* 4 are gray and striped.

The probability of drawing any one ball from this box is 0.1, since there are 10 balls, each with equal probability of being drawn.

**Example 1:** ***Suppose someone draws a coloured ball from the box. What is the probability that it is dotted? What is the probability it is striped?***

**Solution:** P(D│C) = = = 0.75

P(S│C) = = = 0.25

**Example 2:** ***On the basis of above example what is the probability of getting dotted ball given the probability of gray ball? What is the probability of getting striped ball given the probability of gray ball?***

**Solution:** P (D│G) = = =

P (S│G) = = =

***Explanation:*** The total probability of gray is 0.6 (6 out of 10 balls). To determine the probability that the ball (which we know is gray) will be dotted, we divide the probability of gray and dotted (0.2) by the probability of gray (0.6), or 0.2/0.6 = 1/3. Similarly, to determine the probability that the ball will be striped, we divide the probability of gray and striped (0.4) by the probability of gray (0.6), or 0.4/0.6 = 2/3.

**Example 3:** ***Calculate* P (G│D) *and* P (C│D*) on the basis of above example.***

**Solution:** P (G│D) ***=***  = = = 0.4

P (C│D) ***=***  = = = 0.6

**Example 4:** ***Calculate* P (C│S) *and* P (G│S) *on the basis of above example.***

**Solution:** P (C│S) ***=***  = = = 0.2

P (G│S) ***=***  = = = 0.8

***Joint Probability under Statistical Dependence:***

The formula for calculation of joint probability for statistical dependence is given as-

P (BA) = P (B│A) × P (A)

i.e., *Joint probability of events B and A happening together or in succession = Probability of event B given that event A has happened × Probability that event A will happen.*

Converting the general formula P (BA) = P (B│A) × P (A) to our example and to the terms of coloured, gray, dotted and striped, we have P(CD) = P (C│D) × P (D) = 0.6×0.5 = 0.3. Here, 0.6 is the probability of coloured; given dotted and 0.5 is the probability of dotted.

The following joint probabilities are computed in the same manner and can also be verified by direct observation:

P (CS) = P (C│S) × P (S) = 0.2×0.5 = 0.1

P (GD) = P (G│D) × P (D) = 0.4×0.5 = 0.2

P (GS) = P (G│S) × P (S) = 0.8×0.5 = 0.4

***Marginal Probabilities under Statistical Dependence:***

Marginal probabilities under statistical dependence are computed by summing up the probabilities of all the joint events in which the simple event occurs. In the example above, we can compute the marginal probability of the event colored by summing the probabilities of the two joint events in which colored occurred:

P(C) = P(CD) + P(CS) = 0.3 + 0.1 = 0.4

Similarly, the marginal probability of the event gray can be computed by summing the probabilities of the two joint events in which gray occurred:

P(G) = P(GD) + P(GS) = 0.2 + 0.4 = 0.6

In the same way, we can compute the marginal probability of the event dotted by summing the probabilities of the two joint events in which dotted occurred:

P(D) = P(CD) + P(GD) = 0.3 + 0.2 = 0.5

And finally, the marginal probability of the event striped can be computed by summing the probabilities of the two joint events in which gray occurred:

P(S) = P(CS) + P(GS) = 0.1 + 0.4 = 0.5

***Summary of three types of probabilities under statistical dependence:***

|  |  |  |  |
| --- | --- | --- | --- |
| **Types of Probability** | **Symbol** | **Formula under Statistical Independence** | **Formula under Statistical Dependence** |
| Marginal | P(A) | P(A) | Sum of the probabilities of the joint events in which A occurs. |
| Joint | P(AB)  Or, P(BA) | P(A)×P(B)  P(B)×P(A) | P(A⃒B)×P(B)  P(B⃒A)×P(A) |
| Conditional | P(B⃒A)  Or, P(A⃒B) | P(B)  P(A) |  |

\*\*\*\*\*

Chapter- 4

**Probability Distributions**

**Objectives:**

* To introduce probability distributions most commonly used in decision making;
* To show which probability distribution to use and how to find its values; and
* To understand the limitations of each of the probability distributions you use.

**Chapter Contents:**

* Basic Terms Introduced in this Chapter
* What is Probability Distribution?
* Random Variables;
* Use of Expected Value in Decision Making;
* The Binomial Distribution;
* The Poisson Distribution;
* The Normal Distribution;
* Choosing the Correct Probability Distribution.

**Basic Terms Introduced in this Chapter**

***Probability Distribution:***

A list of the outcomes of an experiment with the probabilities we would expect to see associated with these outcomes is called probability distribution.

***Discrete Probability Distribution:***

A probability distribution in which the variable is allowed to take on only a limited number of values, which can be listed, is called discrete probability distribution.

***Random Variable:***

A variable whose values are determined by chance is called random variable.

***Continuous Random Variable:***

A random variable allowed to take on any value within a given range is called continuous random variable.

***Discrete Random Variable:***

A random variable that is allowed to take on only a limited number of values, which can be listed is called discrete random variable.

***Expected Value:***

A weighted average of the outcomes of an experiment is called expected value.

***Binomial Distribution:***

A discrete distribution describing the results of an experiment is known as binomial distribution.

***Poisson Distribution:***

A discrete distribution in which the probability of the occurrence of an event within a very small time period is a very small number, the probability that two or more such events will occur within the same time interval is effectively 0, and the probability of the occurrence of the event within one time period is independent of where that time period is.

***Normal Distribution:***

A distribution of a continuous random variable with a single- peaked, bell- shaped curve. The mean lies at the center of the distribution, and the curve is symmetrical around a vertical line erected at the mean. The two tails extend indefinitely, never touching the horizontal axis.

***Standard Normal Probability Distribution:***

A normal probability distribution, with mean μ = 0 and standard deviation σ = 1 is called standard normal probability distribution.

**Theoretical or Expected Frequency Distributions**

Following are various types of theoretical or expected frequency distributions:

1. Binomial Distribution,
2. Multinomial Distribution,
3. Negative Binomial Distribution,
4. Poisson Distribution,
5. Hypergeometric Distribution, and
6. Normal Distribution.

Amongst these the first five distributions are of discrete type and the last one is of continuous type. In these six distributions binomial, poisson and normal distributions have much more wider application in practice. So we shall discuss these three.

**Binomial Distribution**

The binomial distribution describes discrete, not continuous, data, resulting from an experiment known as a Bernoulli process, after the 17th century Swiss mathematician Jacob Bernoulli. The tossing of a fair coin a fixed number of times is a Bernoulli process, and the outcomes of such tosses can be represented by the binomial probability distribution. The success or failure of interviewees on an aptitude test may also be described by a Bernoulli process.

**Use of the Bernoulli Process:**

We can use the outcomes of a fixed number of tosses of a fair coin as an example of a Bernoulli process. We can describe this process as follows:

1. Each trial has only two possible outcomes: heads or tails, yes or no, success or failure.
2. The probability of the outcomes of any trial remains fixed over time. With a fair coin, the probability of heads remains 0.5 for each toss regardless of the number of times the coin is tossed.
3. The trials are statistically independent; that is, the outcome of one toss does not affect the outcome of any other toss.

**Binomial Formula:**

Probability of r successes in n trials =

Where p = characteristic probability or probability of success

q = (1-p) = probability of failure

r = number of successes desired

n = number of trials undertaken.

**Example:** ***Calculate the chances (probability) of getting exactly two heads (in any order) on three tosses of a fair coin.***

**Solution:** We can use the above binomial formula to calculate desired probability. For this we can express the values as follows:

p = characteristic probability or probability of success = 0.5

q = (1-p) = probability of failure = 0.5

r = number of successes desired = 2

n = number of trials undertaken = 3

Probability of 2 successes (heads) in 3 trials =

=

= = 0.375

Thus, there is a 0.375 probability of getting two heads on three tosses of a fair coin.

**Mean of a Binomial Distribution, *μ = np***

Where

n = number of trials

p = probability of success

**Standard Deviation of Binomial Distribution, *σ =***

Where

n = number of trials

p = probability of success

q = probability of failure = 1- p

**Example:** ***A packaging machine that produces 20 percent defective packages. If we take a random sample of 10 packages, what is the mean and standard deviation of the binomial distribution?***

**Solution:** Mean, μ = np = 10×0.2 = 2

Standard Deviation, σ = = = 1.265

**The Poisson Distribution**

It is a discrete probability distribution developed by a French mathematician Simeon Denis Poisson. It may be expected in cases where the chance of any individual event being a success is small. This distribution is used to describe the behaviour of rare events such as the number of accidents on road, number of printing mistakes in a book, etc., and has been called “*the law of improbable events*”.

**Poisson Formula:**

Probability of exactly X occurrences, ***P(X) =***

Where

lambda (the mean number of occurrences per interval of time) raised to the power x

e, or 2.71828 (the base of the Napierian, or natural, logarithm system), raised to the power negative lambda,

= x factorial.

**Example:** ***Suppose that we are investigating the safety of a dangerous intersection. Past police records indicate a mean of five accidents per month at this intersection. The number of accidents is distributed according to a Poisson distribution, and the Highway Safety Division wants us to calculate the probability in any month of exactly 0, 1, 2, 3, or 4 accidents.***

Solution: Using the Poisson formula, we can calculate the probability of no accidents:

P(0) = = = = 0.00674

For exactly one accident:

P(1) = = = = 0.03370

For exactly two accidents:

P(2) = = = = 0.08425

For exactly three accidents:

P(3) = = = = 0.14042

For exactly four accidents:

P(4) = = = = 0.17552

Our calculations will answer several questions. If we want to know the probability of 0, 1, or 2 accidents in any month, we can add these probabilities as:

P(0,1, or 2) = P(0) + P(1) + P(2)

= 0.00674 + 0.03370 + 0.08425 = 0.12469

For, P(3 or fewer) = P(0, 1, 2, or 3) = P(0) + P(1) + P(2) + P(3)

= 0.00674 + 0.03370 + 0.08425 + 0.14042 = 0.26511

If we want to calculate the probability of more than three then we must be 0.73489 (1- 0.26511).

**Important Point:**

*The poisson distribution is a good approximation of the binomial distribution when n is greater than or equal to 20 and p is less than or equal to 0.05.*

**The Normal Distribution**

It is a continuous probability distribution developed by Karl Gauss. The normal probability distribution is often called Gaussian distribution.

The normal curve is represented in several forms. The following is the basic form relating to the curve with mean μ and standard deviation σ:

**The Normal Distribution, *P(X) = .***

Where

*X = values of the continuous random variable*

*μ = mean of the normal random variable*

*e = mathematical constant (= 2.7183)*

*π = mathematical constant (= 3.1416)*

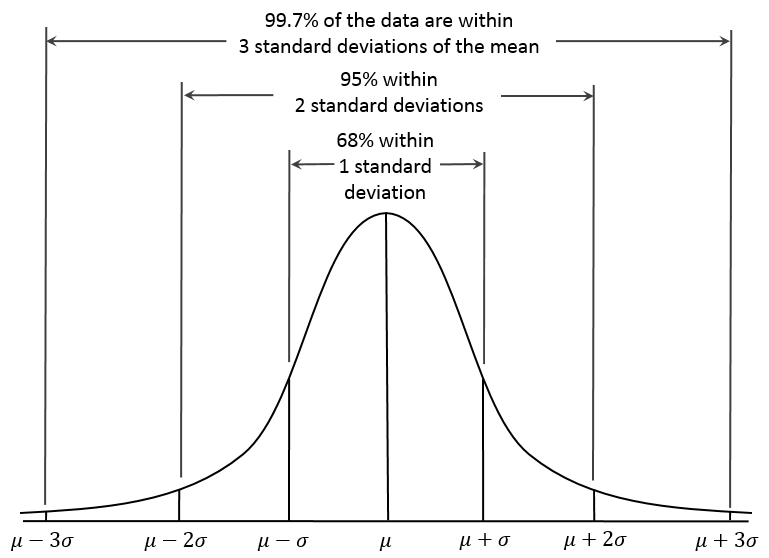
**Characteristics (Graph) of Normal Probability Distribution:**

* The curve has a single peak; thus it is unimodal.
* The normal curve is” bell- shaped” and symmetric.
* For a normal probability distribution, mean median and mode all are equal.
* The two tails of the normal probability distribution extend indefinitely and never touch the horizontal axis.

**Areas under the Normal Curve:**

No matter what the values of mean (π) and standard deviation (σ) are for a normal probability distribution, the total area under the normal curve is 1.00. Mathematically it is true that-

* Approximately 68% of all the values in a normally distributed population lie within ±1σ from mean (π);
* Approximately 95.5% of all the values in a normally distributed population lie within ±2σ from mean (π); and
* Approximately 99.7% of all the values in a normally distributed population lie within ±3σ from mean (π). These are shown in the following graph:



**Formula for measuring distances under normal curve:**

**Standardizing a Normal Random Variable,**

*Z =*

Where

*x = value of the random variable with which we are concerned;*

*μ = mean of the distribution of this random variable;*

*σ = standard deviation of this distribution;*

*z = number of standard deviations from x to the mean of this distribution.*

**Example 1: *What is the probability that a participant selected at random will require more than 500 hours to complete the training program?***

**Solution:** We can see that half of the area under the curve is located on either side of the mean of 500 hours. Thus, we can deduce that the probability that the random variable will take on a value higher than 500 is half, or 0.5.

**Example 2: *What is the probability that a candidate selected at random will take between 500 and 650 hours to complete the training program?***

**Solution:** The probability that will answer this question is the area between the mean (π = 500 hours) and the x value in which we are interested (650 hours). Using equation, we get a z value of

*Z = = = =1.5 standard deviation*

If we look up z = 1.5 in Z- table, we find a probability of 0.4332. Thus, the chance that a candidate selected at random would require between 500 and 650 hours to complete the training program is slightly higher than 0.4.

**Example 3: *What is the probability that a candidate selected at random will take more than 700 hours to complete the training program?***

**Solution:** This situation is different from the above example 2. We are interested in the area to the right of the value 700 hours. So, first we will find out z value by using the formula-

Z = = = =2.0 standard deviation

Looking in the Z- table for z value of 2.0, we find a probability of 0.4772. That represents the probability the program will require between 500 and 700 hours. But, we have to find out the probability that take more than 700 hours. Because the right half of the curve (between the mean and the right- hand tail) represents a probability of 0.5, we can get our answer (the area to the right of the 700- hour point) if we subtract 0.4772from 0.5; 0.5000- 0.4772 = 0.0228. Therefore, there are just over 2 percent chances (or 2 out of 100) that a participant chosen at random would take more than 700 hours to complete the course.

**Example 4:** ***Suppose the training program director wants to know the probability that a participant chosen at random would require between 550 and 650 hours to complete the required work.***

**Solution:** First calculate a z value for the 650 hour point, as follows:

Z = = = =1.5 standard deviation

When we look up a z of 1.5 in z table, we see a probability value of 0.4332 (the probability that the random variable will fall between the mean and 650 hours). Now we calculate a z value for 550 hours as follows:

Z = = = =0.5 standard deviation

When we look up a z of 0.5 in z table, we see a probability value of 0.1915 (the probability that the random variable will fall between the mean and 550 hours). To answer our question, we must subtract as follows to get probability that the random variable will lie between 550 and 650 hours = 0.4332 – 0.1915 = 0.2417.

Thus, the chance of a candidate selected at random taking between 550 and 650 hours to complete the program is 24 in 100.

**Example 5: *What is the probability that a candidate selected at random will require fewer than 580 hours to complete the program?***

**Example 6: *What is the probability that a candidate chosen at random will take between 420 and 570 hours to complete the program?***

Chapter- 5

**Introduction to Estimation**

**Objectives:**

* To learn how to estimate certain characteristics of a population from samples;
* To learn the strengths and shortcomings of point estimates and interval estimates;
* To calculate how accurate our estimates really are;
* To learn how to use the t distribution to make interval estimates in some cases when the normal distribution cannot be used; and
* To calculate the sample size required for any desired level of precision in estimation.

**Chapter Contents:**

* Some Important Terminology Used in this Chapter;
* Point Estimates;
* Interval Estimates: Basic Concepts;
* Interval Estimates and Confidence Intervals;
* Calculating Interval Estimates of the Mean from Large Samples;
* Calculating Interval Estimates of the Proportion from Large Samples;
* Interval Estimates Using the t Distribution; and
* Determining the Sample Size in Estimation.

**Basic Terminology**

***Estimate:***

A specific observed value of an estimator is called estimates. The objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic. E.g., the sample mean (x) is employed to estimate the population mean (μ).

***Estimator:***

A sample statistic used to estimate a population parameter is called estimator. A good estimator should be unbiased, efficient, consistent and sufficient.

***Point Estimate:***

A single number used to estimate an unknown population parameter.

***Interval Estimate:***

A range of values to estimate an unknown population parameter is called interval estimate.

***Confidence Interval:***

A range of values that has some designated probability of including the true population parameter value.

***Confidence Level:***

The probability that statisticians associate with an interval estimate of a population parameter, indicating how confident they are that the interval estimate will include the population parameter.

***Confidence Limits:***

The upper and lower boundaries of a confidence interval are known as confidence limits.

***Degrees of Freedom:***

The number of values in a sample we can specify freely once we know something about that sample.

**Types of Estimates**

We can make two types of estimates about population: a *point estimate* and an *interval estimate*. A point estimate is a single number that is used to estimate an unknown population parameter. A point estimate is often insufficient, because it is either right or wrong. An interval estimate is a range of values used to estimate a population parameter. It indicates the error in two ways: by the extent of its range and by the probability of the true population parameter lying within that range.

**Point Estimates**

A ***point estimator*** draws inferences about a population by estimating the value of an unknown parameter using a single value or point.

The sample mean *x̅* is best estimator of the population mean μ. It is unbiased, consistent, the most efficient estimator, and, as long as the sample is sufficiently large, its sampling distribution can be approximated by the normal distribution.

*Sample mean,* x̅ =

*Sample Variance, s2* =

**Interval Estimates: Basic Concepts**

An interval estimate describes a range of values within which a population parameter is likely to lie.

An ***interval estimator*** draws inferences about a population by estimating the value of an unknown parameter using an interval.

That is we say (with some \_\_\_% certainty) that the population parameter of interest is between some lower and upper bounds.

*For example* suppose we want to estimate the mean summer income of a class of business students. For n=25 students, sample mean (x̅ = point estimate- certainty) is calculated to be 400 $/week.

***An alternative statement is:*** The mean income is between 380 and 420 $/week (interval estimate- an uncertainty).

To provide such an uncertain statement, we need to find out the *standard error of the mean (S.E.)*. The formula for S.E. is given as-

**σx̅ =**

Where

σx̅ = Standard Error of Mean

Standard deviation of the population

n = no of sample observations.

**Confidence Interval Estimator for μ:**

The probability 1- α is called the ***confidence level (1- α).*** Usually represented as-

x̅±zα/2 = x̅±zα/2 σx̅ = x̅±zα/2SE

**Lower Confidence Level (LCL)** = x̅- zα/2 = x̅- zα/2 σx̅ = x̅- zα/2SE

**Upper Confidence Level (UCL)** = x̅+ zα/2 = x̅ + zα/2 σx̅ = x̅ + zα/2SE

**Note:** The probability is 0.955 that the mean of a sample will be within ± 2 standard errors of the population mean. In other words, 95.5 percent of all the sample means are within ± 2 standard errors (±2σx̅) from μ, and hence μ is within ±2 standard errors of 95.5 percent of all the sample means. So,

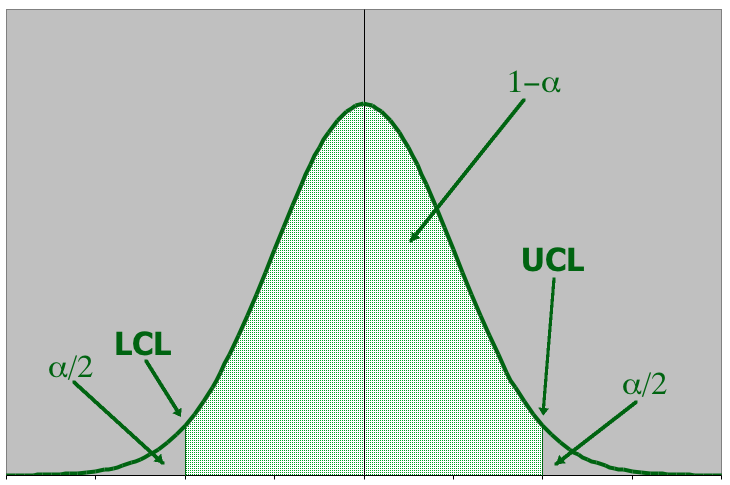
x̅ ± 1σx̅ covers 68.3 percent confidence level;

x̅ ± 2σx̅ covers 95.5 percent confidence level; and

x̅ ± 3σx̅ covers 99.7 percent confidence level.

As far as any particular interval is concerned, it is either contains the population mean or it does not, because the population mean is a fixed parameter.

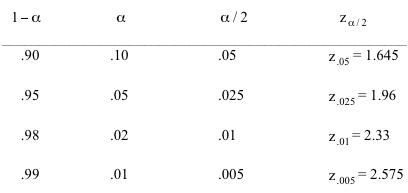
**The *actual* location of the population mean:**



|  |  |  |
| --- | --- | --- |
| … may be here | ….. or here | .or possibly here |

The population mean is a fixed but ***unknown*** quantity. It is incorrect to interpret the confidence interval estimate as a probability statement about. The interval acts as the lower and upper limits of the interval ***estimate*** of the population mean.

**Four commonly used confidence levels:**



***Example 1:***

For a population with a known variance (σ2) of 185, a sample of 64 individuals leads to 217 as an estimate of the mean.

1. Find the standard error of the mean.
2. Establish an interval estimate that should include the population mean 68.3 percent of the time.

**Solution: The Standard Error of Mean, σx̅** =  = = = 1.70

Interval Estimate at 68.3 percent = x̅ ± 1σx̅ = 217±1.70

Lower Confidence Level (LCL) = 217- 1.70 = 215.3

Upper Confidence Level (UCL) = 217+ 1.70 = 218.7

***Example 2:***

Muhammad is a frugal undergraduate management student at King Saud University who is interested in purchasing a used car. He randomly selected 125 want ads and found that the average price of a car in this sample was SR 10,000. He knows that the standard deviation of used- car prices in Riyadh is SR 200.

(a) Establish an interval estimate for the average price of a car so that Muhammad can be 68.3 percent certain that the population mean lies within this interval.

(b) Establish an interval estimate for the average price of a car so that Muhammad can be 95.5 percent certain that the population mean lies within this interval.

**Solution:** (a) **Interval Estimate at 68.3 percent** = x̅ ± 1σx̅

= 10,000± = 10,000± =10,000± = 10,000±17.88

Lower Confidence Level (LCL) = 10,000- 17.88 = 9982.12

Upper Confidence Level (UCL) = 10,000+ 17.88 = 10,017.88

(b) **Interval Estimate at 95.5 percent** = x̅ ± 2σx̅

= 10,000± = 10,000± =10,000± = 10,000±35.77

Lower Confidence Level (LCL) = 10,000- 35.77 = 9964.22

Upper Confidence Level (UCL) = 10,000+ 35.77 = 10,035.77

***Example 3:***

From a population known to have a standard deviation of 1.4, a sample of 60 individuals is taken. The mean for this sample is found to be 6.2.

1. Find the standard error of mean.
2. Establish an interval estimate around the sample mean, using one standard error of the mean.

**Solution: (a) The Standard Error of Mean, σx̅** =  = = = 0.18

(b) **Interval Estimate at 1 S.E.** = x̅ ± 1σx̅ = 6.2±0.18

Lower Confidence Level (LCL) = 6.2- 0.18 = 6.02

Upper Confidence Level (UCL) = 6.2+ 0.18 = 6.38

***Example 4:***

A computer company samples demand during lead time over 25 time periods:



It is known that the standard deviation of demand over lead time is 75 computers. We want to estimate the ***mean*** demand over lead time with 95% confidence in order to set inventory levels…

“We want to estimate the ***mean*** demand over lead time with 95% confidence in order to set inventory levels…”

**Solution:** Thus, the parameter to be estimated is the population mean: μ

And so our confidence interval estimator will be: x̅±zα/2

In order to use our confidence interval estimator, we need the following pieces of data:

|  |  |  |
| --- | --- | --- |
| x̅ | 370.16 | Calculated from the data |
| zα/2 | 1.96 | 1- α = 95% = 0.95, hence α/2 = 0.025, so zα/2 = z0.025 = 1.96 |
| σ | 75 | Given |
| n | 25 | Given |

Therefore:

x̅±zα/2 = : 370.16±z0.025 = 370.16 ± 1.96 = 370.16 ± 29.40

The **lower** and **upper** confidence limits are 340.76 and 399.56.

**When the Population Standard Deviation is Unknown**

***Example 5:***

When we are interested in estimating the mean annual income of 700 families living in four-square- block section of a community. We take a simple random sample and find these results:

n = 50 (sample size)

x̅ = SR 11,800 (Sample mean)

s = SR 950 (Sample standard deviation)

You have to calculate an interval estimate of the mean annual income of all 700 families so that it can be 90 percent confident that the population mean falls within that interval.

**Solution:** The sample size is over 30, so the central limit theorem enables us to use the normal distribution as the sampling distribution. Here, we do not know the population standard deviation, and so we will use the sample standard deviation to estimate the population standard deviation:

σ̂ = s =;

Where σ̂ = Estimate of the population standard deviation

Now we can estimate the standard error of the mean. Because we have a finite population size of 700, and because our sample is more than 5 percent of the population, we will use the formula for deriving the standard error of the mean of finite populations:

σx**̅** =  ×;

But because we are calculating the standard error of the mean using estimate of the standard deviation of the population, we must rewrite this equation so that it is correct symbolically:

σ̂x**̅**= ×

= × = 129.57 (Estimate of the standard error of the mean of a finite population)

Now, we consider the 90 percent confidence level, which would include 45 percent of the area on either side of the mean of the sampling distribution. Looking in the Z- table (area under the standard normal probability distribution between the mean and positive values of z) for the 0.45 value, we find that about 0.45 of the area under the normal curve is located between the mean and a point 1.64 standard errors away from the mean. Therefore, 90 percent of the area is located between plus and minus 1.64 standard errors away from the mean, and our confidence limits are:

x̅ ± 1.64 σ̂x**̅**= 11,800±1.64(129.57)

= 11,800±212.50

Lower Confidence Level (LCL) = 11,800- 212.50 = 11,587.50

Upper Confidence Level (UCL) = 11,800+ 212.50 = 12,012.50

So, we can report with 90 percent confidence that the average annual income of all 700 families living in the four-square- block section falls between SR 11,587.50 and SR 12,012.50.

**Calculating Interval Estimates of the Proportion from Large Samples**

Mean of the Sampling Distribution of the Proportion, μp̅ = p

Standard Error of the Proportion, σp̅ =

Estimated Standard Error of the Proportion, σ̂p̅ =

***Example 6:***

When a sample of 70 retail executives was surveyed regarding the poor November performance of the retail industry, 66 percent believed that decreased sales were due to unseasonably warm temperatures, resulting in consumers’ delaying purchase of cold- weather items.

1. Estimate the standard error of the proportion of retail executives who blame warm weather for low sales.
2. Find the upper and lower confidence limits for this proportion, given a 95 percent confidence level.

**Solution:** n = 70; p̅ = 0.66; so q̅ = 0.34

1. Standard Error of the Proportion, σ̂p̅ = = = 0.0566
2. p̅±1.96 σ̂p̅ = 0.66 ± 1.96 (0.0566) = 0.66 ± 0.111 = (0.549, 0.771)

**Interval Width…**

***A wide interval provides little information***.

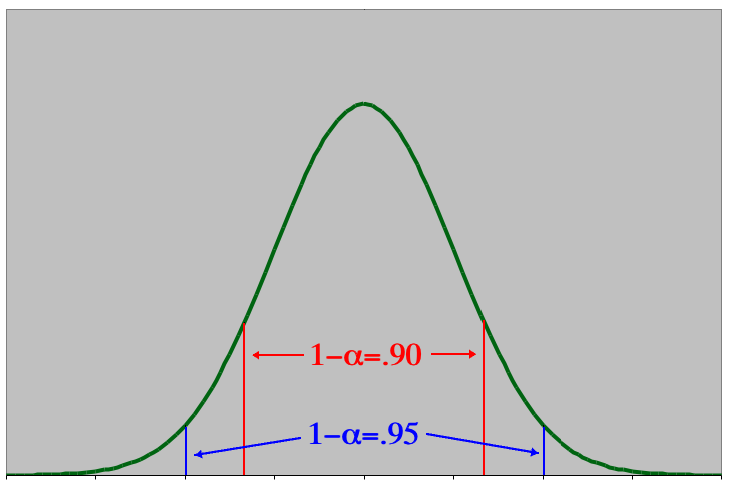
*For example*, suppose we estimate with 95% confidence that an accountant’s average starting salary is between $15,000 and $100,000.

***Contrast*** this with: a 95% confidence interval estimate of starting salaries between $42,000 and $45,000.

The second estimate is much narrower, providing accounting students more precise information about starting salaries.

The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size; *x̅±zα/2*where *zα/2*indicates confidence level, ***σ*** shows population standard deviation and ***n*** denotes sample size.

A larger confidence level produces a **wider** confidence interval.



Increasing the sample size decreases the width of the confidence interval while the confidence level can remain unchanged.

**Note:** this also increases the ***cost*** of obtaining additional data

**Selecting the Sample Size:**

We can control the width of the interval by determining the sample size necessary to produce narrow intervals.

Suppose we want to estimate the mean demand “to within 5 units”; i.e. we want to the interval estimate to be: *x̅±5*

Since: *x̅±zα/2*

It follows that *zα/2* = 5

Solving this equation for n, we get

n = (*zα/2*)2 = ()2 = 865

That is, to produce a 95% confidence intervals estimate of the mean (±5 units), we need to sample 865 lead time periods (vs. the 25 data points we have currently).

**Sample Size to Estimate a Mean:**

The general formula for the sample size needed to estimate a population mean with an interval estimate of: *x̅±W*

Requires a sample size of at least this large: ***n = (zα/2)2***

***Question 1:*** A lumber company must estimate the mean diameter of trees to determine whether or not there is sufficient lumber to harvest an area of forest. They need to estimate this to within 1 inch at a confidence level of 99%. The tree diameters are normally distributed with a standard deviation of 6 inches. How many trees need to be sampled?

***Solution:***

**Interval Estimates Using the t Distribution**

How can we handle estimates where the normal distribution is not the appropriate sampling distribution, in other words when we are estimating the population standard deviation and the sample size is 30 or less? In that case we will use t distribution to solve these types of questions.

W. S. Gosset (his pen name was *student*) gave the concept of *t distribution*. It is also known as *student’s t distribution* or simply *student’s distribution*.

*Use of the t distribution for estimating is required whenever the sample size is 30 or less and the population standard deviation is not known. Furthermore, in using the t distribution, we assume that the population is normal or approximately normal.*

**Characteristics of the t Distribution:**

|  |  |
| --- | --- |
| * Like normal distribution, t distribution is also bell- shaped symmetric distribution. * In general, the t distribution is flatter than the normal distribution, and there is a different t distribution for every possible sample size. As the sample size gets larger, the shape of the t distribution loses its flatness and becomes approximately equal to the normal distribution. * A t distribution is lower at the mean and higher at the tails than normal distribution. | http://image.slidesharecdn.com/samplingexcel-13332144521611-phpapp02-120331125837-phpapp02/95/sampling-excel-11-728.jpg?cb=1333199544 |

**Degrees of Freedom (n – 1):**

It can be define as the number of values we can choose freely. For example, assume that we are dealing with two sample values, *a* and *b*, and we know that they have a mean of 20. Symbolically, the situation is = 20. How can we find what values a and b can take on this situation? The answer is that *a* and *b* can be any values whose sum is 40, because 40÷2= 20.

Suppose we learn that *a* has a value of 15. Now b is no longer free to take on any value but must have the value of 25, because

If a = 15, then = 20

= 20 = 15+b = 40 => b = 40 – 15 = 25.

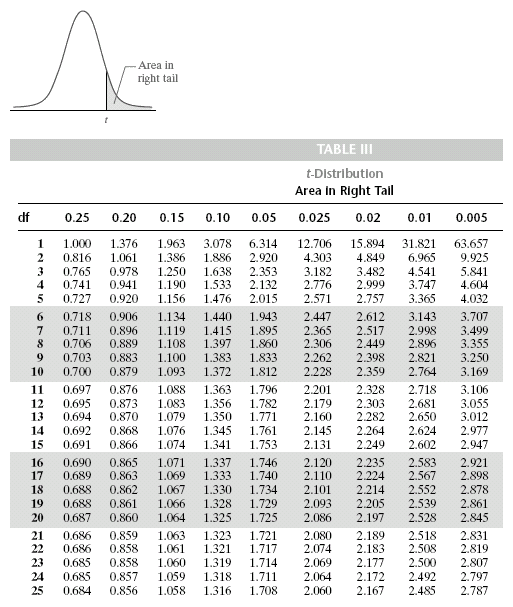
So, we can say that the degree of freedom, or the number of variables we can specify freely, is (n-1) = 2-1 = 1.

**Using the t Distribution Table:**

The t table is more compact and shows areas and t values for only a few percentages (10, 5, 2 and 1 percent). Because there is a different t distribution for each number of degrees of freedom, a more complete table would be quite lengthy.

A second difference in the t table is that it does not focus on the chance that the population parameter being estimated will fall within our confidence interval. Instead, it measures the chance that the population parameter we are estimating will not be within our confidence interval (i.e., it will lie outside it).

A third difference in using the t table is that we must specify the degrees of freedom with which we are dealing. Suppose we make an estimate at 90 percent confidence level with a sample of 15, which is 14 degrees of freedom (n-1). Look in the following table under the 0.10 column until you encounter the row labeled 14. Like a z value, the t value there of 1.345 shows that if we mark off plus and minus 1.345 σ̂x̄’s (estimated standard error of x̅) on either side of the mean, the area under the curve between these two limits will be 90 percent, and the area outside these limits (the chance of error) will be 10 percent.



**Summary of Confidence Limits under Various Conditions**

|  |  |  |
| --- | --- | --- |
|  | When the Population is Finite  (and n/N > 0.05) | When the Population is Infinite  (or n/N > 0.05) |
| *Estimating μ (the population mean):* when σ (the population standard deviation) is known | Confidence Limits = x̅ ± z × | Confidence Limits = x̅ ± z |
| *When σ (the population standard deviation) is not known (σ̂ = s):* when n > 30 | Confidence Limits = x̅ ± z × | Confidence Limits = x̅ ± z |
| *When n is 30 or less and the population is normal or approximately normal* | This case is beyond your course. | Confidence Limits = x̅ ± t |
| *Estimating p (the population proportion): when n > 30*  *σ̂p̅ =* | This case is beyond your course. | Confidence Limits = p̅ ± z |

Chapter- 6

**The Chi-Square Statistic**

**Objectives:**

* To introduce the chi- square and F distributions and learn how to use them in statistical inferences;
* To use the chi- square distribution to see whether two classifications of the same data are independent of each other;
* To use a chi- square test to check whether a particular collection of data is well described by a specified distribution;
* To use the chi- square distribution for confidence intervals and testing hypotheses about a single population variance;
* To compare more than two population means using analysis of variance (ANOVA); and
* To use the F distribution to test hypotheses about two population variances.

**Contents:**

* Basic Terminology Used in this Chapter;
* Chi- Square as a Test of Independence;
* Chi- Square as a Test of Goodness of Fit: Testing the Appropriateness of Distribution;
* Analysis of Variance (ANOVA);
* Inferences about a Population Variance; and
* Inferences about Two Population Variances.

**Basic Terminology**

***Chi- Square Distribution:***

A family of probability distributions, differentiated by their degrees of freedom, used to test a number of hypotheses about variances, proportions and distributional goodness of fit.

***Goodness- of- Fit Test:***

A statistical test for determining whether there is a significant difference between an observed frequency distribution and a theoretical probability distribution hypothesized to describe the observed distribution.

***Test of Independence:***

A statistical test of proportions of frequencies to determine whether membership in categories of one variable is different as a function of membership in the categories of a second variable.

***Expected Frequencies:***

The frequencies we would expect to see in a contingency table or frequency distribution if the null hypothesis is true.

***Analysis of Variance (ANOVA):***

A statistical technique used to test the equality of three or more sample means and thus make inferences as to whether the samples come from populations having the same mean.

***F- Distribution:***

A family of distributions differentiated by two parameters (df- numerator, df- denominator), used primarily to test hypotheses regarding variances.

***R- Ratio:***

A ratio used in the analysis of variance, among other tests, to compare the magnitude of two estimates of the population variance to determine whether the two estimates are approximately equal; in ANOVA, the ratio of between – column variance to within- column variance is used.

***Between- Column Variance:***

An estimate of the population variance derived from the variance among the sample means.

***Within- Column Variance:***

An estimate of the population variance based on the variances within the k samples, using a weighted average of k sample variances.

***Contingency Table:***

A table having R rows and C columns. Each row corresponds to a level of one variable, each column to a level of another variable. Entries in the body of the table are the frequencies with which each variable combination occurred.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Introduction:**

This chapter introduces two non-parametric hypothesis tests using the chi-square statistic: the chi-square test for goodness of fit and the chi-square test for independence.

The term "non-parametric" refers to the fact that the chi‑square tests do not require assumptions about population parameters nor do they test hypotheses about population parameters.

The t- tests and analysis of variance are **parametric tests** and they do include assumptions about parameters and hypotheses about parameters.

The most obvious difference between the chi‑square tests and the other hypothesis tests we have considered (t and ANOVA) is the nature of the data.

Chi-square (χ2) procedures measures the differences between observed (O) and expected (E) frequencies of nominal variables, in which subjects are grouped in categories or cells. There are three basic uses of chi-square analysis, *the Goodness of Fit Test* (used with a single nominal variable), *the Test of Independence (*used with two nominal variables) and the test of homogeneity. These types of chi-square use the same formula.

The chi-square formula is as follows:

χ2 = ∑ (O-E)2 ÷ E

Where O = observed frequency (the actual count -- in a given cell);

E = expected frequency (a theoretical count -- for that cell). Its value must be computed.

For chi‑square, the data are frequencies rather than numerical scores.

**Conditions or Assumptions for Applying χ2 Test:**

1. Large number (generally not less than 50) of observations or frequencies;
2. Expected frequency should not be small (less than 5). If it is less than 5, then frequencies taken from adjacent items or cells are pooled in order to make it 5 or more than 5.Yate’s correction may also be applied in such case;
3. Data should be in original units such as percentage or proportion;
4. Random sampling; and
5. Events should be mutually exclusive.

**The Chi-Square Test for Goodness-of-Fit**

The Goodness of Fit Test is applied to a single nominal variable and determines whether the frequencies we observe in k categories fit what we might expect. Some textbooks call this procedure *the Badness of Fit Test* because a significant χ2 value means that Observed counts do not fit what we expect. The Goodness of Fit Test can be applied with *equal or proportional expected frequencies* (EE, PE).

***Equal Expected (EE) Frequencies:***

Equal expected frequencies are computed by dividing the number of subjects (N) by the number of categories (k) in the variable. A classic example of equal expected frequencies is testing the fairness of a die. If a die is fair, we would expect equal tallies of faces over a series of rolls.

**The Example of a Die:**

Let’s say I roll a real die 120 times (N) and count the number of times each face (k = 6) comes up. The number “1” comes up 17 times, the number “2” 21 times, “3” 22 times, “4” 19 times, “5” 16 times, and “6” 25 times. Results are listed under the “O” column below. We would Expect a count of 20 (E=N/k) for each of the six faces (1-6). This E value of 20 is listed under the “E” column below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | O | E | O- E | (O-E)2 | (O-E)2 ÷ E |
| 1 | 17 | 20 | -3 | 9 | 9/20 = 0.45 |
| 2 | 21 | 20 | 1 | 1 | 1/20 = 0.05 |
| 3 | 22 | 20 | 2 | 4 | 4/20 = 0.20 |
| 4 | 19 | 20 | -1 | 1 | 1/20 = 0.05 |
| 5 | 16 | 20 | -4 | 16 | 16/20 = 0.80 |
| 6 | 25 | 20 | 5 | 25 | 25/20 = 1.25 |
|  | 120 | 120 | 0 |  | χ2 = 2.80 |

The table above shows the step-by-step procedure in computing the chi-square formula. Notice that both **O** and **E** columns add to the same value (N=120).

**Testing the Chi Square Value:**

The computed value of χ***2*** is compared to the appropriate critical value. The critical value is found in the Chi-square Table. **Using** α **and df**, locate the critical value from the table. For the Goodness of Fit Test, the degrees of freedom (df) equal the number of categories (k) minus one (**df = k-1**). In our example above, the critical value (α=0.05, df =5) is **11.07**. Since the computed value (**2.80**) is less than the critical value (11.07), we declare the χ***2 not significant***.

What does this non-significant χ***2*** mean in English? The observed frequencies of the six categories of die rolls do not significantly differ from the expected frequencies. The observed frequencies have a “good fit” with what was expected. Or, simply stated, **“The die is fair.”** Had the computed value been greater than 11.07, the χ***2*** would have been declared significant. This would mean that the difference between observed and expected values is greater than we expect by chance. The observed frequencies would have a “bad fit” with what was expected. Or simply stated, “The die is loaded.”

Equal E is usually an unrealistic assumption of the break-down of categories. A better approach is to compute proportional expected frequencies (PE).

***Proportional Expected (PE) Frequencies:***

With proportional expected frequencies, the expected values are derived from a known population. Suppose you are in an Advanced Greek class of 100 students. You notice a large number of women in the class, and wonder if there are more women in the class than one might expect, given the student population. Using equal E’s, you would use the value (E=N/k) of 50. But you know that women make up only 15% of the student population. This gives you expected frequencies of 15 women (.15 x 100) and 85 men (.85 x 100). This latter design is far more accurate than the EE value of 50.

**The Example of Political Party Preference**

Suppose you want to study whether political party preference has changed since the last Presidential election. A poll of 1200 voters taken four years before showed the following breakdown: 500 Republicans, 400 Democrats, and 300 Independents. The ratio equals 5:4:3. In your present study, you poll 600 registered voters and find 322 Republicans, 184 Democrats, and 94 Independents. The null hypothesis for this study is that **party preference has not changed in four years.** That is, your hypothesis is that the present observed preferences are in a ratio of 5:4:3.

**Computing the Chi Square Value**

Compute the expected frequencies as follows. The ratio of 5:4:3 means there are 5+4+3=12 parts. Twelve parts divided into 600 voters yield *50 voters per part* (600/12=50).

The first category, Republicans, has 5 parts (**5**:4:3), or 5x50=***250* E**xpected voters. The second, Democrats, has 4 (5:**4**:3) parts, or 4x50=**200 E**xpected voters. The third, Independents, has 3 parts (5:4:**3**), or 3x50=**150 E**xpected voters. Putting this in a table as before, we have the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Party | O | E | (O-E) | (O-E)2 | (O-E)2 ÷ E |
| Republic | 322 | 250 | 72 | 5184 | 5184/250 = 20.74 |
| Democratic | 184 | 200 | -16 | 256 | 256/200 = 1.28 |
| Independent | 94 | 150 | -56 | 3136 | 3136/150 = 20.91 |
|  | 600 | 600 | 0 |  | **χ2 = (O-E)2/E= 42.93** |

Notice that both **O** and **E** columns add to 600 (N). Notice that the **O-E** column adds to zero. Notice that the E values are unequal, reflecting the 5:4:3 ratio derived from the earlier poll. The resulting χ***2*** value equals **42.93**.

**Testing the Chi Square**

The critical value (α=0.05, df = 2) is **5.991**. Since the computed value of **42.93** is greater than the critical value of 5.991, we declare the **chi-square value significant**. The observed values **do not fit** the expected values.

Since the recent poll does not fit the ratio of 5:4:3 found in the earlier poll, we can say that ***party preference has changed over the last four years****.*

**The Chi-Square Test for Independence**

The second chi-square test, the **chi-square test for independence**, can be used and interpreted in two different ways:

1. Testing hypotheses about the relationship between two variables in a population, or
2. Testing hypotheses about differences between proportions for two or more populations.

Although the two versions of the test for independence appear to be different, they are equivalent and they are interchangeable.

The first version of the test emphasizes the relationship between chi-square and a correlation, because both procedures examine the relationship between two variables.

The second version of the test emphasizes the relationship between chi-square and an independent-measures t- test (or ANOVA) because both tests use data from two (or more) samples to test hypotheses about the difference between two (or more) populations.

The first version of the chi-square test for independence views the data as one sample in which each individual is classified on two different variables.

The data are usually presented in a matrix with the categories for one variable defining the rows and the categories of the second variable defining the columns.

The data, called **observed frequencies**, simply show how many individuals from the sample are in each cell of the matrix.

The null hypothesis for this test states that there is no relationship between the two variables; that is, the two variables are independent.

The second version of the test for independence views the data as two (or more) separate samples representing the different populations being compared.

The same variable is measured for each sample by classifying individual subjects into categories of the variable.

The data are presented in a matrix with the different samples defining the rows and the categories of the variable defining the columns.

The data, again called **observed frequencies**, show how many individuals are in each cell of the matrix.

The null hypothesis for this test states that the proportions (the distribution across categories) are the same for all of the populations.

Both chi-square tests use the same statistic. The calculation of the chi-square statistic requires two steps:

1. The null hypothesis is used to construct an idealized sample distribution of **expected frequencies** that describes how the sample would look if the data were in perfect agreement with the null hypothesis.

For the goodness of fit test, the expected frequency for each category is obtained by expected frequency = fe = pn (p is the proportion from the null hypothesis and n is the size of the sample)

For the test for independence, the expected frequency for each cell in the matrix is obtained by

Expected frequency = fe =

Where;

Row total = sum of all frequencies in the row

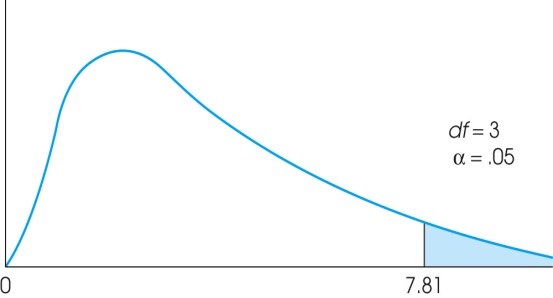
Column total = sum of all frequencies in the column

n = overall sample size

**Decision Rule:**

If χ2 > χ2U, reject H0, otherwise, do not reject H0

Where χ2U is from the chi-squared distribution with (r – 1)(c – 1) degrees of freedom



1. A chi-square statistic is computed to measure the amount of discrepancy between the ideal sample (expected frequencies from H0) and the actual sample data (the observed frequencies = fo).

A large discrepancy results in a large value for chi-square and indicates that the data do not fit the null hypothesis and the hypothesis should be rejected.

The calculation of chi-square is the same for all chi-square tests:

Chi-square = χ2 =

The fact that chi‑square tests do not require scores from an interval or ratio scale makes these tests a valuable alternative to the t- tests, ANOVA, or correlation, because they can be used with data measured on a nominal or an ordinal scale.

***Example:***

**The meal plan selected by 200 students is shown below:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class  Standing | **Number of meals per week** | | | **Total** |
| 20/week | 10/week | none |
| Fresh. | 24 | 32 | 14 | **70** |
| Soph. | 22 | 26 | 12 | **60** |
| Junior | 10 | 14 | 6 | **30** |
| Senior | 14 | 16 | 10 | **40** |
| **Total** | **70** | **88** | **42** | **200** |

**The hypothesis to be tested is:**

H0: Meal plan and class standing are independent (i.e., there is no relationship between them)

H1: Meal plan and class standing are dependent (i.e., there is a relationship between them)

**Example: Expected Cell Frequencies**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class  Standing | **Number of meals**  **per week** | | | **Total** |
| 20/wk | 10/wk | none |
| Fresh. | 24.5 | 30.8 | 14.7 | **70** |
| Soph. | 21.0 | 26.4 | 12.6 | **60** |
| Junior | 10.5 | 13.2 | 6.3 | **30** |
| Senior | 14.0 | 17.6 | 8.4 | **40** |
| **Total** | **70** | **88** | **42** | **200** |



**The test statistic value is:**



χ2U = 12.592 for α = 0.05 from the chi-squared distribution with (4 – 1)(3 – 1) = 6 degrees of freedom

**Decision and Interpretation:**



Decision Rule:

If χ2 > 12.592, reject H0, otherwise, do not reject H0

Here,

χ2 = 0.709 < χ2U = 12.592, so do not reject H0

Conclusion: there is not sufficient evidence that meal plan and class standing are related at α = 0.05

**Analysis of Variance (ANOVA)**

To test the significance of mean of one sample or significance of difference of means of two samples, t- test or χ2- test are very useful. But if there are more than two samples, the method of analysis of variance is used.

**Components of Total Variance:**

The total variation is split into two components- (a) Variance between samples and (b) Variance within samples, i.e.,

*Total Variance = Variance between samples + Variance within samples.*

**Assumptions:**

1. Normality;
2. Independence;
3. Additive Property.

**Uses or Applications or Importance of ANOVA:**

1. To test the significance of differences between means of more than two samples;
2. To test the significance of differences between variances;
3. Use in two- way classification;
4. To test the significance of correlation and regression.

**Chapter- 7**

**Linear Regression**

# **Introduction:**

After knowing the relationship between two variables we may be interested in estimating (predicting) the value of one variable given the value of another. The variable predicted on the basis of other variables is called the “dependent” or the ‘explained’ variable and the other the *independent* or the *predicting variable*. The prediction is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise, is called the regression equation or the explaining equation.

For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

The relationship between two variables can be considered between, say, rainfall and agricultural production, price of an input and the overall cost of product, consumer expenditure and disposable income. Thus, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

# **Definition:** Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

# **Types of Regression:** The regression analysis can be classified into:

1. Simple and Multiple
2. Linear and Non –Linear
3. Total and Partial

# **Simple and Multiple:**

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationships, more than two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones.

For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as y = f (x,z).

# **Linear and Non-linear:**

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predictive value, a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

# **Total and Partial:**

In the case of total relationships all the important variables are considered. Normally, they take the form of a multiple relationships because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

# **Linear Regression Equation:**

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of Y on X is Y = *a* + *b*X ……. (1) and

X on Y is

X = *a* + *b*Y……. (2) where *a, b* are constants.

From (1) We can estimate Y for known value of X.

(2) We can estimate X for known value of Y.

# **Regression Lines:**

For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables.

For perfect correlation, positive or negative i.e., r = +1, the two lines coincide i.e., we will find only one straight line. If r = 0, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axes.

**Y** r = - 1

r = + 1

Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X- axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y- axis will touch the mean value of Y.

# **Principle of ‘Least Squares’:**

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of “least squares”. This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

* + - 1. The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.,

(X – Xc) = 0 or (Y- Yc ) = 0

Where Xc and Yc are the values obtained by regression analysis.

* + - 1. The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e.,

(Y – Yc)2 < (Y – Ai)2

Where Ai = corresponding values of any other straight line.

* + - 1. The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is

*x*, *y* .

# **Methods of Regression Analysis:**

There are two methods of regression analysis-

1. Graphic Method through Scatter Diagram; and
2. Algebraic Method through regression equations (normal equation and through regression coefficient).

# **Graphic Method:**

# **Scatter** **Diagram:**

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

# **Algebraic Methods:**

* + - 1. *Regression Equation.*

The two regression equations for X on Y; X = *a + b*Y

And for Y on X; Y = *a* + bX

Where X, Y are variables, and a,b are constants whose values are to be determined

For the equation, X = *a + b*Y, The normal equations are

X = n*a + b* Y and

XY = *a*Y + *b*Y2

For the equation, Y= *a* + *b*X, the normal equations are

Y = n*a* + *b*X and

XY = *a*X + *b*X2

From these normal equations the values of *a* and *b* can be determined.

# **Example 1:**

***Find the two regression equations from the following data:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***X:*** | ***6*** | ***2*** | ***10*** | ***4*** | ***8*** |
| ***Y:*** | ***9*** | ***11*** | ***5*** | ***8*** | ***7*** |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | X2 | Y2 | XY |
| 6 | 9 | 36 | 81 | 54 |
| 2 | 11 | 4 | 121 | 22 |
| 10 | 5 | 100 | 25 | 50 |
| 4 | 8 | 16 | 64 | 32 |
| 8 | 7 | 64 | 49 | 56 |
| 30 | 40 | 220 | 340 | 214 |

Regression equation of Y on X is Y = *a* + *b*X and the normal equations are

Y = n*a* + *b*X

XY = *a*X + *b*X2

**Substituting the values, we get**

40 = 5*a* + 30*b* …… (1)

214 = 30*a* + 220*b* ……. (2)

**Multiplying (1) by 6**

240 = 30*a* + 180*b*……. (3)

(2) – (3) - 26 = 40*b*

b = = - 0.65

Now, substituting the value of ‘b’ in equation (1) 40 = 5*a* – 19.5

5*a* = 59.5

*a* = = 11.9

Hence, required regression line Y on X is Y = 11.9 – 0.65 X. Again, regression equation of X on Y is

X = *a* + *b*Y and

**The normal equations are**

X = n*a* + *b*Y and

XY = *a*Y + *b*Y2

Now, substituting the corresponding values from the above table, we get

30 = 5*a* + 40*b* …. (3)

214 = 40*a* + 340*b* …. (4)

**Multiplying (3) by 8, we get**

240 = 40*a* + 320 *b* …. (5)

(4) – (5) gives

-26 = 20*b*

b = = - 1.3

Substituting *b* = - 1.3 in equation (3) gives 30 = 5*a* – 52

5*a* = 82

*a = =* 16.4

**Hence, required regression line of X on Y is**

X = 16.4 – 1.3Y

* + - 1. *Regression Co-efficient:*

The regression equation of Y on X is y = *ȳ+r(x- x)*

Here, the regression Co-efficient of Y on X is

b1 = byx = *r*

The regression equation of X on Y is x = *x +r(x- x)*

Here, the regression Co-efficient of X on Y is

b2 = bxy = *r*

If the deviation are taken from respective means of x and y

b1 = byx =

b2 = bxy =

# **Properties of Regression Co-efficient:**

1. Both regression coefficients must have the same sign, ie either they will be positive or negative.
2. Correlation coefficient is the geometric mean of the regression coefficients i.e., r =
3. The correlation coefficient will have the same sign as that of the regression coefficients.
4. If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
5. Regression coefficients are independent of origin but not of scale.
6. Arithmetic mean of b1 and b2 is equal to or greater than the coefficient of correlation i.e.,
7. If r=0, the variables are uncorrelated, the lines of regression become perpendicular to each other.
8. If r= +1, the two lines of regression either coincide or parallel to each other.
9. Angle between the two regression lines is

Where m1 and m2 are the slopes of the regression lines X on Y and Y on X respectively

1. The angle between the regression lines indicates the degree of dependence between the variables.

# **Difference between Correlation and Regression:**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Correlation** | **Regression** |
| 1. | Correlation is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction. | Regression means going back and it is a mathematical measure showing the average relationship between two variables |
| 2. | Both the variables X and Y are random variables | Here X is a random variable and Y is a fixed variable. Sometimes both the variables may be random variables. |
| 3. | It finds out the degree of relationship between two variables and not the cause and effect of the variables. | It indicates the causes and effect relationship between the variables and establishes functional relationship. |
| 4. | It is used for testing and verifying the relation between two variables and gives limited information. | Besides verification it is used for the prediction of one value, in relationship to the other given value. |
| 5. | The coefficient of correlation is a relative measure. The range of relationship lies between –1 and +1 | Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable. |
| 6. | There may be spurious correlation between two variables. | In regression there is no such spurious regression. |
| 7. | It has limited application, because it is confined only to linear relationship between the variables. | It has wider application, as it studies linear and non- linear relationship between the variables. |
| 8. | It is not very useful for further mathematical treatment. | It is widely used for further mathematical treatment. |
| 9. | If the coefficient of correlation is positive, then the two variables are positively correlated and vice-versa. | The regression coefficient explains  that the decrease in one variable is associated with the increase in the other variable. |

**Example 2: *If 2 regression coefficients are b1= and b2 = What would be the value of r?***

**Solution:** The correlation coefficient, r = = = = = 0.6

# **Example 3: *Given b1= and b2 = Find the value of r.***

**Solution:** The correlation coefficient, r = = = = 1.06

It is not possible since *r*, cannot be greater than one. So the given values are wrong.

# **Example 4:**

***Compute the two regression equations from the following data.***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***X*** | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** |
| ***Y*** | ***2*** | ***3*** | ***5*** | ***4*** | ***6*** |

***If x =2.5, what will be the value of y?***

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y | *x* *X* *X̅* | *y* *Y* *Y* | x2 | y2 | xy |
| 1 | 2 | -2 | -2 | 4 | 4 | 4 |
| 2 | 3 | -1 | -1 | 1 | 1 | -1 |
| 3 | 5 | 0 | 1 | 0 | 1 | 0 |
| 4 | 4 | 1 | 0 | 1 | 0 | 0 |
| 5 | 6 | 2 | 2 | 4 | 4 | 4 |
| 15 | 20 | 20 |  | 10 | 10 | 9 |

Mean of X (X̅) = = = 3

Mean of Y (Y̅) = = = 4

Regression Co efficient of Y on X

byx = = = 0.9

Hence regression equation of Y on X is

*Y* *Y̅* *byx* ( *X* *X̅* )

= 4 + 0.9 ( *X* – 3 )

= 4 + 0.9X – 2.7

=1.3 + 0.9X

when X = 2.5

Y = 1.3 + 0.9 × 2.5

= 3.55

Regression co efficient of X on Y

bxy = = = 0.9

So, regression equation of X on Y is

*X* *X̅* *bxy* (*Y* *Y̅* )

= 3 + 0.9 ( Y – 4 )

= 3 + 0.9Y – 3.6

= 0.9Y - 0.6

# **Example 5:**

***Obtain the equations of the two lines of regression for the data given below:***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***X*** | ***45*** | ***42*** | ***44*** | ***43*** | ***41*** | ***45*** | ***43*** | ***40*** |
| ***Y*** | ***40*** | ***38*** | ***36*** | ***35*** | ***38*** | ***39*** | ***37*** | ***41*** |

# **Example 6: *In a correlation study, the following values are obtained***

|  |  |  |
| --- | --- | --- |
|  | ***X*** | ***Y*** |
| ***Mean*** | ***65*** | ***67*** |
| ***S.D*** | ***2.5*** | ***3.5*** |

***Co-efficient of correlation = 0.8***

***Find the two regression equations that are associated with the above values.***

**Solution:**