Industrial Statistics

D. K. Jain & R. Malhotra

Industrial Statistics

Course Developers

D. K. Jain & R. Malhotra

Dairy Economics, Statistics and Management Division

NDRI, Karnal

Index

Lesson 1

DEFINITION AND SCOPE OF STATISTICS

1.1 Introduction

Statistics is a field of mathematics that pertains to data analysis. For the last few centuries, statistics has remained a part of mathematics as the original work was done by mathematicians like Pascal, James Bernoulli, De-Moivre, Laplace, Gauss and others. Till early nineteenth century, statistics was mainly concerned with official statistics needed for the collection of information on revenue, population etc. of a state or kingdom. The science of statistics developed gradually and its field of application widened day by day. In fact, the term statistics is generally used to mean numerical facts and figures.

1.2 Meaning of the Word Statistics

The word 'statistics' seems to have been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'Statistik' each of which means a 'political state'. In ancient times the governments used to collect the information regarding the 'population' and 'property of wealth' of the country- the former enabling the government to have an idea of the manpower of the country (to safeguard itself against external aggression, if any) and the latter providing it a basis for introducing new taxes and levies.

Seventeenth Century saw the origin of 'vital statistics'. Captain John Graunt of London known as the father of vital statistics was the first man to study the statistics of birth and death. Computation of mortality table and the calculation of expectation of life at different ages led to the idea of life insurance and the first life insurance institution was founded in London in 1698.

The theoretical development of the so called modern statistics came during the mid seventeenth century with the introduction of Theory of Probability and Theory of Games and chance. The chief contributors being Pascal, De-Moivre, James Bernoulli, Laplace, Gauss, Sir Francis Galton, Karl Pearson, W.S. Gosset, Helmert, Sir R.A. Fisher.

1.3 Indian History of Statistics

In India, an efficient system of collecting official and administrative statistics existed even more than 2000 years ago, in particular during the reign of Chandragupta Maurya (324-300B.C). From Kautilya's Arthashastra it is known that even before 300 B.C a very good system of collecting vital statistics and registration of births and deaths was in vogue. During Akbar's reign Raja Todarmal, the then land and revenue minister maintained good records of land and Agricultural Statistics. In Aina-e-Akbari written by Abul Fazal we find detailed accounts of administrative and statistical surveys conducted during Akbar's reign.

1.4 Definition of Statistics

Statistics has been defined differently by different authors from time to time. In ancient times statistics was confined only to the affairs of the state but now it embraces almost every sphere of human activity.

Webster defines statistics as "classified facts representing the conditions of the people in a state-especially those facts which can be stated in numbers or in any other tabular or classified arrangement". This definition confines

statistics only to the data pertaining to the state is inadequate as the domain of statistics is much wider.

Bowley defines statistics as "Numerical statements of the facts in any department of enquiry placed in relation to each other". He himself defines statistics in three different ways:

(i) Statistics may be called as the science of counting.

(ii) Statistics may rightly be called as the science of averages.

(iii) Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations.

The above definitions are inadequate. The first because statistics is not merely confined to the collection of data as other aspects like presentation, analysis and interpretation etc. are also covered by it. The second because averages are only a part of the statistical tools used in the analysis of data. These are not only the tools but others being Dispersion, Skewness, Kurtosis, Correlation & Regression etc. The third because it restricts the application of statistics to sociology while today the statistics has found its application in almost every field of science.

Perhaps the best definition seems to be one given by Croxton and Cowden, according to whom statistics may be defined as 'the science which deals with the collection, analysis and interpretation of numerical data'.

Statistics, therefore is defined as the science of collection, compilation, tabulation, analysis and interpretation of quantitative data. It is essentially a branch of applied mathematics i.e. mathematics applied to the observational data. Statistics essentially mean the procedure by which we understand data.

Statistics used as a singular word i.e. 'statistics' means as particular kind of estimate compiled from set of observations, usually according to some algebraic formulae. 'Statistics' used as singular is a name given for the body of scientific methods (the statistical methods) which are meant for the collection, compilation, analysis and interpretation of numerical data.

'Statistics' used as a plural noun means numerical data which result from a host of uncontrolled, and mostly unknown causes acting together. It is in this sense that the term is used when our daily newspaper give vital statistics, crime statistics, sports statistics, agriculture and dairy statistics, food production statistics etc.

Statistics has two broad functions:

- Descriptive Statistics - The first function is to describe and summarize the information in such a way so as to make it more usable
- Inductive Statistics - The second function is to draw inferences about the population from the information contained in the sample which is only a part of the population and thus we pass from the particular to the general. Here the induction has to be achieved within a probabilistic frame work.

1.5 Application of Statistics

Early applications of statistics were mainly concerned with reduction of large amounts of observed data to the point where general trends become apparent. At the same time, emphasis in many sciences turned from the study of individuals to the study of the behavior of aggregates of individuals. Statistical methods were found suitable for such studies, aggregate data fitting consistently with the concept of a population.

Statistical Science has wide applications in Dairy production, processing and management. In dairy production, the productive and reproductive performance of various breeds/species of animals is carried out through various statistical measures. For example, age at first calving, body weight, lactation length, dry period, inter-calving period etc. are closely monitored for best production performance of female animals. In the field of Animal Nutrition, many experiments have been devised to discover the significance of various vitamins, proteins, diets in the different phases of animal production. Similarly, production parameters like daily/monthly lactation yield, fat, SNF, protein and other minerals as well as microbiological parameters in milk are closely monitored for getting best quality and safe milk for human consumption. In industry, Statistics is very widely used in 'Quality Control'. In production engineering, to find whether the product is conforming to specifications or not, Statistical techniques viz., control charts and Acceptance Sampling, plans etc. are of extreme importance which will be discussed in modules of Statistical Quality Control. In dairy processing, various value added dairy products are developed for which proportions of ingredients are required so as to get least cost product mix which fulfills certain minimum requirements. The chemical, microbiological and sensory attributes of such developed dairy products are also monitored over different periods of storage time. Various statistical techniques are employed in order to fulfill such requirements. Statistics is also playing an important role in Engineering. For example, such topics as the study of heat transfer through insulating materials per unit of time, performance guarantee testing programs, production control, inventory control, standardization of fits and tolerances of machine parts, job analyses of technical personnel, time and motion studies and many other specialized problems in research and development make great use of probabilistic and statistical methods. Agricultural engineering, which combines the practices of engineering and agriculture has also benefited greatly from the use of statistical methods. In dairy management cost of calf rearing maintenance of animals is required to be worked out. Similarly cost of milk production for various categories of animals is also required to be computed across different seasons/regions etc. taking into consideration various fixed and variable costs that enter into cost. There is also requirement for computing cost of processing of milk into various dairy products. Similarly there is also requirement to monitor milk production, utilisation and marketed surplus across various categories of producers and consumers and also assess the demand and supply of milk and milk products. All these aspects require employment of various statistical techniques to achieve the desired objectives.

1.6 Distrust of Statistics

We often hear the following interesting comments on Statistics:

(i) 'An ounce of truth will produce tons of Statistics',
(ii) 'Statistics can prove anything',
(iii) 'Figures do not lie. Liars figure',
(iv) 'If figures say so it can't be otherwise',
(v) 'There are three types of lies – lies, damn lies and Statistics – written in the order of their naming and so on.'

Some of the reason for the existence of such divergent views regarding the nature and function of Statistics are as follows:

- Figures are innocent, easily believable and more convincing. The facts supported by figures are psychologically more appealing.

- Figures put forward for arguments may be inaccurate or incomplete and thus might lead to wrong inferences.
- Figures, though accurate, might be moulded and manipulated by selfish persons to conceal the truth and present a distorted picture of facts to the public to meet their selfish motives. When the skilled talkers, writers or politicians through their forceful writings and speeches or the business and commercial enterprises through advertisements in the press mislead the public or belie their expectations by quoting wrong statistical statements or manipulating statistical data for personal motives, the public loses its faith and belief in the science of Statistics and starts condemning it. We cannot blame the layman for his distrust of Statistics, as he, unlike statistician, is not in a position to distinguish between valid and invalid conclusions from statistical analysis.

It may be pointed out that Statistics neither proves anything nor disproves anything. It is only a tool which when rightly used may prove extremely useful and if misused, might be disastrous. According to Bowley, "Statistics only furnishes a tool, necessary though imperfect, which is dangerous in the hands of those who do not know its use and its deficiencies". It is not the subject of Statistics that is to be blamed but those people who twist the numerical data and misuse them either due to ignorance or deliberately for personal selfish motives. As king points out, "Science of Statistics is the more useful servant but only of great value to those who understand its proper use."

1.7 Limitations of Statistics:

- It does not deal with individual measurements.
- It deals with quantitative characteristics.
- Statistical results are true only on an average
- It is only one of the methods of studying a problem

1.8 Statistical Agencies

The responsibility of collection, processing and tabulation and their dissemination lies with statistical agencies. Following are the major agencies at national level:

1. Central Statistical Organisation (Department of Statistics, Ministry of Planning and Programme Implementation), New Delhi.
2. National Sample Survey Organisation (Department of Statistics, Ministry of Planning and Programme Implementation), New Delhi.
3. Registrar General of India (Ministry of Home Affairs), New Delhi.
4. Directorate General of Commercial Intelligence and Statistics. (Ministry of Commerce), Calcutta.
5. Directorate of Economics and Statistics (Department of Agriculture and Cooperation, Ministry of Agriculture), New Delhi.
6. Labour Bureau (Ministry of Labour), Shimla and Chandigarh.
7. Department of Economic Analysis and Policy, Reserve Bank of India, Mumbai.
8. Office of the Economic Advisor, Department of Industrial Development, New Delhi.
9. Directorate General of Employment and Training, (Ministry of Labour), New Delhi.

10. Ministry of Food Processing Industries, New Delhi
11. Agricultural and Processed Food Products, Export Development Authority (APEDA), New Delhi
12. National Dairy Development Board (NDDB), Anand

Apart from these, each Government of India ministry has either a full-fledged statistical division or section. Public sector organizations have their own arrangements for collection and maintenance of statistics. In states and Union Territories (UTs), there are State Statistical Bureaus. On the whole, statistical system in India is a decentralized one; the responsibility of collection and dissemination of statistics is divided between the union and state governments. Statistics is collected by other bodies are All India Statistical Operations such as Census of India, Annual Survey of Industries (ASI), National Sample Survey etc.

Official statistical websites are:

- http://www.nic.in/stat
- http://www.mospi.nic.in
- http://www.nddb.org
- http://www.apeda.gov.in
- http://www.mofpi.nic.in
- http://www.censusindia.net
- rgindia@hub.nic.in
- http://www.rbi.org.in

At last it would be worthwhile quoting our Hon'ble Director General, Indian Council of Agricultural Research and President, Indian Society of Agricultural Statistics, Dr. S. Ayappan that "Statistic is like a salt in food, no one recognizes its importance when it is there, everyone feels its importance when it is not there". This clearly emphasizes the importance of Statistics in all branches of Agriculture and Dairy Sciences.

Lesson 2

CLASSIFICATION OF DATA AND FREQUENCY DISTRIBUTION

2.1 Introduction

As discussed in Lesson 1, statistics are a set of numerical statements and facts collected from any field of enquiry for drawing valid inferences. Data collection is in fact, the most important aspect of a research experiment/statistical survey. After data have been collected, the next step is to present the data in some orderly and logical form so that their essential features may become explicit. The need for proper presentation of data arises because the mass of collected data in their raw form is often so voluminous which cannot be easily comprehended and analysed. Therefore, after the collection of data, it is imperative that data are classified and presented in such a way so as to bring out points of similarities and dissimilarities in the data.

2.2 Collection of Data

To study any problem by means of statistical methods first, the relevant data are collected. Sometimes the data is to be collected from some research experiment or the primary sampling units (households). Sometimes, the relevant data may exist in a published or unpublished form, being collected by a private body or by the Government agency or by some research organization, for its own use or for supplying popular information. In making use of such data (called secondary data), one has to be particularly careful about the definitions of terms and concepts used by the collecting authority and also about the method of collection and the reliability of the data. More often, one has to collect data directly from the field of enquiry. The data are then said to be of the primary type. The collection of primary data may be done by interviewing a number of persons and filling in questionnaires relevant to the problem e.g., in the family income and expenditure survey, one will generally interview the head of each family. The data collected should be carefully scrutinized before they are subjected to statistical treatment.

2.3 Classification of Data

Classification is the process of arranging the data into different groups or classes according to some common characteristics. According to Connor “Classification may be defined as the process of arranging things in groups or classes according to their resemblances and affinities”. The functions of classification may be summarized as follows:

- It condenses the data.
- It facilitates comparisons.
- It helps to study the relationships.
- It facilities the statistical treatment of data.

The classification of data is generally done on geographical, chronological, qualitative or quantitative basis on the following lines:

a) In geographical classification, data are arranged according to places, areas or regions.

b) In chronological classification, data are arranged according to time i.e. weekly, monthly, quarterly, half-yearly, annually, etc.

c) In qualitative classification the data are arranged according to attributes like sex, marital status, educational standard, region, farm, breed, disease etc.

d) Quantitative classification means arranging data according to certain characteristic that has been measured e.g. according to height, weight or milk yield, fat contents in a dairy product etc. In this type of classification, certain classes are formed and the units belonging to these classes are attached to them. The quantitative phenomenon under study is known as variable and hence this classification is also sometimes called classification by variables.

Variable: The quantitative phenomenon under study, wages, barometer readings, rainfall records, heights, weights, milk yield, fat, SNF, age at first calving, first lactation production etc. is termed a variable or a variate. In other words a quantity which can vary from one individual to another individual is called a variable. Variables are of two kinds

a) **Continuous variable:** Quantities which can take any numerical value within a certain range are called continuous variables e.g., the height of a calf at various ages is a continuous variable since as the calf grows from 200 cm to 300 cm his height assumes all possible values within this limit.

b) **Discrete variable :** Quantities which are incapable of taking all possible values are called discontinuous or discrete variables e.g. the number of animals in a herd can take only integer values such as 2, 3, 4 etc.

2.4 Frequency Distribution

The frequency distribution is a statistical table which shows the value of a variable in order of magnitude, either individually or in groups, along with the corresponding frequencies side by side. The data pertaining to a quantitative phenomenon can be classified in four ways:

- The set or series of individual observations- ungrouped (raw) or arranged (arrayed) data.
- Discrete or ungrouped frequency distribution.
- Grouped frequency distribution.
- Continuous frequency distribution.

Example 1: The following data pertain to first lactation milk yield (in kg) of 100 Karan Swiss cows

1630	1648	1663	1665	1671	1677	1680	1687	1690	1695
1787	1788	1790	1800	1862	1855	1815	1835	1845	1818
1974	1998	2000	2000	2005	2031	2045	2045	2050	2056
2168	2171	2180	2187	2200	2218	2245	2323	2372	2397
2063	2069	2085	2098	2100	2100	2100	2105	2117	2131
1736	1743	1760	1765	1763	1767	1775	1775	1776	1780
1695	1754	1698	1700	1742	1732	1711	1713	1718	1728
1854	1850	1855	1856	1857	1860	1863	1863	1875	1880
1890	1900	1910	1912	1915	1918	1928	1916	1915	1947
1950	1958	1951	1960	1963	1968	1965	1967	1970	1969

The data given in example 1 are called the raw or ungrouped data which does not give us any useful information. Our objective will be to express the huge data in a suitable condensed form which will highlight the significant

facts and comparisons and furnish more useful information without sacrificing any information of interest about the important characteristics of the distribution.

2.4.1 Array

A better presentation of above raw data would be to arrange them in an ascending or descending order of magnitude which is called arraying of data. However, this method is better than raw data but does not reduce the volume of the data.

2.4.2 Discrete or ungrouped frequency distribution

A much better way of the presentation of the data is to express in the form of a discrete or ungrouped frequency distribution, where we count the number of times each value of the variable occurs in the data. The number of times a variate value is repeated is called frequency of the variate value e.g. suppose there are seven Karan Fries cows having first lactation milk yield equal to 1900 kg, 7 is the frequency of first lactation yield of 1900 kg.

2.4.3 Grouped frequency distribution

It is a statistical table which shows the values of the variable in groups and also the corresponding frequencies side by side. In this type of set up, the condensation of data consists in classifying the data into different classes (or class intervals) by dividing the entire range of the values of the variable into a suitable number of groups, called classes and then recording the number of observations in each group. The type of such representation of data is called a grouped frequency distribution. The groups are called the classes and the boundary ends are called class limits e.g. for a class interval 0 – 10, 0 is the lower limit and 10 is the upper limit. The difference between upper and lower limit is called magnitude of the class. The number of observations falling within a particular or defined class is called its frequency or class frequency. The variate value which lies midway between the upper and lower limits is called mid value or midpoint of that class.

While preparing the frequency distribution the following points must be kept in mind:

1. The class interval should be uniform i.e. it should be of equal width. A comparison of different frequency distributions is facilitated if the same class interval is used for all. The class interval should be an integer as far as possible.
2. The class interval should be so chosen that all the observations should be reflected by the frequency distribution.
3. The class interval should be continuous open end classes less than ‘a’ or greater than ‘b’ should be avoided. These classes create difficulty in analysis and interpretation.
4. The observations corresponding to the common point between two classes should always be put in the higher class e.g. a number corresponding to the variate 30 is to be put in the class 30-40 and not in 20-30.
5. There should not be too many or too small number of classes. The number of classes should never be less than 6 and not more than 30 i.e. the number of class intervals should lie between 6 and 30. With less number of classes the accuracy may be lost, and with more number of classes the computations become tedious. The optimum number of classes is generally considered as 15.

2.4.3.1 Number of classes

The following formula due to Sturges may be used to determine the number of classes k = 1+3.322log$_{10}$N where k is the number of classes and N is the total frequency.

2.4.3.2 Size of class intervals

The choice of class interval depends on the number of classes for a given distribution and size of the data. As far as possible the class intervals should be of equal size. Prof. Sturges has given the following formula for determining the size of class intervals

$$\text{Size of class interval (i)} = \frac{\text{Largest value} - \text{Smallest value}}{1+3.322\log_{10}N} = \frac{\text{Range}}{k}$$

Example 2: If we consider the data given in example 1 let us find its size of class interval and prepare its frequency distribution

Solution: The size of class interval is given by

N=100 Largest value =2397 and Smallest value =1630, Range =767

Number of classes k = 1+3.322 log$_{10}$(100)=7.644

Hence, size of class interval = $\frac{767}{7.644} = 100.3401 \sim 100$

Taking class intervals as 1630-1730, 1730-1830, ----, 2330-2430 the frequency distribution of first lactation milk yield of 100 Karan Swiss cows is given below in Table 2.1

Table 2.1 Frequency distribution of First Lactation milk yield of Karan Swiss cows

Class Interval (in kg)	frequency (fi)
1630-1730	17
1730-1830	19
1830-1930	23
1930-2030	16
2030-2130	14
2130-2230	7
2230-2330	2
2330-2430	2

Advantages of grouping

(i) First advantage of grouping is that in subsequent calculations, much labour is saved in numerical computation by treating all individuals in a class interval as having the value at the centre of that interval.

(ii) The second advantage of grouping is where the observed sample is of moderate size and from a large population. In such a case the frequency table is more likely to exhibit a rise or fall of frequency against class interval.

2.5 Cumulative Frequency Distribution

The cumulative frequency of a class is the total frequency up to and including that class. The table of cumulative frequencies is called a cumulative frequency distribution table. There are two types of cumulative frequency

distribution. The cumulative frequency distribution of all values greater than or equal to the lower limit of each class is called more than cumulative frequency distribution. The cumulative frequency of all values less than or equal to the upper limit of each class is called less than cumulative frequency distribution. Let us illustrate this through example 3

Example 3: Prepare the cumulative frequency distribution of the frequency distribution of first lactation milk yield of Karan Swiss cows given in table 2.1.

Solution: The less than cumulative frequency and more than cumulative frequency distribution are shown in table 2.2

Table 2.2 Cumulative frequency distribution of first lactation milk yield

Class Interval (in kg)	frequency (f_i)	Cumulative frequency (c.f.)	
		Less than	More than
1630-1730	17	17	100
1730-1830	19	36	83
1830-1930	23	59	64
1930-2030	16	75	41
2030-2130	14	89	25
2130-2230	7	96	11
2230-2330	2	98	4
2330-2430	2	100	2

Lesson 3

MEASURES OF CENTRAL TENDENCY

3.1 Introduction

The collected data as such are not suitable to draw conclusions about the mass from where it has been taken. Some inferences about the population can be drawn from the frequency distribution of the observed values. One of the important objectives of statistical analysis is to determine various numerical measures which describe the inherent characteristics of a frequency distribution. The averages are the measures which condense a huge unwieldy set of numerical data into single numerical value, are representative of the entire distribution. Hence, in finding a central value, the data are condensed into a single value around which most of the values tend to cluster. Commonly, such a value lies in the centre of the distribution and is termed as central tendency.

3.2 Measure of Central Tendency or Average

One of the most important objectives of the statistical analysis is to get one single value that describes the characteristic of the entire mass of unwieldy data. Such a value is called the 'Central Value' or 'an average'. The word 'Average' is very commonly used in everyday conversation. For example we talk of average milk yield of a cow, average fat content of milk , average height or life of an Indian, average income etc. When we say 'he is an average student' what it means is that he is neither very good nor very bad, just a mediocre type of student. Similarly, when we talk of average size of butter or cheese packet being sold through a retail outlet what we mean is that the size of packet which is being sold to maximum number of individuals by the retail outlet that means it is the modal size. However, in statistics the term average has a different meaning. According to Croxton and Cowden "An average value is a single value within the range of the data and is used to represent all of the values in the series. Since an average is within the range of the data, it is sometimes called a measure of central value." It may be defined as that value of a distribution which is considered as the most representative of the series or typical value for a group. Such a value is of great significance because

- it depicts the characteristic of the whole group
- it facilitates comparison

Averages are sometime referred to as a measure of location since they enable us to locate the position or place of the distribution in question.

Requisites of a good average

- It should be rigidly defined.
- It should be easy to understand and calculate.
- It should be based on all the observations.
- It should not be unduly affected by extreme observations.
- It should be suitable for further mathematical treatment.
- It should be least affected by fluctuations of sampling.

The various measures of central tendency or averages are discussed below:

3.3 Arithmetic Mean

Its value is obtained by adding together all the items and dividing it by the total number of observations. If X_1, X_2, ... , X_n are n values of a variable X, then the arithmetic mean (A.M.) in case of raw data, is defined as

$$\overline{X} = \frac{X_1 + X_2 + - - - + X_n}{n} = \frac{\sum_i^n X_i}{n}$$

In case of frequency distribution

X_i	X_1	X_2	X_3	---	X_n
f_i	f_1	f_2	f_3	---	f_n

If the value X_1 occurs f_1 times, the value X_2 occurs f_2 times, …, the value X_n occurs f_n times, then the arithmetic mean is given by

$$\overline{X} = \frac{f_1X_1 + f_2X_2 + - - - + f_nX_n}{f_1 + f_2 + - - - + f_n} = \frac{\sum_i^n f_iX_i}{\sum_i^n f_i} = \frac{\sum_i^n f_iX_i}{N}$$

where, $N = \sum_i^n f_i$

3.3.1 Arithmetic mean in case of grouped data

In case of grouped data, arithmetic mean can be calculated by applying any of the following methods:

i) Direct Method ii) Short cut Method iii) Step-deviation Method

3.3.1.1 Direct method

- Multiply each value of X_i (the mid value of the class) by the corresponding frequency f_i.
- Obtain the sum of the products $\sum_i^n f_iX_i$.
- Divide this sum of products by the total frequency (N) so as to get mean.

Example 1: Find the mean from frequency distribution given in example 2 of Lesson 2.

Solution: Prepare the following table and calculate arithmetic mean as follows:

Class Interval	Mid-value (X_i)	frequency (f_i)	$f_i X_i$
1630-1730	1680	17	28560
1730-1830	1780	19	33820
1830-1930	1880	23	43240

1930-2030	1980	16	31680
2030-2130	2080	14	29120
2130-2230	2180	7	15260
2230-2330	2280	2	4560
2330-2430	2380	2	4760
Total		100	191000

$$\bar{X} = \frac{\sum_i^n f_i X_i}{N} = \frac{191000}{100} = 1910\text{Kg}$$

3.3.1.2 Short–cut method (Change of Origin)

If the values of X or/and f are large, the calculation of mean by direct method is quite tedious and time consuming. In such a case the calculations can be reduced to a great extent by using short cut method. This method consists in taking deviations of the given observations from any arbitrary value A. The formula for calculation of the arithmetic mean is

$$\bar{X} = A + \frac{\sum_{i=1}^n f_i d'_i}{N}$$

where A is arbitrary mean , $d_i' = X_i - A$ i.e. deviation from the arbitrary or assumed mean.

3.3.1.3 Step-deviation method (Change of origin and scale)

In case of grouped frequency distribution, with class intervals of equal magnitude, the calculations are further simplified by taking; $d_i = \frac{X_i - A}{h}$ where X_i is the mid value of the class and h is the common magnitude or width of the class intervals. So the formula for calculating mean is $\bar{X} = A + \left(\frac{\sum_{i=1}^n f_i d_i}{N}\right) \times h$. The procedure is illustrated in the example 2. It will be seen that the answer in each of the three cases is the same. The step-deviation method is the most convenient method on account of simplified calculations.

Example 2: Solve example 1 with short-cut and step-deviation method.

Solution: Prepare the following table and calculate arithmetic mean as follows:

Class Interval	Mid-value (X_i)	frequency (f_i)	$d_i=X_i-A$ A=2080	$f_i d_i$	$d_i = \frac{X_i - A}{h}$	$f_i d_i$
1630-1730	1680	17	-400	-6800	-4	-68
1730-1830	1780	19	-300	-5700	-3	-57
1830-1930	1880	23	-200	-4600	-2	-46
1930-2030	1980	16	-100	-1600	-1	-16
2030-2130	2080	14	0	0	0	0

2130-2230	2180	7	100	700	1	7
2230-2330	2280	2	200	400	2	4
2330-2430	2380	2	300	600	3	6
Total		100		-17000		-170

Short-cut method: A=2080 $\bar{X} = A + \frac{\sum_{i=1}^{n} f_i d'_i}{N} = 2080 + \frac{-17000}{100} = 2080 - 170 = 1910Kg$

Step-Deviation Method: A=2080, h=100

$\bar{X} = A + \frac{\sum_{i=1}^{n} f_i d'_i}{N} \times h = 2080 + \frac{-170}{100} \times 100 = 2080 - 170 = 1910Kg$

3.3.2 Mathematical properties of arithmetic mean

i) The algebraic sum of the deviations of the given set of observations from their arithmetic mean is zero i.e. $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

ii) If n_1 and n_2 are the sizes and $\bar{X}_1, \bar{X}_2$ are the respective means of two series then the pooled mean $\bar{\bar{X}}$ of the combined series of size (n_1+n_2) observations is given by:

$$\bar{\bar{X}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

iii) The sum of the squares of deviations of the given set of observations, when taken from their arithmetic mean, is minimum.

3.3.3 Merits of arithmetic mean

- The A.M. is rigidly defined
- It is based on all the observations
- It is easily calculated from the given data
- It is least affected by fluctuations of sampling
- It is suitable for further mathematical treatment. The average of two or more series can be obtained from the averages of the individual series.

3.3.4 Demerits of arithmetic mean

- The strongest drawback of arithmetic mean is that it is very much affected by extreme observations.
- In a distribution with open end classes the value of mean cannot be computed without making assumptions regarding the size of class.
- It can neither be located by inspection nor graphically.
- It cannot be used for qualitative type of data such as intelligence, honesty, flavor, overall acceptability of dairy product etc.

- Arithmetic mean cannot be obtained if a single observation is missing or lost.
- In a skewed distribution, usually arithmetic mean is not representative of the distribution.

3.4 Geometric Mean

The geometric mean (usually denoted by G.M.) of a set of n observations is the n^{th} root of their product. If X_1, X_2, . . . , X_n are n values of a variable X, none of them being zero, then the geometric mean, G.M. is defined by

$$G.M. = \sqrt[n]{X_1 . X_2 . X_3 \dots \dots \dots X_n} = (X_1 . X_2 . X_3 . \dots \dots \dots X_n)^{1/n}$$

$$\log G.M. = \left[\frac{1}{n}\sum_{i=1}^{n} \log X_i\right]$$

$$G.M. = \text{Anti log}\left[\frac{1}{n}\sum_{i=1}^{n} \log X_i\right]$$

Thus logarithm of G.M. of a set of observations is the arithmetic mean of their logarithms.
If X_1, X_2, . . . , X_n occurs f_1, f_2 . . . , f_n times respectively then

$$G.M. = \sqrt[N]{X_1^{f_1} . X_2^{f_2} . X_3^{f_3} \dots \dots \dots X_n^{f_n}} = \left(X_1^{f_1} . X_2^{f_2} . X_3^{f_3} \dots \dots \dots X_n^{f_n}\right)^{1/N}$$

$$G.M. = \text{Anti log}\left[\frac{1}{N}\sum_{i=1}^{n} \log X_i\right] \text{ where } N = \sum_{i=1}^{n} f_i$$

3.4.1 Merits of geometric mean

- It is rigidly defined
- The G.M. is based on all observations of a series.
- It is not much affected by fluctuations of sampling.
- It is suitable for further mathematical treatment.
- Unlike arithmetic mean which has a bias for higher values, geometric mean has bias for smaller observations.
- As compared with Arithmetic mean, Geometric mean is affected to a lesser extent by extreme observations.

3.4.2 Demerits of geometric mean

- Computations are difficult
- It is not simple to understand

- It does not give equal weight to every item.
- It cannot be calculated if the number of negative values is odd as well as some value is zero.

3.4.3 Use of geometric mean

It is most appropriate average when dealing with ratios, percentages and rate of increase between two periods. It is applied when increase or decrease in time is proportional e.g. growth of population is proportional to the time, increase in bacterial population is proportional to the time and rate of interest. Geometric Mean is used in the construction of Index numbers.

3.5 Harmonic Mean

If $X_1, X_2, \ldots, X_n$ are n values of a variable X, then their Harmonic Mean, abbreviated as H.M. is defined by

$$H.M. = \frac{1}{\frac{1}{n}\left[\frac{1}{X_1}+\frac{1}{X_2}+---+\frac{1}{X_n}\right]} = \frac{1}{\frac{1}{n}\left[\sum_{i=1}^{n}\frac{1}{X_i}\right]}$$

In other words Harmonic Mean is the reciprocal of the arithmetic mean of the reciprocals of the given observations. In case of grouped frequency distribution, harmonic mean is given by

$$H.M. = \frac{1}{\frac{1}{N}\left[\frac{f_1}{X_1}+\frac{f_2}{X_2}+---+\frac{f_n}{X_n}\right]} = \frac{1}{\frac{1}{N}\left[\sum_{i=1}^{n}\frac{f_i}{X_i}\right]} = \frac{N}{\left[\sum_{i=1}^{n}\frac{f_i}{X_i}\right]}, \text{where } N = \sum_{i=1}^{n} f_i$$

3.5.1 Merits of harmonic mean

- It is rigidly defined
- The H.M. is based on all observations of a series.
- It is not much affected by fluctuations of sampling.
- It is suitable for further mathematical treatment.
- Since the reciprocals of the values of the variable are involved, it gives greater weightage to smaller observations and as such is not very much affected by one or two big observations.

3.5.2 Demerits of harmonic mean

- Computations are difficult and not simple to understand.
- It cannot be calculated if any one of the observations is zero.
- It is not a representative figure of the distribution unless the phenomenon requires greater weightage to be given to smaller values.

3.5.3 Use of harmonic mean

H.M. is used in finding averages involving speed, time, price and ratios. It is useful for computing the average rate of increase of profits of a concern or average speed at which a journey has been performed or the average price at which an article has been sold. The rate usually indicates the relation between two different types of measuring units that can be expressed reciprocally. The H.M. is used for the problems about work, time and rate, where the amount of work is held constant and the average rate is required, or in problems about total cost, number of persons and per capita cost is called for or in problems of similar nature involving rates.

The arithmetic mean (A.M.), the geometric mean (G.M.) and the harmonic mean (H.M.) of a series of n observations are connected by the relation $A.M. \geq G.M. \geq H.M.$

The computation of G.M. and H.M. is illustrated in example 3.

Example 3: Find G.M. and H.M. of data given in example 1.

Solution: Prepare the following table and calculate G.M. and H.M. as follows:

Class Interval	Mid-value (X_i)	frequency (f_i)	$\log X_i$	$f_i \log X_i$	$\frac{1}{X_i}$	$\frac{f_i}{X_i}$
1630-1730	1680	17	3.2253	54.8303	0.0006	0.0101
1730-1830	1780	19	3.2504	61.7580	0.0006	0.0107
1830-1930	1880	23	3.2742	75.3056	0.0005	0.0122
1930-2030	1980	16	3.2967	52.7466	0.0005	0.0081
2030-2130	2080	14	3.3181	46.4529	0.0005	0.0067
2130-2230	2180	7	3.3385	23.3692	0.0005	0.0032
2230-2330	2280	2	3.3579	6.71587	0.0004	0.0009
2330-2430	2380	2	3.3766	6.75315	0.0004	0.0008
Total		100		327.9316		0.0528

$$G.M. = \text{Antilog}\left[\frac{1}{N}\sum_{i=1}^{n} f_i \log X_i\right] = \text{Antilog}\left[\frac{1}{100}(327.9316)\right] = \text{Antilog}(3.2793) = 1902.5$$

$$H.M. = \frac{1}{\frac{1}{N}\left[\sum_{i=1}^{n}\frac{f_i}{X_i}\right]} = \frac{1}{\frac{1}{100}[0.0528]} = \frac{100}{0.0528} = 1893.9394$$

From example 2 and example 3 one can verify that $A.M. \geq G.M. \geq H.M.$

3.6 Median (Positional Average)

The median is defined the measure of the central value when arranged in ascending or descending order of magnitude. According to L.R. Connor "The median is that value of the variate which divides the group in two equal parts, one part comprising all the values greater and the other, all values less than median". Thus, as against arithmetic mean which is based on all the items of the distribution, the median is only positional average i.e. its value depends on the position occupied by a value in the frequency distribution.

3.6.1 Calculation of median

3.6.1.1 Ungrouped data

When the total numbers of observations are odd, then the median is the middle value after the observations are arranged in ascending or descending order of magnitude. If the number of observations is equal to n, then the value of $((n+1)/2)^{th}$ item gives the value of median e.g. the median of 5 observations 65,69,52,58,45 i.e. 45,52,58,65,69 is 58. When the total number of observations is even then median is obtained as the arithmetic mean of the two middle observations after they are arranged in ascending or descending order of magnitude. If number of observations are say 2n, then the arithmetic average of n^{th} and $(n+1)^{st}$ (central items) gives the value of median. If it is n then median is the arithmetic average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ values e.g. the median of 6 observations 65,69,52,58,45,67 i.e. 45,52,58,65,67,69 is arithmetic mean of 58 and 65 which is equal to 61.5.

3.6.1.2 Grouped data

Steps involved for its computation are:

1) Prepare less than cumulative frequency(c.f.) distribution table
2) Find N/2.
3) Find cumulative frequency just greater than N/2
4) The class corresponding to step 3 contains the median value and is called the median class.

The median for a grouped series is given by the following formula:

$$\text{Median} = 1 + \frac{\frac{N}{2} - cf}{f} \times h$$

Where : l is the lower limit of the median class

f is the frequency of the median class

h is width of the class interval

c.f. is the cumulative frequency of the class preceding the median class.

$$N = \sum_{i=1}^{n} f_i = \text{Total frequency}$$

The computational procedure is illustrated in Example 3.4.

3.6.2 Merits of median

- It is rigidly defined.
- It is easily understood, very readily calculated and can exactly be located.
- It is readily obtained without the necessity of measuring all the objects.

- It is not affected by abnormally large or small values of the variable.
- Median can be computed while dealing with a distribution with open end classes.
- It can be determined by mere inspections and can be computed graphically.
- The median gives the best results in a study of direct qualitative measurements such as intelligence, honesty etc.

3.6.3 Demerits of median

- The median does not lend itself to algebraic treatment. The median of several series by combining the medians of the component series cannot be computed.
- Median being positional average is not based on each and every item of the observations.
- Median is relatively less stable than mean, particularly for small samples since it is affected more by fluctuations of sampling as compared to arithmetic mean.

3.7 Quartiles

Quartiles are those values of the variate which divide the total frequency into four equal parts. Obviously there will be three such points Q_1, Q_2 and Q_3 such that $Q_1 \leq Q_2 \leq Q_3$ termed as the three quartiles. Q_1 is known as the lower or first quartile and is the value which has 25% of the items of the distribution below it and consequently 75 percent of the items are greater than it. Q_3 is known as the upper or third quartile and has 75percent of the observations below it and consequently 25 percent of the observations above it.

$$Q_i = 1 + \left(\frac{iN}{4} - c.f.\right) \times \frac{h}{f}, i = 1, 2 \text{ and } 3$$

3.8 Deciles

Deciles are those values of the variate which divide the total frequency into 10 equal parts. The formula for obtaining j^{th} Decile (D_j) in case of grouped frequency distribution is given as $D_j = 1 + \left(\frac{jN}{10} - c.f.\right) \times \frac{h}{f}$ j=1, 2, 3, ---, 9

3.9 Percentiles

Percentiles are the values of the variates which divide the total frequency into 100 equal parts. The formula for obtaining k^{th} Percentile (P_k) in case of grouped frequency distribution is given as $P_k = 1 + \left(\frac{kN}{100} - c.f.\right) \times \frac{h}{f}$

k=1, 2, 3, ---, 99.

3.9.1 Graphical method of locating position values

The various partition values viz., median, quartiles, deciles and percentiles can be located graphically with the help of curve called the cumulative frequency curve or ogive. Draw a perpendicular from the point of the two ogives i.e. more than ogive and less than ogive on the x-axis, the foot of the perpendicular gives the value of

median. The points corresponding to N/4, 3N/4, N/10,….., 9N/10, N/100,….., 99N/100 on y-axis with the foot values of the perpendicular on x-axis provide the value of Q_1, Q_3, D_1, ……, D_9, P_1, ……., P_{99}.

3.10 Mode

Mode is the value which occurs most in a set of observations and around which the other items of the set cluster densely. It is defined to be size of the variable which occurs most frequently or the point of maximum frequency or the point of greatest density. In other words mode is that value of observation for which the height of the ordinate is maximum. Modal value of the distribution is that value of the variate for which frequency is maximum. In the words of Croxton and Cowden "The mode of a distribution is value at the point around which the items tend to be heavily concentrated. It may be regarded as the most typical value of a series of values. "

3.10.1 Computation of mode

In case of a frequency distribution, mode is the value of the variable corresponding to the maximum frequency. For a continuous frequency distribution, the class corresponding to maximum frequency is called the modal class. The mode is computed by the formula:

$$\text{Mode} = 1 + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2} \times h$$

Where l = lower limit of the modal class

f_1 is the frequency of the modal class.

f_o is the frequency of the class just preceding the modal class (pre-modal class).

f_2 is the frequency of the class just succeeding the modal class (post-modal class).

h is the magnitude of the modal class.

The computational procedure is illustrated in example 4.

3.10.2 Merits of mode

- It is easily understood.
- It is the most typical value and it is the most descriptive average.
- It is a positional average.
- It can be easily located by mere inspection of certain items.
- It can be easily determined from the graph.
- The extreme items have no effect provided they are not in the modal class.

3.10.3 Demerits of mode

- It is ill defined. A clearly defined mode does not always exist. The value of mode cannot always be determined. A distribution can be bimodal or multimodal.
- It is not based on all the observations of a series.

- It is not suitable of further mathematical treatment.
- As compared to mean, mode is affected to a greater extent by the fluctuations of sampling.

Graphically mode can be located from the histogram of frequency distribution by making use of the rectangles erected on the modal, pre-modal and post modal classes. The method consists of following steps:

a) Join the top right corner of the rectangle erected on the modal class with top left corner of the rectangle erected on the preceding class by means of a straight line.

b) Join the top last corner of the rectangle erected on the modal class with top right corner of the rectangle erected on the succeeding class by means of a straight line.

c) From the point of intersection of the lines in step (i) and (ii) above, draw a perpendicular to the X-axis. The abscissa of the point where this perpendicular meets the X-axis gives the modal value.

Example 4: Find median, first quartile, third quartile and mode of the frequency distribution given in example 1 and obtain them graphically.

Solution: Prepare the following table to calculate median, first quartile, third quartile and mode.

Class Interval (in kg)	frequency (f_i)		Less than Cumulative frequency (c.f.)
1630-1730	17		17
1730-1830	19	→ f_0	36
1830-1930	23	→ f_1	59
1930-2030	16	→ f_2	75
2030-2130	14		89
2130-2230	7		96
2230-2330	2		98
2330-2430	2		100

Here N/2=50. Cumulative frequency greater than 50 is 59. Hence the median class is 1830-1930.

$$\text{Median} = 1830 + \left(\frac{50-36}{23}\right) \times 100 = 1830 + 60.8696 = 1890.8696$$

N/4=25. Cumulative frequency greater than 25 is 36. Hence the first quartile class is 1730-1830.

$$Q_1 = 1730 + (25\text{-}17) \times \frac{100}{19} = 1730 + 42.1053 = 1772.1053$$

3N/4=75. Cumulative frequency greater than 75 is 89. Hence the third quartile class is 2030-2130.

$$Q_3 = 2030 + (75-75) \times \frac{100}{14} = 2030$$

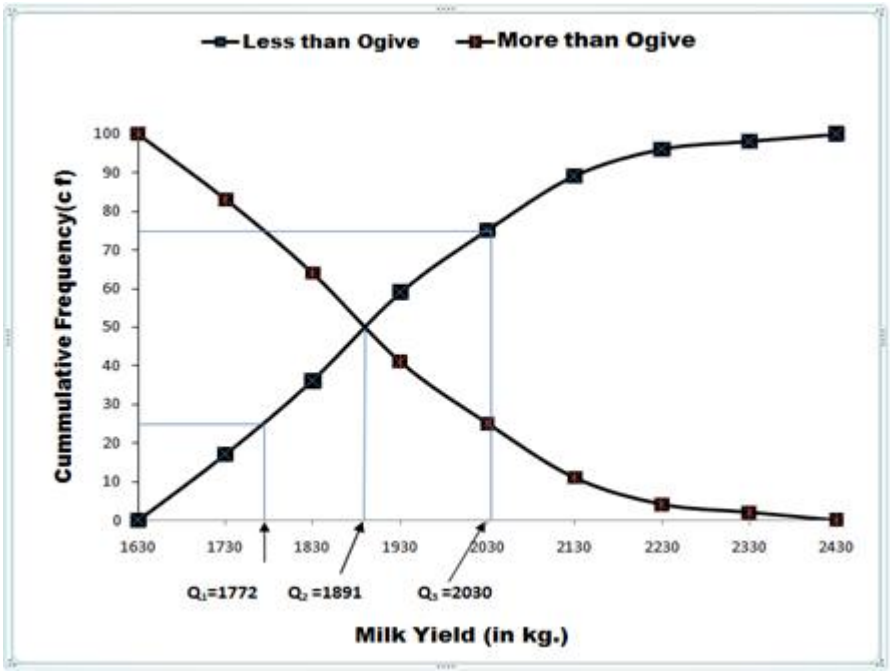


Fig 3.1 Graphical method to find first quartile, median and third quartile

For computing Mode, the maximum frequency (23) occurs in the class interval 1830-1930, which is called modal class.f_1=23, f_0=19 and f_2=16. Using formula

$$\text{Mode} = 1830 + \frac{100(23-19)}{2(23)-19-16} = 1830 + \frac{400}{11} = 1866.3636$$

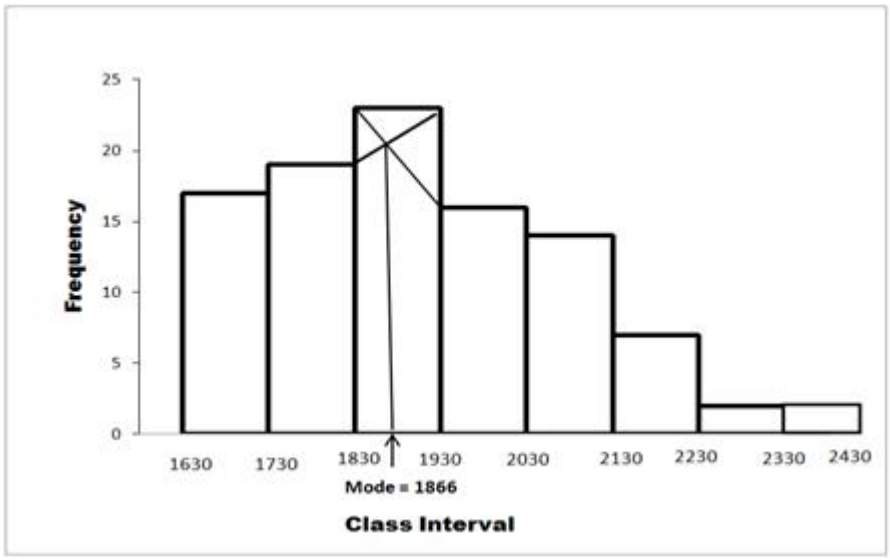


Fig 3.2 Graphical method to find mode

Empirical relation between Mean, Median and Mode

In case of symmetrical distribution mean, mode and median coincide while for asymmetrical distribution the empirical relationship is Mode = 3 Median -2 Mean.

Lesson 4

MEASURES OF DISPERSION

4.1 Introduction

In the preceding lesson, we have seen different measures of central tendency and learnt how they can be calculated for varying types of distributions. The measures of central tendency are just different types of averages and do not indicate the extent of variability in a distribution. Averages or the measures of central tendency give us an idea of the concentration of the observations about the central part of the distributions. If we are given the average of a series of observations, we cannot form complete idea about the distribution since there may exist a number of distributions whose averages are same but they may differ widely from each other in a number of ways. Let us consider two series I and II of 6 items each

Series							Total	Mean
I	20	20	25	25	30	30	150	25
II	15	20	25	25	30	35	150	25

We notice that there is no difference as far as the average is concerned. But we notice that in the first case the observations vary from 20 to 30 and in the second case, the observations vary from 15 to 35 i.e. we notice that the greatest deviation from the mean in the first case is 5 and in the second case it is 10. Clearly this indicates a difference in the two series. Such a variation is called scatter or dispersion. Thus, the measures of central tendency must be supported and supplemented by some other measures. One such measure is dispersion. Measures of dispersion help us to study variability of the items i.e. the extent to which the items vary from one another and also from the central value.

4.2 Meaning of Dispersion

The term dispersion is generally used in two senses. Firstly, dispersion refers to the variation of the items among themselves. If the value of all the items of a series is the same, there will be no variation among the various items and the dispersion will be zero. On the other hand, the greater the variation among different items of a series, the more will be the extent of dispersion. Secondly, dispersion refers to the variation of the items about an average. If the difference between the value of items and the average is large, the dispersion will be high and on the other hand if the difference between the values of items and average is small, the dispersion will be low. Thus, dispersion is defined as scatteredness around central value or the spread of the individual items in a given series. According to A.L. Bowley “Dispersion is the measure of the variation of the items”. Spiegel defined dispersion as “The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data”.

4.3 Objectives of Measuring Dispersion

The measures of dispersion are helpful in statistical investigation. Some of the main objectives of dispersion are:

- To determine the reliability of an average.
- To compare the variability of two or more series.
- For facilitating the use of other statistical measures.
- Basis of Statistical Quality Control.

4.4 Characteristics for an Ideal Measure of Dispersion

The following are the essential requisites for an ideal measure of dispersion:

- It should be rigidly defined.
- It should be based on all observations.
- It should be readily comprehensive.

- It should be easily calculated.
- It should be amenable to further mathematical treatment.
- It should be affected as little as possible by fluctuations of sampling.
- It should not be affected much by extreme observations.

4.5 Absolute and Relative Measures of Dispersion

The measures of dispersion which are expressed in terms of the original units of a series are termed as Absolute Measures. Such measures are not suitable for comparing the variability of the two distributions which are expressed in different units of measurement. On the other hand, relative measures of dispersion are obtained as ratios or percentages and are thus pure numbers independent of the units of measurement. These measures are used to compare two series expressed in different units.

4.6 Measures of Dispersion

Various measures of dispersion in common use are:

4.6.1 Range

The simplest possible measure of dispersion is the range which is nothing but the difference between the greatest and the smallest observation of the distribution Thus

$$\text{Range} = X_{max} - X_{min}$$

where X_{max} is the greatest observation and X_{min} is the smallest observation of the variable value. In case of the grouped frequency distribution range is defined as the difference between upper limit of the highest class and the lower limit of the smallest class. In order to compare the variability of the two or more distributions given in different units of measurement, the relative measure , called coefficient of range is used and this is defined as follows:

$$\text{Coefficient of range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$$

In other words coefficient of range is the ratio of the difference between two extreme observations of the distribution to their sum.

4.6.1.1 Merits and demerits of range

Range is the simplest though crude measure of dispersion. It is rigidly defined, readily comprehensible and easiest to compute. It got the following drawbacks

- It is not based on all the observations.
- It is very much affected by fluctuations of sampling.
- It is unreliable measure of the dispersion.
- It cannot be used if we are dealing with open end classes.
- Range is not suitable for mathematical treatment.

4.6.1.2 Uses of range

In spite of above limitations range as a measure of dispersion, has following applications

- In a number of fields where the data have small variations like in stock market fluctuations, the variations in money rates and rate of exchange .
- It is used in industry for the statistical quality control of the manufactured products by the construction of R chart i.e. the control chart for range.
- It is also used as a very convenient measure by meteorological department for weather forecasts.

4.6.2 Quartile deviation or semi-inter-quartile range

The difference between the upper and lower quartiles i.e. $Q_3 - Q_1$ is known as the inter-quartile range and half of this difference i.e. $\frac{1}{2}(Q_3 - Q_1)$ is called the semi-inter-quartile range or the quartile deviation denoted by Q.D. For comparative studies of variability of two distributions the relative measure which is known as Coefficient of Quartile deviation which is given by

$$\text{Coefficient of Quartile Deviation} = \frac{X_3 - X_1}{X_3 + X_1}$$

4.6.2.1 Merits of quartile deviation

- The quartile deviation is easy to compute and understand.
- It is a better measure of dispersion than range because it makes use of 50% of the data.
- It is not affected at all by extreme observations.
- It can be computed from the frequency distribution with open end classes.

4.6.2.2 Demerits of quartile deviation

- It is not based on all the observations.
- It is affected considerably by fluctuations of sampling.
- It is not suitable for further mathematical treatment.

4.6.3 Mean deviation or average deviation

This measure of dispersion is obtained by taking the arithmetic mean of the absolute deviations of the given values from a measure of central tendency. According to Clark and Schkade: "Average deviation is the average amount of scatter of the items in a distribution either the mean or the median, ignoring the signs of deviations. The average that is taken of the scatter is an arithmetic mean, which accounted for the fact that this measure is often called the mean deviation".

4.6.3.1 Calculation of mean deviation

If $X_1, X_2, ---, X_n$ are n given observations then mean deviation (M.D.) about an average A is given by:

$$\text{M.D. (about an average A)} = \frac{1}{n}\sum_{i=1}^{n}|X_i - A| = \frac{1}{n}\sum_{i=1}^{n}|d_i|$$

where $|d_i| = |X_i - A|$ read as mod $(X_i - A)$ is the modulus value or absolute value of the deviation and A is one of the averages viz., Mean (M), Median (M_d) and Mode (M_o)

In case of grouped frequency distribution, mean deviation about an average A is given by:

M.D. (about an average A) = $\frac{1}{N}\sum_{i=1}^{n} f_i|X_i - A| == \frac{1}{N}\sum_{i=1}^{n} f_i|d_i|$ where X_i is the mid value of the class interval, f_i is the corresponding frequency, $N = \sum_{i=1}^{n} f_i$ is the total frequency.

Mean deviation is minimum when it is calculated from median. In other words, mean deviation calculated about median will be less than the mean deviation about mean or mode. The relative measures of mean deviation is called coefficient of mean deviation is given by

$$\text{Coefficient of M.D.} = \frac{\text{Median deviation}}{\text{Average about which it is calculated}}$$

$$\text{Coefficient of M.D. about mean} = \frac{\text{Mean deviation}}{\text{Mean}}$$

$$\text{Coefficient of M.D. about median} = \frac{\text{Mean deviation}}{\text{Median}}$$

$$\text{Coefficient of M.D. about mode} = \frac{\text{Mean deviation}}{\text{Mode}}$$

The coefficients of mean deviations defined above are pure numbers independent of units of measurement and are useful for comparing the variability of different distributions. The calculation of various measures is illustrated in example 1.

Example 1: Find mean deviation from mean, median and mode using the data given in example 1 of Lesson 2. Also find the coefficient of mean deviation about mean, median and mode.

Solution : Using the values of Mean (M) =1910 Median (M_d) = 1890.8696 and Mode (M_o) = 1866.3636, calculated in Lesson 3and then prepare the following table:

Class Interval	Mid-value (X_i)	frequency (f_i)	X_i-M	$f_i\|X_i - M\|$	X_i-M_d	$f_i\|X_i - M_d\|$	X_i-M_o	$f_i\|X_i - M_o\|$
1630-1730	1680	17	-230	3910	-210.87	3584.7832	-186.364	3168.1812
1730-1830	1780	19	-130	2470	-110.87	2106.5224	-86.3636	1640.9084
1830-1930	1880	23	-30	690	-10.8696	250.0008	13.6364	313.6372
1930-2030	1980	16	70	1120	89.1304	1426.0864	113.6364	1818.1824
2030-2130	2080	14	170	2380	189.1304	2647.8256	213.6364	2990.9096
2130-2230	2180	7	270	1890	289.1304	2023.9128	313.6364	2195.4548
2230-2330	2280	2	370	740	389.1304	778.2608	413.6364	827.2728
2330-2430	2380	2	470	940	489.1304	978.2608	513.6364	1027.2728

Total		100		14140		13795.6528		13981.8192

$$\text{M. D. (about mean)} = \frac{1}{N}\sum_{i=1}^{n} f_i\,|X_i - M| = \frac{14140}{100} = 141.40$$

$$\text{M. D. (about median)} = \frac{1}{N}\sum_{i=1}^{n} f_i\,|X_i - M_d| = \frac{13795.6528}{100} = 137.9565$$

$$\text{M. D. (about mode)} = \frac{1}{N}\sum_{i=1}^{n} f_i\,|X_i - M_o| = \frac{13981.8192}{100} = 139.8182$$

From above calculations we can verify that mean deviation calculated about median (137.9565) is less than mean deviation about mean (141.10) or mode (139.8182).

$$\text{Coefficient of M. D. about mean} = \frac{\text{Mean deviation}}{\text{Mean}} = \frac{141.40}{1910} = 0.0740$$

$$\text{Coefficient of M. D. about median} = \frac{\text{Mean deviation}}{\text{Median}} = \frac{137.9565}{1890.8696} = 0.0729$$

$$\text{Coefficient of M. D. about mode} = \frac{\text{Mean deviation}}{\text{Mode}} = \frac{139.8182}{1866.3636} = 0.0749$$

4.6.3.2 Merits of mean deviation

- It is rigidly defined, easy to understand and calculate.
- It is based on all observations and is better than range and quartile deviation.
- The averaging of the absolute deviations from an average irons out the irregularities in the distribution and thus provides an accurate measure of dispersion.
- It is less affected by extreme observations.

4.6.3.3 Demerits of mean deviation

- Ignoring the signs is not correct from mathematical point of view.
- It is not an accurate method when it is calculated from mode.
- It is not capable of further mathematical treatment.
- It cannot be used if we are dealing with open end classes.

4.6.4 Standard deviation

Standard deviation, usually denoted by the Greek alphabet σ was first suggested by Karl Pearson as a measure of dispersion in 1893. It is defined as the positive square root of the mean of the square of the deviations of the given observations from their arithmetic mean. If $X_1, X_2, ---, X_n$ is a set of n observations then its standard deviation is given by :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad \text{where } \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$ is the arithmetic mean

is the arithmetic mean.

In case of a grouped data, the standard deviation is given by:

Thus $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \left(X_i - \overline{X}\right)^2}$ where X_i is the value of the variable or mid value of the class in case of grouped frequency distribution; f_i is the corresponding frequency of the value X_i, $N = \sum_{i=1}^{n} f_i$ is the total frequency and

$\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$ is the arithmetic mean of the distribution.

The square of the standard deviation viz., σ^2 is called variance or second moment about mean.

4.6.4.1 Computation of variance (Direct method)

Other formulae for calculating variance is

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

and in case of grouped data is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i X_i^2 - \left(\overline{X}\right)^2 = \frac{1}{N}\sum_{i=1}^{n} f_i X_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i X_i\right)^2$$

4.6.4.2 Short–cut method (Change of origin)

This method consists in taking deviations of the given observations from any arbitrary value A. The formula for calculation of the arithmetic mean is

$$\bar{X} = A + \frac{\sum_{i=1}^{n} f_i d_i'}{N}, \sigma^2 = \left[\frac{1}{N}\sum_{i=1}^{n} f_i d_i'^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i d_i'\right)^2\right]$$

The variance and consequently the standard deviation of a distribution is independent of the change of origin. Thus, if we add (subtract) a constant to (from) each observation of the series, its variance remains same.

4.6.4.3 Step- deviation method (Change of origin and scale)

In case of grouped frequency distribution, with class intervals of equal magnitude, the calculations are further simplified by taking; $d_i = \frac{X_i - A}{h}$ where X_i is the mid value of the class and h is the common magnitude of the class intervals. So the formula for calculating mean and variance is

$$\overline{X} = A + \left(\frac{\sum_{i=1}^{n} f_i d_i}{N}\right) \times h \text{ and } \sigma^2 = \left[\frac{1}{N}\sum_{i=1}^{n} f_i d_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i d_i\right)^2\right] \times h^2$$

which shows that the variance or standard deviation is not independent of change of scale. Thus, if we multiply (divide) each observation of the series by a constant h, its variance will be multiplied (divided) by h^2.Hence variance and consequently the

standard deviation of a distribution is independent of the change of origin but not of the scale. The procedure is illustrated in the example 2.It will be seen that the answer in each of the three cases is the same. The step-deviation method is the most convenient on account of simplified calculations.

Example 2: Find variance of the data given in example 1 of Lesson 3 with short-cut and step-deviation method.

Solution: Prepare the following table to calculate variance by different methods.

Class Interval	Mid-value (X_i)	freq (f_i)	$f_i X_i$	$f_i X_i^2$	$d_i'=X_i-A$ A=2080	$f_i d_i'$	$f_i d_i'^2$	$d_i = \frac{X_i - A}{h}$	$f_i d_i$	$f_i d_i^2$
1630-1730	1680	17	28560	47980800	-400	-6800	2720000	-4	-68	272
1730-1830	1780	19	33820	60199600	-300	-5700	1710000	-3	-57	171
1830-1930	1880	23	43240	81291200	-200	-4600	920000	-2	-46	92
1930-2030	1980	16	31680	62726400	-100	-1600	160000	-1	-16	16
2030-2130	2080	14	29120	60569600	0	0	0	0	0	0
2130-2230	2180	7	15260	33266800	100	700	70000	1	7	7
2230-2330	2280	2	4560	10396800	200	400	80000	2	4	8
2330-2430	2380	2	4760	11328800	300	600	180000	3	6	18
Total		100	191000	367760000		-17000	5840000		-170	584

Direct Method

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i X_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i X_i\right)^2 = \frac{1}{100}(367760000) - \left(\frac{191000}{100}\right)^2 =$$
$$3677600 - 3648100 = 29500$$

Short–cut Method

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i d_i'^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i d_i'\right)^2 = \frac{1}{100}(5840000) - \left(\frac{1}{100}(-17000)\right)^2$$
$$= 58400 - (170)^2 = 29500$$

Step –Deviation Method

$$\sigma^2 = \left[\frac{1}{N}\sum_{i=1}^{n} f_i d_i{}^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i d_i\right)^2\right] \times h^2 = \left[\frac{1}{100}(584) - \left(\frac{1}{100}(-170)\right)^2\right] \times (100)^2$$

$$= [5.84 - (1.7)^2] \times 1000 = 29500$$

4.6.4.4 Merits of standard deviation

- It is rigidly defined.
- It is based on all observations and is the best measure of dispersion.
- The squaring of the deviations from mean removes the drawback of ignoring the signs of deviations in computing the mean deviation. This makes it suitable for further mathematical treatment. The variance of the combined series can also be computed.
- It is least affected by fluctuations of sampling and therefore, it widely used in sampling theory and tests of significance.

4.6.4.5 Demerits of standard deviation

- As compared to the quartile deviation and range etc., it is difficult to understand and difficult to calculate.
- It gives more importance to extreme observations.

4.6.4.6 Variance of the combined series

As pointed earlier variance is suitable for algebraic treatment i.e. if we are given the averages, the sizes and the variances of a number of series, then we can obtain the variance of the resultant series obtained by combining different series. Thus if

$\sigma_1{}^2, \sigma_2{}^2, \ldots\ldots\ldots, \sigma_k{}^2$ are the variances; $\overline{X_1}, \overline{X_2}, \ldots\ldots\ldots\ldots, \overline{X_k}$ and $n_1, n_2, \ldots\ldots\ldots, n_k$ are the arithmetic means and sizes of k series respectively. Then the variance of the combined series of size $N = n_1 + n_2 + \cdots\ldots\ldots + n_k$ is given by the formula

$$N\sigma^2 = [n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + ---- + n_k(\sigma_k^2 + d_k^2)]$$

where $d_1 = \bar{X}_1 - \bar{\bar{X}}, d_2 = \bar{X}_2 - \bar{\bar{X}}, ---, d_k = \bar{X}_k - \bar{\bar{X}}$ *and* $\bar{\bar{X}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 --- n_k\bar{X}_k}{n_1 + n_2 --- n_k}$ is the mean of combined series. In particular, for two series the combined variance is given by

$$(n_1 + n_2)\sigma^2 = \left[n_1(\sigma_1{}^2 + d_1{}^2) + n_2(\sigma_2{}^2 + d_2{}^2)\right]$$

where $d_1 = \bar{X}_1 - \bar{\bar{X}}, d_2 = \bar{X}_2 - \bar{\bar{X}}$ and $\bar{\bar{X}} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$

Substituting the values of $d_1 = \frac{n_2(\bar{X}_1 - \bar{X}_2)}{n_1 + n_2}$ and $d_2 = \frac{n_1(\bar{X}_1 - \bar{X}_2)}{n_1 + n_2}$, combined variance is

$$\sigma^2 = \left[\frac{n_1\sigma_1{}^2 + n_2\sigma_2{}^2}{n_1 + n_2} + \frac{n_1 n_2(\overline{X_1} - \overline{X_2})^2}{n_1 + n_2}\right]$$

4.6.5 Coefficient of variation

Standard deviation is an absolute measure of dispersion. The relative measure of dispersion based on standard deviation is called the coefficient of standard deviation and is given by

$$\text{Coefficient of standard deviation} = \sigma/\bar{X}$$

This is a pure number independent of the units of measurement and thus, is suitable for comparing the variability, homogeneity or uniformity of two or more distributions.

100 times the coefficient of dispersion based on standard deviation is called the coefficient of variation (C.V.) expressed in percentage. Thus,

Coefficient of Variation = $\sigma/\bar{X} \times 100$

This measure was suggested by Prof. Karl Pearson and according to him "Coefficient of variation is the percentage variation in mean, standard deviation being considered as the total variation in the mean". For comparing the variability of two distributions we compute the coefficient of variation for each distribution. A distribution with relatively smaller C.V. is said to be more homogeneous or uniform or less variable or more consistent than the other and the series with relatively greater C.V. is said to be more heterogeneous or more variable or less consistent than the other.

Lesson 5

MEASURES OF SKEWNESS AND KURTOSIS

5.1 Introduction

In the preceding lessons, we have discussed the measures of central tendency and dispersion in case of frequency distribution. The measures of central tendency tell us about the concentration of the observations about the middle of the distribution and the measure of dispersion gives us an idea about the spread or scatter of the observations about some measure of central tendency. These measures, however, don't adequately describe a frequency distribution in the sense that there could be two or more distributions with the same mean and standard deviation but still different from each other with regard to shape or pattern of distribution. Thus these two measures of central tendency and dispersion are inadequate to characterize a distribution completely and must be supported and supplemented by two more measures viz. skewness and kurtosis which we shall discuss in this lesson.

5.2 Skewness

Literal meaning of skewness is "lack of symmetry". It measures the degree of departure of a distribution from symmetry and reveals the direction of scatterdness of the items.

A frequency distribution is said to be symmetrical when values of the variables equidistant from their mean have equal frequencies. If a frequency distribution is not symmetrical, it is said to be asymmetrical or skewed. Any deviation from symmetry is called skewness.

According to *Morris Humberg* "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution".

According to *Croxton & Cowden* "When a series is not symmetrical it is said to be asymmetrical or skewed".

According to *Simpson & Kafka* "Measures of skewness tell us the direction and the extent of skewness. In a symmetrical distribution the mean, median and mode are identical. The more we move away from the mode, the larger the asymmetry or skewness".

In the words of *Riggleman and Frisbee* "Skewness is the lack of symmetry when a frequency distribution is plotted on a chart, skewness present in the items tends to be dispersed more on one side of the mean than on the other".

Thus the above definitions make it clear that the word skewness refers to the lack of symmetry. If a distribution is normal there would be no skewness in it and the curve drawn from the distribution would be symmetrical. In case of skewed distributions the curve drawn would be elongated either to the left or to the right. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution. A measure of skewness is, therefore necessary in order to guard against the consequences of the assumption. The following three figures would give an idea about the shape of symmetrical and asymmetrical curves.

5.2.1 Symmetrical curve

The figure 5.1, given below, presents the shape of a symmetrical curve which is bell shaped having no skewness. The value of mean (M), median (M_d) and mode (M_o) for such a curve would be identical.

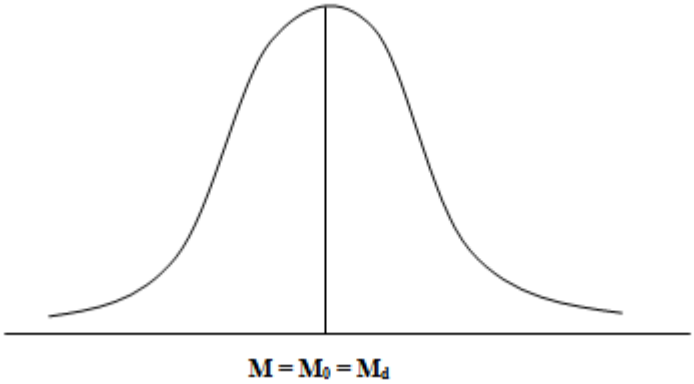


Fig. 5.1 Symmetrical distribution

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 centre point of the curve. For a symmetrical distribution Mean = Median = Mode.

5.2.2 Positively skewed curve

A positively skewed curve has a longer tail towards the higher values of X i.e. the frequency curve gradually slopes down towards the higher values of X. In a positively skewed distribution the mean is greater than the median and then mode and the median lies in between mean and mode. The frequencies are spread over a greater range of values on the high value end of the curve (the right hand side) as is clear from the Fig 5.2.

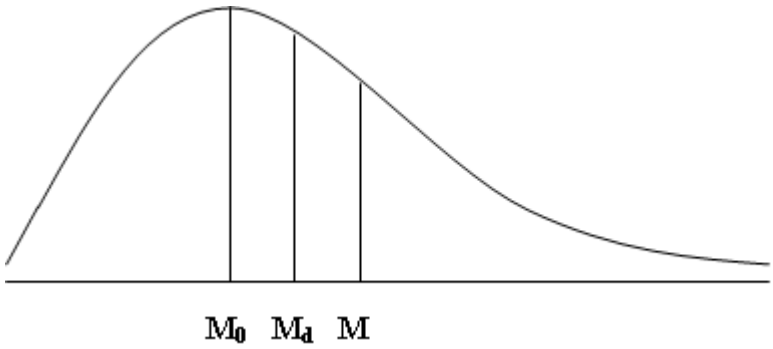


Fig. 5.2 Positively skewed distribution

For a positively skewed distribution Mean > Median > Mode.

5.2.3 Negatively skewed curve

A negatively skewed curve has a longer tail towards the lower values of X i.e. the frequency curve gradually slopes down towards the lower values of X as shown in Fig. 5.3.

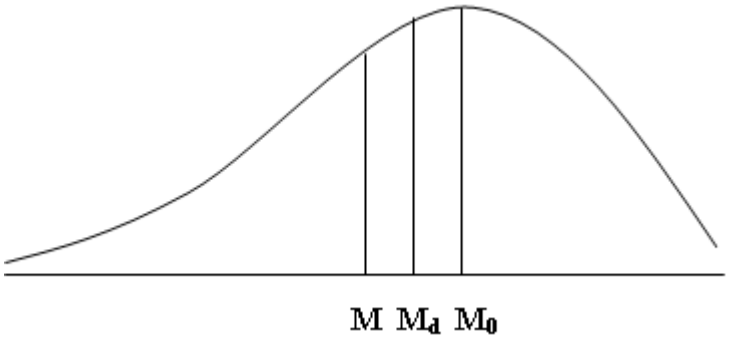


Fig. 5.3 Negatively skewed distribution

In the negatively skewed distribution the mode is the maximum and mean is the least. The median lies in between mean and mode. The elongated tail in negatively skewed distribution is on the left hand side as would be clear from Fig 5.3. For a negatively skewed distribution, Mean < Median < Mode.

5.3 Measures of Skewness

Measures of skewness are meant to give an idea about the extent of asymmetry in a series. A distribution is said to be skewed if

- The frequency curve of the distribution is not a symmetric bell shaped curve but stretched more to one side than to the other.
- The values of mean (M), median (M_d) and mode (M_o) fall at different points i.e they don't coincide.
- Quartiles Q_1 and Q_3 are not equidistant from the median.
- The corresponding pairs of deciles and percentiles are not equidistant from the median.

If a particular distribution is found to be skewed, the next problem that arises is to measure the extent of skewness. To find out the direction and extent of asymmetry in a series statistical measures of skewness are employed.

These measures can be absolute or relative. The absolute measures of skewness tell us the extent of asymmetry and whether it is positive or negative. The absolute skewness is based on the difference between mean and mode. Symbolically,

Absolute skewness = Mean – Mode

Skewness will be positive, if the value of mean is greater than the mode and skewness will be negative, if the value of mean is less than the mode. The difference between the mean and the mode, whether positive or negative, indicates the distribution is positively skewed or negatively skewed. However, such an absolute measure of skewness is not adequate because it cannot be used for comparison of skewness in two distributions, if they are in different units, since difference between the mean and mode will be in terms of the units of distribution. Thus for comparison purposes we use the relative measures of skewness known as co-efficient of skewness.

5.3.1 Karl Pearson's Coefficient of Skewness

The first coefficient of skewness as defined by Karl Pearson is

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Std. deviation}} = \frac{M - M_0}{\sigma}$$

This measure is based on the fact that the mean and the mode are drawn widely apart. Skewness will be positive if mean > mode and negative if mean < mode. There is no limit to this measure in theory and this is a slight drawback. But in practice the value given by this formula is rarely very high and its value usually lies between -1 and +1.

It may also be written as $\frac{3(\text{Mean} - \text{Median})}{\sigma}$ as Mode = 3 Median – 2 Mean

This coefficient is a pure number without units since both numerator and denominator have the same dimensions. The value of this coefficient lies between -3 and +3.

5.3.2 Bowley's Coefficient of Skewness

Prof. A.L. Bowley's Coefficient of Skewness is based on quartiles and is given by:

$$\text{Coefficient of Skewness} = \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{(Q_3 - \text{Median}) + (\text{Median} - Q_1)}$$
$$= \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

This is also known as Coefficient of Skewness based on quartiles and is especially useful in situations where quartiles and median are used viz.

- When the mode is ill-defined and extreme observations are present in the data.
- When the distribution has open end classes or unequal class intervals.

This coefficient is a pure number without units since both numerator and denominator have the same dimensions. The value of this coefficient lies between -1 and +1.

5.3.3 Kelly's Coefficient of Skewness

The drawback of Bowley's Coefficient of Skewness is that it ignores the 50% of the data which can be partially removed by taking two deciles or percentiles equidistant from the median value. The refinement was suggested by Kelly.

$$\text{Coefficient of Skewness} = \frac{P_{90} + P_{10} - 2\,\text{Median}}{P_{90} - P_{10}} = \frac{D_9 + D_1 - 2\,\text{Median}}{D_9 - D_1}$$

5.3.4 Coefficient of Skewness based on moments

5.3.4.1 Moments

Moments are the general statistical measure used to describe and analyse the characteristics of a frequency distribution viz. central tendency, dispersion, skewness and kurtosis. Let us consider the variable X having a frequency distribution as given below:

X	X_1	X_2	X_3	----	---	X_n

f	f_1	f_2	f_3	----	---	f_n

Then, $\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$ is the arithmetic mean.

5.3.4.2 Moments about Mean

The r^{th} moment about Mean $\overline{X}$ is denoted by μ_r and also known as central moment and is defined as

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - \overline{X})^r, r = 0,1,2,3 - - - \qquad (5.1)$$

Putting r=0 in equation (5.1) we get $\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - \overline{X})^0 = 1$ (5.2)

Putting r=1 in equation (5.1) we get $\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - \overline{X})^1 = 0$ (5.3)

because the algebraic sum of deviations of a given set of observations from their mean is zero. Thus the first moment about mean is always zero.

Again taking r=2 we get $\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - \overline{X})^2 = \sigma^2$. (5.4)

Hence second moment about mean gives the variance of the distribution.

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - \overline{X})^3 \qquad (5.5)$$

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - \overline{X})^4 \qquad (5.6)$$

5.3.4.3 Moments about an arbitrary point A

The r^{th} moment about any point A denoted by μ_r' are also known as raw moment and is defined as

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - A)^r, r = 0,1,2,3 - - - \qquad (5.7)$$

Putting r=0 and r=1 in equation (5.7) we get respectively

$$\mu_0' = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - A)^0 = 1 \qquad (5.8)$$

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - A)^1 = \frac{1}{N}\left(\sum_{i=1}^{n} f_i X_i - A\sum_{i=1}^{n} f_i\right) = \frac{1}{N}\left(\sum_{i=1}^{n} f_i X_i\right) - A = \bar{X} - A \qquad (5.9)$$

$$\Rightarrow \bar{X} = A + \mu_1' \quad \text{where } \mu_1' \text{ is the first moment about the arbitrary point A} \qquad (5.10)$$

Taking r=2, 3, 4 in (5.7) we get respectively

Second moment about the point A $\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - A)^2$ (5.11)

Third moment about the point A $\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - A)^3$ (5.12)

Fourth moment about the point A $\mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i (X_i - A)^4$ (5.13)

5.3.4.4 Relation between moments about mean and moments about an arbitrary point A

In this section we shall try to obtain the expression for μ_r, the r[th] moment about mean as defined in (5.1) in terms of μ_r', the r[th] moment about any arbitrary point A as defined in (5.7).

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i \left(X_i - \bar{X}\right)^r = \frac{1}{N}\sum_{i=1}^{n} f_i \left(X_i - A + A - \bar{X}\right)^r \qquad (5.14)$$

From (5.9) we get $\mu_1' = \bar{X} - A \Rightarrow A - \bar{X} = -\mu_1'$, substituting in (5.14) we get

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i [(X_i - A) - \mu_1']^r$$

Now by binomial theorem

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\left[(X_i - A)^r - r_{C_1}(X_i - A)^{r-1}\mu_1' + r_{C_2}(X_i - A)^{r-2}\mu_1'^2 - r_{C_3}(X_i - A)^{r-3}\mu_1'^3 + - - - + (-1)^r\mu_1'^r\right]$$

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^r - r_{C_1}\mu_1'\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^{r-1} + r_{C_2}\mu_1'^2\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^{r-2} - r_{C_3}\mu_1'^3\frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^{r-3} + - - - + (-1)^r\mu_1'^r$$

$$\mu_r = \mu_r' - r_{C_1}\mu_{r-1}'\mu_1' + r_{C_2}\mu_{r-2}'\mu_1'^2 - r_{C_3}\mu_{r-3}'\mu_1'^3 + - - - + (-1)^r\mu_1'^r$$

Putting r = 2, 3, 4 respectively, we get

$$\mu_2 = \mu_2' - 2_{C_1}\mu_1'\,\mu_1' + 2_{C_2}\mu_1'^2\mu_0'$$

$$\mu_2 = \mu_2' - 2\mu_1'\,\mu_1' + \mu_1'^2 \qquad \because \mu_0' = 1$$

$$= \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3_{C_1}\mu_2' + 3_{C_2}\mu_1'^2\mu_1' - 3_{C_3}\mu_1'^3\mu_0'$$

$$= \mu_3' - 3\mu_2'\mu_1' + 3(\mu_1')^2\mu_1' - (\mu_1')^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 3(\mu_1')^3 - (\mu_1')^3$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3$$

$$\mu_4 = \mu_4' - 4_{C_1}\mu_3'\mu_1' + 4_{C_2}\mu_2'\mu_1'^2 - 4_{C_3}\mu_1'^3\mu_1' + 4_{C_4}\mu_1'^4\mu_0' = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

5.3.4.5 Pearson's β and γ coefficients

Karl Pearson gave the following four coefficients calculated from the moments about mean and defined them as follows:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \beta_2 = \frac{\mu_4}{\mu_2^2} \quad \text{and } \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} \text{ and } \gamma_2 = \beta_2 - 3$$

These coefficients are pure numbers independent of units of measurement and as such can be conveniently used for comparative studies.

In terms of these coefficients, the coefficient of skewness is given exactly by

$$\text{Skewness} = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

The importance of these coefficients lies in the fact that they give some idea about the shape of the curve obtained from the frequency distribution. For symmetrical distribution, all the moments of odd order about the mean vanish or (are zero) and therefore $\mu_3 = 0$ and hence $\beta_1 = 0$. Thus, β_1 gives a measure of departure from symmetry i.e. of skewness. Thus, the sign of skewness depends upon μ_3 .If μ_3 is positive we get positive skewness and if μ_3 is negative, we get negative skewness.

5.4 Kurtosis

The expression Kurtosis is used to describe the peakedness of a curve. Kurtosis is a Greek word means 'bulginess'. In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of Kurtosis of a distribution is measured relative to the peakedness of normal curve. If we know the measures of central tendency, dispersion and skewness, we cannot still form a complete idea about the distribution as is clear from the figure 5.4.

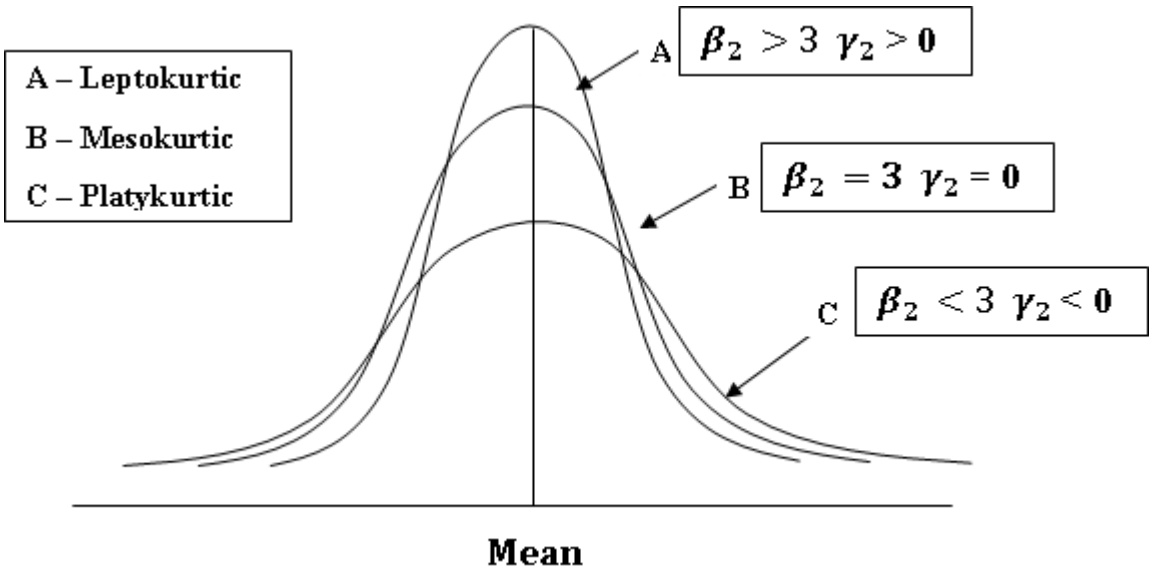


Fig. 5.4 Type of kurtosis

All the three curves are symmetrical about mean and have same variation (range). In order to identify a distribution completely we need one more measure which Prof. Karl Pearson called 'convexity of the curve' or 'kurtosis'. It is measured by β_2 and γ_2 given as under

$\beta_2 = {}^{\mu_4}/_{\mu_2{}^2}$, and $\gamma_2 = \beta_2 - 3$

Curve of type B which is neither flat nor peaked is known as normal curve and shape of its hump is accepted as a standard one. Curves with humps of the form of normal curve are said to have normal kurtosis and are termed

as **Mesokurtic** ($\beta_2 = 3, \gamma_2 = 0$).The curve of type A , which are more peaked than the normal curve are known as **Leptokurtic** ($\beta_2 > 3, \gamma_2 > 0$) and are said to lack kurtosis or to have negative kurtosis. Curves of type C , which are flatter than the normal curve are called **Platykurtic** ($\beta_2 < 3, \gamma_2 < 0$) and they are said to posses kurtosis in excess or have positive kurtosis.

Example 1: Find different measures of skewness and kurtosis taking data given in example 1 of Lesson 3, using different methods.

Solution: Prepare the following table to calculate different measures of skewness and kurtosis using the values of Mean (M) = 1910, Median (M_d) = 1890.8696, Mode (M_o) = 1866.3636, Variance (σ^2) = 29500, Q1 = 1772.1053 and Q_3 = 2030 as calculated earlier.

Calculation of moments about an arbitrary constant 2080

Class Interval	Mid-value (X_i)	frequency (f_i)	$X_i - A$	$f_i(X_i - A)$	$f_i(X_i - A)^2$	$f_i(X_i - A)^3$	$f_i(X_i - A)^4$
1630-1730	1680	17	-400	-6800	2720000	-1088000000	435200000000
1730-1830	1780	19	-300	-5700	1710000	-513000000	153900000000
1830-1930	1880	23	-200	-4600	920000	-184000000	36800000000
1930-2030	1980	16	-100	-1600	160000	-16000000	1600000000
2030-2130	2080	14	0	0	0	0	0
2130-2230	2180	7	100	700	70000	7000000	700000000
2230-2330	2280	2	200	400	80000	16000000	3200000000
2330-2430	2380	2	300	600	180000	54000000	16200000000
Total		100	-400	-17000	5840000	-1724000000	647600000000

First moment about the point A=2080

$$\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^1 = \frac{-17000}{100} = -170 \Rightarrow \bar{X} = A + \mu_1' \quad = 2080 - 170 = 1910$$

Second moment about the point A

$$\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^2 = \frac{5840000}{100} = 58400$$

Third moment about the point A

$$\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^3 = \frac{-1724000000}{100} = -17240000$$

Fourth moment about the point A

$$\mu_4' = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - A)^4 = \frac{647600000000}{100} = 6476000000$$

Compute the central moments using raw moments as follows:

$\mu_2 = \mu_2' - \mu_1'^2$=58400-(-170)^2 =29500

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$= (-17240000)-3(58400)(-170)+2(-170)^3=2718000

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$=6476000000-4(-17240000)(-170)+6(58400)(-170)^2-3(-170)^4

= 2373730000

Calculation of moments about mean

Class Interval	Mid-value (X_i)	frequency (f_i)	$X_i - \bar{X}$	$f_i(X_i - \bar{X})$	$f_i(X_i - \bar{X})^2$	$f_i(X_i - \bar{X})^3$	$f_i(X_i - \bar{X})^4$
1630-1730	1680	17	-230	-3910	899300	-206839000	47572970000
1730-1830	1780	19	-130	-2470	321100	-41743000	5426590000
1830-1930	1880	23	-30	-690	20700	-621000	18630000
1930-2030	1980	16	70	1120	78400	5488000	384160000

2030-2130	2080	14	170	2380	404600	68782000	11692940000
2130-2230	2180	7	270	1890	510300	137781000	37200870000
2230-2330	2280	2	370	740	273800	101306000	37483220000
2330-2430	2380	2	470	940	441800	207646000	97593620000
Total		100		0	2950000	271800000	237373000000

First moment about mean :

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \overline{X}) = \frac{1}{100}(0) = 0$$

Second moment about mean:

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \overline{X})^2 = \frac{1}{100}(2950000) = 29500.$$

which is equal to variance as calculated in example 2.

Third moment about mean:

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \overline{X})^3 = \frac{1}{100}(271800000) = 2718000$$

Fourth moment about mean:

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i(X_i - \overline{X})^4 = \frac{1}{100}(237373000000) = 2373730000$$

Karl Pearson's Coefficient of skewness (S_K):

$$S_K = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{M - M_0}{\sigma} == \frac{1910 - 1866.3636}{\sqrt{29500}} = 0.2541$$

Bowley's Coefficient of skewness (S_K):

$$S_K = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1} = \frac{2030 + 1772.1053 - 2x1890.8696}{2030 - 1772.1053} = 0.0789$$

From the above calculation the coefficient of skewness and kurtosis can be calculated as under:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(2718000)^2}{(29500)^3} = 0.2878 \text{ and } \gamma_1 = \sqrt{\beta_1} = \sqrt{0.2878} = 0.5364$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2373730000}{(29500)^2} = 2.7276 \text{ and } \gamma_2 = \beta_2 - 3 = -0.2724$$

$$S_K = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} = 0.5277$$

Hence this frequency distribution is positively skewed and platykurtic in nature.

Lesson 6

ELEMENTARY NOTIONS OF PROBABILITY

6.1 Introduction

If an experiment is repeated under essentially homogenous and identical conditions then we may come across two types of situations viz., (i) the result or outcome is unique or certain (ii) the result or outcome may be one of the several possible outcomes. The phenomena covered by the former situation (i) is known as 'deterministic' or 'predictable' while the phenomena covered by situation (ii) is known as 'Unpredictable' or 'Probabilistic'. By a deterministic phenomenon we mean that the result can be predicted with certainty e.g. Boyle's law stating that PV = RT = Constant provided temperature remains the same; Newton's first law of motion stating that v = u + at, where u is the initial velocity and a is the acceleration; Ohm's law stating that C = E / R, where C is the flow of current, E is the potential difference between two ends of the conductor and R is the resistance. On the other hand by the probabilistic phenomena we mean that result is not sure e.g. in tossing a coin we are not sure as to whether head or tail will be obtained. Similarly, if any electronic equipment has worked for certain number of hours, nothing can be said about its further performance as to it may fail to function any moment. In such situations, the word 'probability' has relevance. In day to day life, we all make use of the word 'probability', but generally people have no definite idea about the meaning of probability. For example, we often hear or talk phrases like "the fat content of milk sample obtained from buffalo is likely to be 5.5 percent", "the daily milk yield of this cow is likely to be more than 15 kg", "there is a chance that it may rain today" or "India may win this match" or "milk production in India is likely to be 200 million tonnes by 2030". In all the above statements, the terms probably, likely, chance etc. convey the same meaning i.e. the events are not certain to take place. In other words, there is involved an element of uncertainty or chance in all these cases. A numerical measure of uncertainty is provided by the theory of probability.' The theory of probability came into existence when problems of games were referred to mathematicians like B. Pascal, P. Fermat, James Bernoulli, De-Moivre, Karl Pearson, Laplace and others. Later, the classical theory of probability was given by R.A. Fisher. Von-Mises introduced the empirical approach to the theory of probability through the notion of sample space. The idea of axiomatic approach was originated by A. Kolmogorov. But in modern times, it has acquired a great importance in decision making.

6.2 Basic Concept

Before the definition of the word probability is given, it is necessary to define the following basic concepts and terms widely used in its study:

6.2.1 Random experiment

An experiment is said to be a random experiment if when conducted repeatedly under essentially homogeneous conditions, the result is not unique but may be anyone of the various possible outcomes. In other words an experiment whose outcomes can't be predicted in advance is called a random experiment. For instance, if a fair coin is tossed three times, it is possible to enumerate all the possible eight sequences of head (H) and tail (T). But it is not possible to predict which sequence will occur at any time.

6.2.2 Sample space

The set of all possible outcomes of a random experiment is known as the sample space and is denoted by S. Each conceivable outcome of a random experiment under consideration is called a sample point. The totality of all conceivable sample points is called a sample space for example sample space of a trial conducted by tossing of three coins is {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}. In the above experiment it is simple to note that anyone sequence of H and/or T is a sample point whereas all the possible eight sample points constitute the sample space.

6.2.3 Trial and event

Performing a random experiment is called a trial and outcome or combinations of outcomes are termed as 'Events' or 'Cases'. Any subset of the sample space is an event. In other words, the set of sample points which satisfy certain requirement(s) is called an event. For example, if a coin is tossed repeatedly, the result is not unique. Tossing of coin is a random experiment and getting a head or tail is an event.

6.2.4 Exhaustive events

It is defined as total number of all possible outcomes of any trial. In other words, if all the possible outcomes of an experiment are taken into consideration, then such events are called exhaustive events e.g. when a coin is tossed three times there are eight exhaustive events and when two dice are thrown then exhaustive events are 36.

6.2.5 Favourable events

The numbers of outcomes of a random experiment which result in the happening of an event are termed as the cases favourable to the event. For example in three tossing of a coin, the cases favourable to the event that there are exactly two heads is 3, viz. HTH, HHT and THH and for getting at least two heads is 4 , viz. HTH,THH, HHT and HHH.

6.2.6 Mutually exclusive events

Two or more events are said to be mutually exclusive if the happening of one of them prevents or precludes the happening of the all others in the same experiment. Two events E_1 and E_2 are said to be mutually exclusive when they can't happen simultaneously in a single trial. In other words if there is no sample point in E_1 which is common to the sample point in E_2 i.e. $E_1 \cap E_2 = \phi$, the events E_1 and E_2 are said to be mutually exclusive. In tossing a die, the events 1, 2, 3, 4, 5 and 6 are mutually exclusive events because all the six events cannot happen simultaneously in a single trial. If it shows 3, then the event of getting 3 precludes the event of getting 1, 2, 4, 5, 6 at the same time.

6.2.7 Complementary event

The complement of an event A, means non-occurrence of an event A and is denoted by $\bar{A}$ or A^c. $A^c/\bar{A}$ contains those points of the sample space which do not belong to A. In tossing a coin, occurrence of Head (H) and Tail (T) are complementary events. In tossing of a die, occurrence of an even number (2, 4, 6) and odd number (1, 3, 5) are complementary events.

6.2.8 Independent events

Events are said to be independent of each other if happening of any one of them is not affected by and does not affect the happening of any one of others. In other words two or more events are said to be independent if the happening (or non happening) of any one does not depend on the happening or non-happening of any other, otherwise they are said to be dependent. For example

a) In tossing an unbiased coin, the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.

b) If we draw a card from a pack of well-shuffled cards and replace it before the second card is drawn, the result of second draw is independent of the first draw. But if the first card drawn is not replaced then the second draw is dependent on the first draw.

c) A bag contains balls of two different colours say red and white. The two balls are drawn successively. First a ball is drawn from one bag and replaced after noting its colour. Let us suppose that it is white and is denoted by event E_1. Another ball is drawn from the same bag E_2. The result of the second draw is independent of the first draw. Hence the events E_1 and E_2 are independent.

6.2.9 Equally likely cases

The events are said to be equally likely if the chance of happening of each event is equal or same. In other words, cases are said to be equally likely when one does not occur more often than the others e.g. if a die is rolled, any face is as likely to come up as any other face. Hence, the six outcomes 1, 2, 3, 4, 5 or 6 appearing up are equally likely.

6.2.10 Simple (Elementary) events

An event which contains only a single sample point is called an elementary event or simple event e.g. in tossing a die, getting a number 5 is called a simple event.

6.2.11 Compound events

When two or more events occur in connection with each other, their simultaneous occurrence is called a compound event. Joint occurrence of two or more events is called a compound event. An event is termed compound if it represents two or more simple events e.g. if a bag contains 4 white and 3 black balls. If we are required to find a chance in which 3 balls drawn are all white is a simple event. However if we are required to find out the chance of drawing 3 white and then 2 black balls, we are dealing with a compound event because it is made up of two events.

6.3 Definition of Probability

The chance of happening of an event when expressed quantitatively is called probability. The probability is defined in the following three different ways:

- Classical, Mathematical or ‘a Priori’ definition.
- Empirical, Relative or Statistical definition
- Axiomatic definition

6.3.1 Classical or Mathematical definition of probability

This is the oldest and simplest definition of probability. This definition is based on the assumption that the outcomes or results of an experiment are equally likely and mutually exclusive. According to James Bernoulli who was the first man to obtain a quantitative measure of uncertainty "If a random experiment results in N exhaustive, mutually exclusive and equally likely cases out of which m are favourable to the happening of an event A", then probability of occurrence of A, usually denoted by P(A) is given by

$$P(A) = \frac{\text{favourable number of cases}}{\text{Exhaustive numberof cases}} = \frac{m}{N}$$

Example 1. Two identical symmetric dice are thrown. Find the probability of obtaining a total score of 8.

The total number of possible outcomes is 6x6=36. There are 5 sample points (2, 6), (3, 5) (4, 4), (5, 3), (6, 2), which are favourable to the event A of getting a total score of 8. Hence, the required probability is 5/36.

Properties:

- The number of cases favourable to the complimentary event $\bar{A}$ i.e. non-happening of event A are (N-m)and by definition of probability of non–occurrence of A is given by:

  $$P(\bar{A}) = \frac{\text{favourable number of cases to } \bar{A}}{\text{Exhaustive numberof cases}} = \frac{N-m}{N} = 1 - \frac{m}{N} = 1 - P(A)$$

  $P(A) + P(\bar{A}) = 1$

- Since m and N are non-negative integers, P(A)≥0. Further, since the favourable number of cases to A are always less than total number of cases N, i.e. m ≤ N, we have P (A) ≤1. Hence the probability of any event is a number lying between 0 and 1 i.e. 0≤ P (A) ≤1. If P(A)=0 then this event is said to be impossible event. If P (A) =1, then A is called a certain event.

The above definition of probability is widely used, but it cannot be applied under the following situations:

- If it is not possible to enumerate all the possible outcomes for an experiment.
- If the sample points (outcomes) are not mutually independent.
- If the total number of outcomes is infinite.
- If each and every outcome is not equally likely.

It is clear that the above drawbacks of a classical approach restrict its use in practical problems. Yet this is still widely used for problems concerning the tossing of coin(s), throwing of dice, game of cards and selection of balls of different colours from the bag etc.

The probability by classical approach cannot be discovered in the cases where situations like an electric bulb will fuse before it is used for 100 hours, a patient will die if operated for an ailment, a student will fail in a particular examination, a rail compartment in which you are travelling will catch fire, a fan will fall on you while sitting under fan etc. under such circumstances another definition can be used.

6.3.2 Statistical definition of probability

If an experiment is performed repeatedly under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event occurs to the number of trials, as the number of trials becomes indefinitely large, is called the probability of happening of the event, assuming that the limit is finite and unique. Let an event A occurs m times in N repetitions of a random experiment. Then the ratio m/N gives the relative frequency of the event A. When N becomes sufficiently large, it is called the probability of A.

$$P(A) = \lim_{N \to \infty} \frac{m}{N}$$

The above definition of probability involves a concept which has long term consequences. This approach was initiated by Von Mises. Moreover N is not equal to infinity. Thus, in this case, the probability is the limit of relative frequency. Whether such a limit always exists, is not definite. Hence statistical definition of probability is also not very sound.

The two definitions of probability are apparently different. In the 'a prior' definition, it is the relative frequency of favourable cases to the total number of cases. Since in the relative frequency approach, the probability is obtained objectively by repetitive empirical observations, hence it is known as empirical probability. The empirical definition provides validity to the classical theory of probability.

6.3.3 Axiomatic approach to probability

The modern theory of probability is based on the axiomatic approach introduced by the Russian Mathematician A.N. Kolmogorov in 1930's. The axiomatic definition of probability includes both the classical and empirical definition of probability and at the same time is free from their drawbacks. It is based on certain properties or postulates, commonly known as axiom, which are defined from these axioms alone the entire theory is developed by logic of deduction. It is defined as given a sample space of a random experiment, the probability of the occurrence of any event A is defined as asset function P(A)satisfying the following axioms

a) P (A) is defined, is real and non-negative i.e. P (A) > 0.
b) The probability of entire sample space is one i.e. P(S) = 1.
c) If A_1, A_2, ---,A_n are mutually exclusive events, then the probability of the occurrence of either A_1or A_2, ---or A_n denoted by $P(A_1 \cup A_2 \cup \text{---} \cup A_n)=P(A_1)+P(A_2)+\text{---}+P(A_n)$.

The above axioms are known as axioms of positiveness, certainty and unity respectively.
Probability in this approach is defined as, let S be the sample space of a random experiment with large number of sample points N i.e. n(S) =N. Let the number of occurrences (sample points) favourable to the event A be denoted by n(A) . Then the probability of an event A is equal to

$$P(A) = \frac{n(A)}{n(S)} = \frac{n(A)}{N}$$

6.4 Calculation of probability of an event

The probability of an event can be calculated by the following methods:
Method I: Find the total number of exhaustive cases (N). Thereafter, obtain the number of favourable cases to the event i.e. m. Divide the number of favourable cases by the total number of equally likely cases. This will give the probability of an event. The following example will illustrate this.

Example 2. Two dice are tossed. Find the probability that the sum of dots on the faces that turn up is a) 8 and b) 11.

Solution: When two dice are tossed total number of possible outcomes =36

a) Number of outcomes to get a sum of 8 are (6, 2) (5, 3), (4, 4), (3, 5) and (2, 6) i.e. the number cases favourable to this event is equal to 5. Hence, probability of getting a sum of 8 when two dice are thrown $=\frac{5}{36}$

b) Number of outcomes to get a sum of 11 are (6, 5) and (5, 6) i.e. the number cases favourable to this event is equal to 2. Hence, probability of getting a sum of 11 when two dice are thrown $=\frac{2}{36}$

Example 3: From a herd containing 5 Karan Fries and 4 Sahiwal cows, a cow is selected at random. What is the probability that it is a Sahiwal Cow?

Solution: Total number of cows in the herd =5+4=9

Number of Sahiwal cows =4

Probability of getting a Sahiwal cow $=\frac{4}{9}$

Method II: The Fundamental Principle or the Fundamental Rule of Counting:

If one operation can be performed in m different ways and another operation can be performed in n different ways, then the two operations when associated together can be performed in m x n ways.

Method III: Use of Permutation and Combination in Theory of Probability:

Permutation:

The word permutation in simple language means arrangement. A permutation denoted by P is an arrangement of a set of objects in a definite order:

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

Combination:

The concept of combination is very useful in understanding theory of probability. It is not always possible that the number of cases favourable to the happening of an event is easily determined. In such cases the concept of combination is used. The different selections that can be made out of a given set of things taking some or all of them at a time are called combinations. The combinations of n things, taking r at a time is denoted by $^{n}C_{r}$, symbolically

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

Example 4: From a pack of 52 cards, two cards are drawn at random. Find the probability that one is a king and the other is queen.

Solution: Two cards can be drawn from 52 cards in $^{52}C_{2}$ ways.

There are 4 kings and 4 queens in a pack of cards. A king can be drawn in $^{4}C_{1}$ ways and a queen can be drawn in $^{4}C_{1}$ways.

The probability of getting a king card and the other a queen card$=\frac{^{4}C_{1}x^{4}C_{1}}{^{52}C_{2}}=\frac{4x4x2}{52x51}=\frac{8}{663}$

Example 5: A herd consists of 5 Karan Fries and 6 Sahiwal cows. Two cows are chosen at random from this herd. What is the probability that

a) One is Karan Fries and other is Sahiwal.

b) Both are Sahiwal.

Solution: Total number of cows in the herd =5+6=11

Two cows can be chosen from 11 cows in

$$^{11}C_2 = \frac{11!}{2!(11-2)!} = \frac{11x10}{2} = 55 \text{ ways}$$

a) In the herd there are 5 Karan Fries and 6 Sahiwal cows. One Karan Fries out of 5 can be chosen in

$^{5}C_1 = \frac{5!}{1!(5-1)!} = 5$ ways

One Sahiwal cow out of 6 can be chosen in $^{6}C_1 = \frac{6!}{1!(6-1)!} = 6$ ways

The probability of choosing one Karan Fries and one Sahiwal cow is

$$\frac{^{6}C_1 \times {}^{5}C_1}{^{11}C_2} = \frac{6 \times 5}{55} = \frac{6}{11}$$

b) Both the cows are Sahiwal can be chosen in $^{6}C_2 = \frac{6!}{2!(6-2)!} = \frac{6x5}{2} = 15$ ways

The probability of choosing both the Sahiwal cows is

$$\frac{^{6}C_2}{^{11}C_2} = \frac{6x5}{11x10} = \frac{3}{11}$$

Lesson 7

ADDITION THEOREM OF PROBABILITY

7.1 Introduction

In the last lesson we have studied the probability of an event in a random experiment as well as axiomatic approach formulated by Russian Mathematician A.N. Kolmogorov and observed that probability as a function of outcomes of an experiment. By now you know that the probability P(A) of an event A associated with a discrete sample space is the sum of the probabilities assigned to the sample points in A as discussed in axiomatic approach of probability. Moreover, in practical problems, writing down the elements of S and counting the number of cases favourable to a given event often become very tedious. However in such situations the computation of probabilities can be facilitated to a great extent by fundamental theorem of addition. In this lesson we will learn Addition Theorem of Probability to find probability of occurrence for simultaneous trials under two conditions when events are mutually exclusive and when they are not mutually exclusive.

7.2 List of Symbols

$A \cup B$: An event which represents the happening of at least one of the events A and B i.e. either A occurs or B occurs or both A and B occur. This is also denoted as A or B

$A \cap B$: An event which represents the simultaneous happening of both A and B i.e. A and B.

$\bar{A}$: A does not happen.

$\bar{A} \cap \bar{B}$: Neither A nor B happens i.e. none of A and B happens.

$\bar{A} \cap B$: A does not happen but B happens.

$(A \cap \bar{B}) \cup (\bar{A} \cap B)$: Exactly one of the two events A and B happens.

7.3 Addition Theorem for Mutually Exclusive Events

Statement: If A and B are two mutually exclusive events, then the probability of occurrence of either A or B is the sum of the individual probabilities of A and B. Symbolically

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$$

Proof : Let N be the total number exhaustive and equally likely cases of an experiment. Let m_1 and m_2 be the number of cases favourable to the happening of events A and B respectively. Then

$P(A) = \frac{n(A)}{n(S)} = \frac{m_1}{N}$ and $P(B) = \frac{n(B)}{n(S)} = \frac{m_2}{N}$.

Since the events A and B are mutually exclusive, the total number of events favourable to either A or B i.e. $n(A \cup B) = m_1 + m_2$, then

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N} = \frac{m_1 + m_2}{N} = \frac{m_1}{N} + \frac{m_2}{N} = P(A) + P(B)$$

Generalisation: This theorem can be extended to three or more mutually exclusive events. The probability of occurrence of any one of the several mutually exclusive events A, B and C is equal to the sum of their individual probabilities given by

$$P(A \cup B \cup C) = P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

In general, if $A_1, A_2, \ldots, A_n$ are mutually exclusive events then

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

i.e., the probability of occurrence of any one of the n mutually disjoint events $A_1, A_2, \ldots, A_n$ is equal to the sum of their individual probabilities.

The following examples illustrate the application of this theorem:

Example 1: A card is drawn at random from a pack of 52 cards. Find the probability that the drawn card is either a club or an ace of diamond.

Solution : Let A : Event of drawing a card of club and

B: Event of drawing an ace of diamond

The probability of drawing a card of club $P(A) = \frac{13}{52}$

The probability of drawing an ace of diamond $P(B) = \frac{1}{52}$

Since the events are mutually exclusive, the probability of the drawn card being a club or an ace of diamond is:

$$P(A \cup B) = P(A) + P(B) = \frac{13}{52} + \frac{1}{52} = \frac{14}{52} = \frac{7}{26}$$

Example 2: A herd contains 30 cows numbered from 1 to 30. One cow is selected at random. Find the probability that number of the selected cow is a multiple of 5 or 8.

Solution: Let A be the event of number being a multiple of 5 within 30 and B be the event of number being a multiple of 8 within 30.

Favourable cases for event A are {5, 10, 15, 20, 25, 30}

Similarly favourable cases for event B are {8, 16, 24}

The probability of the number being a multiple of 5 within 30 is P(A)= $\frac{6}{30}$

The probability of the number being a multiple of 8 within 30 is P(B)= $\frac{3}{30}$

Since A and B are mutually exclusive, the probability that number of the cow is a multiple of 5 or 8 is:

$$P(A \cup B) = P(A) + P(B) = \frac{6}{30} + \frac{3}{30} = \frac{9}{30} = \frac{3}{10}$$

7.4 Addition Theorem for Non-mutually Exclusive Events

The addition theorem discussed above is not applicable when the events are not mutually exclusive. For example, if one card is drawn at random from a pack of 52 cards then in order to find the probability of either a spade or a king card, it cannot be calculated by simply adding the probabilities of spade and king card because the events are not mutually exclusive as there is one card which is a spade as well as a king. Thus, the events are not mutually exclusive, therefore, the addition theorem is modified as:

Statement: If A and B are not mutually exclusive events, the probability of the occurrence of either A or B or both is equal to the probability that event A occurs, plus the probability that event B occurs minus the probability of occurrence of the events common to both A and B. In other words the probability of occurrence of at least one of them is given by

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

Proof: Let us suppose that a random experiment results in a sample space S with N sample points (exhaustive number of cases). Then by definition

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N}$$

where $n(A \cup B)$ is the number of occurrences (sample points) favourable to the event $(A \cup B)$

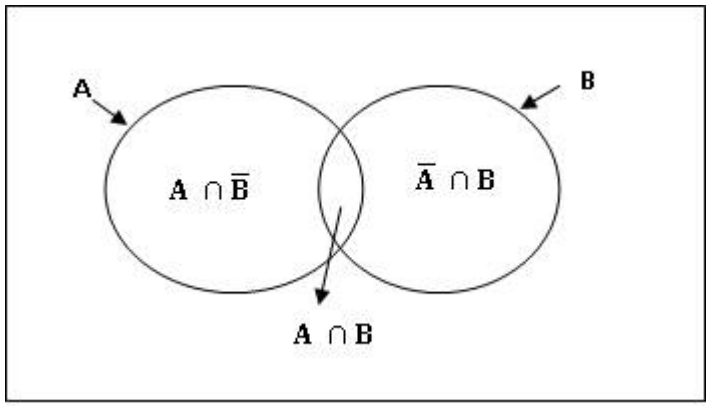


Fig. 7.1

From the above diagram, we get:

$$\begin{aligned} P(A \cup B) &= \frac{[n(A) - n(A \cap B)] + n(A \cap B) + [n(B) - n(A \cap B)]}{N} \\ &= \frac{n(A) + n(B) - n(A \cap B)}{N} \\ &= \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N} \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Generalisation The above theorem can be extended to three or more events. If A, B and C are not mutually exclusive events then the probability of the occurrence of at least one of them is given by

$$P(A \cup B \cup C) = \frac{n(A \cup B \cup C)}{N}$$

$$= \frac{1}{N}[n(A) + n(B) + n(C) - n(A \cap B) - n(B \cap C) - n(A \cap C) + n(A \cap B \cap C)]$$

$$= \frac{n(A)}{N} + \frac{n(B)}{N} + \frac{n(C)}{N} - \frac{n(A \cap B)}{N} - \frac{n(B \cap C)}{N} - \frac{n(A \cap C)}{N} + \frac{n(A \cap B \cap C)}{N}$$

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

If n mutually exclusive events A_1, A_2 … A_n are exhaustive also, so that probability of at least one of the n events to materialize is a certainty then the probability of the constituent events.

$$P(A_1 + A_2 + \ldots\ldots + A_n) = 1$$

$$P(A_1) + P(A_2) + \ldots\ldots + P(A_n) = 1$$

The following examples illustrate the application of this theorem:

Example 3. A card is drawn at random from a pack of 52 cards. Find the probability that the drawn card is either a spade or a king.

Solution: Let A: Event of drawing a card of spade and
B: Event of drawing a king card

The probability of drawing a card of spade $P(A) = \frac{13}{52}$

The probability of drawing a king card $P(B) = \frac{4}{52}$

Because one of the kings is a spade card also therefore, these events are not mutually exclusive. The probability of drawing a king of spade is $P(A \cap B) = \frac{1}{52}$

So, the probability of the drawing a spade or king card is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

Example 4. A herd contains 30 cows numbered from 1 to 30. One cow is selected at random. Find the probability that the number of the selected cow is a multiple of 5 or 6.

Solution: Let A be the event of number being a multiple of 5 within 30 and B be the event of number being a multiple of 6 within 30.

Favourable cases for event A are {5, 10, 15, 20, 25, 30}
Similarly favourable cases for event B are {6, 12, 18, 24, 30}

The probability of the number being a multiple of 5 within 30 is P(A)= $\frac{6}{30}$

The probability of the number being a multiple of 6within 30 is P(B)= $\frac{5}{30}$

Since 30 is a multiple of 5 as well as 6, therefore the events are not mutually exclusive

$$P(A \cap B) = P(A \text{ and } B) = \frac{1}{30}$$

The probability that the number of the selected cow is a multiple of 5 or 6 is :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{6}{30} + \frac{5}{30} - \frac{1}{30} = \frac{10}{30} = \frac{1}{10}$$

Example 5. A number was drawn at random from the number 1 to 50. What is the probability that it will be a multiple of 2 or 3 or 10?

Solution: Probability of getting a multiple of 2: $P(A) = \frac{25}{50}$

Probability of getting a multiple of 3: $P(B) = \frac{16}{50}$

Probability of getting a multiple of 10: $P(C) = \frac{5}{50}$

Common Probability of getting a multiple of 2 and 3: $P(A \cap B) = \frac{8}{50}$

Common Probability of getting a multiple of 3 and 10: $P(B \cap C) = \frac{1}{50}$

Common Probability of getting a multiple of 2 and 10: $P(A \cap C) = \frac{5}{50}$

Common Probability of getting a multiple of 2, 3 and 10: $P(A \cap B \cap C) = \frac{1}{50}$

Probability that it is a multiple of 2 or 3 or 10:

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

$$= \frac{25}{50} + \frac{16}{50} + \frac{5}{50} - \frac{8}{50} - \frac{1}{50} - \frac{5}{50} + \frac{1}{50} = \frac{33}{50}$$

Lesson 8

MULTIPLICATION THEOREM OF PROBABILITY

8.1 Introduction

In the previous lesson we have studied the addition theorem of probability for mutually exclusive events as well as for those events which are not mutually exclusive. In many situations we want to find the probability of simultaneous occurrence of two or more events. Sometimes the information is available that an event A has occurred and one is required to find the probability of occurrence of another event B utilizing the information about event A. Such a probability is known as conditional probability. In this lesson we shall discuss the important concept of conditional probability of an event which will be helpful in understanding the concept of multiplication theorem of probability as well as independence of events.

8.2 Multiplication Theorem for Independent Events

Statement: This theorem states that if two events A and B are independent then the probability that both of them will occur is equal to the product of their individual probabilities.

Symbolically P (AB) = $P(A \cap B) = P(A \text{ and } B)$ = P (A).P (B)

Proof

If an event A can happen in n_1 ways out of which a_1 are favourable and the event B can happen in n_2 ways out of which a_2 are favourable, we can combine each favourable event in the first with each favourable event in the second case. Thus, the total number of favourable cases is $a_1 \times a_2$. Similarly, the total number of possible cases is $n_1 \times n_2$. Then by definition the probability of happening of both the independent events is

$$P(A \cap B) = P(A \text{ and } B) = \frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2} = P(A) \times P(B)$$

$$\text{as } P(A) = \frac{a_1}{n_1} \;\&\; P(B) = \frac{a_2}{n_2}$$

Similarly we can extend the theorem to three events

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

The following examples illustrate the application of this theorem:

Example 1. From a pack of 52 cards, two cards are drawn at random one after the other with replacement. What is the probability that both cards are kings?

Solution: The probability of drawing a king P (A) = $\frac{4}{52}$

The probability of drawing again the king after replacement P (B) = $\frac{4}{52}$

Since the two events are independent, the probability of drawing two kings is:

$$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B) = \frac{4}{52} \times \frac{4}{52} = \frac{1}{169}$$

Example 2. A bag contains 4 red balls, 3 white balls and 5 black balls. Two balls are drawn one after the other with replacement. Find the probability that first is red and the second is black.

Solution: Probability of red ball in the first draw = $\frac{4}{12}$

The probability of a black ball in the second draw = $\frac{5}{12}$

Since the events are independent, the probability that first is red and the second is black will be:

$$P(1R).P(1B) = \frac{4}{12} \times \frac{5}{12} = \frac{20}{144} = \frac{10}{72} = \frac{5}{36}$$

8.3 Conditional Probability

In many situations we have the information about the occurrence of an event A and are required to find out the probability of the occurrence of another event B. Two events A and B are said to be dependent when event A can occur only when event B is known to have occurred (or vice versa). The probability attached to such an event is called the conditional probability and is denoted by P (A|B) or in other words, probability of A given that B has occurred. For example, if we want to find the probability of an ace of spade if we know that card drawn from a pack of cards is black. Let us consider another problem relating to dairy plant. There are two lots of full cream pouches A and B, each containing some defective pouches. A coin is tossed and if it turns up with its head upside lot A is selected and if it turns with tail up, lot B is selected. In this problem we are interested to know the probability of the event that a milk pouch selected from the lot obtained in this manner is defective.

Definition: If two events A and B are dependent, then the conditional probability of B given that event A has occurred is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{if } P(A) \geq 0$$

Let us consider the experiment of throwing of a die once. The sample space of this experiment is {1, 2, 3, 4, 5, 6}.

Let E_1: 'an even number shows up' and E_2: 'multiple of 3 shows up'.

Then E_1: {2, 4, 6} and E_2: {3,6}. Hence, $P(E_1) = \frac{3}{6} = \frac{1}{2}$ and $P(E_2) = \frac{2}{6} = \frac{1}{3}$

In order to find the probability of occurrence of E_2 when it is given that E_1 has occurred we know that in a single throw of die '2'or '4'or '6'has come up. Out of these only '6' is favourable to E_2. So the probability of occurrence of E_2 when it is given that E_1 has occurred is equal to 1/3. This probability of E_2 when E_1 has occurred is written as $P(E_2|E_1)$. Here we find that $P(E_2|E_1) = P(E_2)$. Let us consider the event E_3: 'a number greater than 3 shows up' then E_3:{4,5,6} and $P(E_3)=3/6=1/2$ Out of 2,4 and 6, two numbers namely 4 and 6 are favourable to E_3. Therefore, $P(E_3|E_1)=2/3$. The events of the type E_1 and E_2 are called independent events as the occurrence or non-occurrence of E_1 does not affect the probability of occurrence or non-occurrence of E_2. The events $E_1$and E_3 are not independent.

8.4 Multiplication Theorem of Probability for Dependent Events

Statement: The probability of simultaneous happening of two events A and B is given by:

$$P(A \cap B) = P(A).P(B|A);\ P(A) \neq 0$$
$$P(B \cap A) = P(B).P(A|B);\ P(B) \neq 0$$

where $P(B|A)$ is the conditional probability of happening of B under the condition that A has happened and $P(A|B)$ is the conditional probability of happening of A under the condition that B has happened.

Proof:

Let A and B be the events associated with the sample space S of a random experiment with exhaustive number of outcomes (sample points) N, i.e., $n(S) = N$. Then by definition

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} \qquad (1)$$

For the conditional event $A|B$ (i.e., the happening of A under the condition that B has happened), the favourable outcomes (sample points) must be out of the sample points of B. In other words, for the event $A|B$, the sample space is B and hence

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

Similarly, we have

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

On multiplying and dividing equation (1) by n(A), we get

$$P(A \cap B) = \frac{n(A)}{n(S)} \times \frac{n(A \cap B)}{n(A)}$$

$$= P(A).P(B|A)$$

Also

$$P(A \cap B) = \frac{n(B)}{n(S)} \times \frac{n(A \cap B)}{n(B)}$$

$$= P(B).P(A|B)$$

Generalisation

The multiplication theorem of probability can be extended to more than two events. Thus, for three events A_1, A_2 and A_3 we have

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$$

For n events $A_1, A_2, \ldots, A_n$ we have

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \times \ldots \times P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

In particular, if $A_1, A_2, \ldots, A_n$ are independent events then

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n)$$

i.e., the probability of the simultaneous happening of n independent events is equal to the product of their individual probabilities.

The following examples illustrate the application of this theorem

Example 3. A bag contains 5 white and 8 red balls. Two successive drawings of 3 balls are made such that (a) the balls are replaced before the second drawing, and (b) the balls are not replaced before the second draw. Find the probability that the first drawing will give 3 white and the second 3 red balls in each case.

Solution:

(a) When balls are replaced.

Total balls in the bag = 8 + 5 = 13

3 balls can be drawn out of total of 13 balls in ${}^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in ${}^{5}C_3$ ways.

Probability of 3 white balls = $P(3W) = \frac{{}^{5}C_3}{{}^{13}C_3} = \frac{10}{286}$

Since the balls are replaced after the first draw so again there are 13 balls in the bag 3 red balls can be drawn out of 8 red balls in ${}^{8}C_3$ ways.

Probability of 3 red balls = $P(3R) = \frac{{}^{8}C_3}{{}^{13}C_3} = \frac{56}{286}$

Since the events are independent, the required probability is:

$$P(3W \text{ and } 3R) = P(3W) \times P(3R) = \frac{{}^{5}C_3}{{}^{13}C_3} \times \frac{{}^{8}C_3}{{}^{13}C_3} = \frac{10}{286} \times \frac{56}{286} = \frac{140}{20,449}$$

(b) When the balls are not replaced before second draw

Total balls in the bag = 8 + 5 = 13

3 balls can be drawn out of 13 balls in ${}^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in ${}^{5}C_3$ ways.

The probability of drawing 3 white balls = $P(3W) = \frac{{}^{5}C_3}{{}^{13}C_3}$

After the first draw, balls left are 10, 3 balls can be drawn out of 10 balls in ${}^{10}C_3$ ways.

3 red balls can be drawn out of 8 balls in ${}^{8}C_3$ ways. Probability of drawing 3 red balls = $\frac{{}^{8}C_3}{{}^{10}C_3}$

Since both the events are dependent, the required probability is:

$$P(3W \text{ and } 3R) = P(3W)xP(3R|3W) = \frac{^5C_3}{^{13}C_3} \times \frac{^8C_3}{^{10}C_3} = \frac{5}{143} \times \frac{7}{15} = \frac{7}{429}$$

Example 4. A bag contains 5 white and 3 red balls and four balls are successively drawn and are not replaced. What is the chance that (i) white and red balls appear alternatively and (ii) red and white balls appear alternatively?

Solution (i) The probability of drawing a white ball = $\frac{5}{8}$

The probability of drawing a red ball = $\frac{3}{7}$

The probability of drawing a white ball = $\frac{4}{6}$ and the probability of drawing a red ball = $\frac{2}{5}$

Since the events are dependent, therefore the required probability is:

$$P(W \text{ and } R \text{ and } W \text{ and } R) = P(W \cap R \cap W \cap R)$$

$$= P(W)xP(R|W)xP(W|WR)xP(R|WRW) \quad = \frac{5}{8} \times \frac{3}{7} \times \frac{4}{6} \times \frac{2}{5} = \frac{1}{14}$$

(ii) The probability of drawing a red ball = $\frac{3}{8}$ and the probability of drawing a white ball = $\frac{5}{7}$

The probability of drawing a red ball = $\frac{2}{6}$ and the probability of drawing a white ball = $\frac{4}{5}$

Since the events are dependent, therefore the required probability is:

$$P(R \text{ and } W \text{ and } R \text{ and } W) = P(R \cap W \cap R \cap W)$$

$$= P(R)xP(W|R)xP(R|RW)xP(W|RWR) \quad = \frac{3}{8} \times \frac{5}{7} \times \frac{2}{6} \times \frac{4}{5} = \frac{1}{14}$$

Example 5. A coin is tossed once. If it shows head, it is tossed again and if it shows tail, then a dice is tossed. Let E_1 be the event: 'the first throw of coin shows tail' and E_2 be the event: 'the dice shows a number greater than 4'. Find $P(E_2|E_1)$

Solution: In this problem the random experiment was carried out in two stages

a) A coin is tossed b) If the first stage shows a head, coin is tossed again and if it shows a tail, a dice is thrown. The sample space is {HH,HT,T1,T2,T3,T4,T5,T6}.

P (HH)=P(HT)=1/2x1/2=1/4

and P(T1)=P(T2)=P(T3)=P(T4)=P(T5)=P(T6)=1/6X1/2=1/12

E_1: the event the first throw of coin shows tail={T1,T2,T3,T4,T5,T6}.

E_2 : the event the dice shows a number greater than 4={ T5,T6}

$P(E_1)=6x1/12=1/2$ and $P(E_2)=2x1/12=1/6$

Also $P(E_1 \cap E_2) = 2x\frac{1}{12} = \frac{1}{6}$

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} = \frac{1/6}{1/2} = \frac{1}{3}$$

8.5 Combined use of Addition and Multiplication Theorem

In some problems in probability both addition and multiplication theorems are used simultaneously. The following examples illustrate the combined use of addition and multiplication theorems.

Example 6. A bag contains 5 white and 4 black balls. A ball is drawn from this bag and is replaced and then second draw of a ball is made. What is the probability that two balls are of different colours.

Solution: There are two possibilities

i) First ball is white and the second ball drawn is black.
ii) First ball is black and the second ball drawn is white.

Since the events are independent, so by using multiplication theorem we have

i) Probability of drawing First ball white and the second ball black $=\frac{5}{9}x\frac{4}{9} = \frac{20}{81}$

ii) Probability of drawing First ball black and the second ball white $=\frac{4}{9}x\frac{5}{9} = \frac{20}{81}$

Since these probabilities are mutually exclusive, by using addition theorem

Probability that two balls are of different colours $=\frac{20}{81} + \frac{20}{81} = \frac{40}{81}$

Example 7. A can hit a target 4 times in 5 shots. B 3 times in 4 shots and C twice in 3 shots. They fire a volley. What is the probability that

a) Two shots hit the target
b) At least two shots hit the target.

Solution:

a) There are three possibilities that two shots hit the target

i) A and B hit the target and C could not hit i.e. $A \cap B \cap \bar{C}$
ii) B and C hit the target and A could not hit i.e. $\bar{A} \cap B \cap C$
iii) A and C hit the target and B could not hit i.e. $A \cap \bar{B} \cap C$

Since the events are independent, so using multiplication theorem

i) $P(A \cap B \cap \bar{C}) = \frac{4}{5}x\frac{3}{4}x\left(1-\frac{2}{3}\right) = \frac{4}{5}x\frac{3}{4}x\frac{1}{3} = \frac{12}{60}$

ii) $P(\bar{A} \cap B \cap C) = \left(1-\frac{4}{5}\right)x\frac{3}{4}\ x\frac{2}{3} = \frac{1}{5}x\frac{3}{4}x\frac{2}{3} = \frac{6}{60}$

iii) $P(A \cap \bar{B} \cap C) = \frac{4}{5}x\left(1-\frac{3}{4}\right)\ x\frac{2}{3} = \frac{4}{5}x\frac{1}{4}x\frac{2}{3} = \frac{8}{60}$

Since the above three possibilities are mutually exclusive, so by using addition theorem, we have

$$P(\text{Two shots hit the target}) = P(A \cap B \cap \bar{C}) + P(\bar{A} \cap B \cap C) + P(A \cap \bar{B} \cap C)$$

$$= \frac{12}{60} + \frac{6}{60} + \frac{8}{60} = \frac{26}{60} = \frac{13}{30}$$

b) There are two possibilities that at least two shots hit the target

i) Two could hit the target

ii) Three could hit the target i.e. A and B and C could hit i.e. $A \cap B \cap C$

Since the events are independent, so by using multiplication theorem

i) P(Two could hit the target)= $\frac{13}{30}$

ii) $P(A \cap B \cap C) = \frac{4}{5} x \frac{3}{4} x \frac{2}{3} = \frac{24}{60}$

Since the above two possibilities are mutually exclusive, so by addition theorem

P (At least two shots hit the target)= $\frac{26}{60} + \frac{24}{60} = \frac{50}{60} = \frac{5}{6}$

Example 8. Three small sized Herds A, B and C consist of 3 cows and 1 buffalo, 2 cows and 2 buffaloes, 1 cow and 3 buffaloes, respectively. Find the probability of selecting one cow and two buffaloes from three Herds.

Solution: This particular example illustrates the combined application of Additive and Multiplicative theorem of probability. There are three Herds whose composition is given as under

Herd	Type of animals	
	Cow	Buffalo
A	3	1
B	2	2
C	1	3

Probability of selection of one cow and two buffaloes from Herds A, B and C can take place in following three possible ways.

(i) One cow from Herd A and one buffalo each from Herd B and C respectively
(ii) One cow from Herd B and one buffalo each from Herd A and C respectively
(iii) One cow from Herd C and one buffalo each from Herd A and B respectively

The probability for Situation (i) is $\frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{18}{64}$

The probability for Situation (ii) is $\frac{2}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{6}{64}$

The probability for Situation (iii) is $\frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{2}{64}$

Since either of the above situations can occur, therefore probability of getting selection either through situation (i) or situation (ii) or situation (iii) will be sum of the probabilities of three different situations which is equal to

$$\frac{18}{64} + \frac{6}{64} + \frac{2}{64} = \frac{26}{64} = \frac{13}{32}$$

Lesson 9

RANDOM VARIABLE AND ITS PROBABILITY DISTRIBUTION

9.1 Introduction

In this lesson we will study the important concept of random variable and its probability distribution. It has been a general notion that if an experiment or a trial is conducted under identical conditions, values so obtained would be similar. But this is not always true. The observations are recorded about a factor or character under study e.g. fat, SNF, TS, moisture content etc. in a dairy product. These can take different values, and the factor or character is termed as variable. These observations vary even though the experiment has been conducted under identical conditions. Therefore, we have a set of results or outcomes (Sample points) of a random experiment. A rule that assigns a real number to each outcome (Sample points) is called a random variable. A random variable is a rule or function that assigns numerical values to observations or measurements. It is called random variable because the number that is assigned to the observation is a numerical value which varies randomly. A random variable takes a numerical value with some probability. Hence,, a list of values of a random variable together with their corresponding probabilities of occurrence is termed as probability distribution. We shall also find mean and variance of a probability distribution.

9.2 Random Variable

A random variable (r.v.) is defined as a real number X connected with the outcome of a random experiment E. For example, if E consists of three tosses of a coin, we may consider random variable X which denotes the number of heads (0, 1, 2 or 3)

Outcome :	HHH	HTH	THH	THH	HTT	THT	TTH	TTT
Value of X :	3	2	2	2	1	1	1	0

Thus, to every outcome there corresponds a real number X(w). Since the points of the sample space corresponds to outcomes, this means that a real number, which we denote by X(w), is defined for each w∈S and let us denote them by w_1, w_2, ---,w_8 i.e. $X(w_1)=3$, $X(w_2)=2$---, $X(w_8)=0$. Thus,, we define a random variable as a real valued function whose domain is the sample space associated with a random experiment and range is the real line. Generally it is denoted by capital letters X,Y, Z, --- etc.

9.2.1 Discrete random variable

If a random variable X assumes only a finite or countable set of values, it is called a discrete random variable. In other words, a real valued function defined on a discrete sample space is called a discrete random variable. In case of discrete random variable we usually talk of values at a point. Generally it represents counted data. For example, number of defective milk pouches in a milk plant, number of accidents taking place in milk plant, number of students in a class, number of milch animals in a herd etc.

9.2.2 Continuous random variable

A random variable is said to be continuous if it can assume infinite and uncountable set of values. A continuous random variable is in which different values cannot be put in one to one correspondence with a set of positive integers. For example, weight of calf at the age of six months might take any possible value in the interval of 160 kg

to 260 kg, say 189 kg or 189.4356 kg; likewise milk yield of cows in a herd etc. In case of continuous random variable we usually talk of values in a particular interval. Continuous random variables represent measured data.

9.3 Probability Distribution of a Random Variable

The concept of probability distribution is equivalent to the frequency distribution. It depicts how total probability of one is distributed among various values which a random variable can take.

9.3.1 Probability Mass Function (Discrete Random Variable)

Suppose X is a one-dimensional discrete random variable taking at most a countable infinite number of values X_1, X_2, ……..with each possible outcome X; we associate a number $p_i = P(X=X_i) = P(X_i)$ called the probability of X_i, the numbers $P(X_i)$, i=1, 2, ……must satisfy the following conditions:

a) $p_i = P(X=X_i) = P(X_i) \geq 0$ i.e. p_i's are all non-negative.

b) $\sum_{i=1}^{\infty} P(X_i) = 1$ i.e. the total probability is one.

This function $p_i = P(X=X_i)$ or p(x) is called the probability function or probability mass function (p.m.f.) of the random variable X and set of all possible ordered pairs {x, p(x)} is called the probability distribution of the random variable X.

9.3.2 Probability density function (Continuous random variable)

In case of a continuous random variable X, we talk of probability in an interval. If f(x) is a continuous function of X, f(x) dx gives the probability that the random variable X, takes value in a small interval of magnitude dx i.e. $\left(x-\frac{1}{2}dx\right)$ *and* $\left(x+\frac{1}{2}dx\right)$, then f(x) is called the probability density function (p.d.f.) of a random variable X. It is also known as frequency function because it also gives the proportion of units lying in the interval $\left(x-\frac{1}{2}dx\right)$ and $\left(x+\frac{1}{2}dx\right)$. If x has range $[\alpha,\beta]$, $f(x) \geq 0\ \forall\ x \in [\alpha,\beta]$ and $\int_{\alpha}^{\beta} f(x)dx = 1$.

The following examples illustrate the concept of probability mass function

Example 1. Two cards are drawn one by one without replacement from a well shuffled pack of 52 cards. Find the probability distribution of the number of aces.

Solution: Let 'X' be the random variable, which is the number of aces.

Here X takes values 0, 1, 2

$$P(X=0) = \frac{48}{52} \times \frac{47}{51} = \frac{188}{221}$$

$$P(X=1) = 2\left(\frac{4}{52} \times \frac{48}{51}\right) = 2\left(\frac{1}{13} \times \frac{16}{17}\right) = 2\left(\frac{16}{221}\right) = \frac{32}{221}$$

$$P(X=2) = \frac{4}{52} \times \frac{3}{51} = \frac{1}{13} \times \frac{1}{17} = \frac{1}{221}$$

Hence, the probability distribution is

X:	0	1	2
P (X):	$\frac{188}{221}$	$\frac{32}{221}$	$\frac{1}{221}$

Example 2. Four defective milk pouches are accidently mixed with sixteen good ones and by looking at them it is not possible to differentiate between them. Three milk pouches are drawn at random from the lot. Find the

probability distribution of X, the number of defective milk pouches.

Solution: Let 'X' be the random variable, which is the number of defective milk pouches.

Here X takes values 0, 1, 2, 3.

Total number of milk pouches = 4 + 16 = 20

Number of defective milk pouches = 4

$$\therefore P(X=0) = P(\text{No defective milk pouche}) = \frac{^{16}C_3}{^{20}C_3} = \frac{16\times15\times14}{20\times19\times18} = \frac{140}{285}$$

$$P(X=1) = P(\text{one defective milk pouche}) = \frac{^{4}C_1\times{}^{16}C_2}{^{20}C_3} = \frac{4\times16\times15\times6}{2\times20\times19\times18} = \frac{120}{285}$$

$$P(X=2) = P(\text{two defective milk pouches}) = \frac{^{4}C_2\times{}^{16}C_1}{^{20}C_3} = \frac{4\times3\times16\times6}{2\times20\times19\times18} = \frac{24}{285}$$

$$P(X=3) = P(\text{three defective milk pouches}) = \frac{^{4}C_3}{^{20}C_3} = \frac{4\times3\times2}{20\times19\times18} = \frac{1}{285}$$

Hence,, the probability distribution is

X:	0	1	2	3
P (X):	$\frac{140}{285}$	$\frac{120}{285}$	$\frac{24}{285}$	$\frac{1}{285}$

9.4 Mean and Variance of a Random variable

Let X denotes the random variable which assumes values $x_1,x_2,\text{---},x_n$ with corresponding probabilities $p_1,p_2,\text{---},p_n$ .Then the probability distribution be as follow:

X:	x_1	x_2	……	x_n
P(X):	p_1	p_2	……	p_n

Then

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1$$

The mean (μ) of the above probability distribution is defined as:

$$\mu = \frac{p_1x_1 + p_2x_2 + \cdots + p_nx_n}{p_1 + p_2 + \cdots + p_n} = \frac{\sum p_ix_i}{\sum p_i} = \sum p_ix_i$$

The variance (σ^2) is defined as:

$$\sigma^2 = \sum (x_i - \mu)^2 p_i = \sum (x_i^2 + \mu^2 - 2x_i\mu)p_i = \sum x_i^2p_i + \mu^2\sum p_i - 2\mu\sum x_ip_i$$

$$= \sum x_i^2p_i + \mu^2(1) - 2\mu(\mu) = \sum x_i^2p_i - \mu^2 = \sum x_i^2p_i - \left(\sum p_ix_i\right)^2$$

Mean of a random variable X is also known as expected value and is denoted by E(X)

$$E(X) = \mu = p_1x_1 + p_2x_2 + \cdots + p_nx_n = \sum p_ix_i$$

$$\text{Variance } (\sigma^2) = E(X^2) - (E(X))^2$$

Example 3. Find the mean and variance of the number of heads in two tosses of a coin.

Solution:

Let X denotes the number of heads obtained in two tosses of a coin. Thus, X takes the values 0, 1, 2.

Now p, the probability of getting a head $= \frac{1}{2}$ and q, the probability of not getting a head $= 1 - \frac{1}{2} = \frac{1}{2}$

$$\therefore P(X = 0) = q \times q = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4},$$

$$P(X = 1) = p \times q + q \times p = 2\left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{2},$$

$$P(X = 2) = p \times p = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4},$$

Thus, we have:

x_i	p_i	$p_i x_i$	x_i^2	$p_i x_i^2$
0	1/4	0	0	0
1	1/2	1/2	1	1/2
2	1/4	2/4	4	1
Total		1		3/2

Hence, the mean

$\mu = \sum x_i p_i = 0 + \frac{1}{2} + \frac{2}{4} = 1$

and the variance

$\sigma^2 = \sum p_i x_i^2 - \mu^2 = \frac{3}{2} - (1)^2 = \frac{3}{2} - 1 = \frac{1}{2}$

Example 4. A die is tossed twice. Getting a number greater than 4 is considered a success. Find the variance of the probability distribution of the number of success.

Solution: Here p, probability of a number greater than $4 = \frac{2}{6} = \frac{1}{3}$ and q, probability of a number not greater than $4 = 1 - \frac{1}{3} = \frac{2}{3}$

$$P(X = 0) = q \times q = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9},$$

$$P(X = 1) = p \times q + q \times p = \frac{1}{3} \times \frac{2}{3} + \frac{2}{3} \times \frac{1}{3} = \frac{4}{9},$$

$$P(X = 2) = p \times p = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9},$$

Thus, we have:

x_i	p_i	$p_i x_i$	x_i^2	$p_i x_i^2$
0	4/9	0	0	0
1	4/9	4/9	1	4/9
2	1/9	2/9	4	4/9
Total		6/9		8/9

Hence, the mean $\mu = \sum p_i x_i = \frac{6}{9} = \frac{2}{3}$

and the variance $\sigma^2 = \sum p_i x_i^2 - \mu^2 = \frac{8}{9} - \left(\frac{6}{9}\right)^2 = \frac{8}{9} - \frac{36}{81} = \frac{72-36}{81} = \frac{36}{81} = \frac{4}{9}$

9.5 Mathematical Expectation

Let X denotes a discrete random variable which assumes values $x_1, x_2, \ldots, x_n$ with corresponding probabilities $p_1, p_2, \ldots, p_n$ where $p_1 + p_2 + \ldots + p_n = 1$, the mathematical expectation of X or simply the expectation of X, denoted by E(X) is defined as:

$$E(X) = \mu = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n = \sum_{i=1}^{n} p_i x_i, \text{where} \sum_{i=1}^{n} p_i = 1$$

i.e. it is sum of the product of different possible values of x and the corresponding probabilities. Hence, mathematical expectation of a random variable is equal to its arithmetic mean.

9.5.1 Some results on expectation

- E (c) = c , where c is a constant.
- E (cX) =c E(X) , where c is a constant .
- E (aX+b) =a E(X)+b , where a and b are constants.
- **Addition law of expectation:** If X and Y are random variables then E(X+Y)=E(X)+E(Y) i.e. expected value of the sum of two random variables is equal to sum of their expected values.
- **Multiplication law of expectation:** If X and Y are independent random variables then E(X.Y)=E(X).E(Y) i.e. expected value of the product of two random variables is equal to product of their expected values.
- **Variance in terms of expectation:**

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

Lesson 10

BINOMIAL DISTRIBUTION

10.1 Introduction

In the first module we have studied the empirical or observed or experimental frequency distribution in which the actual data were collected and tabulated in the form of a frequency distribution. In the present lesson we will study theoretical frequency distribution which are not obtained by actual observations or experiments but distributed according to some definite probability law which can be expressed mathematically. Such distributions as are expected on the basis of previous experience or theoretical considerations are known as theoretical distribution or probability distribution. Thus, the theoretical frequency distribution are not based on actual observations but are mathematically deducted under certain assumptions. In this lesson we shall study one of the most popular discrete distributions, the origin of which lies in Bernoullian trials.

10.2 Binomial Distribution

Binomial distribution is a discrete probability distribution. This distribution was discovered by a Swiss Mathematician James Bernoulli (1654-1705). A Bernoullian trial is an experiment having only two possible outcomes i.e. success or failure. In other words the result of the trial are dichotomous e.g. in tossing of a coin either head or tail, the sex of a calf can be either male or female, a manufactured milk product or an engineering equipment or spare part will be either defective or non defective etc. This distribution can be used under the following conditions:

a) The random experiment is performed repeatedly a finite and fixed number of times i.e. n, the number of trials is finite and fixed.
b) The outcome of a trial results in the dichotomous classification of events i.e. each trial must result in two mutually exclusive outcomes –success or failure.
c) Probability of success (or failure) remains same in each trial i.e. in each trail the probability of success, denoted by p remains constant. q=1-p, is then termed as the probability of failure (non-occurrence).
d) Trials are independent i.e. the outcome of any trial does not affect the outcomes of the subsequent trials.

10.3 Probability Mass Function of Binomial Distribution

Statement

If X denotes the number of successes in n trials satisfying the above conditions, then X is a random variable which can take values 0,1,2,---,n i.e. no success, one success, two successes,---, or all the n successes. The general expression for the probability of r successes is given by:

$$P(r) = P(X = r) = {}^{n}C_r\, p^r\, q^{n-r} \quad \text{for } r=0,1,2,\ldots\ldots,n$$

Proof : By the theorem of compound probability, the probability that r trials are success and the remaining (n-r) are failures in a sequence of n trials in a specified order say S,F,S,F,S,,---,S is given by

$$\begin{aligned} P(S \cap F \cap S \cap F \cap F \cap --- \cap S) &= P(S)P(F)P(S)P(F)P(F) --- P(S) \\ &= p.q.p.q.q --- p \\ &= (pxpxpx --- r \text{ times })x(qxqxq— (n-r) \text{times}) = p^r q^{(n-r)} \end{aligned}$$

But we are interested in any r trials being successes and since r trials can be chosen out of n trials in ${}^nC_r$ (mutually exclusive) ways. Therefore, by the theorem of total probability, the chance P (r) of r successes in a series of n independent trials is given by

$$P(r) = {}^nC_r\, p^r\, q^{n-r} \qquad 0 \le r \le n$$

r can take only positive integer values.

Thus, the chance variate i.e. the number of successes, can take the values 0,1,2,…..,r,……..,n with corresponding probabilities $q^n, {}^nC_1\, p\, q^{n-1}, \ldots\ldots, {}^nC_r\, p^r\, q^{n-r}, \ldots\ldots, p^n$

- The probability distribution of the number of successes so obtained is called the binomial probability distribution for the obvious reason that the probabilities are the various terms of the binomial expansion of $(q+p)^n$.
- The sum of probabilities

$$\sum_{r=0}^{n} p(r) = p(0) + p(1) + p(2) + - - - p(r)$$

$$= q^n + {}^nC_1 pq^{n-1} + - - - + {}^nC_r\, p^r q^{n-r} + - - - + p^n = (q+p)^n$$

$$= 1$$

- The expression for P (X = r) is known as probability mass function of the Binomial distribution with parameter n and p. The random variable X following this probability law is called binomial variate with parmeter n and p denoted as X~B(n,p).Hence binomial distribution can be completely determined if n and p are known .

Example 1. It is known that 40 percent cows affected by tuberculosis die every year. Six cows are admitted to a veterinary hospital suffering from tuberculosis. What is the probability that

(i) three cows will die.
(ii) at least five cows will die
(iii) all cows will be cured
(iv) no cow will be saved.

Solution

In this exercise we have p = 0.4 , q = 1- 0.40 = 0.6 and n=6

In binomial distribution we have $P(r) = {}^nC_r \,.\, p^r \,.\, q^{n-r}$

(i) Prob. [Three cows will die] = P[r = 3] = P(3) = ${}^6C_3 \,.\, (0.4)^3\, (0.6)^3$

$$P(3) = \frac{6!}{3!\,3!}\,(0.4)^3\,(0.6)^3 = 20(0.4)^3(0.6)^3 = 0.2765$$

(ii) Prob. (at least five cows will die) = P(5) + P(6)

$$= {}^6C_5\,(0.4)^5\,(0.6)^1 + {}^6C_6\,(0.4)^6\,(0.6)^0$$

$$= 6\,(0.4)^5\,(0.6)^1 + (0.4)^6 = 0.0369 + 0.0041 = 0.0410$$

(iii) Prob. (all cows will be cured) =1 – P (no cow will die) = 1- P(0)

$$=1 - {}^{6}C_0\,(0.4)^0\,(0.6)^6 = 1 - (0.6)^6 = 1 - 0.0467 = 0.9533$$

(iv) Prob. (no cow will be saved) = P (all cows will die) = P(6)

$$= {}^{6}C_6\,(0.4)^6\,(0.6)^0 = (0.4)^6 = 0.0041$$

Example 2. Ten consumers were asked to state their preferences between two types of ice-cream. Assuming that there is no difference between two types of ice–cream, calculate the probability that

a) 3 or less consumers will prefer ice-cream A.

b) 7 or more consumers will prefer ice-cream B.

Solution: In this exercise p = 0.5, q = 0.5 and n = 10

a) Prob. [Three or less consumers will prefer Ice Cream A] = P(0) + P(1) + P(2) + P(3)

$$= {}^{10}C_0\,(0.5)^0\,(0.5)^{10} + {}^{10}C_1\,(0.5)^1(0.5)^9 + {}^{10}C_2\,(0.5)^2\,(0.5)^8 + {}^{10}C_3\,(0.5)^3(0.5)^7$$

$$= (0.5)^{10}({}^{10}C_0 + {}^{10}C_1 + {}^{10}C_2 + {}^{10}C_3)$$

$$= 0.00098\,(1 + 10 + 45 + 120) = 0.00098\,(176) = 0.1725$$

b) Prob. [Seven or more consumers will prefer Ice Cream B] = P(7) + P(8) + P(9) + P(10)

$$= {}^{10}C_7\,(0.5)^7\,(0.5)^3 + {}^{10}C_8\,(0.5)^8(0.5)^2 + {}^{10}C_9\,(0.5)^9\,(0.5)^1 + {}^{10}C_{10}\,(0.5)^{10}$$

$$= (120+45+10+1)\,(0.5)^{10} = 0.1725$$

10.4 Example of Binomial distribution

- The problem relating to tossing of a coin or throwing of dice or drawing cards from a pack of cards with replacement.
- The problems relating to distribution for the preference for a dairy product among families.
- The problem relating to distribution of coli-forms in sterilized milk.
- The problem relating to distribution of number of stables in farm households.
- The problem relating to distribution of number of lactations completed by the milch animals in a dairy farm.

10.5 Properties of Binomial Distribution

i) Mean of binomial distribution is np.

Proof: First raw moment

$$\mu_1' = E(r) = \sum_{r=0}^{n} r\; n_{C_r} p^r q^{n-r} = \sum_{r=0}^{n} r \frac{n!}{r!(n-r)!} p p^{r-1}\, q^{n-r}$$

$$= \sum_{r=0}^{n} \frac{r.n.(n-1)!}{r.(r-1)![(r-1)-(r-1)]!} p p^{r-1} q^{(n-1)-(r-1)} = np \sum_{r=0}^{n} n-1_{C_{r-1}}\, p^{r-1} q^{(n-1)-(r-1)}$$

$$\sum_{r=0}^{n} np\; n-1_{C_{r-1}} p^{r-1} q^{(n-1)-(r-1)} =$$

$$np \sum_{r=0}^{n} n-1_{C_{r-1}} p^{r-1} q^{(n-1)-(r-1)} = np(q+p)^{n-1} = np$$

ii) Variance of binomial distribution is npq

Proof: Second raw moment

$$\mu_2' = E(r^2) = \sum_{r=0}^{n} r^2 \, n_{C_r} p^r \, q^{(n-r)} = \sum_{r=0}^{n} \{r + r(r-1)\} \, n_{C_r} p^r q^{n-r}$$

$$= \sum_{r=0}^{n} r \, n_{C_r} p^r q^{n-r} + \sum_{r=0}^{n} r(r-1) \; n_{C_r} p^r q^{n-r} = np + n(n-1)p^2 = np + n^2p^2 - np^2$$

Variance = $\mu_2 = \mu_2' - (\mu_1')^2 = np + n^2p^2 - np^2 - n^2p^2 = np(1-p) = npq$

For the binomial distribution if mean and variance are known, we can arrive at the frequency distribution and variance is less than mean.

iii) The third and fourth central moment μ_3 and μ_4 can be obtained on the same lines.

$$\mu_3 = npq(q-p)$$

$$\mu_4 = npq[1 + 3(n-2)pq]$$

iv) Pearson's constants β_1 & β_2 as well as γ_1 and γ_2 are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{[npq(1-2p)]^2}{(npq)^3} = \frac{[(1-2p)]^2}{npq} \qquad \gamma_1 = \sqrt{\beta_1} = \frac{(1-2p)}{\sqrt{npq}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{npq[1+3(n-2)pq]}{(npq)^2} = 3 + \frac{1-6pq}{npq}, \quad \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq}$$

γ_1 shows that the binomial distribution is positively skewed if q>p or p<1/2 and it is negatively skewed if q<p or p>1/2 and it is symmetrical if p=q=1/2.The binomial distribution is leptokurtic if pq<1/6 and platykurtic if pq>1/6.

v) Mode of binomial distribution is determined by the value (n+1)p. If this value is an integer equal to k then the distribution is bi-modal, the two modal values being X=k and X=k-1.When this value is not an integer then the distribution has unique mode at X=k_1, the integral part of (n+1)p.

vi) Additive property: If X_1 is B(n_1,p)and X_2 is B(n_2,p) and they are independent then their sum $X_1 + X_2$ is also a binomial variate B(n_{1+} n_2,p).

Example 3. If the mean and variance of a Binomial Distribution are respectively 9 and 6, find the distribution.

Solution: Mean of Binomial Distribution is np and variance is npq

$\therefore np = 9 \; and \; npq = 6$

$Now\ \frac{npq}{np} = \frac{6}{9} \Rightarrow q = \frac{2}{3}$

$\therefore p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$

$\therefore np = 9 \Rightarrow n.\frac{1}{3} = 9 \Rightarrow n = 3 \times 9 = 27$

Hence, the Binomial Distribution is $\left(\frac{2}{3} + \frac{1}{3}\right)^{27}$ i.e. $^{27}C_r(1/3)^r(2/3)^{27-r}$

Example 4. An unbiased dice is thrown 5 times and appearance of face on the dice 2 or 3 is considered as success. Find the probability of (i) exactly one success (ii) at least 4 successes and find mean and variance.

Solution: Here $n = 5, p = \frac{2}{6} = \frac{1}{3}, q = \frac{2}{3}$

$P(X = r) = {}^nC_r p^r q^{n-r}$

$(i) P(\text{exactly one sucess}) = P(X = 1) = {}^5C_1 \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^4 = \frac{80}{243}$

$(ii) P(\text{at least 4 successes}) = P(4 \text{ or } 5) = P(4) + P(5)$

$$= {}^5C_4 \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^1 + {}^5C_5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^0 = 5 \times \frac{1}{81} \times \frac{2}{3} + 1 \times \frac{1}{243} \times 1$$

$$= \frac{10}{243} + \frac{1}{243} = \frac{11}{243}$$

$\text{Mean} = np = 5 \times \frac{1}{3} = \frac{5}{3}$

$\text{Variance} = npq = 5 \times \frac{1}{3} \times \frac{2}{3} = \frac{10}{9}$

Example 5. A binomial variate X satisfies the relation 9 P(X=4)=P(X=2) when n = 6. Find the value of the parameter p.

Solution: Since the binomial probability distribution is

$P(X = r) = {}^nC_r p^r q^{n-r}, r = 0,1,2, \ldots, n$

For $n = 6$, equation becomes

$P(X = r) = {}^6C_r p^r q^{6-r}, r = 0,1,2, \ldots, 6$

Considering the given relation,

9 P(X = 4) = P(X = 2), we have

$$\frac{P(X=4)}{P(X=2)} = \frac{1}{9} \Rightarrow \frac{{}^6C_4 p^4 q^{6-4}}{{}^6C_2 p^2 q^{6-2}} = \frac{1}{9}$$

$$\Rightarrow \frac{15.p^4q^2}{15.p^2q^4} = \frac{1}{9} \Rightarrow \frac{p^2}{q^2} = \frac{1}{9}; \frac{p}{q} = \frac{1}{3}$$

$$\Rightarrow 3p = (1-p) \Rightarrow 4p = 1$$

$$\text{Thus, } p = \frac{1}{4}.$$

10.6 Fitting of Binomial Distribution

Let the n independent trials constitute one experiment and let this experiment be repeated N times. Then we expect r successes to occur N. ${}^nC_r\ p^r\ q^{n-r}$ times. This is called expected frequency of r successes in N experiments and the possible number of successes together with the expected frequencies will constitute binomial (expected) frequency distribution

$$Nxp(r) = Nx{}^nC_r\ p^r q^{n-r}\ ; \quad r=0,1,2,\ldots\ldots,n$$

Putting r=0,1,2,……,n we get the expected or theoretical frequencies of the Binomial distribution , which are given in the following table .

No. of successes (r)	Expected or theoretical Frequencies N.P(r)
0	Nq^n
1	$N{}^nC_1 pq^{n-1}$
2	$N{}^nC_2 pq^{n-2}$
:	:
n	Np^n

Case I: If p the probability of success which is constant for each trial is known , then the expected frequencies can be obtained from the above table.

Case II: If p is not known and if we want to fit a binomial distribution to a given frequency distribution, then first find mean of the given frequency distribution by the formula $m = \frac{\Sigma f_i X_i}{N}$ and equate it to np which is mean of the binomial distribution. Hence, p can be estimated by the relation $m = np \Rightarrow p = \frac{m}{n}$, q = 1-p, with the values of p and q the expected theoretical binomial frequencies can be obtained by using the above table. The expected frequencies can also be computed by using the following recurrence formula

$$\frac{f(r+1)}{f(r)} = Nx\frac{{}^nC_{r+1}\ p^{r+1}q^{n-r-1}}{{}^nC_r\ p^r q^{n-r}} = Nx\left(\frac{n-r}{r+1}\right)\left(\frac{p}{q}\right)$$

The procedure is illustrated through the following example.

Example 6. The following table gives the number of coliforms per ml in thousand pouches of milk:

No of coliforms (X_i)	0	1	2	3	4	5	6	7	8	9	10
No. of pouches (f_i)	2	8	46	116	211	243	208	119	40	7	0

Fit a binomial distribution to the above data.

Solution: In the usual notations we have:

n= 10, N=1000, $\sum f_i X_i$=4971, $\bar{X} = 4.971$, $p = \frac{4.971}{10} = 0.4971, q = 0.5029, \frac{p}{q} = \frac{0.4971}{0.5029}$=0.9985 putting r=0,1,2,3,---,10 in $E(r) = Nx^nC_r\, p^r q^{n-r}$ to get the expected frequencies as given in the following table:

No. of coliforms (X_i)	No. of bottles (f_i)	$f_i X_i$	Expected Frequency E (r)
0	2	0	1.0347
1	8	8	10.2277
2	46	92	45.4939
3	116	348	119.9179
4	211	844	207.4360
5	243	1215	246.0524
6	208	1248	202.6788
7	119	833	114.4808
8	40	320	42.4352
9	7	63	9.3213
10	0	0	0.9214
Total	1000	4971	1000.00

Different expected frequencies are also computed by using recurrence formula

E(0)=1000(0.5029)^{10} =1.0347 ; $E(1) = n\frac{p}{q}E(0) = 10(0.9985)$(1.0347)=10.2277

$E(2) = \frac{9}{2}(0.9985)$(10.2277)=45.4939; $E(3) = \frac{8}{3}(0.9985)$(45.4939)=119.9179;

$E(4) = \frac{7}{4}(0.9985)$(119.9179)= 207.4360; $E(5) = \frac{6}{5}(0.9985)$(207.4360)= 246.0524;

$E(6) = \frac{5}{6}(0.9985)$(246.0524)= 202.6788; $E(7) = \frac{4}{7}(0.9985)$(202.6788)= 114.4808;

$E(8) = \frac{3}{8}(0.9985)$(114.4808)= 42.4352; $E(9) = \frac{2}{9}(0.9985)$(42.4352)= 9.3213;

$E(10) = \frac{1}{10}(0.9985)$(9.3213)= 0.9214

Lesson 11

POISSON DISTRIBUTION

11.1 Introduction

Having discussed the binomial distribution in the preceding lesson, we now turn to Poisson distribution, which is also a discrete probability distribution. Before we know the distribution, it becomes necessary to understand what a Poisson variable is. A variable which can take only one discrete value in an interval of time, however small, is known as Poisson variable. It was given by French Mathematician S. D. Poisson (1781-1840) and hence named after him. When n is very very large $(n \to \infty)$ and p is very very small $(p \to 0)$ then binomial distribution can't be applied. As binomial distribution, Poisson distribution is also one of the most widely used distributions .It is used in quality control to count the number of defective items or in insurance problems to count the number of causalities.

11.2 Poisson Distribution

Poisson distribution is a limiting case of Binomial distribution under the following conditions:

(i) n, the no. of trials is indefinitely large i.e., $n \to \infty$

(ii) p, the constant probability of success for each trial is indefinitely small i.e. $p \to 0$

(iii) np= m (say) is finite. Thus, $p = \frac{m}{n}, q = 1 - \frac{m}{n}$ where m is a positive real number.

Under, the above three conditions the probability mass function of binomial distribution tends to the probability mass function of the Poisson distribution whose definition and derivation given below:

Definition: A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by

$$p(r,m) = P(x = m) = \frac{e^{-m}m^r}{r!}; \qquad r = 0,1,2,\dots\dots\dots,\infty$$

where m is known as the parameter of the distribution .

e = 2.7183(the base of the natural logarithm)

$$e^x = 1 + \frac{X}{1!} + \frac{X^2}{2!} + \frac{X^3}{3!} \dots\dots\dots \infty$$

$$e^{-x} = 1 - \frac{X}{1!} + \frac{X^2}{2!} - \frac{X^3}{3!} + \cdots (-1)^n \frac{X^r}{r!} + \cdots \infty$$

Proof: As $n \to \infty$ and $np = m \Rightarrow p = \frac{m}{n}$ and $q = 1 - \frac{m}{n}$

Probability function of binomial distribution is

$$P(r) = {}^nC_r\, p^r\, q^{n-r} = \frac{n!}{r!(n-r)!}\, p^r\, q^{n-r}$$

$$= \frac{n(n-1)(n-2)---[n-(r-1)]}{r!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{m^r}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)--\left(1-\frac{r-1}{n}\right)x\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-r}$$

Taking limit as $n\to\infty$

$$= \frac{m^r}{r!}(1-0)(1-0)--(1-0)x\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{-r}$$

We know that

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = e^{-m}$$

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^a = 1 \quad \text{a is not a function of n}$$

$$\text{Thus, } P(r) = \frac{m^r}{r!}\frac{e^{-m}.1}{.1} = \frac{e^{-m}m^r}{r!}; r = 0,1,2,---\infty$$

Putting $r = 0,1,2,---$ in above equation, we obtain the probabilities of $r = 0,1,2,----$ successes respectively we get e^{-m}, $\frac{e^{-m}m^1}{1!}$, $\frac{e^{-m}m^2}{2!}$,---, .

Total probability is 1:

$$\sum_{r=0}^{\infty} p(r) = \sum_{r=0}^{\infty} P(x=r) = \sum_{r=0}^{\infty}\frac{e^{-m}m^r}{r!} = \sum_{r=0}^{\infty} P(x=r) = e^{-m} + me^{-m} + \frac{e^{-m}m^2}{2!} + \text{----}$$

$$= e^{-m}\left(1+m+\frac{m^2}{2!}+\frac{m^3}{3!}\pm--\right) = e^{-m}\sum_{r=0}^{\infty}\frac{m^r}{r!} = e^{-m}.e^{m} = 1$$

If we know m, all the probabilities of the Poisson distribution can be obtained, therefore m is the only parameter of the Poisson distribution. The application of this distribution in solving problems is illustrated through following examples.

Example 1. A manufacturer of screws knows that 5% of his product is detective. If he sells his product in a carton of 100 items and guarantees that not more than 10 items will be defective. What is the probability that the carton will fail to meet the guaranteed quality?

Solution:

In this example p = 0.05, n = 100. Therefore, m = n. p = 100 (0.05) = 5

Prob. [That the carton will fail to meet the guaranteed quality] = 1 – Prob. [The carton will meet the guaranteed quality] = Prob. [Not more than 10 items will be defective] = 1 – P [r ≤ 10]

= 1- [P(0) + P(1) + P(2) + P(3)+……+ P(10)]

In case of Poisson distribution $P(r) = \frac{e^{-m}m^r}{r!}$

Therefore, we have $P(r > 10) = 1 - P(r \le 10) = 1 - \left(\frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^1}{1!} + \frac{e^{-5}5^2}{2!} + \cdots + \frac{e^{-10}5^{10}}{10!}\right)$

$$= 1 - e^{-5}\left[1+5+\frac{5^2}{2!}+\frac{5^3}{3!}+\cdots+\frac{5^{10}}{10!}\right] = 1 - 0.9865 = 0.0135$$

Example 2. A milk plant manufacturing ghee pouches; there are small chances of 1/500 for any ghee pouch to be defective. These pouches are supplied in packet of 10. What will be the approximate number of pouches containing no defective, one defective, two defective and three defective ghee pouches in a consignment of 5,000 pouches?

Solution:

In this exercise p = 1/500, n = 10 therefore m = 10/500 = 0.002

Prob. [Packet containing no defective] = $P(0) = \frac{e^{-m} m^0}{0!} = \frac{e^{-0.002}(0.002)^0}{0!} = 0.9802$

Prob. [Packet containing one defective] = $P(1) = \frac{e^{-m} m^{-1}}{1!} = \frac{e^{-0.002}(0.002)^1}{1!} = 0.0196$

Prob. [Packet containing two defective] = $P(2) = \frac{e^{-m} m^2}{2!} = \frac{e^{-0.002}(0.002)^2}{2!} = 0.0002$

Prob. [Packet containing three defective] = $P(3) = \frac{e^{-m} m^3}{3!} = \frac{e^{-0.002}(0.002)^3}{3!} = 0$

Therefore, expected number of pouches containing no defective, one defective, two defective, and three defective ghee pouches in a consignment of 5000 pouches.

E(0) = N P(0) = 5000 (0.9802) = 4901

E(1) = N P(1) = 5000 (0.0196) = 98

E(2) = N P(2) = 5000 (0.0002) = 1

E(3) = N P(3) = 5000 (0) = 0

11.3 Examples of Poisson Distribution

- The number of defective milk pouches per lot.
- The number of deaths of cattle in big dairy farm in one year by a rare disease.
- Number of bottles broken in different months or bottle breakage during different months.
- The number of bacterial colonies in a given culture per unit area of microscope.
- Number of suicides reported in a particular city.
- Number of defective material in a packaging manufactured by a good concern.
- Number of printing mistakes at each page of the book.
- Number of deaths from a disease (not in the form of epidemic) such as heart attack or cancer or snake bite.
- The emission of radioactive (alpha) particles. A small mass of radian contain many millions of atoms.
- Number of fragments received by a surface area 't' from a fragment of atom bomb

11.4 Properties of Poisson Distribution

i) Mean of the Poisson distribution is m

$$\mu_1' = \text{Mean} = \sum_{r=0}^{\infty} r \frac{e^{-m} m^r}{r!} = \sum_{r=0}^{\infty} r \frac{e^{-m} m^r}{(r-1)!} = m \sum_{r=0}^{\infty} \frac{e^{-m} m^{r-1}}{(r-1)!}$$

$$= me^{-m}\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + - - -\right] = m\, e^{-m}e^{m} = m$$

ii) Variance of the Poisson distribution is

$$\text{Variance} = \sum_{r=0}^{\infty} r^2 p(r) - \left(\sum_{r=0}^{\infty} r\, p(r)\right)^2 = \sum_{r=0}^{\infty} r^2 p(r) - (m)^2$$

where,

$$\mu_2' = \sum_{r=0}^{\infty} r^2 \frac{e^{-m}m^r}{r!} = \sum_{r=0}^{\infty} [r + r(r-1)]\frac{e^{-m}m^r}{r!} = \sum_{r=0}^{\infty} r\frac{e^{-m}m^r}{r!} + \sum_{r=0}^{\infty} r(r-1)\frac{e^{-m}m^r}{r!}$$

$$= m + e^{-m}m^2 \sum_{r=0}^{\infty} \frac{m^{r-2}}{(r-2)!} = m + e^{-m}m^2\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + - - -\right]$$

$$= m + e^{-m}m^2 e^{m} = m + m^2$$

$\text{Variance} = \mu_2' - (\mu_1')^2 = m + m^2 - (m)^2 = m$

Hence, for Poisson distribution with parameter m mean is equal to variance.

iii) Third and fourth central moments μ_3 and μ_4

$$\mu_3 = m, \quad \mu_4 = 3m^2 + m$$

iv) Pearson's constants β_1 & β_2 as well as γ_1 and γ_2 are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(m)^2}{(m)^3} = \frac{1}{m}, \quad \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3m^2 + m}{(m)^2} = 3 + \frac{1}{m}, \quad \gamma_2 = \beta_1 - 3 = \frac{1}{m}$$

It may be noted that the first three central moments of the Poisson distribution are identical and are equal to the value of parameter itself namely 'm'. Hence Poisson distribution is always a positively skewed distribution as m>0 as well as leptokurtic. As the value of m increases γ_1 decreases and the thus skewness is reduced for increasing values of m. As $m \to \infty$, γ_1 and γ_2 tend to zero. So we conclude that as $m \to \infty$, the curve of the Poisson distribution tends to be symmetrical curve for large values of m.

v) Mode of Poisson distribution is determined by the value m. If m is an integer then the distribution is bi-modal, the two modal values being X=m and X=m-1.When m is not an integer then the distribution has unique modal value being integral part of m.

vi) Additive property: If X_1 and X_2 are two independent Poisson variate with parameters m_1 and m_2 then their sum $X_1 + X_2$ is also a Poisson variate with parameter $m_1 + m_2$.

Example 3. The mean of the Poisson distribution is 2.25. Find the other constants of the distribution.

Solution: We have $m = 2.25$

$\sigma = \sqrt{m} = \sqrt{2.25} = 1.5$

$\mu_1 = 0$

$\mu_2 = m = 2.25$

$\mu_3 = m = 2.25$

$\mu_4 = m + 3m^2 = 2.25 + 3(2.25)^2 = 2.25 + 15.1875 = 17.4375$

$\beta_1 = \frac{1}{m} = \frac{1}{2.25} = 0.444$

$\beta_2 = 3 + \frac{1}{m} = 3 + 0.444 = 3.444$

$\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}} = \frac{1}{1.5} = 0.67$

$\gamma_2 = \beta_2 - 3 = 3 + \frac{1}{m} - 3 = \frac{1}{2.25} = 0.444$

This curve is positively skewed and leptokurtic.

Example 4. In a Poisson distribution 3P (X = 2) = P (X = 4). Find P (X = 3)

Solution: $p(X = r) = \frac{e^{-m}.m^r}{r!}$

Since 3P(X = 2) = P(X = 4)

$\Rightarrow 3.\frac{e^{-m}.m^2}{2!} = \frac{e^{-m}.m^4}{4!} \Rightarrow 3 = \frac{m^2}{12}$

$\Rightarrow m^2 = 36 \Rightarrow m = 6$ [because Mean is always positive]

Hence, $P(X = 3) = \frac{e^{-m}.m^3}{3!} = \frac{e^{-6}.6^3}{6!} = 36\, e^{-6}$

Example 5. A discrete random variable X follows a Poisson distribution. Find i) P(X ≥ 3) and ii) P (X is at most 2), if it is given E (X) = 3 and $e^{-3} = 0.0498$

Solution: Here m = 3, $e^{-3} = 0.0498$

i) P (X ≥ 3) = 1 - [P (X = 0) + P (X = 1) + P(X = 2)]

$$P(X \geq 3) = 1 - \frac{e^{-m}.m^0}{0!} - \frac{e^{-m}.m}{1!} - \frac{e^{-m}.m^2}{2!} = 1 - e^{-m}\left(1 + m + \frac{m^2}{2}\right)$$

$$= 1 - 0.0498\left(1 + 3 + \frac{9}{2}\right) = 1 - \frac{0.0498 \times 17}{2} = 0.5767$$

ii) $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{e^{-m}.m^0}{0!} - \frac{e^{-m}.m}{1!} - \frac{e^{-m}.m^2}{2!}$

$$= e^{-m}\left(1 + m + \frac{m^2}{2}\right) = 0.0498\left(1 + 3 + \frac{9}{2}\right) = \frac{0.0498 \times 17}{2} = 0.4233$$

11.5 Fitting of Poisson Distribution

If we want to fit a Poisson distribution to a given frequency distribution, we compute mean of the given frequency distribution by the formula $\bar{X} = \frac{\Sigma f_i X_i}{N}$ and equate it to m which is mean of the Poisson distribution. Once m is known, the various probabilities can be calculated by the formula

$p(r,m) = P(x = r) = \frac{e^{-m}m^r}{r!}$; $r = 0,1,2, \ldots\ldots\ldots, \infty$

If N is the total observed frequency, then the expected or theoretical frequencies of the Poisson distribution are given in the following table.

No. of successes (r)	Expected or theoretical Frequencies. N.P(r)
0	Ne^{-m}
1	$N\, m\, e^{-m}$
2	$N\frac{m^2}{2!}e^{-m}$
:	:
r	$N\frac{e^{-m}m^r}{r!}$
:	:

The expected frequencies can also be computed by using the following recurrence formula

$$\frac{f(r+1)}{f(r)} = Nx\frac{\frac{e^{-m}m^{r+1}}{(r+1)!}}{\frac{e^{-m}m^r}{r!}} = Nx\frac{m}{(r+1)}$$

The procedure is illustrated through following examples.

Example 6. The following table gives the number of lactations completed by 1000 cows of Tharparkar breed:

No. of lactations (X_i)	0	1	2	3	4	5	6	7	8	9	10
No. of cows	300	205	155	126	90	47	35	18	13	8	3

Fit a Poisson distribution to the above data.

Solution: In the usual notations we have:

N=1000, $\sum f_i X_i$=2030, $\bar{X} = 2.03 = m$

putting r=0,1,2,3,---,10 in $E(r) = Nx\frac{e^{-m}m^r}{r!}$ we get the expected as given in the following table:

No. of lactations completed (X_i)	No. of cows (f_i)	$f_i X_i$	Expected Frequency E(r)
0	300	0	131.3355
1	205	205	266.6111
2	155	310	270.6103
3	126	378	183.1130
4	90	360	92.9298
5	47	235	37.7295
6	35	210	12.7652
7	18	126	3.7019
8	13	104	0.9394
9	8	72	0.2119

10	3	30	0.0430
Total	1000	2030	999.9905

Different expected frequencies are also computed by using recurrence formula

E(0)=1000x$e^{-2.03}$ =131.3355;

E(1)=2.03x131.3355=266.6111;

$E(2) = \frac{2.03}{2}(266.6111)$ =270.6103;

$E(3) = \frac{2.03}{3}(270.6103)$=183.1130;

$E(4) = \frac{2.03}{4}(183.1130)$=92.9298;

$E(5) = \frac{2.03}{5}(92.9298)$=37.7295;

$E(6) = \frac{2.03}{6}(37.7295)$=12.7652;

$E(7) = \frac{2.03}{7}(12.7652)$=3.7019;

$E(8) = \frac{2.03}{8}(3.7019) =$ 0.9394;

$E(9) = \frac{2.03}{9}(0.9394)$= 0.2119;

$E(10) = \frac{2.03}{10}(0.2119)$= 0.0430

Lesson 12

NORMAL DISTRIBUTION

12.1 Introduction

In preceding two lessons discrete probability distributions viz., Binomial and Poisson distribution were discussed. We shall now take up another distribution, which is an important continuous probability distribution known as the normal distribution. Normal distribution is probably the most important and widely used theoretical distribution. Normal distribution unlike the Binomial and Poisson is a continuous probability distribution. It has been observed that a vast number of variables arising in studies of agricultural and dairying, social, psychological and economic phenomena tend to follow normal distribution. The normal distribution was first discovered by French Mathematician Abraham De-Moivre in 1733, who obtained this continuous distribution as a limiting case of the Binomial distribution. But it was later rediscovered and applied by Laplace and Karl Gauss. It is also known as Gaussian distribution after the name of Karl Friedrich Gauss.

12.2 Definition of Normal Distribution

A continuous random variable X is said to have a normal distribution with parameters μ (mean) and σ (standard deviation), if its density function is given by the probability law

$$f(x) = \frac{1}{\sqrt{2\pi}.\sigma}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

where π and e are given by $\pi = \frac{22}{7}$ and e=2.7183 (base of natural logarithms).

Remarks

1) A random variable X with mean μ and variance σ^2 following the normal law given above is represented as $X \sim N(\mu, \sigma^2)$.

2) If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$, is defined as a standard normal variate with E(Z)=0 and Var (Z)=1and we write Z~N(0, 1)

3) The p.d.f. of a standard normal variate Z is given by

$$\emptyset(z) = \frac{1}{\sqrt{2\pi}}.e^{-\frac{z^2}{2}}, -\infty < z < \infty, where\ Z = \frac{X-\mu}{\sigma}$$

4) Normal distribution is a limiting form of the binomial distribution when
   a)n, the number of trials is indefinite large, i.e. n→∞ and
   b) neither p nor q is very small.

5) Normal distribution is a limiting form of Poisson distribution when its mean m is large and n is also large.

12.3 Chief Characteristics (Properties) of Normal Distribution

It has the following properties:

1. The graph of $f(x)$ is bell shaped unimodal and symmetric curve as shown in the Fig. 12.1. The top of the bell is directly above the mean (μ).

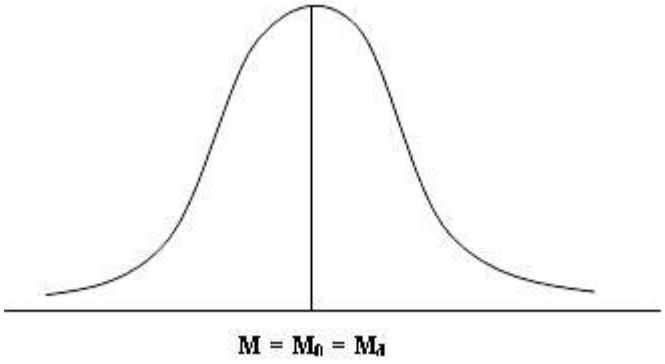


Fig. 12.1 : Normal Probability Curve

2. The curve is symmetrical about the line $X = \mu, (Z = 0)$ i.e., it has the same shape on either side of the line $X = \mu$ (or $Z = 0$).This is because the equation of the curve $\emptyset(z)$ remains unchanged if we change z to $-z$.
3. Since the distribution is symmetrical, mean, median and mode coincide. Thus, $\text{Mean} = \text{Median} = \text{Mode} = \mu$
4. Since $\text{Mean} = \text{Median} = \text{Mode} = \mu$, the ordinate at $X = \mu, (Z = 0)$ divides the whole area into two equal parts. Further, since total area under normal probability curve is 1, the area to the right of the ordinate as well as to the left of the ordinate at $X = \mu$ (or $Z = 0$) is 0.5
5. Also, by virtue of symmetry the quartiles are equidistant from median (μ), i.e,

$$Q_3 - M_d = M_d - Q_1$$

6. Since the distribution is symmetrical, all moments of odd order about the mean are zero. Thus $\mu_{2n+1} = 0\ ;\ (n = 0,1,2, \dots)$ i.e., $\mu_1 = \mu_3 = \mu_5 = \cdots = 0$
7. The moments (about mean) of even order are given by

$$\mu_{2n} = 1.3.5 \dots (2n-1)\sigma^{2n}, (n = 1,2,3 \dots)$$

Putting n=1 and 2 we get

$\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$

$\therefore \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$

8. Since the distribution is symmetrical, the moment coefficient of skewness based on moments is given by

$$\beta_1 = 0 \Rightarrow \gamma_1 = 0$$

9. The coefficient of kurtosis is given by

$$\beta_2 = 3 \Rightarrow \gamma_2 = 0$$

10. No portion of the curve lies below the x-axis, since $f(x)$ being the probability can never be negative.
11. Theoretically, the range of the distribution is from $-\infty < to < \infty$. But practically, $\text{range} = 6\sigma$
12. As x increases numerically [i.e. on either side of $X = \mu$], the value of $f(x)$ decreases rapidly, the maximum probability occurring at $x = \mu$ and is given by

$$[f(x)]_{max} = \frac{1}{\sqrt{2\pi}\sigma}$$

Thus maximum value of $f(x)$ is inversely proportional to the standard deviation. For large values of σ, $f(x)$ increases, i.e., the curve has a normal peak.

13. Distribution is unimodal with the only mode occurring at $X = \mu$.
14. X-axis is an asymptote to the curve i.e., for numerically large value of X (on either side of the line ($X = \mu$)), the curve becomes parallel to the X-axis and is supposed to meet it at infinity.
15. A linear combination of independent normal variates is also a normal variate. If $X_1, X_2, \ldots, X_n$ are independent normal variates with mean $\mu_1, \mu_2, \ldots, \mu_n$ and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively then their linear combination

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

where $a_1, a_2, \ldots, a_n$ are constants, is also a normal variate with

Mean = $a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$ and Variance = $a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$

In particular, if we take $a_1 = a_2 = \cdots = a_n = 1$ then we get

"$X_1 + X_2 + \cdots + X_n$ is a normal variate with mean $\mu_1 + \mu_2 + \cdots + \mu_n$ and variance $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$" .Thus, the sum of independent normal variates is also a normal variate with mean equal to sum of their means and standard deviation equal to square root of sum of the squares of their standard deviations. This is known as the 'Re-productive or Additive Property' of the Normal distribution.

16. Mean Deviation (M.D.) about mean or median or mode is given by

$$\text{M. D.} = \sqrt{\frac{2}{\pi}}.\sigma \cong \frac{4}{5}\sigma$$

17. Quartiles are given (in terms of μ and σ) by

$$Q_1 = \mu - 0.6745\sigma \text{ and } Q_3 = \mu + 0.6745\sigma$$

18. Quartile deviation (Q.D.) is given by

$$\text{Q. D.} = \frac{Q_3 - Q_1}{2} = 0.6745\sigma \cong \frac{2}{3}\sigma$$

Also

$$\text{Q. D.} = \frac{2}{3}\sigma = \frac{4}{6}\sigma = \frac{5}{6} \times \frac{4}{5}\sigma = \frac{5}{6}\ \text{M.D.}$$

$$\therefore \text{Q. D.} = \frac{5}{6}\ \text{M. D.}$$

19. We have (approximately):

$$\text{Q. D.} : \text{M. D} : \text{S. D.} :: \frac{2}{3}\sigma : \frac{4}{5}\sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1 \Rightarrow \text{Q. D.} : \text{M. D.} : \text{S. D.} :: 10 : 12 : 15$$

From property 18 we also have $4\text{S. D.} = 5\text{M. D.} = 6\text{Q. D.}$

20. Points of inflexion of the normal curve are at $X = \mu \pm \sigma$ i.e. they are equidistant from mean at a distance of σ and are given by :

$$X = \mu \pm \sigma\ , \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-1/2}$$

21. Area property: One of the most fundamental property of the normal probability curve is the area property. If $X \backsim N(\mu, \sigma^2)$, then the probability that random value of X will lie between X= μ and X=x_1 is

given

$$P(\mu < X < x_1) = \int_{\mu}^{x_1} f(x).dx = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{\mu}^{x_1} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

$$\text{put} \quad z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + \sigma z \;;\; \therefore \text{at } x = \mu, z = 0; \text{ and at } x = x_1, z = \frac{x_1-\mu}{\sigma} = z_1$$

$$\therefore P(\mu < x < x_1) = P(0 < z < z_1) \quad = \frac{1}{\sqrt{2\pi}}\int_0^{z_1} e^{-\frac{1}{2}z^2}\,dz = \int_0^{z_1} \emptyset(z)dz$$

where $\emptyset(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ is the probability function of standard normal variate. The definite integral $\int_0^{z_1} \emptyset(z)dz$ is known as **Normal Probability integral** and gives the area under standard normal curve between the ordinate z=0 and z = z_1. These areas have been provided in the form of table for different values of z_1 at the intervals of 0.01 which are available in any standard text books of statistics.

Particular Cases:

21.1 In particular, the probability that a random variable X lies in the interval ($\mu - \sigma$, $\mu + \sigma$) is given by

$$P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)dx.$$

$$P(-1 < Z < 1) = \int_{-1}^{1} \emptyset(z)dz = 2\int_0^1 \emptyset(z)dz = 2(0.3413) = 0.6826$$

The area under the normal probability curve between the ordinates at X= $\mu - \sigma$ and X= $\mu + \sigma$ is 0.6826.In other words, the range $X = \mu \pm \sigma$ covers 68.26% of the observations (as shown in Fig.12.2). This is known as 1σ limit of normal distribution

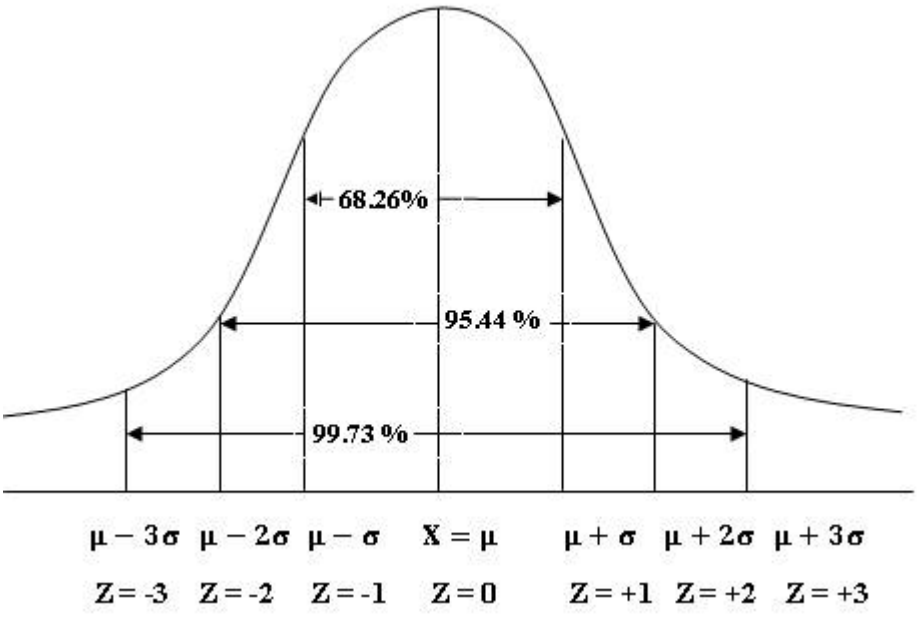


Fig. 12.2 1σ, 2σ and 3σ under Normal Probability Curve

21.2 The probability that random variable X lies in the interval (μ −2σ, μ +2σ) is given by

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \int_{\mu-2\sigma}^{\mu+2\sigma} f(x)dz. \Rightarrow P(-2 < Z < 2) = \int_{-2}^{2} \emptyset\ (z)dz$$

$$= 2\int_{0}^{2} \emptyset\ (z)dz = 2(0.47725) = 0.95445$$

The area under the normal probability curve between the ordinates at X= $\mu - 2\sigma$ and X= $\mu + 2\sigma$ is 0.95445. In other words, the range $X = \mu \pm 2\sigma$ covers 95.445% of the observations (as shown in Fig. 12.2). This is known as 2σ limits of normal distribution and is considered as warning limit in case of statistical quality control which implies that it is a warning to the manufacturer that the manufacturing process is going out of control.

21.3 The probability that random variable X lies in the interval (μ −3σ, μ +3σ) is given by

$$P(\mu - 3\sigma < X < \mu + 3\sigma = P(-3 < Z < 3)$$

$$\int_{-3}^{3} \emptyset\ (z)dz = 2\int_{0}^{3} \emptyset\ (z)dz = 2(0.49865) = 0.9973$$

The area under the normal probability curve between the ordinates at X= $\mu - 3\sigma$ and X= $\mu + 3\sigma$ is 0.9973.In other words, the range $X = \mu \pm 2\sigma$ covers 99.73% of the observations (as shown in Fig. 12.2). This is known as 3σ limits of normal distribution and it implies the manufacturing process is out of control in case of statistical quality control.

Thus, the probability that a normal variate X lies outside the range μ ± 3σ is given as

$$P(|X - \mu| > 3\sigma) = P(|z| > 3) = 1 - P(-3 < z < 3) = 1 - 0.9973 = 0.0027$$

Thus, in all probability, we should expect a normal variate to lie within the range μ ± 3σ though theoretically may range from −∞ to ∞.

12.4 Examples of Normal Distribution

(i) The age at first calving of cows belonging to the same breed and living under similar environmental conditions tend to normal frequency distribution.

(ii) The milk yield of cows in a large herd tends to follow a normal frequency distribution.

(iii) The chemical constituents of milk like fat, SNF, protein etc. for large samples follow normal distribution.

12.5 Computation of Area Under Normal Probability Curve

Probability that a continuous random variable X in any value between a and b is

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

which is the area bounded by the curve p(x), X-axis and the ordinates at X=a and X=b and is shown in Fig. 12.3.

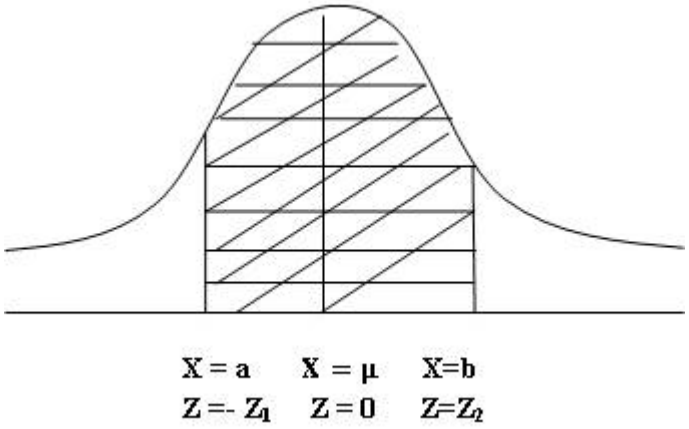


Fig. 12.3

Similarly the area to the right of the ordinates at $X=x_1$ and x_1 is less than mean (as shown in Fig.12.4)

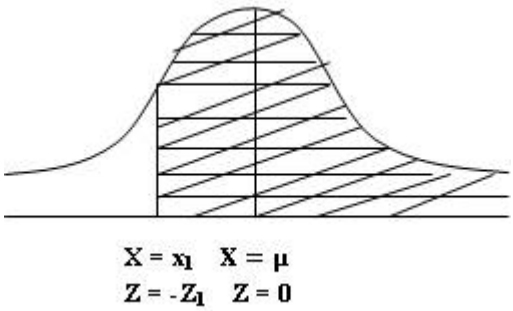


Fig.12.4

At $X=X_1$ $Z=-Z_1$ so $P(X>X_1)=0.5+P(-Z_1<Z<0)=0.5+ P(0<Z<Z_1)$, where $P(0<Z<Z_1)$ can be read from the normal tables. The application of this distribution in solving problems is illustrated through following examples.

Example 1. Average lactation yield for 1000 cows maintained at a farm is 1700 kg and their standard deviation is 85 kg. A cow is considered as high yielder if it has a lactation yield greater than 1900 kg and poor yielder if it has lactation yield less than 1600 kg. Find the number of high yielding and poor yielding cows.

Solution:

Here, μ=1700 kg and $\sigma = 85$ kg and let X denote the lactation milk yield

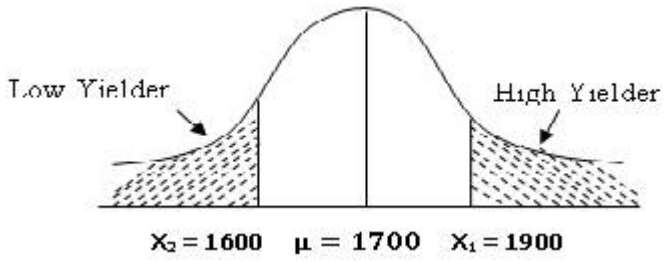


Fig. 12.5

a) To find number of high yielder cows we first find the probability of cows yielding more than 1900 kg. i.e. $P(X > 1900 \text{ kg.})$ (Fig. 12.5). So, we first compute the value standard normal variate i.e. Z_1 and then find

area under shaded region using normal tables

At X_1 = 1900 kg. $Z_1 = \frac{1900-1700}{85} = 2.3529 \sim 2.353$

$P(X_1 > 1900) = P(Z_1 > 2.353) = 0.5 - P(0 \leq Z_1 \leq 2.353) = 0.5 - 0.49069 = 0.00931$

Number of high yielder cows = $N \times P(z_1 > 2.353) = 0.00931 \times 1000 = 9.31$= 9 cows

b) To find number of low yielder cows, we first find the probability of cows yielding less than 1600 kg i.e. $P(X_2 < 1600\ kg.)$. So, we first compute the value standard normal variate i.e. Z_2 and then find area under shaded region using normal tables

$Z_2 = \frac{1600-1700}{85} = -1.18$

$P(X_2 < 1600) = P(Z_2 < -1.18) = 0.5 - P(0 \leq Z_2 \leq 1.18) = 0.5 - 0.38109 = 0.119$

Number of low yielder cows = $N \times P(X_2 < 1600) = 0.119 \times 1000$ =119 cows

Conclusion: Total number of high yielding & low yielding cows are 9 and 119 respectively.

Example 2. An Intelligence test was administrated to 1000 students. The average score of students was 42 with standard deviation of 24. Find

(a) Number of students exceeding a score of 50
(b) Number of students scoring between 30 & 58
(c) Value of score exceeded by top 100 students.

Solution: In this problem μ=42 and $\sigma = 24$ and let X denote the score obtained

(a) Number of students exceeding score 50

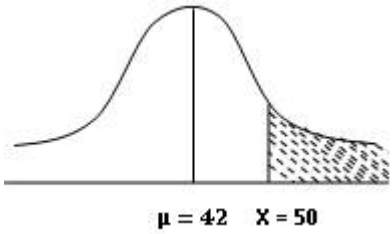


Fig. 12.6

As shown in figure 12.6 we want to find P(X>50) i.e. probability of shaded portion

At X=50, $Z = \frac{50-42}{24} = \frac{8}{24} = 0.334$

P(X>50) =P(Z > 0.334) = 0.5 – P(0≤ Z ≤ 0.334)= 0.5 – 0.1308== 0.3692

No of students = 1000 * 0.3692= 369.2 ~ 369 students

(b) Number of students scoring between 30 and 58

As shown in figure 12.7 we want to find P(30<X<58) i.e. probability of shaded portion

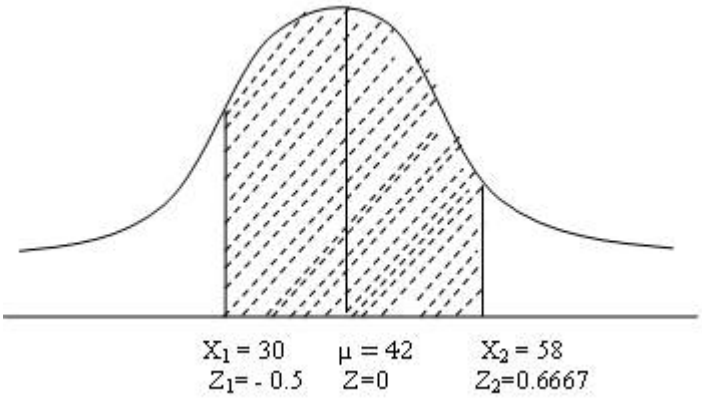


Fig.12.7

$$\text{At } X_1 = 30 \quad Z_1 = \frac{30-42}{24} = -0.5$$

$P(Z_1 > -0.5) = P(0 \le Z_1 \le 0.5) = 0.1915$

$$\text{At } X_2 = 58 \quad Z_2 = \frac{58-42}{24} = 0.6667$$

$P(Z_2 < 0.6667) = P(0 \le Z_2 \le 0.6667) = 0.2476$

$P(30 < X < 58) = P(-0.5 \le Z \le 0.6667) = 0.1915 + 0.2476 = 0.4391$

No of students = 1000 * .4391 = 439.1 ~ 439 students

(c) Value of score exceeded by top 100 students. Let x_1 be the value of score exceeded by top 100 students, the probability of top 100 students = 100/N = 100/1000 = 0.1 such that $P(X > x_1) = 0.1$

At $X = x_1$ $Z = \frac{x_1-42}{24} = Z_1$. From Fig. 12.8 the $P(X > x_1)$ shown as shaded region

$P(X > x_1) = P(Z > Z_1) = 0.1 \Rightarrow P(0 \le Z \le Z_1) = 0.4 \quad \Rightarrow \frac{x_1-42}{24} = 1.286$

$x_1 = 72.86 \sim 73$

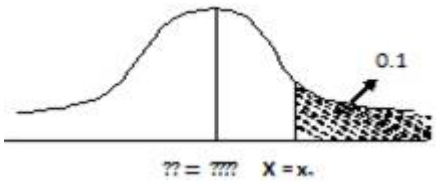


Fig. 12.8

Conclusion

(a) 369 students scored more than 50.

(b) 439 students scored between 30 & 58.

(c) Minimum score of top 100 students is 73.

Example 3. Tins are filled by an automatic filling machine with ghee. Average quantity filled in tin is 1000g. It is found that 5% of tins had ghee less than 980 grams. Find the standard deviation.

Solution : Here μ=1000g and let X be quantity of ghee filled in a tin

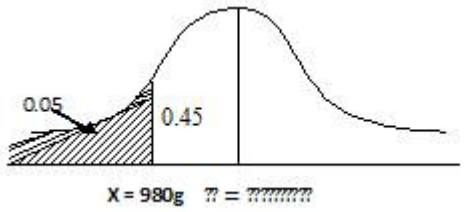


Fig. 12.9

From fig. 12.9 P(X <980)=0.05⇒ $P(Z < -Z_1)$ = -1.645

$\Rightarrow \frac{980-1000}{\sigma}$ =-1.645 $\Rightarrow -20$ =-1.645 σ , Hence σ =12.1840~12.18 gm

Conclusion :

Standard derivation of the ghee filled in tin is 12.18 gm.

Example 4. In a normal distribution, 31% of items are under 45 and 8% are over 64. Find the mean & standard deviation.

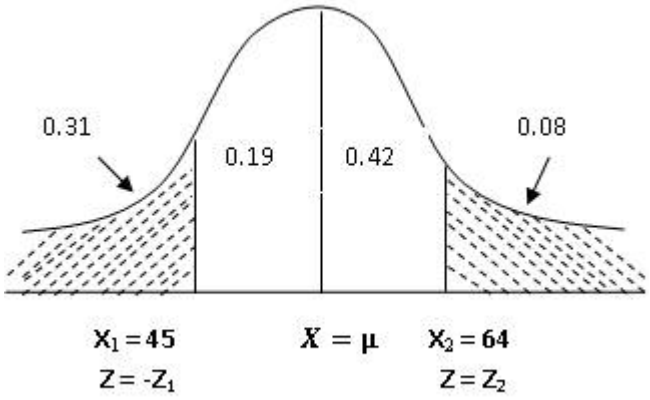


Fig.12.10

Solution:

Let X denotes the variable under consideration. We are given that P (X_1 < 45) = 0.31 and P (X_2 > 64) = 0.08. If X has normal distribution with mean μ and standard deviation σ. Then standard normal variates corresponding to X_1 = 45 and X_2 = 64 (from Fig 12.10) are

When X_1 = 45, $Z = \frac{45-\mu}{\sigma} = -Z_1$

When X_2 = 64, $Z = \frac{64-\mu}{\sigma} = Z_2$

From the fig. 12.10, P (0 <Z < Z_2) = 0.42 ⇒ Z_2 = 1.405 (from normal tables)

$\frac{64-\mu}{\sigma} = 1.405 \Rightarrow 64 - \mu = 1.405\sigma$ ----------(1)

P (-Z_1 <Z < 0) = 0.19 ⇒ P (0 <Z < Z_1) = -0.496 (from normal tables)

$\frac{45-\mu}{\sigma} = -0.496 \Rightarrow 45 - \mu = -0.496\sigma$ ----------(2)

Solving equations (1) & (2) we get

$\mu = 49.95 \sim 50$ and $\sigma = 9.99 \sim 10$

Conclusion

Mean & standard deviation of given distribution are 50 & 10 respectively.

Example 5. Average net weight of coffee complete powder is 250 g with a standard deviation of 3g. The powder is packed in polypack with an average weight of 5g with a standard deviation of 0.2g. Average weight of tin in which polypack is packed is 100g with a standard deviation of 1.5g. Individual weights of all items follow normal distribution. If 5% tins are classified as underweight tins then what would be the weight of filled in tin. Filled in tins are classified as overweight if their weight exceeds a weight of 360g. What proportion of tins are overweight tin?

Solution:

Let X_1 be the normal variate for weight of coffee with mean(μ_1)= 250g and s.d.(σ_1)=3g

X_2 be the normal variate for poly pack with mean (μ_2)= 5g and s.d.(σ_2)=0.2 g

X_3 be the normal variate for tin with mean (μ_3)= 100g and s.d.(σ_3)=1.5 g

By using the reproductive property of normal distribution, the composite weight of tin comprising of coffee complete, polypack and tin will follow a normal distribution with mean ($\mu=\mu_1+\mu_2+\mu_3$)=250+5+100=355g and standard deviation ($\sigma = \sqrt{(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)}$) = $\sqrt{(3)^2 + (0.2)^2 + (1.5)^2}$ = $\sqrt{11.29}$=3.36 g i.e. ($X=X_{1+} X_{2+} X_3$)~N(355g,3.36 g)

If 5% tins are classified as underweight, Let x_1 be the weight of tin considered as underweight, then we have :

P($X<x_1$)=0.05. Then the standard normal variate corresponding to x_1 is $Z_1 = \frac{x_1-355}{3.36} = -Z_1$

From fig.12.11 $P(-Z_1 < Z < 0) = 0.45 \Rightarrow P(0<Z<Z_1) = 0.45 \Rightarrow \frac{x_1-355}{3.36} = -1.6415 \Rightarrow x_1 = 349.67g$

Hence weight of underweight tins is 349.47 g.

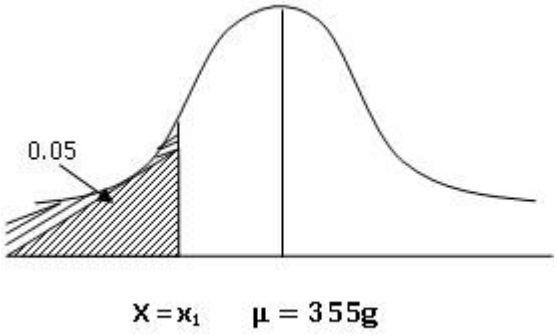


Fig. 12.11

Filled in tins exceeding a weight of 360g are classified as overweight. To find the proportion of tins which are overweight we proceed as follows:

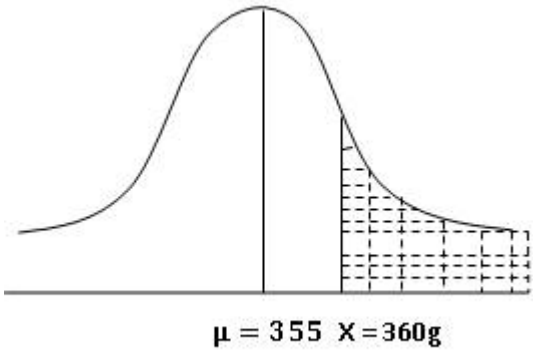


Fig. 12.12

As shown in figure 12.12, we want to find P(X>360) i.e. probability of shaded portion

At X=360, $Z = \frac{360-355}{3.36} = 1.488$

P (X > 360) = P (Z > 1.488) = 0.5 – P (0≤ Z ≤ 1.488) = 0.5 – 0.4316== 0.0684

Hence, 6.84 % tins are overweight

Conclusion:

Weight of underweight tins is 349.47g & 6.84% tins are overweight

12.6 Importance of Normal Distribution

Normal distribution plays a very important role in statistics because

(i) Most of the discrete probability distributions occurring in practice e.g., Binomial and Poisson can be approximated to normal distribution as n number of trials tends to increase.

(ii) Even if a variable is not normally distributed, it can be sometimes be brought to normal by a simple mathematical transformation, if the distribution of X is skewed, the distribution of $\sqrt{x}$ or log x might come out to be normal.

(iii) If $X \sim N(\mu,\sigma^2)$ then $P[\mu-3\sigma < x < \mu+3\sigma] = 0.9973 \Rightarrow P[|Z| > 3] = 1 - 0.9973 = 0.0027$. Thus the probability of standard normal variate going outside the limits ± 3 is practically zero. This property of normal distribution forms the basis of entire large sample theory.

(iv) Many of the sampling distribution e.g., student's t, Snedecor's F, Chi square distributions etc tend to normality for large samples. Further, the proof of all the tests of significance in the sample is based upon the fundamental assumptions that the populations from which the samples have been drawn are normal.

(v) The whole theory of exact sample (small sample) tests viz. t , χ^2, F etc, is based on the fundamental assumption that the parent population from which the samples have been drawn follows normal distribution.

(vi) Normal distribution finds large applications in statistical quality control in industry for setting up of control limits.

(vii) Theory of normal curves can be applied to the graduation of the curve which is not normal.

Lesson 13

SAMPLING THEORY AND SAMPLING DISTRIBUTION

13.1 Introduction

The science of statistics may broadly be studied under the two heads descriptive and inductive. So far we have confined ourselves to descriptive statistics which help us in describing the characteristics of numerical data. In other part i.e. inductive statistics also known as statistical inference which is termed as logic of drawing valid statistical conclusions about population in any statistical investigation on the basis of examining a part of population known as sample. It is drawn from population in a scientific manner. In all the spheres of life (such as economic, social, scientific, industry etc.) the need for statistical investigation and data analysis is increasing day by day. There are two methods of collection of statistical data i.e. census and sample method**.** Under census method, information related to the entire field of investigation or units of population is collected; whereas under sample method, rather than collecting information about all the units of population, information relating to only selected units is collected. Before we make a detailed study of both the methods, we will explain some basic concepts related to them.

13.2 Some Basic Concepts

13.2.1 Universe or population

In any statistical investigation interest lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. The group of individuals under study is called population or universe. In statistics a universe or population means the entire field under investigation about which knowledge is sought. “It is the totality of persons, objects, items or anything conceivable pertaining to certain characteristics”. In statistical usage, the term population is applied to any finite or infinite collection of individuals as per the statistical dictionary definition given by Kendall and Buckland. It is obvious that for any statistical investigation, complete enumeration of the population is rather impracticable. For example if we want to have an idea of the average per capita monthly income of the people in India, we will have to enumerate all the earning individuals in the country which is rather a very difficult task because of administrative and financial implications. A population can be of two kinds (i) Finite and (ii) Infinite. In a finite population, number of items is definite such as, number of students or teachers in a college, daily milk yield of 500 milch animals in a livestock farm. On the other hand, an infinite population has infinite number of items e.g. the population of pressures at various points in the atmosphere, the population of real numbers between 0 and 1, the population of all integers, number of water drops in an ocean, number of leaves on a tree or number of hairs on the head etc.

13.2.2 Sample

A finite subset of the population, selected from it by using scientific procedure with the objective of investigating its properties is called a sample In other words, selected or sorted units from the population are known as a sample. Thus, sample means some units selected out of a population which represent it. For example, if an investigator selects 100 animals from 2000 animals in a herd then these 100 animals will be termed as a sample and number of the individuals in the sample is called sample size.

13.2.3 Sampling

The process of selecting a sample is called sampling. It is a tool which enables us to draw conclusions about the characteristics of the population after studying only those items which are included in the sample. The main objective of sampling is

- To obtain the maximum information about the characteristics of population with the available sources e.g. time, money, manpower etc.
- To obtain best estimates of population parameter

13.2.4 Parameter and statistic

The statistical constants of the population like mean (μ), variance (σ^2), skewness (β_1), kurtosis (β_2), correlation coefficient (ρ) etc. are known as parameters. Similar statistical measures computed from the sample observations alone e. g. mean ($\bar{x}$) variance (s^2), skewness (b_1), kurtosis (b_2), correlation coefficient(r) etc. have been termed by Prof. R. A. Fisher as statistics. Let us consider a finite population of N units and let $Y_1, Y_2, Y_3, ---, Y_N$ be the observations on the N units in the population.

$$\text{Mean } (\mu)=\frac{1}{N}(Y_1+Y_2+Y_3+--+Y_N)=\frac{1}{N}\sum_{i=1}^{N}Y_i$$

$$\sigma^2=\frac{1}{N}[(Y_1-\mu)^2+(Y_2-\mu)^2+(Y_3-\mu)^2+--+(Y_N-\mu)^2]=\frac{1}{N}\sum_{i=1}^{N}(Y_i-\mu)^2$$

Suppose we draw a sample of size n from this population. Let $X_1, X_2, X_3, ---, X_n$ be the observations on the sample units. Then we can compute sample mean ($\bar{X}$) and sample variance (s^2) as given below:

$$\text{Mean } (\bar{X})=\frac{1}{n}(X_1+X_2+X_3+--+X_n)=\frac{1}{n}\sum_{i=1}^{n}X_i$$

$$s^2=\frac{1}{n}[(X_1-\bar{X})^2+(X_2-\bar{X})^2+(X_3-\bar{X})^2+--+(X_n-\bar{X})^2]=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

In practice the parameter values are not known and their estimates based on sample values are generally used. Thus statistics which may be regarded as an estimate of the parameter obtained from the sample is a function of sample values only and vary from sample to sample. If t is any general statistic which is a function of the sample observations $X_1, X_2, X_3, ---, X_n$ then a statistic $t=f(X_1, X_2, X_3, ---, X_n)$ is said to be unbiased estimate of population parameter θ if $E(t)=\theta$.

13.2.5 Sampling distribution

If we draw a sample of size n from a given finite population of size N, then the total number of possible samples is ${}^{N}C_n = k$ (say). For each of these samples we can compute some statistic $t = t(X_1, X_2, \ldots\ldots, X_n)$ e.g. mean $\bar{X}$, the variance s^2 etc. as given below.

Table 13.1

Sample number	Statistic		
	t	$\bar{X}$	s^2

1	t_1	$\bar{X}_1$	s_1^2
2	t_2	$\bar{X}_2$	s_2^2
3	t_3	$\bar{X}_3$	s_3^2
-	-	-	-
-	-	-	-
-	-	-	-
k	t_k	$\bar{X}_k$	s_k^2

The set of values of the statistic so obtained one for each sample constitutes what is called the sampling distribution of the statistic. For example, the values t_1, t_2, …, t_k determine the sampling distribution of the statistic t. In other words, the statistic t can be regarded as a random variable which can take values t_1, t_2, …, t_k and we can compute various statistical constants like mean, variance, skewness, kurtosis etc. for its distribution e.g. the mean and variance of the sampling distribution of the statistic are given by

$$\bar{t} = \frac{1}{k}\sum_{i=1}^{n} t_i \; ; \bar{\bar{X}} = Mean(\bar{X}) = \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i$$

$$v(t) = \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2 ; \; v(\bar{X}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{X}_i - \bar{\bar{X}})^2$$

13.2.6 Standard error

The standard deviation of the sampling distribution of a statistic is known as its Standard Error (S.E.). The standard Error of a statistic t is given by:

$$S.E.(t) = \sqrt{var(t)} = \sqrt{v(t) = \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2}$$

The standard errors of some of the well known statistics, for large samples, are given below where n is the sample size, σ^2 the population variance, P the population proportion and Q= 1 – P, n_1 and n_2 represent sizes of two independent random samples.

Table 13.2

Sr. No.	Statistic	Standard Error
1.	Sample mean $\bar{X}$	$\sigma/\sqrt{n}$
2.	Sample proportion p	$\sqrt{PQ/n}$
3.	Sample standard deviation	$\sqrt{\sigma^2/2n}$
4.	Sample variance (s^2)	$\sigma^2\sqrt{2/n}$
5.	Sample correlation coefficient(r)	$(1-\rho^2)/\sqrt{n}$
6.	Difference between two sample means $(\bar{X}_1 - \bar{X}_2)$	

		$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
7.	Difference between two sample Standard deviation $(s_1\text{-}s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$
8.	Difference between two sample proportion $(p_1\text{-}p_2)$	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

13.2.6.1 Utility of Standard Error

1) It plays a very important role in large sample theory and forms the basis of the testing of hypothesis. Thus, if the discrepancy between the observed and expected (hypothetical value of a statistic) is greater than or equal to Z_α times S.E., the hypothesis is rejected at α level of significance otherwise the deviation is not regarded as significant and is considered as due to fluctuations of sampling or chance causes.
2) The magnitude of S.E. gives an index of the precision of the estimate of parameter. The reciprocal of the S.E. is taken as the measure of reliability or precision of the sample e.g. S.E. of sample mean and sample proportion are $\frac{\sigma}{\sqrt{n}}$ and $\sqrt{\frac{PQ}{n}}$ respectively which vary inversely as the square root of the sample size. Thus in order to double the precision which amounts to reducing the S.E. to one half, the sample size has to be increased four times.
3) S.E. enables us to determine the probable limits/confidence limits within which the population parameter may be expected to lie.

13.3 Census and Sample Method

There are two methods to collect data

a) Census Method

b) Sample Method

13.3.1 Census method

Census method is that method in which information or data is collected from each and every unit of the population relating to the problem under investigation and conclusions are drawn on their basis. This method is also called as Complete Enumeration Method**.** For example, suppose some information (like monthly average milk yield, average lactation length, average fat contents in milk samples etc.) is to be collected from 2000 milking cows in a village. For that purpose if we collect data by inquiring each and every household of that village having milking animals then this method will be called as Census method. In this example, the whole village consisting of milking cows i.e. all 2000 cows will be considered as a population and every cow as an individual will be called the unit of the population. Population census and livestock census in India is conducted after every ten years and five years respectively by using census method. The census method seems to provide more accurate and exact information as compared to sample enumeration as the information is collected from

each and every unit of the population. Moreover, it affords more extensive and detailed study. This method has limitations and drawbacks given below:

- It requires lot of time and resources in terms of money, manpower and administrative personnel.
- This method can be adopted only by the government and big organization that have resources at their disposal.
- It is very time consuming process.

13.3.2 Sampling method

Sampling method is that method in which data is collected from the sample of items selected from population and conclusions are drawn from them. For example, if a study is to be made regarding the monthly average milk yield of 2000 milking cow in a village, then instead of inquiring each and every household having milking animals of that village, if we collect information by selecting some households say 100 ,then this will be called sampling method. On the basis of sampling method, it is possible to study the monthly average milk yield of the entire population of milking cows in a village. Sampling method has three main stages

- To select a sample
- To collect information from it
- To make inferences regarding the population.

Prof. R.A. Fisher sums up the advantages of sampling techniques over complete enumeration in just four words: speed, economy, adaptability and scientific approach. A properly design and carefully executed sampling plan yields fairly good results, often better than those obtained by the census method.

13.3.3 Importance of sampling method

In modern times sampling method is an important and popular method of statistical inquiry. Besides economic and business world, this method is widely used in daily life. For example, a researcher while preparing paneer wants to evaluate the whole tray of paneer by evaluating a part of this viz. sample of paneer. In the same way, we learn about of a commodity while buying the items of daily use like wheat, rice and pulses, etc. by observing the sample or specimen. In industries, statistical quality control manager inspects the quality of items by examining a few units produced.

13.3.4 Difference between census and sample method

The main difference between the census method and the sampling method are as follows

- In census method, all items relating to a universe are investigated whereas in sampling method only a few sub set of items are inquired.
- Census method is expensive from the point of view of time, money and labour whereas sampling method economizes on them.
- In such fields where study of each and every unit of the universe is necessary, census method is more appropriate. On the contrary, when population is infinite or vast or liable to be destroyed as a result of complete enumeration, then sampling method is considered to be more appropriate.

Lesson 14

SIMPLE RANDOM SAMPLING

14.1 Introduction

In the selection of a sample, it is desired to have the sample a true representative of the population. A large number of sampling schemes are available to achieve this objective. Various methods of sampling can be grouped under two broad heads; probability sampling (also known as random sampling) and non-probability (or non-random) sampling. Probability sampling methods are those in which every item in the universe has a known chance or probability of being chosen in the sample. This implies that the selection of sample item is independent of the person making the study-that is the sampling operation is controlled so objectively that the items will be chosen strictly at random.

14.2 Methods of Sampling

There are various methods of sampling that may be used singly or along with others. The choice of an appropriate sampling design is of paramount importance in the execution of a sample survey and is generally made keeping in view the objectives and scope of the enquiry and the nature of the population to be sampled. The sampling techniques may be broadly classified as follows.

(i) Purposive or Subjective or Judgment Sampling
(ii) Probability Sampling
(iii) Mixed Sampling or Restricted Sampling

14.2.1 Purposive or subjective or judgment sampling

In this method of sampling the choice of sample items depends exclusively on the judgment of the investigator. In this method, a desired number of sample units are selected deliberately or purposively depending upon the object of the enquiry so that only the important items representing the true characteristics of the population are included in the sample. Purposive sampling is one in which the sample units are selected with definite purpose in view. This type of sampling suffers from the drawback of favouritism and nepotism depending upon the beliefs and prejudices of the investigator and thus does not give a true representation to the population.

14.2.2 Probability sampling

Probability sampling provides a scientific technique of drawing samples from the population according to some laws of chance in which each unit in the universe has some definite pre-assigned probability of being selected in the sample. The selection of the sample based on the theory of probability is also known as random selection and the probability sampling is also called Random Sampling. Different types of sampling are:

- Each sample unit has an equal chance of being selected
- Sampling units have varying probability of being selected
- Probability of selection of a unit is proportional to the sample size.

14.2.3 Mixed sampling

Sampling design in which the sample units are selected partly according to some probability laws and partly according to a fixed sampling rule (no use of chance), is known as Mixed Sampling.

14.3 Simple Random Sampling

Simple random sampling (S.R.S.) is the technique in which sample is drawn in such a way that each and every unit in the population has an equal and independent chance of being included in the sample. Suppose we take a sample of size n from a finite population of size N. Then there are ${}^{N}C_{n}$ possible samples. A S.R.S. is the technique of selecting the sample in which each of ${}^{N}C_{n}$ samples has an equal chance or probability $p = \frac{1}{{}^{N}C_{n}}$ of being selected.

14.3.1 Simple random sampling without replacement

If the unit selected or drawn one by one in such a way that a unit drawn in any draw is not replaced in the population before making the next draw, then it is known as simple random sampling without replacement (srswor). A very important and interesting feature of simple random sampling without replacement (srswor) is that, "the probability of selecting a specified unit of population at any given draw is equal to the probability of its being selected at the first draw." This implies that in srswor from a population of size N, the probability that any sampling unit is included in the sample is 1/N and this probability remains constant throughout all the drawings. Mathematically, if E_r is the event that any specified unit is selected at r th draw, then

$$P(E_r) = \frac{1}{N} \quad \text{where } (r = 1,2, \dots, n)$$

where n is the sample size. In particular it implies

$$P(E_r) = \frac{1}{N} = P(E_1)$$

i.e. the chance of selection of any specified item is same at any subsequent draw as it was in the first draw, viz. 1/N. Alternatively srswor can be defined as "If a sample of size n is drawn without replacement from a population of size N then there are ${}^{N}C_{n}$ possible samples. Simple Random Sampling is the technique of selecting the sample so that each of these ${}^{N}C_{n}$ samples has an equal chance or probability $p = \frac{1}{{}^{N}C_{n}}$ of being selected in the sample."

14.3.2 Simple random sampling with replacement

If the unit selected is drawn one by one in such a way that a unit drawn at a time is replaced in the population before the subsequent draw, then it is known as simple random sampling with replacement (srswr).In this type of sampling, a unit can be included more than once in a sample. Therefore, if the required sample size is n, the effective sample size is sometimes less than n due to inclusion of one or more units more than once. Thus, simple random sampling with replacement always amounts to sampling from an infinite population, even though the population is finite. If sampling is done with replacement, then there are N^n possible samples of size n. In this case, simple random sampling (srswr) gives equal chance $p = \frac{1}{N^n}$ for each of the N^n samples to be selected.

Remark: With the idea that effective sample size should be adhered to, the simple random sampling without replacement is adopted**.**

14.4 Selection of Simple Random Sample

Generally, the method of selection should be independent of the properties of sampled population. Proper care should be taken to ensure that the selected sample is random. Human bias, which varies from individual to individual, is inherent in any sampling scheme administered by human beings. Random samples can be obtained by any of the following two methods:

(i) Lottery method

(ii) Use of tables of Random numbers

14.4.1 Lottery method

The simplest method of drawing a random sample is the lottery system. Suppose we want to select 'r' individuals out of n. We assign the numbers (1 to n), one number to each individual and write these (numbers) 1 to n on n slips which are made as homogenous as possible in shape and size etc. These slips are then put in a bag and thoroughly shuffled and then 'r' slips are drawn one by one. The 'r' individuals corresponding to numbers on these slips drawn constitute the random sample. This method of selection is quite independent of the properties of population and is one of the most reliable methods of selecting a random number.

14.4.2 Use of tables of random numbers

The lottery method described above is quite time consuming and cumbersome to use if the population is sufficiently large. The most practical and inexpensive method of selecting a random sample consists in the use of random number tables. These tables have been constructed in such a way that each of the digits 0,1,2---, 9 appear with approximately same frequency and independent of each other. If we have to select a sample from a population of size N($\leq$ 99) then the numbers can be combined two by two to give pairs from 00 to 99.Simliarly if N($\leq$ 999) or N($\leq$ 9999) and so on, then combine the digits three by three (or four by four and so on) .The method of drawing the random sample consists of the following steps :

(i) Identify the N units in the population with numbers from 1 to N.

(ii) Select at random, any page of the random number tables and pick up the numbers in any row or column or diagonal at random.

(iii) The population units corresponding to the numbers selected in step (ii) constitutes the random sample.

The following random number tables are commonly used in practice:

1) Tippet's (1927) Random Numbers Tables: These tables consist of 10400 four-digited numbers giving in all 10400×4 i.e., 41600 digits taken from British Census reports.
2) Fisher and Yates (1938) Tables: These tables comprise 15,000 digits arranged in twos.
3) Kendall and Babington Smith's (1939) Random Tables: These tables consist of 1,00,000 digits grouped into 25, 000 sets of 4 digited random numbers.
4) Rand Corporation (1955) random number tables consist of one million random digits consisting of 2,00,000 random numbers of 5 digits each.

14.5 Some Important Results on Simple Random Sampling

Let us consider a simple random sample (without replacement) of size n from a population of size N. Let the observations on the population units be denoted by $Y_1, Y_2, Y_3, \ldots, Y_N$ and the observations on the sample units be denoted by $X_1, X_2, X_3, \ldots, X_n$. Then, in the usual notation we have:

Population mean (μ) is given by:

$$\mu = \bar{Y} = \frac{1}{N}(Y_1 + Y_2 + Y_3 + \cdots + Y_N) = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Population variance (σ^2) is given by:

$$\sigma^2 = \frac{1}{N}[(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_N - \bar{Y})^2] = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \qquad (14.1)$$

Population mean squares (S^2) is given by

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \qquad (14.2)$$

From equations 14.1 and 14.2 we get,

$$N\sigma^2 = (N-1)S^2 \quad \Rightarrow \quad S^2 = \frac{N\sigma^2}{N-1}$$

Sample mean is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (14.3)$$

Sample variance is given by

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (14.4)$$

14.5.1 In simple random sampling without replacement (srswor), the sample mean is an unbiased estimate of the population mean

$$E(\bar{x}) = \bar{y} \qquad (14.5)$$

14.5.2 In srswor, the sample mean square is an unbiased estimate of the population mean square

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = S^2 \qquad (14.6)$$

14.5.3 In srswor, the variance of the sample mean is given by

$$Var(\bar{x})_{srswor} = \frac{N-n}{N}.\frac{S^2}{n} = \frac{N-n}{Nn}.S^2 = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \qquad (14.7)$$

14.5.4 Comparison of srswor with srswr

Simple random sampling with replacement (srswr) can be regarded as sampling from an infinite population and its variance is given by:

$$\text{Var}(\bar{x})_{srswr} = \frac{\sigma^2}{n} = \frac{N-1}{N}.\frac{S^2}{n} \qquad (14.8)$$

Comparing it with equation 14.7 we get $\text{Var}(\bar{x})_{srswor} < \text{Var}(\bar{x})_{srswr}$ i.e. the variance of the sample mean (as an estimate of $\mu = \bar{y}$) is less in srswor as compared with its variance in the case of srswr. This implies that srswor provides a better (more efficient) estimator of the population mean μ relative to srswr.

14.6 Merits and Demerits of Simple Random Sampling

Merits

- Since it is a probability sampling, it eliminates the bias due to personal judgment or discretion of the investigator. Accordingly, the sample selected is more representative of the population than in the case of judgment sampling.
- Because of its random character, it is possible to ascertain the efficiency of the estimates by considering the standard errors of their sampling distributions.
- The theory of random sampling is highly developed so that it enables us to obtain the most reliable and maximum information at the least cost, and results in savings in time, money and labour.

Demerits

- Simple random sampling requires an up-to-date frame, i.e., a complete and up-to-date list of the population units to the sampled. In practice, since this is not readily available in many enquiries, it restricts the use of this sampling design.
- In field surveys if the area of coverage is fairly large, then the units selected in the random sample are expected to be scattered widely geographically and thus it may be quite time consuming and costly to collect the requisite information or data.
- If the sample is not sufficiently large, then it may not be representative of the population and thus may not reflect the true characteristic of the population.
- The numbering of the population units and the preparation of the slips is quite time consuming and uneconomical particularly if the population is large therefore, this method cannot be used effectively to collect most of the data in social sciences.

Lesson 15

ELEMENTARY CONCEPTS OF OTHER SAMPLING TECHNIQUES

15.1 Introduction

In the preceding lesson we have seen the nature of simple random sampling and how a random sample can be drawn. The use of simple random sampling requires an up-to-date frame, i.e. a complete and up-to-date list of the population units to the sampled. In practice, since this is not readily available in many enquiries, it restricts the use of this sampling design. Moreover, in field surveys if the area of coverage is fairly large, then the units selected in the random sample are expected to be scattered widely/geographically and thus it may be quite time consuming and costly to collect the requisite information or data. If each element is drawn individually from the population at large, it is an unrestricted sample. Restricted sampling is where additional controls are imposed, in other words it covers all other forms of sampling. Because of a more effective distribution of the sampling units, restricted random sampling is generally more efficient than unrestricted random sampling which is possible using simple random sampling. In the restricted random sampling, the cost of sampling are minimised for a given precision of the population estimate. As an alternative to the simple random sampling design, several complex probability sampling designs are available which can be used that are more viable and effective. Efficiency is improved because more information can be obtained for a given sample size using some of the complex probability sampling procedures than the simple random sampling design. In this lesson we will discuss some of these sampling techniques.

15.2 Stratified Random Sampling

In simple random sampling without replacement, sampling variance of the sampling distribution of mean is given by $\mathrm{Var}(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$ .This implies that variance of the sample estimate of the population mean is

- inversely proportional to the sample size
- directly proportional to the variability of the sampling units in the population .

Since the precision of an estimate is reciprocal of sampling variance, so apart from increasing the sample size n, the other way of increasing the precision is to devise a sampling technique which will reduce S^2, i.e. population heterogeneity. A sampling plan which will effectively reduce the variability in the population is called stratified sampling. When the population is heterogeneous with respect to the variable or characteristic under study, then the technique of stratified random sampling is used to obtain more efficient results. Stratification means division into layers or groups. A stratified random sampling is one where the population is divided into mutually exclusive and mutually exhaustive strata or sub-groups and then a simple random sample is selected within each strata or sub-group e.g. cows in a big herd can divided into different strata on the basis of breed, age groups, body weight groups, lactation length, lactation order, daily/lactation milk yield groups etc. Stratified random sampling involves the following steps:

1. Stratify the given population into number of sub-groups or sub-populations known as strata such that:
   a) The units within each stratum (sub-group) are as homogeneous as possible.

b) The difference between various strata are as marked as possible, i.e., the stratum means differ as widely as possible.

c) Various strata are non-overlapping. This means each and every unit in the population belongs to one and only one stratum.

The criterion used for the stratification of the universe into various strata is known as stratifying factor. In general geographical, sociological or economic characteristics form the basis of stratification of the given population. Some of the commonly used stratifying factors are age, sex, income, occupation, educational level, geographical area, economic status, etc. Thus in stratified sampling the given population of size N is divided into, say k relatively homogeneous strata of sizes $N_1, N_2,\ldots,N_k$ respectively such that

$$N = \sum_{i=1}^{k} N_i.$$

2. Draw simple random sample (without replacement) from each of the strata. Let random sample of size n_i , be drawn from the i^{th} strata, (i = 1,2,…,k) such that

$$\sum_{i=1}^{k} n_i = n,$$

Where n is the total sample size from a population of size N. The sample of n units is known as stratified random sample (without replacement) and the technique of drawing such a sample is known as stratified random sampling.

The size of the sample from each stratum can either be proportional, optimum or disproportional to the size of each stratum.

15.2.1 Proportional allocation

In this, the items are selected from each stratum in the same proportion as they exist in the population. The allocation of sample sizes is termed as proportional if the sample fraction, i.e. the ratio of the sample size to the population size, remains the same in all the strata. Mathematically, the principle of proportional allocation gives

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_k}{N_k}$$

15.2.2 Optimum allocation

In this case the size of the samples to be drawn from the various strata is determined by the principle of optimization, i.e., obtaining best results at minimum possible cost. In optimum allocation, number of units in i^{th} sample n_i's, (i = 1,2,…,k) are determined so that

(i) Variance of sample estimate of the population mean is minimum (i.e., its precision is maximum) for a fixed total sample size n. (Neyman's Allocation)

(ii) Variance of the estimate is minimum for a fixed cost of the plan.

15.3 Systematic Sampling

Systematic sampling is slightly different than the simple random sampling in which only the first sample unit is selected at random and the remaining units are automatically selected in a definite sequence at equal spacing or equal time interval from one another. This technique of drawing samples is usually recommended if the complete and up-to-date list of the sampling units, *i.e.,* the frame is available and the units are arranged in some systematic order such as alphabetical, chronological, geographical order, time interval etc.

Let us suppose that N sampling units in the population are arranged in some systematic order and serially numbered from 1 to N and we want to draw a sample of size n from it such that $N = nk \Rightarrow k = \frac{N}{n}$, where k is usually called the sample interval. Systematic sampling consists in selecting any unit at random from the first k units numbered from 1 to k and then selecting every k^{th} unit in succession subsequently. Thus, if the first unit selected at random is i^{th} unit, then the systematic sample of size n will consist of the units numbered $i, i+k, i+2k, \ldots, i+(n-1)k$. The random number 'i' is called the random start and its value, in fact, determines the whole sample. As an example, let us suppose that we want to select 50 cows from a list of cows containing 2,000 identification numbers tags arranged systematically. Here

$$n = 50 \text{ and } N = 2{,}000$$

$$\therefore k = \frac{N}{n} = \frac{2{,}000}{50} = 40$$

We select any number from 1 to 40 at random and the corresponding serial number cow is selected. Suppose the random number selected is 25 .Then, the systematic sample will consist of 50 cows in the list at identification number 25, 65, 105, ---, 1905, 1945, 1985.

15.4 Cluster Sampling

When the population size is very large, the previously mentioned sampling methods lead to several difficulties. The sampling frame is not available and it is too expensive and time consuming to prepare it. The other difficulties are firstly the high cost and administrative difficulty of surveying widely scattered sampling units and secondly the elementary units may not be easily identifiable and locatable. In such cases cluster sampling is useful. In this case the total population is divided, depending upon on problem under study, into some recognizable sub-divisions which are termed as clusters. A specified number of clusters are selected at random, and the observation is made on all the units in the sampled clusters. We then observe, measure and interview each and every unit in the selected clusters. The clusters are called the primary units. Cluster sampling is also known as area sampling. For example, cluster may be consisting of all households in a village and hence there are as many clusters as the number of villages in a district. It may be noted that the cluster is a heterogeneous sub-population whereas stratum is a homogeneous sub-population. Certain precautions should be taken while using cluster sampling which are given below:

- Each elementary unit should belong to one and only one cluster.
- The cluster constituting the population should include each and every elementary unit belonging to the population.
- All clusters should be distinct, meaning thereby that there should neither be overlapping nor omission of units.

- Clusters should be as small as possible consistent with the cost and limitations of the survey.

15.5 Multistage Sampling

Instead of enumerating all the sampling units in the selected clusters one can obtain better and more efficient estimators by resorting to sub sampling within the clusters. The technique is called two-stage sampling, clusters being termed as primary units and the units within the clusters as secondary units. The above technique may be generalised to what is called multistage sampling. As the name suggests, multistage sampling refers to a sampling technique which is carried out in various stages. Here the population is regarded as consisting of a number of primary units each of which is further composed of a number of secondary stages unit in which we are interested. For example, if we are interested in obtaining a sample of, say, n households from a particular state the first stage units may be districts and second stage units may be villages in the districts and third stage units will be households in the villages. Each stage thus results in a reduction of the sample size.

Multistage sampling consists in sampling first stage units by some suitable method of sampling. From among the selected first stage units, a sub-sample of secondary stage units is drawn by some suitable method of sampling which may be same or different from the method used in selecting first stage units. Further stages may be added to arrive at a sample of desired sampling units. Multistage sampling is more flexible as compared to other methods of sampling.

Lesson 16

TESTING OF HYPOTHESIS

16.1 Introduction

A researcher or experimenter has always some fixed ideas about certain population(s) vis-à-vis population parameters(s) based on prior experiments/sample surveys or past experience. It is therefore desirable to ascertain whether these ideas or claims are correct or not by collecting information in the form of data through conduct of experiment or survey. In this manner, we come across two types of problems, first is to draw inferences about the population on the basis of sample data and other is to decide whether our sample observations have come from a postulated population or not. In this lesson we would be dealing with the second type of problem. In the ensuing section we would provide concepts and definitions of various terms used in connection with the testing of hypothesis.

16.2 Testing of Hypothesis

The inductive inference is based on deciding about the characteristics of the population on the basis of a sample. Such decisions involve an element of risk, the risk of wrong decisions. In this endeavor, modern theory of probability plays an important role in decision making and the branch of statistics which helps us in arriving at the criterion for such decisions is known as testing of hypothesis. The theory of testing of hypothesis was initiated by J. Neyman and E.S. Pearson. Thus, theory of testing of hypothesis employs various statistical techniques to arrive at such decisions on the basis of the sample theory. We first explain some fundamental concepts associated with testing of hypothesis.

16.3 Basic Concepts of Testing of Hypothesis

16.3.1 Hypothesis

According to Webster Hypothesis is defined as tentative theory or supposition provisionally adopted, explain certain facts and guide in the investigation of others e.g. looking at the cloudy weather, the statement that “It may rain today” is considered as hypothesis.

16.3.2 Statistical hypothesis

A statistical hypothesis is some assumption or statement, which may or may not be true about a population, which we want to test on the basis of evidence from a random sample. It is a definite statement about population parameter. In other words, it is a tentative conclusion logically drawn concerning any parameter of the population. For example, the average fat percentage of milk of Red Sindhi Cow is 5%, the average quantity of milk filled in the pouches by an automatic machine is 500 ml.

16.3.3 Null hypothesis

According to Prof. R. A. Fisher “A hypothesis which is tested for possible rejection under the assumption that it is true is usually called Null Hypothesis” and is denoted by H_0.The common way of stating a hypothesis is that there is no difference between the two values, namely the population mean and the sample mean. The term ‘no difference’ means that the difference, if any, is merely due to sampling fluctuations. Thus, if the statistical test

shows that the difference is significant, the hypothesis is rejected. A statistical hypothesis which is stated for the purpose of possible acceptance is called Null Hypothesis. To test whether there is any difference between the two populations we shall assume that there is no difference. Similarly, to test whether there is relationship between two variates, we assume there is no relationship. So a hypothesis is an assumption concerning the parameter of the population. The reason is that a hypothesis can be rejected but cannot be proved. Rejection of no difference will mean a difference, while rejection of no relationship will imply a relationship. For example if we want to test that the average milk production of Karan Swiss cows in a lactation is 3000 litres then the null hypothesis may be expressed symbolically as H_o: μ = 3000 litres.

16.3.4 Alternative hypothesis

Any hypothesis which is complementary to the null Hypothesis is called an alternative hypothesis. It is usually denoted by H_1 or H_A. For example if we want to test the null hypothesis that the population has a specified mean μ_o i.e. H_o: $\mu=\mu_o$ then the alternative hypothesis could be

(i) H_1: $\mu \neq \mu_o$ ($\mu > \mu_o$ or $\mu < \mu_o$)

(ii) H_1: $\mu > \mu_o$

(iii) H_1: $\mu < \mu_o$

The alternative hypothesis in (i) is known as two tailed alternative and the alternatives in (ii) and (iii) are known as right tailed and left tailed alternatives. The setting of alternative hypothesis is very important since it enables us to decide whether to use a single tailed (right or left) or two tailed test. The null hypothesis consists of only a single parameter value and is usually simple while alternative hypothesis is usually composite.

16.3.5 Simple and composite hypothesis

If the statistical hypothesis completely specifies the population or distribution, it is called a simple hypothesis, otherwise it is called a composite hypothesis. For example, if we consider a normal population N (μ, σ^2) where σ^2 is known and we want to test the hypothesis H_0: μ=25 against H_1: μ=30. From these hypotheses, we know that μ can take either of the two values, 25 or 30. In this case H_0 and H_1 are both simple. But generally H_1 is composite, i.e., of the form H_1: $\mu\neq$25, viz, H_1: μ<25 or H_1: μ>25. In sampling from a normal population N (μ, σ^2), the hypothesis H: $\mu=\mu_0$ and $\sigma^2=\sigma_0^2$ is a simple hypothesis because it completely specifies the distribution. On the other hand (i) $\mu=\mu_0$ (σ^2 is not specified) (ii) $\sigma^2=\sigma_0^2$ (μ is not specified) (iii) $\mu<\mu_0$, $\sigma^2=\sigma_0^2$ etc. are composite hypothesis.

16.3.6 Types of errors in testing of hypothesis

The main objective in sampling theory is to draw a valid inference about the population parameters on the basis of the sample results. In practice we decide to accept or reject a null hypothesis (H_0) after examining a sample from it. As such we are liable to commit errors. The four possible situations that arise in testing of hypothesis are expressed in the following dichotomous table:

Table 16.1

Decision from sample	True Situation	
	Hypothesis is true	Hypothesis is false
Accept the hypothesis	No error	Type II error
Reject the hypothesis	Type I error	No error

In testing hypothesis, there are two possible types of errors which can be made. The error of rejection of a hypothesis H_0 when H_0 is true is known as Type I error and error of acceptance of a hypothesis H_0 when H_0 is false is known as type II error. When setting up an experiment to test a hypothesis it is desirable to minimize the probabilities of making these errors. But practically it is not possible to minimize both these errors simultaneously. In practice, in most decision making problems, it is more risky to accept a wrong hypothesis than to reject a correct one. These two types of errors can be better understood with an example where a patient is given a medicine to cure some disease and his condition is examined for some time. It is just possible that the medicine has a positive effect but it is considered that it has no effect or adverse effect. Therefore it is the Type I error. On the other hand if the medicine has an adverse effect but it is considered to have had a positive effect, it is called Type II error. Now let us consider the implications of these two types of error. If type I error is committed, the patient will be given another medicine, which may or may not be effective. But if type II error is committed i.e., the medicine is continued inspite of an adverse effect, the patient may develop some other complications or may even die. This clearly shows that the type II error is much more serious than the type I error. Hence in drawing inference about the null hypothesis, generally type II error is minimized even at the risk of committing type I error which is usually chosen as 5 per cent or 1 per cent.

Probability of committing type I error and type II error are denoted by α and β and are called size of type I and type II error respectively. In Industrial Quality Control, while inspecting the quality of a manufactured lot, the Type I error and type II error amounts to rejecting a good lot and accepting a bad lot respectively. Hence α=P(Rejecting a good lot) and β=P(Accepting a bad lot). The sizes of type I and type II errors are also known as producer's risk and consumer's risk respectively. The value of (1-β) is known as the power of the test.

16.3.7 Level of significance

It is the amount of risk of the type I error which a researcher is ready to tolerate in making a decision about H_0. In other words, it is the maximum size of type I error, which we are prepared to tolerate is called the level of significance. The level of significance denoted by α is conventionally chosen as 0.05 or 0.01. The level of 0.01 is chosen for high precision and the level 0.05 for moderate precision. Sometimes this level of risk is further brought down in medical statistics where the efficiency of life saving drug on the patient is tested. If we adopt 5% level of significance, it means that on 5 out of 100 occasions, we are likely to reject a correct H_0 .In other words this implies that we are 95% confident that our decision to reject H_o is correct. That is, we want to make the significance level as small as possible in order to protect the null hypothesis and to prevent, as far as possible, the investigator from inadvertently making false claims. Level of significance is always fixed in advance before collecting the sample information.

16.3.8 P-value concept

Another approach followed in testing of hypothesis is to find out the P-value at which H_0 is significant i.e., to find the smallest level α at which H_0 is rejected. In this situation, it is not inferred whether H_0 is accepted or rejected at a level of 0.05 or 0.01 or any other level. But the researcher only gives the smallest level α at which H_0 is rejected. This facilitates an individual to decide for himself as to how much significant the research results are. This approach avoids the imposition of a fixed level of significance. About the acceptance or rejection of H_0, the experimenter can himself decide the level of α by comparing it with the P-value. The criterion for this is that if the P-value is less than or equal to α, reject H_0 otherwise accept H_0.

16.3.9 Degrees of freedom

For a given set of conditions, the number of degrees of freedom is the maximum number of variables which can freely be designed (i.e., calculated or assumed) before the rest of the variates are completely determined. In other words, it is the total number of variates minus the number of independent relationships existing among them. It is also known as the number of independent variates that make up the Statistic. In general, degree of freedom is the total number of observations (n) minus the number of independent linear constraints (k) i.e. n-k.

16.3.10 Critical region

The total area under a standard curve is equal to one representing probability distribution. In test of hypothesis the level of significance is set up in order to know the probability of making type I error of rejecting the hypothesis which is true. A statistic is required to be used to test the null hypothesis H_0. This test is assumed to follow some known distribution. In a test, the area under the probability density curve is divided into two regions, viz, the region of acceptance and the region of rejection. If the value of test statistics lies in the region of rejection, the H_0 will be rejected. The region of rejection is also known as a critical region. The critical region is always on the tail of the distribution curve. It may be on both sides of the tails or on one side of the tail depending upon alternative hypothesis H_1.

16.3.10.1 One –tailed test

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right-tailed or left- tailed) is called a one tailed test. For example, a test for testing the mean of a population H_0: $\mu=\mu_0$ against the alternative hypothesis H_1: $\mu>\mu_0$ (Right –tailed) or H_1: $\mu<\mu_0$ (Left–tailed) is a single –tailed test. If the critical region is represented by only one tail, the test is called one-tailed test or one–sided test. In right tailed test (H_1: $\mu>\mu_0$) the critical region lies entirely on the right tail of the sampling distribution of $\bar{x}$ as shown in Fig, 16.2 , while for the left tail test (H_1: $\mu<\mu_0$), the critical region is entirely in the left tail of the distribution of $\bar{x}$ as shown in Fig. 16.1.

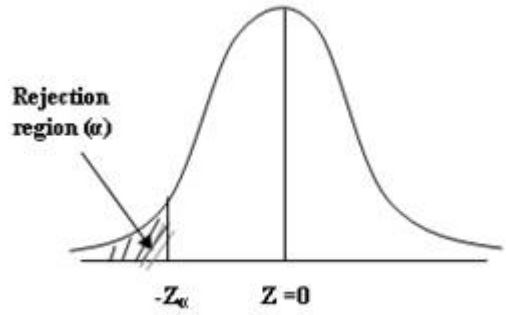


Fig. 16.1 Left tailed Test

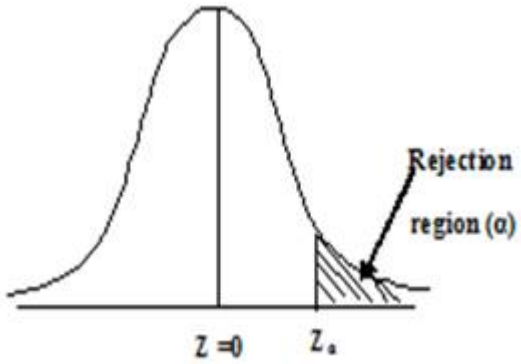


Fig. 16.2 Right tailed Test

16.3.10.2 Two–tailed test

A test of statistical hypothesis where the alternative hypothesis is two–sided such as: H_0: $\mu=\mu_0$ against the alternative hypothesis H_1: $\mu\neq\mu_0$ ($\mu>\mu_0$ or $\mu<\mu_0$) is known as a two–tailed test and in such a case the critical region is given by the portion of the area lying on both the tails of the probability curve of the test statistic as shown in Fig. 16.3

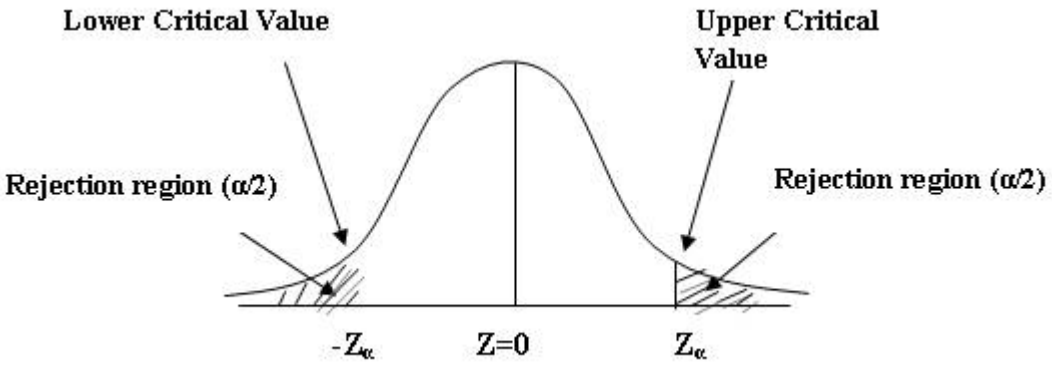


Fig. 16.3 Two Tailed Test

In a particular problem, whether one-tailed or two-tailed test is to be applied depends entirely on the nature of the alternative hypothesis.

16.3.11 Critical values or significant values

The value of a test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depends upon

- the level of significance and
- the alternative hypothesis, whether it is two tailed or single tailed.

The critical value of the test statistic at α level of significance for a two tailed test is given by Z_α where Z_α is determined by the equation $P(|Z| > Z_\alpha) = \alpha$, where Z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is symmetric curve, we get $P[Z > Z_\alpha] = \alpha/2$ i.e., the area of each tail is $\alpha/2$. Thus Z_α is the value such that area to the right of Z_α is $\alpha/2$ and to the left of - Z_α is $\alpha/2$ as shown in fig. 16.3 .In case of single tail alternative, the critical value of Z_α is determined so that total area to the right of it (for right tailed test) is α (as shown in fig. 16.2) and for left tailed test the total area to the left of - Z_α is α (as shown in Fig. 16.1).Thus the significant or critical value of Z for a single tailed test (left or right) at level of significance 'α' is same as the critical value of Z for a two tailed test at level of significance 2α.

Table 16.2 Critical values of (Z_α) of Z

Critical values (Z_α)	Level of significance		
	1%	5%	10%
Two tailed test	$\|Z_\alpha\| = 2.58$	$\|Z_\alpha\| = 1.96$	$\|Z_\alpha\| = 1.645$
Right tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

16.4 Test of Significance

The tests of significance which are dealt hereafter pertain to parametric tests. A statistical test is defined as a procedure governed by certain rules, which leads to take a decision about the hypothesis for its acceptance or rejection on the basis of sample observations. Test of significance enables us to decide on the basis of sample results if

(i) deviation between observed sample statistic and the hypothetical parameter value or

(ii) the deviation between two sample statistics is significant.

Test of significance is a procedure of either accepting or rejecting a Null hypothesis. The tests are usually called tests of significance since here we test whether the difference between the sample values and the population values or between the values given by two samples are so large that they signify evidence against the hypothesis or these differences are small enough to be accounted for as due to fluctuations of sampling, i.e. they may be regarded as due only to the fact that we are dealing with a sample and not with the whole population. Statistical tests play an important role in biological sciences, dairy industry, social sciences and agricultural sciences etc. The use of these tests is made clear through a number of practical examples:

1. An automatic machine is filling 500 ml. of milk in the pouch. Now to make sure whether the claim is correct or not one has to take a random sample of the filled in pouches and note the actual quantity of milk in the pouches. From these sample observations it would be decided whether the automatic machine is filling the right quantity of milk in the pouches. This is done by performing test of significance.
2. There is a process A which produced certain items. It is considered that a new process B is better than process A. Both the processes are put under operation and then items produced by them are sampled and

observations are taken on them. A statistical based test on sample observations will help the investigator to decide whether the process B is better than A or not.

3. Psychologists are often interested in knowing whether the level of IQ of a group of students is up to a certain standard or not. In this case some students are selected and an intelligence test is conducted. The scores obtained by them are subjected to certain statistical test and a decision is made whether their IQ is up to the standard or not.

There is no end to such types of practical problems where statistical tests can be applied. Here one important point may be noted. Whatever conclusions are drawn about the population (s), they are always subjected to some error.

16.5 Steps of Test of Significance

Various steps in test of significance are as follows:

(i) Set up the Null hypothesis H_o.

(ii) Set up the alternative hypothesis H_1 .This will decide whether to go for single tailed test or two tailed test.

(iii) Choose the appropriate level of significance depending upon the reliability of the estimates and permissible risk. This is to be decided before sample is drawn.

(iv) Compute the test statistic $Z = \frac{t - E(t)}{S.E.(t)}$.

(v) Compare the computed value of Z in previous step with the significant value Z_α at a given level of significance.

(vi) Conclusion :

a) If $|Z| < Z_\alpha$ i.e. if calculated value of Z (test statistic) is less than Z_α, we say it is not significant, null hypothesis is accepted at level of significance α.

b) If $|Z| > Z_\alpha$ i.e. if calculated value of Z (test statistic) is greater than Z_α, we say it is significant and null hypothesis is rejected at level of significance α.

Lesson 17

Z-TEST AND ITS APPLICATIONS

17.1 Introduction

In the previous lesson we encountered a problem to decide whether our sample observations have come from a postulated population or not. On the basis of sample observations, a test is performed to decide whether the postulated hypothesis is accepted or rejected and this involves certain amount of risk. The amount of risk is termed as a level of significance. When the hypothesis is rejected, we consider it as a significant result and when a reverse situation is encountered, we consider it as a non-significant result. We have seen that for large values of n, the number of trials, almost all the distributions e.g., Binomial, Poisson etc. are very closely approximated by Normal distribution and in this case we apply Normal Deviate test (Z-test). In cases where the population variance (s) is/are known, we use Z-test. The distribution of Z is always normal with mean zero and variance one. In this lesson we shall be studying the problem relating to test of significance for large samples only. In statistics a sample is said to be large if its size exceeds 30.

17.2 Test of Significance for Large Samples

In cases where the population variance(s) is/are known, we use Z-test. Moreover when the sample size is large, sample variance approaches population variance and is deemed to be almost equal to population variance. In this way, the population variance is known even if we have sample data and hence the normal deviate test is applicable. The distribution of Z is always normal with mean zero and variance one. Thus, if $X \sim N(\mu, \sigma^2)$

$$\text{then,} \qquad Z = \frac{X - \mu}{\sqrt{v(X)}} = \frac{X - E(X)}{\sigma} \sim N(0,1)$$

From normal probability tables, we have

$P[-3 \le Z \le 3] = P[|Z| \le 3] = 0.9973 \Rightarrow P[|Z| > 3] = 1 - P[|Z| \le 3] = 0.0027$. Thus, the value of Z=3 is regarded as critical or significant value at all levels of significance. Thus if $|Z| > 3$, Ho is always rejected. If $|Z| < 3$, we test its significance at certain level of significance usually at 5% and sometimes at 1% level of significance. Also $P[|Z| > 1.96] = 0.05$ and $P[|Z| > 2.58] = 0.01$.Thus, significant values of Z at 5% and 1% level of significance are 1.96 and 2.58 respectively. If $|Z| > 1.96$, H_o is rejected at 5% level of significance if $|Z| < 1.96$, H_o may be retained at 5% level of significance. Similarly $|Z| > 2.58$, H_o is rejected at 1% level of significance and if $|Z| < 2.58$, H_o is retained at 1% level of significance. In the following sections we shall discuss the large sample (normal) tests for attributes and variables.

17.3 Applications of Z Test

17.3.1 Test for single proportion

If the observations on various items or objects are categorized into two classes c_1 and c_2 (binomial population), viz. defective or not defective item, we often want to test the hypothesis, whether the proportion of items in a particular class, viz., defective items is P_0 or not. For example, the management of a dairy plant is interested in

knowing that whether the population of leaked pouches filled by automatic milk filling machine is one percent. Thus for binomial population, the hypothesis we want to test is whether the sample proportion is representative of the Population proportion P = P_0 against H_1: P≠P_0 or H_1: P>P_0 or H_1: P<P_0 can be tested by Z-test where P is the actual proportion of items in the population belonging to class c_1. Proportions are mostly based on large samples and hence Z-test is applied.

If X is the number of successes in n independent trials with constant probability P of success for each trial then E (X) = P and V (X) = n P Q where Q = 1−P. It is known that for large n, the Binomial distribution tends to Normal distribution. Hence , for large n, X~N (nP, nPQ). Therefore, Z statistic for single proportion is given by

$$Z = \frac{X - E(X)}{SE(X)} = \frac{X - E(X)}{\sqrt{V(X)}}$$

$$Z = \frac{(X - nP)}{\sqrt{nPQ}} \sim N(0,1)$$

and we can apply a normal deviate test.

If in a sample of size n, X be the number of persons possessing the given attributes then observed proportion of successes $^{x}/_{n} = p$

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} nP = P$$

$$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} nPQ = \frac{PQ}{n}$$

$$S.E.(p) = \sqrt{\frac{PQ}{n}}$$

Since X and consequently $\frac{X}{n}$ is asymptotically normal for large n, the normal deviate test for the proportion of success becomes.

$$Z = \frac{p - E(p)}{SE(p)} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

Example 1. In a large consignment of baby food packets, a random sample of 100 packets revealed that 5 packets were leaking. Test whether the sample comes from the population (large consignment) containing 3 percent leaked packets.

Solution: In this example n=100, X=5, P=0.03, $p = {}^{x}/_{n} = \frac{5}{100} = 0.05$

H_0: P = 0.03 .i.e., the proportion of the leaked pouches in the population is 3 per cent

H_1: P ≠ 0.03.

Here, we shall use standard normal deviate (Z) test for single proportion as under

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.05 - 0.03}{\sqrt{\frac{(0.03)(0.97)}{100}}} = \frac{0.02}{0.01706} = 1.17$$

Since calculated value of Z statistic is less than 1.96 therefore H_0 is not rejected at 5% level of significance which implies that the sample is representative of the population (large consignment) of packets containing 3% leaked packets.

17.3.2 Test of Significance for difference of proportions

If we have two populations and each item of a population belong to either of the two classes C_1 and C_2. A person is often interested to know whether the proportion of items in class C_1 in both the populations is same or not that is we want to test the hypothesis.

H_0: $P_1=P_2$ against H_1: $P_1\neq P_2$ or $P_1>P_2$ or $P_1<P_2$ where P_1 and P_2 are the proportions of items in the two populations belonging to class C_1.

Let X_1, X_2 be the number of items belonging to class C_1 in random samples of sizes n_1 and n_2 from the two populations respectively. Then the sample proportion

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

If $P_1$and P_2 are the proportions then $E(p_1) = P_1, E(p_2) = P_2$

$$V(p_1) = \frac{P_1Q_1}{n_1},\quad V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for the large sample, p_1, and p_2 are asymptotically normally distributed, (p_1-p_2) is also normally distributed. Therefore, the Z-statistic for difference between two proportions is given by

$$Z = \frac{(p_1-p_2) - E(p_1-p_2)}{\sqrt{V(p_1-p_2)}} \sim N(0,1)$$

$$\text{Since, } E(p_1-p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0$$

$$V(p_1-p_2) = V(p_1) + V(p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$$

$$Z = \frac{(p_1-p_2)}{\sqrt{V(p_1-p_2)}} = \frac{(p_1-p_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

Since $P_1 = P_2 = P$ and $Q_1 = Q_2 = Q$, therefore

$$Z = \frac{p_1-p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

If the population proportion P_1 and P_2 are given to be distinctly different that is $P_1 \neq P_2$, then

$$Z = \frac{(p_1-p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

In general P, the common population proportion (under H_o) is not known, then an unbiased estimate of population proportion 'P' based on both the samples is used and is given by

$$\bar{P} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Example 2. Before an increase in excise duty on tea, 400 people out of a sample of 500 persons were found to be tea drinkers. After an increase in excise duty, 400 people were observed to be tea drinkers in a sample of 600 people. Test whether there is a significant change in the number of tea drinkers after increase in excise duty on tea.

Solution: In this example $X_1 = 400$, $n_1 = 500$, $X_2 = 400$, $n_2 = 600$

H_0: $P_1 = P_2$ i.e., there is no change in the number of tea drinkers after increase in excise duty on tea

H_1: $P_1 \neq P_2$

Here we shall use standard normal deviate (Z) test for difference of proportions as under:

In our example $p_1 = 400/500 = 0.8$, $p_2 = 400/600 = 0.6667$

$q_1 = 1 - p_1 = 0.2$, $q_2 = 1 - p_2 = 0.333$

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{0.8 - 0.6667}{\sqrt{\frac{(0.8)(0.2)}{500} + \frac{(0.6667)(0.333)}{600}}} = \frac{0.1333}{0.0263} = 5.07$$

Since calculated value of Z statistic is greater than 3, therefore H_0 is rejected at all levels of significance which implies that there is a significant change in the number of tea drinkers after increase in excise duty on tea. It is further observed that the number of tea drinkers have significantly declined after increase in excise duty on tea which is due to decrease in thevalue of p_2 (0.667) from the value of p_1 (0.8).

Example 3. A machine turns out 16 imperfect articles in a sample of 500. After overhauling it turns 3 imperfect articles in a batch of 100. Has the machine improved after overhauling?

Solution : We are given n_1=500and n_2=100

p_1= Proportions of defective items before overhauling of machine =16/500=0.032

p_2= Proportions of defective items after overhauling of machine =3/100=0.03

H_0: $P_1 = P_2$ i.e. the machine has not improved after overhauling.

H_1: $P_1 > P_2$

$$Z = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{0.032 - 0.03}{\sqrt{\frac{(0.032)(0.968)}{500} + \frac{(0.3)(0.7)}{100}}} = \frac{0.002}{0.01878} = 0.106$$

Since Z<1.645 (Right –tailed test), it is not significant at 5% level of significance. Hence we may accept the null hypothesis and conclude that the machine has not improved after overhauling.

17.3.3 Test for significance of single mean

We have seen that if X_i (i=1, 2, …., n) is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean $\bar{X}$ is distributed normally with mean μ and variance σ^2/n i.e.,$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ .Thus for large samples normal variate corresponding to $\bar{X}$ is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

In test of significance for a single mean we deal the following situations

1) To test if the mean of the population has a specified value (μ_0) and null hypothesis in this case will be H_0: $\mu=\mu_0$ i.e., the population has a specified mean value.
2) To test whether the sample mean differs significantly from the hypothetical value of population mean with null hypothesis as there is no difference between sample mean ($\bar{X}$) and population mean (μ).
3) To test if the given random sample has been drawn from a population with specified mean μ_0 and variance σ^2 with null hypothesis the sample has been drawn from a normal population with specified mean μ_0 and variance σ^2

In all the above three situations the test statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

If $|Z|$<1.96, H_o is not rejected at 5% level of significance which implies that there is no significant difference between sample mean and population mean and whatever difference is there, it exists due to fluctuation of sampling.

$|Z|$>1.96, H_o is rejected at 5% level of significance which implies that there is a significant difference between sample mean and population mean. The above situations are illustrated by following examples:

Example 4. A random sample of 100 students gave a mean weight of 64 kg with a standard deviation of 16 kg. Test the hypothesis that the mean weight in the population is 60 kg.

Solution: In this example, n=100, μ=60 kg., $\bar{X}$=64 kg., σ=16

H_0: μ=60 kg. , i.e. the mean weight in the population is 60 kg.

We shall use standard normal deviate (z) test for single mean as under:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{64 - 60}{16/\sqrt{100}} = 2.5$$

Since calculated value of Z statistic is more than 1.96, it is significant at 5% level of significance. Therefore, H_0 is rejected at all levels of significance which implies that mean weight of population is not 60 kg.

Example 5. A sample of 50 cows in a herd has average lactation yield 1290 litres. Test whether the sample has been drawn from the population having herd average lactation yield of 1350 litres with a standard deviation of 65 litres.

Solution: In this example, n=50, μ=1350 litres, $\bar{X}$ =1290, σ=65

H_0: μ=1350 litres i.e., the mean lactation milk yield of the cows in the population is 1350

H_1: $\mu \neq$1350 litres

We shall use standard normal deviate (Z) test for single mean as under:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{1290-1350}{65/\sqrt{50}} = -6.53 \Rightarrow |Z| = 6.53$$

Since calculated value of Z statistic is more than 3, it is significant at all levels of significance. Therefore, H_0 is rejected at all levels of significance which implies that the sample has not been drawn from the population having mean lactation milk yield as 1350 litres or there is a significant difference between sample mean and population mean.

17.3.4 Test of significance for difference of means

Let $\bar{X}_1$ be the mean of a sample of size n_1 drawn from a population with mean μ_1 and variance ${\sigma_1}^2$ and let $\bar{X}_2$ be the mean of an independent sample of size n_2 drawn from another population with mean μ_2 and variance ${\sigma_2}^2$. Since sample sizes are large.

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ and } \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Also $(\bar{X}_1 - \bar{X}_2)$, being the difference in means of two independent normal variates is also a normal variate. The standard normal variate corresponding to $\bar{X}_1 - \bar{X}_2$ is given by

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - E(\bar{X}_1 - \bar{X}_2)}{\sqrt{v(\bar{X}_1 - \bar{X}_2)}} \sim N(0,1)$$

Under the null hypothesis H_o: $\mu_1 = \mu_2$ i.e., the two population means are equal, we get

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2 = 0$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The covariance terms vanish, since the sample means $\bar{X}_1$ and $\bar{X}_2$ are independent.

Thus under H_o: $\mu_1 = \mu_2$, the Z statistic is given by

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Here ${\sigma_1}^2$ and ${\sigma_2}^2$ are assumed to be known. If they are unknown then their estimates provided by corresponding sample variances ${s_1}^2$ and ${s_2}^2$ respectively are used, i.e., $\hat{\sigma}_1^2 = {s_1}^2$ and $\hat{\sigma}_2^2 = {s_2}^2$, thus, in this case the test statistic becomes

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

Remarks: If we want to test whether the two independent samples have come from the same population i.e., if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (with common S.D. σ), then under $H_o : \mu_1 = \mu_2$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

If the common variance σ^2 is not known, then we use its estimate based on both the samples which is given by

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

Example 6. In a certain factory there are two independent processes manufacturing the same item. The average weight in a sample of 100 items produced from one process is found to be 50g with a standard deviation of 5g while the corresponding figures in a sample of 75 items from the other process are 52g and 6g respectively. Is the difference between two means significant?

Solution: In this example, $n_1 = 100, \bar{X}_1 = 50\ g, s_1 = 5\ g;\ n_2 = 75, \bar{X}_2 = 52\ g, s_2 = 6\ g.$

Let μ_1 and μ_2 be the population mean of the weight of items manufactured by two independent processes.

H_0: $\mu_1 = \mu_2$, i.e., mean weights of the items manufactured by two independent processes in the population is same.

H_0: $\mu_1 \neq \mu_2$

Here, we shall use standard normal deviate test (Z-test) for calculating difference between two means as under

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{50 - 52}{\sqrt{\frac{25}{100} + \frac{36}{75}}} = \frac{-2}{0.8544} = -2.34 \Rightarrow |Z| = 2.34$$

Since calculated value of Z statistic is more than 1.96, therefore, H_0 is rejected at 5% level of significance which implies that there is a significant difference between mean weights of the items obtained from two manufacturing processes.

Lesson 18

t-TEST AND ITS APPLICATIONS

18.1 Introduction

The various tests of significance discussed in the previous lesson were related to large samples. The large sample theory was based on the application of 'Normal deviate test'. However if sample size n is small (n<30), the distribution of the various statistics, e.g., $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ are far from normality and as such 'Normal deviate test' cannot be applied if n is small. Hence to deal with small samples, new techniques and tests of significance known as 'exact sample tests' were developed which were pioneered by W. S. Gosset (1908) who wrote under the pen name of Student and later on developed and extended by Professor R. A. Fisher (1926). From practical point of view, a sample is small if its size is less than 30. In this lesson we shall discuss Student's t-test. In exact sample tests, the basic assumption is that "the population(s) from which sample(s) are drawn is (are) normal i.e., the parent population(s) is (are) normally distributed and sample(s) is (are) random and independent of each other. The exact sample tests can be used even for large samples but large sample theory cannot be used for small samples.

18.2 Student's t

Definition

Let X_i (i=1,2,…,n) be a random sample of size n drawn from a normal population with mean μ and variance σ^2, then student's t is defined by the statistic.

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$\text{where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

where S^2 is an unbiased estimate of the population variance σ^2, and it follows student's t distribution with (n-1) degrees of freedom.

Therefore (n-1) S^2=n s^2

18.3 Applications of t-test

The t-test has number of applications in statistics which are discussed in following sections

- t-test for significance of single mean, population variance being unknown
- t-test for the significance of the difference between two means, the population variances being equal
- t-test for significance of an observed sample correlation coefficient.

18.3.1 t-Test for single mean

Suppose we want to test

(i) If the given normal population has a specified value of the population mean μ_0.

(ii) If the sample mean differs from the specified value μ_0 of the population mean.

(iii) If a random sample of size n viz., X_i (i=1,2,…, n) has been drawn from a normal population with specified mean μ_0.

Basically all the above three problems are same with corresponding null hypothesis H_o as follows

(i) $\mu = \mu_0$ i.e., the population mean is μ_0

(ii) There is no difference between the sample mean $\bar{x}$ and the population mean μ.

(iii) The given sample has been drawn from the population with mean μ_0.

The test statistic is given by

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$\text{where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

follows student's t distribution with (n-1) degrees of freedom. If calculated |t| > tabulated value of t at 5 percent level of significance viz., $t_{0.05}$; (n-1) d.f. then H_o is rejected at 5 per cent level of significance which implies that there is a significant difference between sample mean and population mean or the sample has not been drawn from the population having specified mean $\mu = \mu_0$. If calculated |t| < tabulated value of t at 5 percent level of significance viz., $t_{0.05}$; (n-1) d.f. then H_o is accepted. This is explained with the help of following illustrations.

Example .1: A random sample of 9 values from a normal population showed a mean of 41.5 and the sum of squares of deviations from the mean equal to 72.Test whether the assumption of mean 44.5 in the population is reasonable.

Solution: In this problem n=9 μ=44.5 , $\bar{X}$=41.5 and $\sum_{i=1}^{9}(X_i - \bar{X})^2 = 72$

H_0: μ=44.5 i.e., population mean is 44.5

H_1: $\mu \neq 44.5$

Applying t-test

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{8}(72) = 9$$

$$t = \frac{41.5 - 44.5}{3/\sqrt{9}} = -3$$

Tabulated value of t at 5% level of significance and 8 d.f. =2.306. Since the calculated value of |t| is greater than tabulated value 2.306, hence it is significant. We reject null hypothesis and conclude that the population mean is not equal to 44.5.

Example 2: An automatic machine was expected to fill 250 ml of flavored milk in the pouches. A random sample of pouches was taken and the actual content of milk was weighed. Weight of flavored milk (in ml.) is

253, 251, 248, 251, 252, 250, 249, 254, 247, 249, 248, 255, 245, 246, 254.

Do you consider that the average quantity of flavored milk in the sample is the same as that of adjusted value?

Solution : In this problem n=15 μ=250 ml.

H_0: μ=250 ml i.e., automatic machine on an average fills 250 ml milk in each pouch

H_1: μ ≠ 250

Prepare the following table

Table 18.1

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
253	2.8667	8.2178
251	0.8667	0.7511
248	-2.1333	4.5511
251	0.8667	0.7511
252	1.8667	3.4844
250	-0.1333	0.0178
249	-1.1333	1.2844
254	3.8667	14.9511
247	-3.1333	9.8178
249	-1.1333	1.2844
248	-2.1333	4.5511
255	4.8667	23.6844
245	-5.1333	26.3511
246	-4.1333	17.0844
254	3.8667	14.9511
3752	0.0000	131.7333

$$\bar{X} = \frac{3752}{15} = 250.1333\,, \qquad S^2 = \frac{131.7333}{14} = 9.4095$$

Applying t-test

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{250.1333 - 250}{0.7920} = 0.1683$$

Tabulated value of t at 5% level of significance for 14 d.f. is 2.15. Since the calculated value of $|t|$ is less than tabulated value 2.15, hence it is not significant. We accept null hypothesis and conclude that the on an average automatic machine fills 250 ml. of flavored milk in pouches.

18.3.2 t-Test for difference of means

Suppose we want to test if two independent samples X_i (i=1,2,…,n_1) and Y_j(j=1,2,…,n_2) of sizes n_1 and n_2 have been drawn from two normal populations with means μ_1 and μ_2 respectively. Under the Null hypothesis H_o: $\mu_1 = \mu_2$ i.e. ,that the samples have been drawn from the populations having same mean .

H_1: μ ≠ μ_0

The t- statistic is given by

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which follows t distribution with $(n_1 + n_2 - 2)$

$$\text{where } \bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \text{ and } \bar{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2\right]$$

is an unbiased estimate of the common population variance σ^2 based on both the samples. By comparing the computed value of t with the tabulated value of t for $(n_1 + n_2 - 2)$ d.f. and at desired level of significance, we reject or retain null hypothesis H_o

18.3.2.1 Assumptions for difference of means test

(i) Parent populations from which the samples have been drawn are normally distributed.

(ii) The two samples are random and independent of each other.

(iii) The population variances are equal $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but unknown.

Thus before applying t-test for testing the equality means, it is theoretical desirable to test the equality of population variances by applying F-test. If the hypothesis H_o: $\sigma_1^2 = \sigma_2^2$ is rejected then we cannot apply t-test and in such situations Behren's d test is applied. This procedure is explained with the help of following illustrations.

Example 3 : The prices of ghee were compared in two cities. For this purpose ten shops were selected at random in each city. The following table gives per kg. prices of ghee in two cities:

City A	361	363	356	364	359	360	362	361	358	357
City B	368	369	370	366	367	365	371	372	366	367

Test whether the average price of ghee is of the same order in two cities.

Solution :

Null hypothesis H_o: $\mu_A = \mu_B$ i.e., average price of ghee is of same order in cities A and B .

H_1: $\mu_A \neq \mu_B$

Prepare the following table:

Table 18.2

City A			City B		
X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	Y_j	$Y_j - \bar{Y}$	$(Y_j - \bar{Y})^2$
361	0.9	0.81	368	-0.1	0.01
363	2.9	8.41	369	0.9	0.81
356	-4.1	16.81	370	1.9	3.61

364	3.9	15.21	366	-2.1	4.41
359	-1.1	1.21	367	-1.1	1.21
360	-0.1	0.01	365	-3.1	9.61
362	1.9	3.61	371	2.9	8.41
361	0.9	0.81	372	3.9	15.21
358	-2.1	4.41	366	-2.1	4.41
357	-3.1	9.61	367	-1.1	1.21
3601		60.9	3681		48.9

and calculate,

$$\bar{X} = \frac{3601}{10} = 360.1 \text{ and } \bar{Y} = \frac{3681}{10} = 368.1$$

$$S^2 = \frac{1}{18}[60.9 + 48.9] = \frac{109.8}{18} = 6.1$$

$$t = \frac{(\bar{X} - \bar{Y})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{360.1 - 368.1}{1.1045} = -7.2428$$

Tabulated value of t at 5% level of significance and 18 d.f. (for two–tail) is 2.10. Since the calculated value of $|t|$ is more than tabulated value (2.10), hence it is significant. We reject null hypothesis at 5 percent level of significance and conclude that average prices of ghee in both the cities are different.

18.3.3 Paired t-test

Let us now consider the case when

(i) Sample sizes are equal i.e., $n_1 = n_2 = n$ and

(ii) The samples are not independent but the sample observations are paired together i.e., the pair of observations (X_i, Y_i) i=1,2,…,n corresponds to the same i^{th} sample unit. The problem is to test if the sample means differ significantly or not.

For example suppose we want to test the efficacy of a particular drug say for inducing sleep or controlling blood pressure or blood sugar among the patients or if we want to test the difference between two analysts or machines with regard to detection of mean fat percentage in milk. Let X_i and Y_i (i=1,2,…,n) be the readings of fat percentage of i^{th} milk sample, detected by two machines A and B respectively. Here instead of applying the difference of the means test discussed in previous section, we apply paired t-test.

Here we consider the difference $d_i = X_i - Y_i$ (i=1,2,…,n)

Under the Null hypothesis H_o difference in fat percent in milk by both the machines is due to fluctuations of sampling i.e., $H_0: \mu_d = 0$

against $H_1: \mu_d \neq 0$

then the test statistic

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

follows t distribution with (n-1) degrees of freedom

$$\text{where } \bar{d} = \frac{1}{n}\sum_{i=1}^{n} di$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i{}^2 - \frac{(\sum_{i=1}^{n} d_i)^2}{n}\right]$$

Different examples of paired t test are:

1. A sample of boys was given a test mathematics. They were given a month's extra coaching and a second test was held at the end of it? Do the marks give evidence that the students have been benefitted by the extra coaching?
2. A sample of patients was examined to know whether a drug tends to reduce the blood pressure. The data give the blood pressure readings before the drug was given and also after it was given. The question is to examine whether the drug is effective in controlling blood pressure.
3. It is desired to test the adoption of a new technology by the farmers. A group of farmers is taken where the knowledge level score is measured before the new technology is infused and after infusion of technology, the knowledge level score is again measured. Do the difference in technology level scores provide the evidence that the farmers have been benefitted by the adoption of new technology.

This procedure is explained with the help of following illustrations.

Example 4: Ten B.Tech. (Dairy Tech.) second year students were selected for a training on quality control on the basis of marks obtained in an examination conducted for this purpose . After one month training they were given a test and marks were recorded out of 50.

Student	A	B	C	D	E	F	G	H	I	J
Before training	25	20	35	15	42	28	26	44	35	48
After training	26	20	34	13	43	40	29	41	36	46

Test whether there is any change in performance after the training.

Solution:

In this problem, the marks obtained by the students before training (X) and after training (Y) are not independent but paired together, hence we shall apply paired t test

Null Hypothesis H_o: $\mu_X=\mu_Y$ or H_0: $\mu_d = 0$ i.e., mean scores before training and after training are same . In other words, the training has no impact on students' performance.

against H_1: $\mu_d \neq 0$.

Preapare the following table

Table 18.3

Before training (X_i)	After training(Y_i)	$d_i = X_i - Y_i$	d_i^2

25	26	-1	1
20	20	0	0
35	34	1	1
15	13	2	4
42	43	-1	1
28	40	-12	144
26	29	-3	9
44	41	3	9
35	36	-1	1
48	46	2	4
Total		$\sum d_i = -10$	$\sum d_i^2 = 174$

and calculate

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} di = -1$$

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^{\,2} - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right] = \frac{1}{9}\left[174 - \frac{(-10)^2}{10}\right] = 18.2222$$

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{-1}{1.3498} = -0.7408$$

Tabulated value of t at 5% level of significance and 9 d.f. (for two–tail) is 2.262. Since the calculated value of |t| is less than tabulated value 2.262, hence it is not significant. We accept null hypothesis and conclude that students have not been benefited from the training.

Example 5: A certain stimulus administered to each of 12 calves resulted in the following changes in the blood sugar levels 5, 2, 8,-1, 3,0, -2, 1, 5, 0, 4, 6 .

Can it be concluded that the stimulus will in general be accompanied by increase in blood sugar level? Test at 5% level of significance.

Solution: In this problem we are given the increments $d_i = X_i - Y_i$ in the blood sugar levels of 12 calves

Null Hypothesis H_o: $\mu_X = \mu_Y$ or $\mu_d = 0$, i.e., there is no difference in blood sugar levels of the calves before and after the administering drug. In other words, the stimulus has no impact on blood sugar levels of calves.

Against H_1: $\mu_X < \mu_Y$ or $\mu_d < 0$ .i.e., the stimulus results in increase in blood sugar level of calves.

Preapare the following table:

d_i	5	2	8	-1	3	0	-2	1	5	0	4	6	$\sum d_i = 31$
d_i^2	25	4	64	1	9	0	4	1	25	0	16	6	$\sum d_i^2 = 185$

and calculate

where $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{31}{12} = 2.5833$

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^{\,2} - \frac{(\sum_{i=1}^{n} d_i)^2}{n}\right] = \frac{1}{11}\left[185 - \frac{(31)^2}{12}\right] = 9.5379$$

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2.5833}{0.8915} = 2.8977$$

Tabulated value of t at 10% level of significance and 14 d.f. is 1.80 [in this problem the alternative hypothesis is right tailed hence to test at 5% level of significance we have to see the t table at 10% level of significance]. Since the calculated value of $|t|$ is greater than tabulated value 1.80, hence it is significant. We reject null hypothesis and conclude that the stimulus is effective in increasing blood sugar in calves.

18.3.4 t-Test for significance of an observed sample correlation coefficient

Let a random sample (x_i, y_i) (i=1,2---,n) of size n has been drawn from a bivariate normal distribution and let r be the observed sample correlation coefficient . In order to test whether sample correlation coefficient r is significant or there is no correlation between the variables in the population. Prof. R. A. Fisher proved that under the null hypothesis Ho: ρ=0 i.e. the population correlation coefficient is zero. The statistic

$$t = \frac{r}{\sqrt{1-r^2}}\, x\sqrt{n-2} \sim t_{n-2}$$

follows student's t distribution with (n-2) d.f., n being the sample size.

Lesson 19

CHI-SQUARE TEST AND ITS APPLICATIONS

19.1 Introduction

In the preceding lesson some tests of significance were discussed which is based on the assumption that the samples were drawn from normally distributed population. These tests are known as parametric tests as the testing procedure involves the assumption about the type of population of parameters. There are, however, many situations where it is not possible to assume a particular type of population distribution from which the samples are drawn. This leads to development of alternative techniques known as non-parametric or distribution –free methods. Chi-square (χ^2) (pronounced as Ki-square) test is one of the important non-parametric test and one of the most commonly used test of significance. The chi-square test dates back to 1900, when Prof. Karl Pearson used it for frequency data classified into k-mutually exclusive categories. It is also a frequently used test in genetics, where one tests whether the observed frequencies in different crosses agree with the expected frequencies or not. The chi-square test is applicable to test the hypothesis of the variance of a normal population, goodness of fit of the theoretical distribution to the observed frequency distribution, in a one way classification having k-categories. It is also applied for the test of independence of attributes, when the frequencies are presented in a two-way classification called the contingency table. In this lesson we give chi-square test of various hypotheses.

19.2 Chi-Square Distribution

If X is a normal variate with mean μ and standard deviation σ viz., X ~ N (μ, σ^2) then $Z = \frac{X-\mu}{\sigma}$ is a standard normal variate .The square of a standard normal variate i.e., $\left(\frac{X-\mu}{\sigma}\right)^2$ is known as Chi-square (χ^2) variate with one degree of freedom (d.f.). If $X_1, X_2, ---, X_n$ are n independent random variate following normal distribution with means $\mu_1, \mu_2, ---, \mu_n$ and standard deviations $\sigma_1, \sigma_2, ---, \sigma_n$ respectively then the variate

$$\chi^2 = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2 - \mu_2}{\sigma_2}\right)^2 + ---+ \left(\frac{X_n - \mu_n}{\sigma_n}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

which is sum of the squares of n independent standard normal variates , follows Chi-square distribution with n d.f.

19.3 Applications of the χ^2 -distribution:

Chi-square distribution has a number of applications which are enumerated below:

a) To test if the population has a specified value of the variance σ^2.

b) Chi-square test of goodness of fit.

c) Chi-square test for independence of attributes.

19.3.1 Chi-Square Test for Population Variance:

Suppose on the basis of previous knowledge, we have a preconceived value of population variance σ_0^2. Suppose we draw a random sample of size n from this population. On the basis of n sample observations ($X_1, X_2, \ldots, X_n$), the population variance value σ_0^2 of the population variance σ^2 is to either be substantiated or refuted with the help of a statistical test. In this case we use Chi-square test. For this the null hypothesis is taken as

$H_0: \sigma^2 = \sigma_0^2$

$H_1: \sigma^2 \neq \sigma_0^2$

and is tested by the statistic

$$\chi^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2} = \frac{n s^2}{\sigma_0^2} \quad i = 1,2,3, \ldots, n$$

follows a χ^2 distribution with $(n-1)$ degrees of freedom

Where $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the variance of the sample. If calculated value of chi-square is more than or equal to tabulated chi-square value at α% level of significance then H_0 is rejected at α level of significance otherwise if calculated χ^2 is less than tabulated χ^2 at α level of significance then H_0 is not rejected at α level of significance.

Example 1: An owner of a big firm agrees to purchase the product of a factory, if the produced items do not have variance of more than 0.5mm^2 in their length. To make sure of the specifications, the buyer selects a sample of 18 items from his lot. The length of each item was measured in mm which are given as under:

18.57	18.10	18.61	18.32	18.33	18.46	18.37	18.64	18.58
18.12	18.34	18.57	18.22	18.63	18.43	18.34	18.43	18.63

Test whether the sample has been drawn from the population having specified variance not more than value of 0.5

Solution : On the basis of the sample data, the hypothesis

$$H_0; \sigma^2 = 0.5 \text{ aganist } H_1: \sigma^2 > 0.5$$

can be tested by the statistic

$$\chi^2 = \frac{\sum_i (X_i - \bar{X})^2}{\sigma_0^2} \quad i = 1,2 \ldots, 18$$

Calculate, $\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}$

For the given data, $\sum_i X_i^2 = 6112.64; \sum_i X_i = 331.69$

$$\therefore \sum_i (X_i - \bar{X})^2 = 6112.64 - \frac{(331.69)^2}{18} = 6112.640 - 6112.125 = 0.515$$

.

Thus, $\chi^2 = \frac{0.515}{0.5}$ = 1.03

Tabulated value of χ^2 at 5% level of significance and 17 d.f. viz., $\chi^2_{0.05,17}$ is 27.587. Since the calculated value of χ^2 (1.03) is less than 27.857, we don't reject the null hypothesis at α = 0.05. i.e. , $\sigma^2 = 0.5$, It means that the buyer should purchase the lot having specified variance not more than value of 0.5 mm^2 length **.**

19.3.2 Chi- square test of goodness of fit

A very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as "chi-square test of goodness of fit". This test is used for testing the significance of discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis. It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

Under the null hypothesis that there is no significant difference between the observed (experimental) values and the theoretical values i.e., there is good compatibility between theory and experiment , Karl Pearson proved that the statistic

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_i} + \frac{(O_2 - E_2)^2}{E_2} + - - - + \frac{(O_n - E_n)^2}{E_n} = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i}\right]$$

follows the $\chi^2 -$ distribution with $(n-1)$ d.f. where O_i (i=1,2,…….,n) is a set of observed (experimental) frequencies and E_i (i=1,2,……..,n) be the corresponding set of expected (theoretical) frequencies obtained under some theory or hypothesis .

If calculated value of χ^2 is less than the corresponding tabulated value at (n-1) d.f. then it is said to be non-significant at the required level of significance. This implies that the discrepancy between experimental (observed) values and the theoretical (expected) values obtained under some theory or hypothesis may be attributed to chance. In other words, data do not provide us any evidence against the null hypothesis and we may conclude that there is good correspondence (fit) between theory and experiment. On the other hand if calculated value of χ^2 is greater than the corresponding tabulated value at (n-1) d.f. then it is said to be significant at the required level of significance. This implies that the discrepancy between experimental (observed) values and the theoretical (expected) values obtained under some theory or hypothesis cannot be attributed to chance and we reject the null hypothesis. In other words we conclude that the experiment does not support the theory.

Remarks:

a) The observed and expected frequencies are subjected to a linear constraint

$\sum_{i=1} O_i = \sum_{i=1} E_i = N$, where N is the total frequency since it does not involve squares and higher powers of the frequencies. $\sum_{i=1}(O_i - E_i) = \sum_{i=1} O_i - \sum_{i=1} E_i$

=N-N =0 i.e. the sum of deviations of the observed and expected frequencies is always zero.

b) Sometimes the following formula is useful for computation of χ^2

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] = \sum_{i=1}^{n}\frac{O_i^2 + E_i^2 - 2O_iE_i}{E_i}\chi^2 = \sum_{i=1}^{n}\frac{O_i^2}{E_i} + \sum_{i=1}^{n}E_i - 2\sum_{i=1}^{n}O_i$$

$$\because \sum_{i=1}^{n}O_i = \sum_{i=1}^{n}E_i = N \qquad \chi^2 = \sum_{i=1}^{n}\frac{O_i^2}{E_i} - N$$

where N is the total frequency.

c) χ^2-test depends only on the observed and expected frequencies and on degree of freedom (n-1) .It does not make any assumption regarding the parent population from which the observations care taken .Since χ^2 does not involve any population parameter it is known as statistic and the test is known as Non-parametric test or Distribution Free test.

***19.3.2.1 Degrees of freedom**:*

The number of independent variates which makes up the statistic (e.g., χ^2) is known as the degrees of freedom (d.f.). The no. of degrees of freedom in general is the total no. of observations minus the no. of independent linear constraints imposed on the observations. e.g., if k is the no. of independent constraints imposed on a set of data of n observations then d.f. = (n−k).Thus in a set of n observations usually, the degrees of freedom for χ^2 are (n – 1), one d.f. being lost because of the linear constraint. $\Sigma_{i=1}O_i = \Sigma_{i=1}E_i = N$. If 'r' independent linear constraints are imposed on the cell frequencies, then the d.f. are reduced by 'r'.

In addition if any of the population parameters (s) is (are) calculated from the given data and used for computing the expected frequencies then in applying χ^2 test of goodness of fit, we have to subtract one d.f. for each parameter estimated.

19.3.2.2 Conditions for the validity of χ^2 test

Following are the conditions which should be satisfied before χ^2 test can be applied.

(i) N the total number of frequencies must be large, greater than50.

(ii) The sample observations should be independent.

(iii) No theoretical cell frequency should be small. Five should be regarded as the very minimum and 10 is better. The chi-square distribution is essentially a continuous distribution but it cannot maintain its character of continuity if the cell frequency is less than 5). If any theoretical cell frequency is less than 5, then for the application of χ^2 test, it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjusts for the d.f. lost in the pooling.

(iv) The constraints on the cell frequencies, if any, should be linear. Constraints which involve linear especially in the cell frequencies are called linear constraints such as $\Sigma_i O_i = \Sigma_i E_i = N$.

The above procedure is explained through following examples:

Example 2 : The following table gives the number of coli forms per ml in thousand bottles of sterilized milk:

No of coli forms (X_i)	0	1	2	3	4	5	6	7	8	9	10
No. of bottles (f_i)	2	8	46	116	211	243	208	119	40	7	0

Fit a binomial distribution to the above data and test the goodness of fit.

Solution

The fitting of this problem is already explained in example 10.6 of lesson 10 (module 3) .In the usual notations we have: n=10, N=1000, $\Sigma f_i X_i$=4971, $\bar{X} = 4.971$,

$p = \frac{4.971}{10} = 0.4971, q = 0.5029, \frac{p}{q} = \frac{0.4971}{0.5029}$=0.9985

putting r=0,1,2,3,---,10 in $f(r) = Nx^nC_r\, p^r q^{n-r}$ we get the expected frequency as given in the following table:

Table 19.1

No. of coliforms	**No. of bottles (f_i) (O_i)**		**Expected Frequency E i**		**(O_i - E_i)**	**(O_i - E_i)^2**	$\frac{(O_i - Ei)^2}{E_i}$
0	2		1				
1	8	10*	10	11*	-1	1	0.0909

2	46		46		0	0	0.0000
3	116		120		-4	16	0.1333
4	211		207		4	16	0.0773
5	243		246		-3	9	0.0366
6	208		203		5	25	0.1232
7	119		115		4	16	0.1391
8	40		42		-2	4	0.0952
9	7	7*	9	10*	-3	9	0.9000
10	0		1				
Total	1000		1000				1.5956

* in the above table since some of expected cell frequencies being less than 5 , therefore, they have been merged with either preceding or succeeding frequencies. Accordingly the corresponding observed frequencies have also been merged

Thereafter the value of χ^2 is calculated as follows

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] = 1.5956$$

Required Degrees of freedom : 11-1-1-2=7 (n=11,one d.f. is lost due to linear constraint $\sum_{i=1} O_i = \sum_{i=1} E_i = N$;one d.f. is lost because the parameter p of binomial distribution is estimated from the given data;2 d.f. are lost due to pooling first and last two frequencies). Tabulated value of χ^2 for 7 d.f. and at 5% level of significance is 14.067.Since calculated value of χ^2 is less than tabulated value it is not significant. Thus we conclude that Binomial distribution is a good fit to the given data or Binomial distribution fits well to the given data.

Example 3: The following table gives the number of lactations completed by 1000 cows of Tharparker breed:

No of lactations (X_i)	0	1	2	3	4	5	6	7	8	9	10
No. of females (f_i)	300	205	155	126	90	47	35	18	13	8	3

Fit a Poisson distribution to the above data and test its goodness of fit ..

Solution :

The fitting of this problem is already explained in example 11.6 of lesson 11 (module 3) . In the usual notations we have : N=1000, $\sum f_i X_i$=2030, $\bar{X} = 2.03 = m$

putting r=0,1,2,3,---,10 in

$$f(r) = nx\frac{e^{-m}m^r}{r!}$$

we get the expected frequency as given in the following table

Table 19.2

No. of lactation (X_i)	No. of Females (f_i) (Oi)		Expected Frequency (E_i)		$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - Ei)^2}{E_i}$
0	300		131		169	28447.71	216.6034
1	205		267		-62	3795.93	14.2377
2	155		271		-116	13365.74	49.3911
3	126		183		-57	3261.89	17.81356
4	90		93		-3	8.58	0.092368
5	47		38		9	85.94	2.277851
6	35	77*	13	18*	59	3481.00	193.3889
7	18		4				
8	13		1				
9	8		0				
10	3		0				
Total	1000		1000				493.8049

* in the above table since some of expected cell frequencies being less than 5 , therefore, they have been merged with either preceding or succeeding frequencies. Accordingly the corresponding observed frequencies have also been merged

Thereafter the value of χ^2 is calculated as follows

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right] = 611.4467$$

Required Degrees of freedom: 11-1-1-4=5 (n=11,one d.f. is lost due to linear constraint $\sum_{i=1} O_i = \sum_{i=1} E_i = N$;one d.f. is lost because the parameter m of Poisson distribution is estimated from the given data; 4 d.f. are lost due to pooling last five frequencies). Tabulated value of χ^2 for 5 d.f. and at 5% level of significance is 11. 07.Since calculated value of χ^2 is more than tabulated value it is significant. Thus, we conclude that Poisson distribution is not a good fit to the given data or Poisson distribution does not fit well to the given data.

19.3.3 Independence of attributes

Let us consider two attributes A and B, A divided into r classes $A_1,A_2,\ldots\ldots\ldots,A_r$ and B divided into s classes $B_1,B_2,\ldots\ldots B_s$. (Such a classification in which attributes are divided into more than two classes is known as manifold classification). The various cell frequencies can be expressed in table known as r×s (manifold) contingency table where (A_i) is the number of persons possessing the attributes A_i (i = 1,2,……,r), (B_j) is the number of persons possessing the attributes B_j, (j=1,2,……..,s) and $(A_i B_j)$ is the number of persons possessing both the attributes A_i and B_j, (i=1,2,…….,r; j=1,2,……,s). Also

$$\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N \text{ is the total frequency}$$

The problem is to test if two attributes A and B under consideration are independent or not.
Under the null hypothesis that the attributes are independent, the theoretical cell frequencies are calculated as

$$E(A_1B_1) = \frac{(A_1)(B_1)}{N} \ldots\ldots\ldots E(A_iB_j) = \frac{(A_i)(B_j)}{N} \qquad \begin{matrix} i = 1,2,\ldots..,r \\ j = 1,2,\ldots..,s \end{matrix}$$

Hence expected frequency for any of the cell frequencies can be obtained by multiplying the row totals and column totals in which the frequency occurs and then dividing the product by the total frequency N.

Table 19.3 rxs Manifold Contingency Table

A_i / B_j	A_1	A_2	---	A_i	---	A_r	Total
B_1	(A_1B_1)	(A_2B_1)	---	(A_iB_1)	---	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	---	(A_iB_2)	---	(A_rB_2)	(B_2)
⋮	⋮	⋮		⋮		⋮	⋮
B_s	(A_1B_s)	(A_2B_s)	---	(A_iB_s)	---	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	---	(A_i)	---	(A_r)	$\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$

Now for rxs observed frequencies $(A_i B_j)$ and the corresponding expected frequencies $E(A_i B_j)$. Applying χ^2 test of goodness of fit, the chi-square statistic is given by

$$\chi^2 = \sum\frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\{A_iB_j - E(A_iB_j)\}^2}{E(A_iB_j)}\right] = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\left\{A_iB_j - \frac{(A_i)(B_j)}{N}\right\}^2}{\frac{(A_i)(B_j)}{N}}\right]$$

follows χ^2 distribution with (r – 1) (s – 1) d.f.

Comparing this calculated value with the tabulated value for (r-1)(s-1) d.f. and at certain level of significance , we reject or accept the null hypothesis of independence of attributes at that level of significance.

19.3.3.1 Degrees of Freedom for rxs contingency table

In (r×s) contingency table in calculation of the expected frequencies, the row totals, the column totals and the grand total remain fixed. Further $\sum A_i = \sum B_j = N$.Further since the total number of cell frequencies is (r×s), the required number of degrees of freedom is

d.f. = rs – (r + s –1) =(r–1) (s–1)

19.3.3.2 2×2 contingency table

Under the null hypothesis of independence of attributes, the value of χ^2 for the 2x2 contingency table

			Total
	a	b	a+b
	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

is given by

$$\chi^2 = \frac{N\,(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

which follows Chi-square distribution with (2-1)(2-1)= 1 degree of freedom .

19.3.3.3 Yate's correction for continuity for 2×2 contingency table

In 2×2 contingency table, the no. of d.f. is (2−1) (2−1) =1. If any one of theoretical cell frequency is less than 5, then the using of pooling method for χ^2 results in χ^2 with 0 d.f. (since 1 d.f. is lost in pooling) which is meaningless. In this case we apply "Yate's correction for continuity" which says that add 1/2 to cell frequency which is less than 5 and then adjust for remaining cell frequencies accordingly. The modified formula forχ^2 is as follows:

$$\chi^2 = \frac{N\left[|ad-bc|-\frac{N}{2}\right]^2}{(a+c)(b+d)(a+b)(c+d)}$$

If N is large then Yate's correction will make a very little difference and this can be applied in 2x2 contingency table only.

19.3.3.4 2x r Contingency table:

Under the hypothesis of independence of attributes, the value of χ^2 for 2xr contingency table:

	2xr Contingency Table					Total
	a_1	a_2	a_3	----	a_r	$n_1 = \sum_{i=1}^{r} a_i$
	b_1	b_2	b_3	----	b_r	$n_2 = \sum_{i=1}^{r} b_i$
Total	m_1	m_2	m_3	----	m_r	$N = n_1 + n_2 \; or \sum_{i=1}^{r} m_i$

can be computed from Brandt and Snedecor formula :

$$\chi^2 = \frac{N^2}{n_1 x n_2}\left[\frac{a_1^2}{m_1}+\frac{a_2^2}{m_2}+\frac{a_3^2}{m_3}+--+\frac{a_r^2}{m_r}-\frac{n_1^2}{N}\right] = \frac{N^2}{n_1 x n_2}\left[\sum_{i=1}^{r}\frac{a_i^2}{m_i}-\frac{n_1^2}{N}\right]$$

Other form of this formula is

$$= \frac{1}{pq}\left[\sum a_i p_i - n_1 p\right] \text{ where } p = \frac{n_1}{N}, q = \frac{n_2}{N} \text{ and } p_i = \frac{a_i}{m_i}$$

which is χ^2 distribution with (2-1)(k-1)=k-1 d.f.

The above procedure is explained through following examples:

Example 4: A milk producer's union wishes to test whether the preference pattern of consumers for its product is dependent on income levels. A random sample of 500 individuals gives the following data:

Table 19.4

Income	Product Preferred			
	Product A	Product B	Product C	Total
Low	170	30	80	280
Medium	50	25	60	135
High	20	10	55	85
Total	240	65	195	500

Can you conclude that the preference patterns are independent of income levels?

Solution: Let us take the hypothesis that preference patterns are independent of income levels. On the basis of this hypothesis, the expected frequencies corresponding to different rows and columns shall be:

$$E_{11} = \frac{240 \times 280}{500} = 134.4 \qquad E_{12} = \frac{65 \times 280}{500} = 36.4$$

$$E_{21} = \frac{240 \times 135}{500} = 64.8 \qquad E_{22} = \frac{65 \times 135}{500} = 17.55$$

and so on.

The expected frequencies would be as follows:

Table 19.5

134.40	36.40	109.20	280
64.80	17.55	52.65	135
40.80	11.05	33.15	85
240	65.00	195.00	500

Applying χ^2-test:

Table 19.6

O_i	E_i	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
170	134.60	1267.36	9.430
50	64.80	219.04	3.380
20	40.80	432.64	10.604
30	36.40	40.96	1.125
25	17.55	55.50	3.162
10	11.05	1.10	0.099
80	109.20	852.64	7.808
60	52.65	54.02	1.026
55	33.15	477.42	14.402
			$\sum (O_i - E_i)^2/E_i = 51.036$

$$\therefore \chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = 51.036$$

Degree of freedom = v =(r - 1) (c - 1) = (3 - 1) (3 - 1) = 4

The tabulated value of χ^2 for 4 d.f. at 5% level of significance i.e., $\chi^2_{0.05.4} = 9.488$

Since the calculated value of χ^2 is greater than the table value, therefore, we reject the null hypothesis and hence conclude that preference patterns are not independent of income levels.

Example 5: In an experiment of cattle from tuberculosis, the following were obtained:

	Affected	Not Affected	Total
Inoculated	4	20	24
Not inoculated	6	50	56
Total	10	70	80

Calculate χ^2 and discuss the effect of vaccine in controlling susceptibility to tuberculosis.

Solution : N=80

H_o: The vaccine is not effective in controlling susceptibility to tuberculosis.

H_1: The vaccine is effective in controlling susceptibility to tuberculosis.

Since one of the observed frequency is less than 5. Applying Yate's correction, we increase the value of that observed frequency by 0.5 and adjust other frequencies. The adjusted observed frequencies after Yate's correction will be as follows:

Affected	Not Affected	Total
Inoculated	4+0.5=4.5	20-0.5=19.5	24
Not inoculated	6-0.5=5.5	50+0.5=50.5	56
Total	10	70	80

$$\chi^2 = \frac{N\left[|ad-bc|-\frac{N}{2}\right]^2}{(a+c)(b+d)(a+b)(c+d)} = \frac{80\left[|4.5x50.5-19.5x5.5|-\frac{80}{2}\right]^2}{(10)(70)(24)(56)} = \frac{512000}{940800} = 0.5442$$

The tabulated value of χ^2 for 1 d.f. at 5% level of significance i.e., $\chi^2_{0.05.1} = 3.84$

Since the calculated value of χ^2 is less than the tabulated value, we accept the null hypothesis and hence conclude that the vaccine is not effective in controlling susceptibility to tuberculosis.

Example 6: A milk product factory is bringing out a new product. In order to map out its advertising campaign, it wants to determine whether product appeals equally to all age groups. The following table gives the number of persons who liked or disliked the product in different age groups.

Table 19.7 No. of persons

	Age group (in years)				
Preference	< 20	20-30	30-40	> 40	Total
Liked	75	70	60	55	260
Disliked	25	30	40	45	140

Can it be reasonably concluded that the new product appeals equally to all age groups?

Solution:

N = 400

H_0: The new product appeals equally to all age groups.

H_1: The new product does not equally appeal to all the age groups.

Table 19.8

	Age group (in years)				
Preference	< 20	20-30	30-40	> 40	Total
Liked	a_1=75	a_2=70	a_3=60	a_4=55	n_1=260
Disliked	b_1=25	b_2=30	b_3=40	b_4=45	n_2=140
	m_1=100	m_2=100	m_3=100	m_4=100	N=400

$$\chi^2 = \frac{400^2}{260x140}\left[\frac{75^2}{100}+\frac{70^2}{100}+\frac{60^2}{100}+\frac{55^2}{100}-\frac{260^2}{400}\right]$$

$$= 4.3956(2.5)=10.989$$

The tabulated value of χ^2 for 3 d.f. at 5% level of significance i.e., $\chi^2_{0.05.3} = 7.815$

Since the calculated value of χ^2 is greater than the tabulated value, we reject the null hypothesis and hence conclude that the new product did not equally appeal to all the age groups.

Lesson 20

F- TEST AND ITS APPLICATIONS

20.1 Introduction

A large number of research experiments are conducted to examine the effect of various factors on the production and quality attributes of milk and milk products. F-test is used either for testing the hypothesis about the equality of two population variances or the equality of two or more population means. The equality of two population means was dealt with t-test. Besides a t-test, we can also apply F-test for testing equality of two population means. Sir Ronald A. Fisher defined a statistic Z which is based upon ratio of two sample variances. In this lesson we will consider the distribution of ratio of two sample variances which was worked out by G.W. Snedecor.

20.2 F-Statistic

Let X_{1i} (i=1,2,…,n_1) be a random sample of size n_1 from the first population with variance σ_1^2 and X_{2j} (j=1,2, …,n_2) be another independent random sample of size n_2 from the second normal population with variance σ_2^2. The F- statistic is defined as the ratio of estimates of two variances as given below:

$$F = \frac{S_1^2}{S_2^2}$$

where, $S_1^2 > S_2^2$ and are unbiased estimates of population variances which are given by:

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$$

It follows Snedecor's F- distribution with (n_1-1, n_2-1) d.f. i.e., F~F$(n_1 - 1, n_2 - 1)$.

Further, if X is a χ^2-variate with n_1 d.f. and Y is another independent χ^2-variate with n_2 d.f., then F-statistic is defined as: $F = \frac{X/n_1}{Y/n_2}$ i.e. F-statistic is the ratio of two independent Chi-square variates divided by their respective degrees of freedom. This statistic follows G.W. Snedecor F distribution with (n_1,n_2) d.f. The sampling distribution of F-statistic does not involve any population parameter and depends only on the degrees of freedom n_1 and n_2.

20.3 Application of F- Distribution

F-distribution has a number of applications in statistics, some of which are given below

a) F-test for equality of population variances
b) F test for testing equality of several population means

20.3.1 F-test for equality of population variances

Suppose we are interested to find if two normal populations have same variance. Let X_{1i} (i=1,2,…,n_1) be a random sample of size $n_1$from the first population with variance σ_1^2 and X_{2j} (j=1,2,…,n_2) be another independent random sample of size n_2 from the second normal population with variance σ_2^2. The two samples are independent of each other.

Under the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e., population variances are equal. In other words the two independent estimates of the common population variance do not differ.
The test statistic

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \quad (S_1^2 > S_2^2)$$

Where S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 and are given by

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$$

It follows Snedecor's F- distribution with (n_1-1, n_2-1) d.f. i.e., F~F$(n_1 - 1, n_2 - 1)$. Since F-test is based on the ratio of two variances it is also known as Variance Ratio Test.
If calculated value of F is greater than tabulated value at α% level of significance H_0 is rejected at α percent level of significance which implies that the two independent estimates of population variances are heterogeneous. On the other hand if calculated value of F is less than tabulated F then H_0 is not rejected at α percent level of significance suggesting that the estimates of population variances are homogenous.

Remarks:

- Since the available tables of the significant values of F are for the right –tail test, i.e., against an alternative $\sigma_1^2 > \sigma_2^2$, in numerical problems we will take greater of the variance S_1^2 or S_2^2 in the numerator and adjust for the degrees of freedom i.e., degree of freedom of the large variance must be taken in the numerator while computing F.
- In numerical problems, usually sample variance s^2 is given from which S^2 can be obtained on using the relation n s^2 = (n-1) S^2.

The procedure is illustrated by following examples

Example 1: In a sample of 8 observations, the sum of squared deviations of items from the mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference is significant at

5% level of significance.

Solution: Let us take the hypothesis that there is no difference in the variances of two samples i.e., H_0: $\sigma_1^2 = \sigma_2^2$, with an alternative H_0: $\sigma_1^2 \neq \sigma_2^2$

we are given $n_1 = 8,\ \sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 = 94.5$

$n_2 = 10,\ \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2 = 101.7$

$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 = \frac{94.5}{7} = 13.5$; $S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2 = \frac{101.7}{9} = 11.3$

Applying F test

$F = \frac{S_1^2}{S_2^2} = \frac{13.5}{11.3} = 1.195$

Tabulated value of F (7, 9) d.f. at 5% level of significance is 3.29.

Calculated value of F is less than the tabulated value, it is not significant. Hence H_0 may be accepted and conclude that the difference in the variances of two samples is not significant at 5% level of significance which implies that the two population variances are homogenous or the sample have been drawn from the population having same variances.

Example 2: Two random samples drawn from normal population are :

Sample I	20	16	26	27	23	22	18	24	25	19		
Sample II	27	33	42	35	32	34	38	28	41	43	30	37

Obtain estimates of the variances of two populations and test whether two populations have same variances.

Solution: Let us take the hypothesis that the two populations have the same variance i.e., H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$. With an alternative H_0: $\sigma_1^2 \neq \sigma_2^2$

Table 20.1

Sample I X_1	$(X_1 - \bar{X_1})$ $\bar{X_1} = 22$	$(X_1 - \bar{X_1})^2$	**Sample II** X_2	$(X_2 - \bar{X_2})$ $\bar{X_2} = 35$	$(X_2 - \bar{X_2})^2$
20	-2	4	27	-8.00	64
16	-6	36	33	-2.00	4
26	4	16	42	7.00	49
27	5	25	35	0.00	0
23	1	1	32	-3.00	9
22	0	0	34	-1.00	1
18	-4	16	38	3.00	9
24	2	4	28	-7.00	49
25	3	9	41	6.00	36
19	-3	9	43	8.00	64

			30	-5.00	25
			37	2.00	4
220		120	420		314

$$n_1 = 10, \bar{X}_1 = \frac{220}{10} = 22 \quad \sum_{i=1}^{n_1}(X_1 - \bar{X}_1)^2 = 120$$

$$n_2 = 12, \bar{X}_2 = \frac{420}{12} = 35 \quad \sum_{i=1}^{n_2}(X_2 - \bar{X}_2)^2 = 314$$

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_1 - \bar{X}_1)^2 = \frac{120}{9} = 13.3333;$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_2 - \bar{X}_2)^2 = \frac{314}{11} = 28.5455$$

Applying F test

$$F = \frac{S_2^2}{S_1^2} = \frac{28.5455}{13.3333} = 2.1409$$

Tabulated value of F(11,9) d.f. at 5% level of significance is 3.10.

Calculated value of F is less than the tabulated value, it is not significant. Hence H_0 may be accepted and conclude that the difference in the variances of two samples is not significant at 5% level of significance which implies that the two populations have the same variance.

20.3.2 F test for equality of several population means

This test is widely used in the technique of analysis of variance which plays a very important and fundamental role in Design of experiments which is discussed in details in next module.

Lesson 21

ONE WAY CLASSIFICATION

21.1 Introduction

The t-test enables us to test the significance of the difference between two sample means but if we have a number of means and we need to test the hypothesis that the means are homogenous or that there is no difference among the means, then the technique known as Analysis of variance developed by Professor R. A. Fisher in 1920 is useful. Initially the technique was used in agricultural experiments but now a days it is widely used in almost all the branches of agricultural and animal sciences. This technique is used to test whether the differences between the means of three or more populations is significant or not. By using the technique of analysis of variance, we can test whether moisture contents of paneer or khoa prepared by different methods or batches differ significantly or not. Analysis of variance thus enables us to test on the basis of sample observations whether the means of three or more populations are significantly different or not. Thus basic purpose of the analysis of variance is to test the homogeneity of several means and the technique consists in splitting up the total variation into component variation due to independent factors where each of the components give us the estimate of population variation. In other words, in this technique, the total sum of squares is decomposed into sum of squares due to independent factors and the remaining is attributed to random causes or commonly called due to error.

21.2 Analysis of Variance

The term 'Analysis of Variance' was introduced by Prof. R.A. Fisher in 1920's to deal with problem in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes, and (ii) Chance causes.

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human being and cannot be accounted for separately.

21.2.1 Definition

According to Prof. R. A. Fisher, Analysis of Variance (ANOVA) is the "Separation of variance ascribable to one group of causes from the variance ascribable to other group." Thus, ANOVA consists in the estimation of the amount of variation due to each of the independent factors (causes) and the remaining due to chance factor (causes), the later being known as experimental error or simply error. The technique of the Analysis of variance consisting in splitting up the total variation into component variation due to independent factors where each of the components gives us the estimate of the population variance. The total sum of squares is broken up into sum of squares due to independent factors and the remaining is attributed to random causes or commonly called due to error. Consider, for instance, an industrial problem such as the following. A factory produces components, many machines being at work on the same operation. The process is not purely mechanical, the machine operators having an influence on the quality of the output. Moreover it is thought that on certain days of the week (e.g. Monday) the output is found to be of poorer quality than on other days (e.g. Friday). The quality

therefore depends on at least three factors, the machine, the operator and the day of the week. There may be other factors in operation and some of the factors mentioned may have no significant effect. It will be possible by the technique of analysis of variance whether any of the above factors, or some combinations of these has an appreciable effect on the quality and also to estimate the contribution made by each factor to the overall variability in the production or quality of product. Thus the purpose of the analysis is to establish relations of 'Cause' and effect.

21.2.2 Assumptions in analysis of variance

For the validity of the F-test in ANOVA, the following assumptions are made:

(i) The samples are drawn from the population randomly and independently.

(ii) The data are quantitative in nature and are normally distributed. Parent population from which observations are taken is normal.

(iii) Various treatments and environmental effects are additive in nature.

(iv) The population from where the samples have been drawn should have equal variance σ^2. This is known as **Homoscedasticity** and can be tested by Barttlet's test.

21.3 One-way Analysis of Variance

The simplest type of analysis of variance is known as one way analysis of variance, in which only one source of variation or factor of interest is controlled and its effect on the elementary units is observed. It is an extension of three or more samples of the t-test procedure for use with two independent samples. In other words t-test for use with two independent samples is a special case of one-way analysis of variance. In typical situation one–way classification refers to the comparison of means of several univariate normal population, having the same unknown variance σ^2, on the basis of random samples selected from each population. The population means are denoted by $\mu_1, \mu_2, ..., \mu_k$, if there are k populations. The one way analysis of variance is designed to test the null hypothesis:

$H_o : \mu_1 = \mu_2 = ... = \mu_k$ i.e. the arithmetic means of the population from which the k samples have been randomly drawn are equal to one another.

Let us suppose that N observations $X_{ij}, (i = 1,2,..,k; j = 1,2,..,n_i)$ of a random variable X are grouped on some basis, into k classes (T_1,T_2,---, T_k) of sizes $n_1, n_2, \dots, n_k$ respectively $N = \sum_{i=1}^{k} n_i$ as exhibited below:

Table 21.1

Treatment						Means	Total
T_1	X_{11}	X_{12}	.....	X_{1n_1}		$\bar{X}_{1\cdot}$	$T_{1\cdot}$
T_2	X_{21}	X_{22}	.....	X_{2n_2}		$\bar{X}_{2\cdot}$	$T_{2\cdot}$
	. .	. .		. .		. .	. .
T_i	X_{i1}	X_{i2}	.....	X_{in_i}		$\bar{X}_{i\cdot}$	$T_{i\cdot}$
	.	.	.			.	.

	.	.	.			.	.
T_k	X_{k1}	X_{k2}	.....	X_{kn_k}		$\bar{X}_{k\cdot}$	$T_{k\cdot}$
							G

The total variation in the observations X_{ij} can be split into the following two components:

(i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.

(ii) The variation within the classes, i.e., the inherent variation of the random variable within the observations of class

The first type of variation is due to assignable causes which can be detected and controlled by human being and the second type of variation is due to chance causes which are beyond the control of human being.

The main object of analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability within the separate classes.

In particular, let us consider the effect of k brands of yoghurt on price of yoghurt of N shops / retail stores (of same type) divided into k brands/classes of sizes $n_1, n_2, \ldots, n_k$ respectively, $N = \sum_{i=1}^{k} n_i$

Here the sources of variations are

(i) Effect of the brands

(ii) Error 'e' produced by numerous causes of such magnitude that they are not detected and identified with the knowledge that we have and they together produce a variation of random nature obeying Gaussian (Normal) law of errors.

21.3.1 Mathematical model

The linear mathematical model will be

$X_{ij} = \mu_{ij} + e_{ij}$

$X_{ij} = \mu + \alpha_i + e_{ij}$ (i=1,2,…,k) (j=1,2,…,n_i)

where X_{ij} is the value of the variate in the j^{th} observation (j=1,2,…,n_i) belonging to i^{th} class (i=1,2,…,k)

μ is the general mean effect

α_i is the effect due to i^{th} class where $\alpha_i = \mu_i - \mu$

e_{ij} is random error which is assumed to be independently and normally distributed with mean zero and variance σ_e^2.

Let the mean of k populations be $\mu_1, \mu_2, \ldots, \mu_k$ then our aim is to test null hypothesis

$H_o : \mu_1 = \mu_2 = \ldots = \mu_k = \mu$ which reduces to $H_o : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$.

H_1 : At least one pair of μ_i's is not equal.

21.3.2 Calculation of different sum of squares

a) **Total Sum of Squares (TSS)** $=\sum_i \sum_j \left(X_{ij} - \bar{X}_{..}\right)^2 = \sum_i \sum_j X_{ij}^2 - \frac{\left(\sum_i \sum_j x_{ij}\right)^2}{N}$

$$= \sum_i \sum_j X_{ij}^2 - \frac{(G)^2}{N}$$

where G is the grand total of all the observations and $N = n_1 + n_2 + \cdots + n_k$

The expression $\sum_i \sum_j X_{ij}^2$ i.e. , sum of squares of all the observations is known as Raw Sum of Squares (R.S.S.) and the expression $\frac{(G)^2}{N}$ is called Correction Factor (C.F.)

b) **Sum of Squares Among Classes (SSC):** To find the SSC, divide the squares of sum of each class by their class size or number of observations in each class and find their sum and thereafter, Subtract the correction factor from this sum i.e.,

$$SSC = \left[\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + - - - + \frac{T_k^2}{n_k}\right] - \frac{(G)^2}{N} = \sum_i \frac{T_i^2}{n_i} - C.F.$$

where $T_{i.}$ is the total of the observations pertaining to the i^{th} class.

c) **Sum of Squares within classes (SSE)**: It is obtained by subtracting sum of squares among the classes from the total sum of squares i.e., SSE=TSS-SSC.

This sum of squares is also called error sum of squares denoted by SSE.

d) **Mean Sum of Squares (M.S.S.):** It is obtained by dividing sum of squares by their respective degrees of freedom.

e) Analysis of Variance Table

The results of the above calculations are presented in a table called Analysis of Variance or ANOVA table as follows:

Table 21.2

Source of variation	Degree of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares (M.S.S.)	F-Ratio
Among Classes	k-1	SSC	$S_C^2 = \frac{SSC}{k-1}$	$\frac{S_C^2}{S_E^2} \sim F(k-1, N-k)$

Within Classes (Error)	N-k	SSE	$S_E^2 = \frac{SSE}{N-k}$	
Total	N-1	TSS		

If the calculated value of F is greater than the tabulated value of F α;(k-1, N-k), where α denotes the level of significance, the hypothesis H_0 , is rejected and can be inferred that the class effects are significantly different from one another.

Standard Error

a) The estimated standard error of any class/treatment mean, say i^{th} treatment/class mean, is given by

b)

$$SE_d = \sqrt{\frac{S_E^2}{n_i}}$$

where S_E^2 is the mean sum of squares within samples or MSS(Error)

c) The estimated standard error of the difference i^{th} and j^{th} treatment mean, is

d)

$$SE_d = \sqrt{S_E^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

where n_i and n_j are the number of observations for i^{th} and j^{th} treatment/class

c) If $n_i = n_j$ =n then S.E. of difference of means is

$$SE_d = \sqrt{\frac{2\,S_E^2}{n}}$$

d) The Critical Difference (C.D.) or Least Significant Difference (L.S.D.) can be calculated as

C.D. = SE_d xt $_{\alpha,(N-k)}$ where α is level of significance and (N-k) is the d.f. for error .

The treatment means are $\overline{X}_i = \frac{T_i}{n_i}$ $\forall$ i=1,2,…,k

These can be compared with the help of critical difference. Any two treatments means are said to differ significantly if the difference is larger than the critical difference (CD).The procedure of one way ANOVA is illustrated through the following example:

Example 1: The following table gives the moisture contents of Paneer prepared by four methods Manual(M_1), Mechanical with pressure 10 pound/inch^2 (M_2);with pressure 12 pound/inch^2 (M_3) and pressure 15 pound/inch^2 (M_4).

Table 21.3

Methods			
M_1	M_2	M_3	M_4
50.3	54.1	57.5	52.3
52.2	53.7	56.3	53.2
52.5	55.5	55.8	53.6
51.7	54.6	56.9	53.4
52.6		55.8	53.8
		59.6	

Analyze the data to find whether the mean moisture content in paneer is different prepared by different methods.

Solution:

H_o : $\mu_1=\mu_2=\mu_3=\mu_4$ i.e., the mean moisture content in paneer prepared by different methods is same.

H_1 : Mean moisture content in paneer prepared by at least two methods are not equal.

Prepare the following table to calculate sum of squares due to different components:

Table 21.4

Methods	**M_1**	**M_2**	**M_3**	**M_4**	**Total**
Total ($T_{i.}$)	259.30	217.90	341.90	266.30	G=1085.40
No. of observations (n_i)	5	4	6	5	20
Mean	51.8600	54.4750	56.9833	53.2600	

$$\text{Correction Factor (CF)} = \frac{(G)^2}{N} = \frac{(1085.4)^2}{20} = 58904.66$$

$$\text{Total Sum of Squares (TSS)} = \sum_i \sum_j X_{ij}^{\ 2} - CF$$

$$= (50.3)^2 + (52.2)^2 + - - - + (53.4)^2 + (53.8)^2 - 58904.66$$

=59000.2200-58904.66=95.5620

Sum of Squares among Classes (SSC) or Sum of Squares between Methods=

$$\sum_i \frac{T_{i.}^2}{n_i} - CF = \frac{(259.30)^2}{5} + \frac{(217.9)^2}{4} + \frac{(341.9)^2}{6} + \frac{(266.30)^2}{5} - 58904.66$$

$$= 58983.14 - 58904.66 == 78.4822$$

Sum of Squares within classes (SSE) or Sum of squares due to error:

SSE=TSS-SSC= 95.5620-78.48217=17.0798

Prepare the following analysis of variance table:

Table 21.5 ANOVA Table

Source of variation	Degree of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Sum of Squares (M.S.S.)	F-Ratio
Among Methods	4-1=3	78.4822	$S_c^2 = \frac{78.4822}{3}$ ==26.1607	$F = \frac{26.1607}{1.0675}$ =24.5068
Within Methods (Error)	20-4=16	17.0798	$S_E^2 = \frac{17.0798}{16}$ =1.0675	
Total	20-1=19	95.5620		

From Fisher and Yate's tables, F value for 3 and 16 d.f. at 5% level of significance is 3.2389 Since the observed value of F in the analysis of variance table is greater than the 5 % tabulated F value, it can be inferred that mean moisture content in paneer prepared by different methods differ significantly from one another.

Calculation of critical differences for comparison among various pairs of methods of preparing paneer

Table 21.6

Methods	M_3	M_2	M_4	M_1
Mean	56.9833	54.4750	53.2600	51.8600
No. of observations	6	4	5	5

C.D.(for comparing mean moisture content prepared by Method 3 and Method 2) =

$\sqrt{1.0675\left(\frac{1}{6}+\frac{1}{4}\right)}\ x\ t_{5\%,16\ d.f.}$ =0.6669x2.12=1.4138

C.D.(for comparing mean moisture content prepared by Method 2 and Method 4) =

$\sqrt{1.0675\left(\frac{1}{4}+\frac{1}{5}\right)}\ x\ t_{5\%,16\ d.f.}$ =0.6931x2.12=1.4693

C.D.(for comparing mean moisture content prepared by Method 4 and Method 1) =

$\sqrt{1.0675\left(\frac{1}{5}+\frac{1}{5}\right)}\ x\ t_{5\%,16\ d.f.}$ =0.6534x2.12=1.3853

Conclusion

It can be concluded the moisture content of paneer prepared by different methods was found to be significantly different from each other. The mean moisture content was found to be maximum in method M_3(56.9833) followed by method M_2(54.4750) which is significantly different from each other. The next mean moisture contents was found for method M_4 (53.26) followed by method M_1(51.86) which is significantly different from each other.

Lesson 22

TWO WAY CLASSIFICATION

22.1 Introduction

In one way classification analysis of variance explained in the previous lesson the treatments constitute different levels of a single factor which is controlled in the experiment. There are , however, many situations in which the response variable of interest may be affected by more than one factor. For example milk yield of cow may be affected by differences in treatments i.e. feeds fed as well as differences in breed of the cows, moisture contents of butter prepared by churning cream may be affected with different levels of fat and churning speed etc. When two independent factors might have an effect on the response variable of interest, it is possible to design the test so that an analysis of variance can be used to test the effect of the two factors simultaneously. Such a test is called two–factor analysis of variance. In a two –way classification the data are classified according to two different criteria or factors. The procedure for analysis of variance is somewhat different than the one followed earlier while dealing with problems of one-way classification.

22.2 Two Way Classification

Let us plan the experiment in such a way so as to study the effect of two factors at a time in the same experiment. For each factor there will be a number of classes or levels. Let us consider the case when there are two factors which may affect the variate values be operators and machines. Suppose the N observations are classified into p categories (or classes) $O_1, O_2, \ldots, O_p$ according to Factor A (Operator) and into q categories $M_1, M_2, \ldots, M_q$ according to factor B(Machine) having pq combinations $A_i B_j$ ($O_i M_j$) i=1,2,---,p ;j=1,2,---,q; often called cells. This scheme of classification according to two factors is called two way classification and analysis is called two way analysis of variance. The number of observations in each cells may be equal or different, but we shall consider the case of one observation per cell so that N=pq. i.e., total number of cells is N.

Let X_{ij} be the observation on the i^{th} level of Operator (O_i) and j^{th} level of Machine (M_j) i=1,2,---,p ; j=1,2,---,q;

$T_{i.} = \sum_{j=1}^{q} X_{ij}$ = Total or Sum of the observations for i^{th} operator O_i

$$\bar{X}_{i.} = \frac{1}{q}\sum_{j=1}^{q} X_{ij} = \frac{X_{i.}}{q} = \text{meanof the observations for } i^{th} \text{ operator } O_i$$

$T_{.j} = \sum_{i=1}^{p} X_{ij}$ = Sum of the observations for jth machine M_j

$\bar{X}_{.j} = \frac{1}{p}\sum_{i=1}^{p} X_{ij} = \frac{X_{.j}}{p}$ = mean of the observations forjth machine M_j

$$G \text{ or } T_{..} = \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij} = \sum_{i=1}^{p} X_{i.} = \sum_{j=1}^{q} X_{.j} = \text{Sum of all the observations or grand total}$$

$$\bar{X}_{..} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij} = \frac{X_{..}}{N} = \text{mean of all the observations or overall mean}$$

These N observations, the marginal totals and their means can be represented in the tabular form as follows:

Table 22.1

Operators	Machines	Total	Mean
	M_1 M_2…….M_j……M_q		
O_1	X_{11} X_{12}..... X_{1j}……X_{1q}	$T_{1.}$	$\overline{X_{1.}}$
O_2	X_{21} X_{22}..... X_{2j}……X_{2q}	$T_{2.}$	$\overline{X_{2.}}$
..	.. .. ……. … ……. … …	…	…
..	.. .. …….. …… … …. ….	…	…
O_i	X_{i1} X_{i2} ..... X_{ij} ……X_{iq}	$T_{i.}$	$\overline{X_{i.}}$
..	.. .. …….. …… … …. ….	…	…
O_p	X_{p1} X_{p2} ..... X_{pj} ……X_{pq}	$T_{p.}$	$\overline{X_{p.}}$
Total	$T_{.1}$ $T_{.2}$ ……$T_{.j}$…… $T_{.q}$	G= $T_{..}$	
Mean	$\bar{X}_{.1}$ $\bar{X}_{.2}$ ..... $\bar{X}_{.j}$…… $\bar{X}_{.q}$		$\bar{X}_{..}$

22.2.1 Assumptions

i. The observations are independent random variables having normal distributions with mean μ_{ij} and common but unknown variance σ^2. Under this assumption model for this problem may be taken as

$$X_{ij} = \mu_{ij} + e_{ij}$$

where e_{ij} vary from observation to observation and are independent random variable values having normal distributions with mean zero and variance $\sigma^2 \Rightarrow E(X_{ij}) = \mu_{ij}$.

ii. The observations in the p rows are independent random samples of size q from p normal populations having mean $\mu_{1.}, \mu_{2.}, \ldots, \mu_{p.}$ and a common variance σ^2.

iii. The observations in the q columns are independent random samples of size p from q normal populations with mean $\mu_{.1}, \mu_{.2}, \ldots, \mu_{.q}$ and a common variance σ^2.

iv. The effects are additive.

Here $\mu_{i.}$ (i=1,2,---,p) are called fixed effect due to factor operators O_i ; $\mu_{.j}$ (j=1,2,---, q) are fixed effect due to the factor machines M_j .

22.2.2 Mathematical model

Here the mathematical model can be written as

$$X_{ij} = \mu_{ij} + e_{ij}$$

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

i) μ is the general mean effect given by $\mu = \sum \mu_{ij}/N$.

ii) α_i (i=1, 2,….,p) is the effect due to i^{th} operator

where $\alpha_i = \mu_{i.} - \mu$; $\mu_{i.} = \frac{1}{p}\sum_{j=1}^{q} \mu_{ij}$ $(i = 1, 2, \ldots, p)$. Obviously $\sum_{i=1}^{p} \alpha_i = 0$.

iii) β_j $(j = 1, 2, \ldots., q)$ is the effect due to j^{th} machine

where $\beta_j = \mu_{.j} - \mu$; $\mu_{.j} = \frac{1}{p}\sum_{i=1}^{p} \mu_{ij}$, $(j = 1, 2, \ldots, q)$ .Obviously $\sum_{j=1}^{q} \beta_j = 0$

iv) e_{ij}'s are independently normally distributed with mean zero and variance σ_e^2 i.e. $e_{ij} \sim N(0, \sigma_e^2)$

$$\sum_i \alpha_i = \sum_j \beta_j = 0$$

22.2.3 Null hypothesis

We set up the null hypothesis, that the operators and machines are homogeneous. In other words, the null hypothesis for operators and machines are respectively:

$H_{01}: \mu_{1.} = \mu_{2.} = \cdots = \mu_{p.}$ or $\alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$
i.e. mean output obatined from different operators is same.

$H_{02}: \mu_{.1} = \mu_{.2} = \cdots = \mu_{.q}$ or $\beta_1 = \beta_2 = \cdots = \beta_q = 0$
i.e. mean output obatined from different machines is same.

Against the corresponding hypothesis
H_{11} : at least two of the means $\mu_{i.}'s$ are not equal
H_{12} : at least two of the means $\mu_{j}'s$ are not equal

22.2.1 Computations of different sum of squares

a) **Total Sum of Squares (TSS)** $= \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_i \sum_j X_{ij}^2 - \frac{(\sum_i \sum_j X_{ij})^2}{N}$

$$= \sum_i \sum_j X_{ij}^2 - \frac{(G)^2}{N}$$

where G is the grand total of all the observations and N=pq. The expression $\sum_i \sum_j X_{ij}^2$ i.e., sum of squares of all the observations is known as Raw Sum of Squares (R.S.S.) and the expression $\frac{(G)^2}{N}$ is called Correction Factor (CF)

b) **Sum of Squares due to factor A (Operators) denoted by SSA**

To find the sum of squares due to factor A (SSA) i.e., sum of squares among the rows (SSR) divide the squares of sum of each row by the number of observations in respective rows and find their sum and thereafter, subtract the correction factor from this sum i.e.,

$$SSA(SSR) = \left[\frac{T_1^2}{q} + \frac{T_2^2}{q} + - - - + \frac{T_k^2}{q}\right] - \frac{(G)^2}{N} = \sum_i \frac{T_{i.}^2}{q} - CF$$

where $T_{i.}$ is the total of the observations pertaining to the i^{th} row.

c) **Sum of Squares due to factor B(Machines) denoted by SSB**

To find the sum of squares due to factor B (SSB) i.e. sum of squares among the columns (SSC) divide the squares of sum of each column by number of observations in respective columns and find their sum and thereafter, subtract the correction factor from this sum i.e.,

$$SSB(SSC) = \left[\frac{T_{.1}^2}{p} + \frac{T_{.2}^2}{p} + - - - + \frac{T_{.k}^2}{p}\right] - \frac{(G)^2}{N} = \sum_i \frac{T_{.j}^2}{p} - CF$$

where $T._j$ is the total of the observations pertaining to the j^{th} column.

d) **Sum of Squares due to residuals or error denoted by SSE**

The sum of squares of the residuals is obtained by subtracting sum of squares due to Factor A (SSA) and sum of squares due to factor B (SSB) from the total sum of squares (TSS) i.e., SSE=TSS-SSA-SSB.

This sum of squares is also called error sum of squares denoted by SSE.

Prepare the analysis of variance table as follows:

Table 22.2 ANOVA Table

Source of variation	d.f.	S.S.	M.S.S.	F-Ratio
Among levels of factor A (Operators)	$(p-1)$	SSA	$S_A^2 = \frac{SSA}{p-1}$	$F_1 = \frac{S_A^2}{S_E^2}$
Among levels of factor B (Machines)	$(q-1)$	SSB	$S_B^2 = \frac{SSB}{q-1}$	$F_2 = \frac{S_B^2}{S_E^2}$
Error	$(p-1)(q-1)$	SSE	$S_E^2 = \frac{SSE}{(p-1)(q-1)}$	
Total	pq – 1			

Interpretation

By comparing the values of F_1 and F_2 with the tabulated value of F for respective d.f. and at α level of significance , the null hypothesis of the homogeneity of various factor A (Operators) and various factor B (Machines) may be rejected or accepted at the desired level of significance .

Standard error

a) The estimated standard error of the difference between means of factor A i.e., between means of two operators is

$$SE_d = \sqrt{\frac{2S_E^2}{q}}$$

b) The estimated standard error of the difference between means of factor B i.e., between means of two machines is

$$SE_d = \sqrt{\frac{2S_E^2}{p}}$$

c) The Critical Difference (C.D.) or Least Significant Difference (L.S.D.) can be calculated as

C.D. = SE_dxt $_{\alpha,(p-1)(q-1)}$ where SE_d is the S.E. of difference between two means, α is level of significance and (p-1)(q-1) is the d.f. for error .

The treatment means are $\bar{X}_{i.} = \frac{T_{i.}}{q}$ ∀ i=1,2,…,p and $\bar{X}_{.j} = \frac{T_{.j}}{p}$ ∀ j=1,2,…,q

These can be compared with the help of critical difference. Any two treatments means are said to differ significantly if the difference is larger than the critical difference (CD).The procedure of two way ANOVA is illustrated through the following example:

Example 1: The average partial size of dried ice-cream mix spray powder dried by varying in-let temperature and automiser speed was measured in an experiment with 6 in-let temperatures and 4 automiser speed. The results obtained from the experiment are given below:

Table 22.3

Automiser Speed	**In-let Temperatures**					
	T_1	**T_2**	**T_3**	**T_4**	**T_5**	**T_6**
S_1	35.7	39.0	42.1	25.1	29.9	27.3
S_2	32.9	33.6	37.7	24.0	23.2	24.3
S_3	35.6	32.5	37.4	21.0	24.9	23.1
S_4	30.7	35.8	40.1	26.3	28.3	26.4

Analyze the data and discuss whether there is any significant difference between in-let temperature and automiser speed on particle size of ice-cream mix powder?

Solution :

H_{oA} : $\mu_{1.}=\mu_{2.}=\mu_{3.}=\mu_{4.}=\mu_{5.}=\mu_{6.}$ i.e., the mean particle size of ice-cream mix powder at different in-let temperature is same.

H_{1A} : At least two of the means $\mu_{i.}'s$ are not equal .

H_{oB} : $\mu_{.1}=\mu_{.2}=\mu_{.3}=\mu_{.4}$ i.e., the mean particle size of ice-cream mix powder at different automiser speeds is same.

H_{1A} : At least two of the means $\mu_j's$ are not equal .

Prepare the following two way table:

Table 22.4 Calculation of Treatments totals, means and the grand total

Automiser Speed	In-let Temperatures						Total	Mean
	T_1	T_2	T_3	T_4	T_5	T_6		
S_1	35.7	39.0	42.1	25.1	29.9	27.3	$T_{1.}$=199.1	$\overline{X_{1.}}$ =33.1833
S_2	32.9	33.6	37.7	24.0	23.2	24.3	$T_{2.}$=175.7	$\overline{X_{2.}}$ =29.2833
S_3	35.6	32.5	37.4	21.0	24.9	23.1	$T_{3.}$=174.5	$\overline{X_{3.}}$ =29.0833
S_4	30.7	35.8	40.1	26.3	28.3	26.4	$T_{4.}$=187.6	$\overline{X_{4.}}$ =31.2667
Total	$T_{.1}$ = 134.9	$T_{.2}$ = 140.9	$T_{.3}$ = 157.3	$T_{.4}$= 96.4	$T_{.5}$= 106.3	$T_{.6}$ = 101.1	G=736.9	
Mean	$\overline{X}_{.1}$ = 33.73	$\overline{X}_{.2}$ = 35.23	$\overline{X}_{.3}$= 39.33	$\overline{X}_{.4}$ = 24.10	$\overline{X}_{.5}$ = 26. 58	$\overline{X}_{.6}$ = 25.28		

$$\text{Correction factor} = \frac{(G)^2}{N} = \frac{(736.9)^2}{24} = 22625.9004$$

$$\text{Total Sum of Squares (TSS)} = \sum_i \sum_j X_{ij}^{\ 2} - \frac{(G)^2}{N}$$

$$= (35.7)^2 + (39.0)^2 + --- + (28.3)^2 + (26.4)^2 - 22625.9004$$

=23513.27-22625.9004=887.3696

Sum of Squares due to factor A (Speed)

$$\text{SSA(SS due to speed)} = \sum_i \frac{T_i^2}{q} - \text{C. F.}$$

$$= \frac{(199.1)^2 + (175.7)^2 + (174.5)^2 + (187.6)^2}{6} - 22625.9004$$

= 22692.5517-22625.9004=66.6512

c) **Sum of Squares due to factor B (In-let temperature)**

$$\text{SSB(SS due to inlet temperature)} = \sum_i \frac{T_j^2}{p} - \text{C. F.}$$

$$= \frac{(134.9)^2 + (140.9)^2 + (157.3)^2 + (96.4)^2 + (106.3)^2 + (101.1)^2}{4} - 22625.9004$$

= 23401.9925-22625.9004=776.0921

d) **Sum of Squares due to residuals (SSE)**

SSE=TSS-SSA-SSB=887.3696-66.6512-776.0921=44.62625

Prepare the following ANOVA table:

Table 22.5 ANOVA Table

Source of variation	d.f.	S.S.	M.S.S.	F-Ratio
Among levels of factor A (Speed)	(4-1)=3	66.65125	$S_A^2 = \frac{66.6512}{3}$ =22.2171	$F_1 = \frac{22.2171}{2.9751}$ =7.4677
Among levels of factor B (Temperature)	(6-1)=5	776.0921	$S_B^2 = \frac{776.0921}{5}$ =155.2184	$F_2 = \frac{155.2184}{2.9751}$ =52.1728
Error	(4-1)(6-1) =15	44.62625	$S_E^2 = \frac{44.62625}{15}$ =2.9751	
Total	24 – 1=23	887.3696		

From Fisher and Yate's tables, tabulated F values for 3 and 15 d.f. and for 5 and 15 d.f. at 5% level of significance are 3.2874 and 2.9013 respectively. Since the observed values of F for factor A (automiser speed) and factor B (in-let temperature) in the analysis of variance table are greater than the respective 5 % tabulated F value , F_1 and F_2 are significant at 5% levelof significance . Hence both the null hypothesis $H_{oA \text{ and }} H_{oB}$ are rejected at 5% level of significance.

Critical difference

C.D. (For comparison of different speed)= $\sqrt{\frac{2X2.9751}{6}} \times t_{0.05,15 \text{ d.f.}}$=0.9958x2.131=2.1221

C.D. (For comparison of different in-let temperature) = $\sqrt{\frac{2X2.9751}{4}} \times t_{0.05,15 \text{ d.f.}}$=1.2196x2.131=2.5991

Conclusion

It can be concluded that mean particle size of ice-cream mix powder differ significantly at various levels of in-let temperature as well as at various automiser speed levels. The mean particle size of ice-cream mix powder was found maximum at different auto miser speed S_1(33.1833) which is at par with speed S_4 (31.2667). Similar argument holds for speed S_4 and S_2 as well as for the S_2 and S_3 speeds. Similarly the mean particle size of ice-cream mix powder was found maximum in temperature T_3 (39.33) followed by temperature T_2(35.23) and T_1(33.73) , both are statistically at par with each other. Similar argument holds for temperature T_5, T_6 and T_4 .

Lesson 23

LINEAR CORRELATION

23.1 Introduction

In the first module we have confined our discussion to univariate distributions only i.e., the distributions involving only one variate and also saw how the various measures of central tendency, dispersion, skewness and kurtosis can be used for the purposes of comparison and analysis. We may, however, come across certain series where each item of the series may assume the values of two or more variables. Such distribution, in which each unit of the series assumes two values, is called a bivariate distribution. Further, if we measure more than two variables on each unit of a distribution, it is called a multivariate distribution. In a series, the units on which different measurements are taken may be of almost any nature such as different individuals, times, places etc.

In our day –to –day life, we find many situations when a mutual relationship exists between two variables i.e., with change (fall or rise) in the value of one variable there may be change (fall or rise) in the value of other variable. At times, an increase in one variable is accompanied by increase or decrease in the other variable. Such changes in variables suggest that there is certain relationship between them. For example, as price of a commodity increases the demand for the commodity decreases. In milk , the content of total solid decreases with decrease in fat level. Similarly with the increase in the levels of pressure, the volume of a gas decreases at a constant temperature. These facts indicate that there is certainly some mutual relationships that exist between the demand of a commodity and its price, total solid level and fat level and pressure and volume. Such association is studied in correlation analysis. The correlation is a statistical tool which measures the degree or intensity or extent of relationship between two variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

23.2 Meaning of Correlation

Correlation is a statistical technique which measures and analyses the degree or extent to which two or more variables fluctuate with reference to one another. It denotes the inter-dependence amongst variables. The degrees are expressed by a coefficient which ranges between -1 to +1. The direction of change is indicated by + or – signs; the former, refers to the movement in the same direction and the later, in the opposite direction. An absence of correlation is indicated by zero. Correlation thus expresses the relationship through a relative measure of change and it has nothing to do with the units in which the variables are expressed.

23.3 Definition of Correlation

Some important definitions are given below:

“If two or more quantities very in sympathy so that movement in the one tend to be accompanied by corresponding movements in the other (s) then they are said to be correlated”.

-L.R. Connor

“Correlation analysis attempts to determine the degree of relationship between variables”.

-Ya-Lun Chou

“When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.”

-Croxton and Cowden

“Correlation is an analysis of the covariation between two or more variables.”- **A.M. Tuttle**

23.4 Types of Correlation

23.4.1 Positive and negative correlation

If the values of the two variables deviate in the same direction i.e., if the increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, correlation is said to be positive or direct e.g., relationship between Total solids and SNF in milk, height and weight of student in the class, income and expenditure, demand and production, time and microbial growth in curd.

On the other hand, correlation is said to be negative or inverse if the variable deviate in the opposite direction i.e., if the increase (decrease) in the values of one variable results, on the average, in a corresponding decrease (increase) in the values of the other variable e.g. relationship between fat and SNF in milk, price and demand of a commodity, pressure and volume of a gas, Temperature and microbe number in milk, pH and acidity in milk.

23.4.2 Linear and non–linear correlation

The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values. In general, two variables x and y are said to be linearly related, if there exists a relationship of the form Y = a + bX between them which is the equation of a straight line with slope 'b' and which makes an intercept 'a' on the y-axis. Hence, if the values of the two variables are plotted as points in the xy-plane, we shall get a straight line. The relationship between two variables is said to be Non-linear or curvilinear if corresponding to a unit change in one variable, the other variable does not change at a constant rate but at fluctuating rate. In such cases if the data are plotted on the XY-plane we do not get a straight line but a curve. Mathematically speaking, the correlation is said to be non-linear if the slope of the plotted curve is not constant.

23.5 Methods of Studying Correlation

We shall confine our discussion to the methods of ascertaining only linear relationship between two variables (series). The commonly used methods for studying the correlation between two variables are

i. Scatter diagram method.
ii. Karl Pearson's coefficient of correlation. (Covariance method).
iii. Rank method.

23.5.1 Scatter diagram method

Scatter diagram is one of the simplest ways of diagrammatic representation of a bivariate distribution and provides us one of the simplest tools of ascertaining the correlation between two variables. Suppose we are given n pairs of values (X_1, Y_1), (X_2, Y_2),…, (X_n, Y_n) of two variables X and Y. For example if the variables X and Y denote the fat and SNF contents in milk respectively then the pairs (X_1, Y_1), (X_2, Y_2),…, (X_n, Y_n) may represent the fat and SNF contents (in pairs) of n samples of milk. These n points may be plotted as dots (.) on the X-axis and Y-axis in the XY-plane. (It is customary to take the dependent variable along the Y –axis and independent variable along the X-axis.) The diagram of dots so obtained is known as scatter diagram. From scatter diagram we can fairly obtain good, though rough, idea about the relationship between the two variables. The following points may be borne in mind in interpreting the scatter diagram (as shown in Fig. 23.1) regarding the correlation between the two variables:

a) If the points are very dense then a fairly good amount of correlation is expected. On the other hand if the points are scattered than it indicates poor correlation.

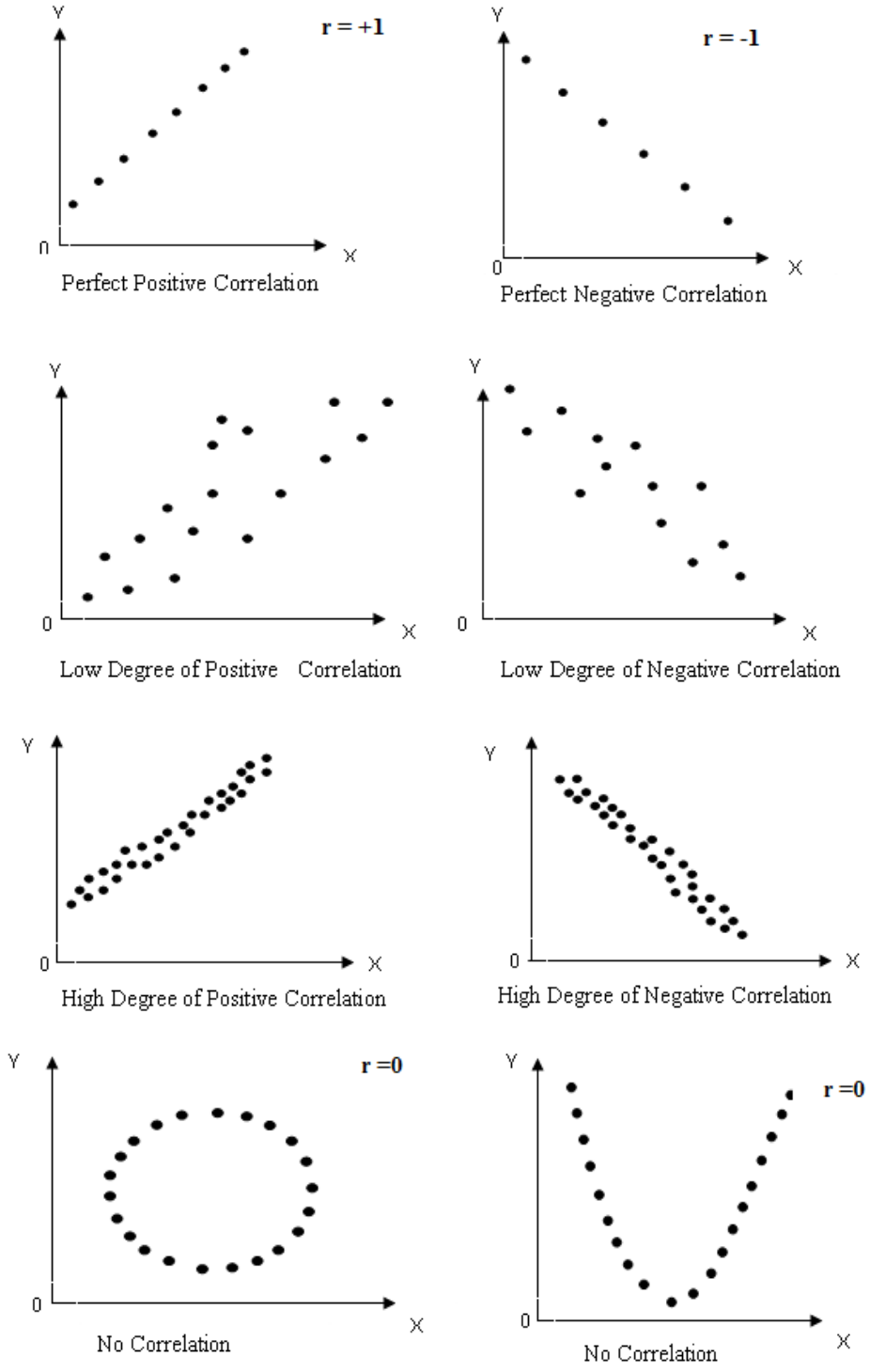


Fig. 23.1 Scatter diagram depicting different forms of correlation

b) If the point on the scatter diagram reveals any trend (either upward or downward) the variables are said to be correlated and if no trend is revealed then variables are uncorrelated.

c) If there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner then correlation is said to be positive. On the other hand the points depict a downward trend from upper left hand corner than correlation is said to be negative.

d) If all the points lie on the straight line starting from the left bottom and going up to right top. The correlation is perfect and positive.

e) On the other hand if all the points lie on a straight line starting from left top and coming down to right bottom, correlation is perfect and negative.

The major limitation of this method is that it only tells whether there exists a correlation between the variables and its direction. It does not give idea about the precise degree of relationship between the two variables.

23.5.2 Karl pearson's coefficient of correlation (Covariance method)

A mathematical method for measuring the intensity or the magnitude of linear relationship between two variable series was suggested by Karl Pearson (1867-1936), a great British Bio-metrician and Statistician and is by far the most widely used method in practice.

Karl Pearson's measure, known as Pearsonian correlation coefficient between two variables (series) X and Y, usually denoted by r(X, Y) or r_{XY} or simply r is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between X and Y, written as Cov (X, Y) to the product of the standard deviations of X and Y. Symbolically,

$$r = \frac{Cov.(X,Y)}{\sigma_X \sigma_Y} \qquad (23.1)$$

If (X_1, Y_1), (X_2, Y_2),---, (X_n, Y_n) are n pairs of observations of the variables X and Y in a bivariate distribution, then

$$Cov.(X,Y) = \frac{1}{n}\sum(X-\overline{X})(Y-\overline{Y}) \quad \sigma_X = \sqrt{\frac{1}{n}\sum(X-\overline{X})^2} \quad \sigma_Y = \sqrt{\frac{1}{n}\sum(Y-\overline{Y})^2}$$

summation being taken over n pair of observations. Substituting in (23.1) we get, (23.2)

$$r = \frac{\frac{1}{n}\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\frac{1}{n}\sum(X-\overline{X})^2 \frac{1}{n}\sum(Y-\overline{Y})^2}} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2 \sum(Y-\overline{Y})^2}} \qquad (23.3)$$

Simplifying equation (23.2)

$$\therefore Cov(X,Y) = \frac{1}{n}\sum XY - \overline{X}\,\overline{Y} - \overline{Y}\overline{X} + \overline{X}\overline{Y} = \frac{1}{n}\sum XY - \overline{X}\,\overline{Y} \qquad (23.4)$$

$$\Rightarrow Cov(X,Y) = \frac{1}{n}\sum XY - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right) =$$

$$\frac{1}{n^2}[n\sum XY - (\sum X)(\sum Y)] \qquad (23.5)$$

$$\sigma_{X^2} = \frac{1}{n}\sum(X-\overline{X})^2 = \frac{1}{n}\sum X^2 - \overline{X}^2 = \frac{1}{n}\sum X^2 - \left(\frac{\sum X}{n}\right)^2 =$$

$$\frac{1}{n^2}[n\sum X^2 - (\sum X)^2] \qquad (23.6)$$

Similarly we have,

$$\sigma_{Y^2} = \frac{1}{n^2}\left[n\sum Y^2 - \left(\sum Y\right)^2\right] \qquad (23.7)$$

Substituting (23.5), (23.6) and (23.7) in (23.1) we get

$$r = \frac{\frac{1}{n^2}[n\sum XY - (\sum X)(\sum Y)]}{\sqrt{\frac{1}{n^2}\{n\sum X^2 - (\sum X)^2\}}\sqrt{\frac{1}{n^2}\{n\sum Y^2 - (\sum Y)^2\}}}$$

$$= \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \qquad (23.8)$$

The above procedure is illustrated by the following example

Example 1 : The following data pertains to spoilage of milk (in %)(X) and the temperature (^{0}C) (Y) of storage of milk in a dairy plant.

Spoilage of Milk (X)	27.3	29.5	26.8	29.5	30.5	29.7	25.6	25.4	24.6	23.6
Temperature(^{0}C) (Y)	33.9	34.6	34.5	36.9	37.1	37.3	28.8	29.6	31.2	30.7

Find Karl Pearson's Correlation Coefficient between spoilage of milk (in %)(X) and the temperature (^{0}C) (Y) using different formulae discussed above.

Solution

Calculate means of two variables X and Y and prepare the following table:

$$\bar{X} = \frac{272.5}{10} = 27.25 \quad \bar{Y} = \frac{334.6}{10} = 33.46$$

Spoilage of Milk (X)	Temperature (^{0}C) (Y)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
27.3	33.9	0.05	0.44	0.0025	0.1936	0.022
29.5	34.6	2.25	1.14	5.0625	1.2996	2.565
26.8	34.5	-0.45	1.04	0.2025	1.0816	-0.468
29.5	36.9	2.25	3.44	5.0625	11.8336	7.7400
30.5	37.1	3.25	3.64	10.5625	13.2496	11.830
29.7	37.3	2.45	3.84	6.0025	14.7456	9.408
25.6	28.8	-1.65	-4.66	2.7225	21.7156	7.689
25.4	29.6	-1.85	-3.86	3.4225	14.8996	7.141
24.6	31.2	-2.65	-2.26	7.0225	5.1076	5.989
23.6	30.7	-3.65	-2.76	13.3225	7.6176	10.074
Total 272.5	334.6			53.3850	91.7440	61.990

Calculate r by using the following formula

$$r = \frac{\frac{1}{n}\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{n}\Sigma(X-\bar{X})^2 . \frac{1}{n}\Sigma(Y-\bar{Y})^2}} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2 . \Sigma(Y-\bar{Y})^2}}$$

$$= \frac{61.99}{\sqrt{(53.3850)(91.7440)}} = \frac{61.99}{69.9839} = 0.8858$$

prepare the following table

Spoilage of Milk (X)	Temperature (^{0}C) (Y)	$(X)^2$	$(Y)^2$	(XY)
27.3	33.9	745.29	1149.21	925.47
29.5	34.6	870.25	1197.16	1020.70
26.8	34.5	718.24	1190.25	924.60
29.5	36.9	870.25	1361.61	1088.55
30.5	37.1	930.25	1376.41	1131.55
29.7	37.3	882.09	1391.29	1107.81

25.6	28.8	655.36	829.44	737.28
25.4	29.6	645.16	876.16	751.84
24.6	31.2	605.16	973.44	767.52
23.6	30.7	556.96	942.49	724.52
Total 272.5	334.6	7479.01	11287.46	9179.84

Calculate r by using the following formula

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

$$= \frac{10x9179.84 - (272.5)(334.6)}{\sqrt{[10x7479.01 - (272.5)^2][10x11287.46 - (334.6)^2]}} = \frac{619.9}{699.8403}$$

$$= 0.8858$$

23.6 Assumptions of Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient is based on the following assumptions:

a) There is linear relationship between the variables X & Y i.e., if the pair observations of both the variables are plotted on a scatter diagram, the plotted points will form a straight line.

b) There is often a cause and effect relationship between the forces affecting the distribution of the observations in the two series. Correlation is meaningless if there is no such relationship.

c) Each of the variables (series) is being affected by a large number of independent contributory causes of such a nature so as to produce a normal distribution. For example, relationships between price and demand, price and supply, total solids and fat contents in milk etc. are affected by several factors such that the series result into a normal distribution.

23.7 Properties of Correlation Coefficient

Correlation coefficient has following properties:

23.7.1 Limits of correlation coefficient

Pearson's Correlation Coefficient can't exceed numerically. In other words correlation coefficient lie between -1 to +1 i.e. $-1 \le r_{(X,Y)} \le +1$

Proof:

$$\sum\left[\frac{X_i - \bar{X}}{\sigma_X} + \frac{Y_i - \bar{Y}}{\sigma_Y}\right]^2 > 0$$

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{{\sigma_X}^2} + \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{{\sigma_Y}^2} \pm \frac{2\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X \cdot \sigma_Y} \ge 0$$

Dividing both sides by n we get;

$$\frac{\frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^2}{{\sigma_X}^2} + \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{{\sigma_Y}^2} \pm \frac{2\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X \cdot \sigma_Y} \ge 0$$

$1+1 \pm 2r \ge 0$

Or, $2(1 \pm r) \ge 0$

$1 \pm r \ge 0$

Either $1+r \ge 0$ i.e. $r \ge -1$

or $1 - r \ge 0$ i.e. $r \ge +1$

Hence $r_{(X, Y)}$ lies between -1 to +1

23.7.2 Correlation coefficient is independent of change of origin and scale

If X_i and Y_i are the given variables and these variables are transformed to new variables U_i and V_i by the change of origin and scale

$$U_i = \frac{X_i - A}{h} \Rightarrow X_i = A + hU_i \quad \text{and} \quad V_i = \frac{Y_i - A}{K} \Rightarrow Y_i = B + kV_i$$

$$\bar{X} = A + h\bar{U} \quad \text{and} \quad \bar{Y} = B + k\bar{V}$$

$$X_i - \bar{X} = h(U_i - \bar{U}) \qquad Y_i - Y = k(V_i - \bar{V})$$

$$r_{(X,Y)} = \frac{\frac{1}{n}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\Sigma(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\Sigma(Y_i - \bar{Y})^2}} = \frac{hk\frac{1}{n}\Sigma(U_i - \bar{U}) \cdot (V_i - \bar{V})}{\sqrt{\frac{1}{n}\Sigma h^2(U_i - \bar{U})^2}\sqrt{\frac{1}{n}\Sigma k^2(V_i - \bar{V})^2}}$$

$$= \frac{hk\frac{1}{n}\Sigma(U_i - \bar{U}) \cdot (V_i - \bar{V})}{h\sqrt{\frac{1}{n}\Sigma(U_i - \bar{U})^2}\; k\sqrt{\frac{1}{n}\Sigma(V_i - \bar{V})^2}} = \frac{\frac{1}{n}\Sigma(U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\frac{1}{n}\Sigma(U_i - \bar{U})^2}\sqrt{\frac{1}{n}\Sigma(V_i - \bar{V})^2}}$$

$$r_{(X_i,Y_i)} = \frac{n\Sigma U_iV_i - (\Sigma U_i)(\Sigma V_i)}{\sqrt{n\Sigma U_i^2 - (\Sigma U_i)^2}\sqrt{n\Sigma V_i^2 - (\Sigma V_i)^2}} = \frac{Cov\,(U_i, V_i)}{\sigma_{U_i} \cdot \sigma_{V_i}} = r(U_i, V_i)$$

Hence correlation coefficient is independent of change of origin and scale. The procedure is illustrated in the following example by taking the data from example 23.1.

Example 2 : Solve example 1 by change of origin and scale:

Solution : Let us define $U_i = X_i - 29.7$ and $V_i = Y_i - 34.5$

prepare the following table

(X)	(Y)	$U_i = X_i - 29.7$	$V_i = Y_i - 34.5$	U_i^2	V_i^2	U_iV_i
27.3	33.9	-2.4	-0.7	5.76	0.49	1.68
29.5	34.6	-0.2	0	0.04	0	0
26.8	34.5	-2.9	-0.1	8.41	0.01	0.29
29.5	36.9	-0.2	2.3	0.04	5.29	-0.46
30.5	37.1	0.8	2.5	0.64	6.25	2
29.7	37.3	0	2.7	0	7.29	0
25.6	28.8	-4.1	-5.8	16.81	33.64	23.78
25.4	29.6	-4.3	-5	18.49	25	21.5
24.6	31.2	-5.1	-3.4	26.01	11.56	17.34
23.6	30.7	-6.1	-3.9	37.21	15.21	23.79
Total 272.5	334.6	-24.5	-11.4	113.41	104.74	89.92

Calculate r by using the following formula

$$r(U_i,V_i)=\frac{n\sum U_iV_i-(\sum U_i)(\sum V_i)}{\sqrt{n\sum U_i^2-(\sum U_i)^2}\sqrt{n\sum V_i^2-(\sum V_i)^2}}$$

$$=\frac{10x89.92-(-24.5)(-11.4)}{\sqrt{10x113.41-(-24.5)^2}\sqrt{10x104.74-(-11.4)^2}}=\frac{619.9}{(23.1052)(30.2893)}$$

$$=0.8858$$

This value is same as obtained in example 1, which shows that correlation coefficient is independent of change of origin.

23.7.3 Two independent variables are uncorrelated but converse is not true

If X & Y are two independent variables then Cov (X, Y) = 0 and hence r = 0 i.e. independent variables are uncorrelated. However, the converse of this result is not true i.e., uncorrelated variables need not necessarily be independent.

23.8 Test of significance of correlation coefficient

As discussed in lesson 18 the correlation coefficient can be tested by t-test. Let a random sample (x_i, y_i) (i=1,2---,n) of size n has been drawn from a bivariate normal population and let r be the observed sample correlation coefficient . In order to test whether sample correlation coefficient r is statistical significant or there is no correlation between the variables in the population. Prof. R. A. Fisher proved that under the null hypothesis Ho: ρ=0 i.e. the variables are uncorrelated in the population, the statistic

$$t=\frac{r}{\sqrt{1-r^2}}\,x\sqrt{n-2}\sim t_{n-2}$$

i.e., t follows student's t distribution with (n-2) d.f. , n being the sample size .

Example 3 Test the significance of sample correlation coefficient obtained in example 1

Solution: Karl Pearson's Correlation Coefficient between spoilage of milk (in %)(X) and the temperature (^{0}C) (Y) in example 1 was found to be 0.8858 and n=10

We test the hypothesis Ho: ρ=0 i.e. the variables are uncorrelated in the population vs. H_1: ρ≠0 by the statistic

$$t=\frac{r}{\sqrt{1-r^2}}\,x\sqrt{n-2}=\frac{0.8858}{\sqrt{1-0.8858^2}}x\sqrt{10-2}=5.3988$$

The tabulated value of t at α=0.05 at 8 d.f. is 2.31.The calculated value of t is more than tabulated value hence it is statistically significant, so we reject H_0.This leads to conclusion that spoilage of milk (in %) and the temperature (^{0}C) is highly and significantly correlated.

23.9 Coefficient of Determination

Coefficient of correlation between two variable series measures the linear relationship between them and indicates the amount of variation of one variable which is associated with or is accounted for by another variable. The coefficient of determination, which is given by r^2, explains to what extent the variation of dependent variable Y is being explained by the explanatory variable X. In other words, the coefficient of determination gives the ratio of the explained variation to the total variation. The coefficient of determination is given by the square of the correlation coefficient i.e. r^2. Thus,

$$\text{Coefficient of determination}=r^2=\frac{\text{Explained Variation}}{\text{Total Variation}}$$

The quantity $(1-r^2)$ is called the coefficient of non-determination or coefficient of alienation. The value of $(1-r^2)$ thus measures the deviation from perfect linear relationship. For example if the value of r=0.9, we cannot conclude that 90% of the variation in the relative series (dependent variable) is due to the variation in the subject series (independent variable). But the coefficient of determination in this case is r^2=0.81which implies that only 81% of the variation in the relative series has been explained by the subject series and the remaining 19% of the variation is due to other factors.

Lesson 24

RANK CORRELATION

24.1 Introduction

There are many occasions in problems related with business and industry where it is not possible to measure the variable under consideration quantitatively or where the statistical series is composed of items, the exact magnitudes of which cannot be ascertained. Many characters are expressed in comparative terms such as beauty, intelligence, flavour and body texture of a milk product etc. In such cases the subjects are ranked pertaining to that particular character instead of taking measurements on them. Sometimes, the units are also ranked according to their quantitative measure. In these types of studies, two situations arise, (i) the same set of units is ranked according to two characters A and B (ii) two judges give ranks to the same set of units independently, pertaining to one character only. In both these situations, we get paired ranks for a set of units. For example, (i) two judges are asked to rank ten value added milk products in terms of taste independently in a sensory evaluation experiment whereas it may be difficult to give them a numerical grade in terms of taste, (ii) The students are ranked according to their marks in Operations Research and Statistics. In all these situations, the usual Pearsonian correlation coefficient cannot be used. Hence, the psychologist, Charles Edward Spearman (1906) developed a formula for correlation coefficient, which is known as rank correlation or Spearman's correlation coefficient.

24.2 Formula of Rank Correlation Coefficient

Suppose we want to find if two characteristics A (flavour) and B (consistency) are related or not. Both the characteristics are incapable of quantitative measurement but can be arranged in order of rank with respect to proficiency of two characteristics. Let X & Y be the random variable denoting the rank of the individuals in the characteristics A & B respectively. It is assumed that there is no tie i.e., no two individuals get the same rank for a characteristic then, obviously X and Y assume numerical values ranging from 1 to n. Then Spearman's rank correlation coefficient is given by the formula.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i = (X_i - Y_i)$ difference between pairs of rank of some individual in the two characters.
n= number of pairs of observations.

Proof

Let $(X_i, Y_i), (X_2, Y_2), \ldots, (X_n, Y_n)$ be the ranks of the n individuals assigned for two characteristics A & B respectively. In general, an individual will not be equally proficient in both the characteristics.
i.e. X_i (the rank of i^{th} individual in character A) will not be equal to Y_i (the rank of the ith individual in character B).
Let $d_i = X_i - Y_i$ difference between the ranks assigned by i^{th} individual for two characters A and B.
X and Y will take values from $1,2,3, - - -. n$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{1+2+3+----n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2} = \bar{Y}$$

$$\sigma_X{}^2 = \frac{1}{n}\sum X_i^2 - (\bar{X})^2 = \frac{1}{n}[1^2 + 2^2 + ---- + n^2] - \left[\frac{n+1}{2}\right]^2$$

$$\sigma_X{}^2 = \frac{n(n+1)(2n+1)}{6-n} - \left(\frac{n+1}{2}\right)^2 = \left(\frac{n+1}{2}\right)\left[\frac{2n+1}{3} - \frac{n+1}{2}\right]$$

$$= \frac{n+1}{2}\left(\frac{n-1}{6}\right) = \frac{n^2-1}{12}$$

$$\sigma_X{}^2 = \sigma_Y{}^2 = \frac{n^2-1}{12}$$

$$d_i = X_i - Y_i$$

$$\sum d_i = \Sigma[(X_i - \bar{X}) - (Y_i - \bar{Y})]$$

$$\Sigma d_i{}^2 = \Sigma(X_i - \bar{X})^2 + \Sigma\left((Y_i - \bar{Y})\right)^2 - 2\Sigma(X_i - \bar{X})(Y_i - \bar{Y}_)$$

$$\frac{1}{n}\Sigma d_i{}^2 = \frac{1}{n}(X_i - \bar{X})^2 + \frac{1}{n}\left((Y_i - \bar{Y})\right)^2 - \frac{2}{n}\Sigma(X_i - \bar{X})(Y_i - \bar{Y}_)$$

$$= \sigma_X{}^2 + \sigma_Y{}^2 - 2\,\mathrm{Cov}\,(X,Y)$$

$$\frac{1}{n}\sum d_i{}^2 = 2\sigma_X{}^2 + 2r_s\sigma_X{}^2 = \sigma_X{}^2(1 - r_s)$$

$$\frac{1}{n}\sum d_i{}^2 = \frac{2(n^2-1)(1-r_s)}{12} = \frac{(n^2-1)}{6}(1-r_s)$$

$$\frac{6\sum d_i{}^2}{n(n^2-1)} = (1-r_s) \Rightarrow r_s = 1 - \frac{6\sum d_i{}^2}{n(n^2-1)}$$

24.3 Features of Spearman's Rank Correlation Coefficient

a) Spearman's rank correlation coefficient lies between -1 to +1 i.e., $-1 \leq r_s \leq +1$.

b) Sum of the difference of ranks between two variables shall be zero i.e. $\sum d_i = 0$

c) Spearman's correlation coefficient is distribution free because no assumptions are made about the form of the population from which the sample observations are drawn.

24.4 Computation of Rank Correlation Coefficient

In rank correlation we have three types of problems;

24.4.1 When ranks are given

When ranks are given the following procedure is adopted to find the rank correlation coefficient:

1) Compute d_i the difference of ranks i.e., $d_i = X_i - Y_i$ difference between the ranks of i^{th} individual in two characters A and B.

2) Compute d_i^2 i.e., square of the rank difference.
3) Obtain the sum of squares of rank difference i.e., $\sum d_i^2$
4) Use the following formula to compute rank correlation

5) $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

The procedure is illustrated by the following example

Example 1: In a sensory evaluation experiment, two judges accorded the following ranks to eight milk products

Judge A	8	7	6	3	1	1	5	4
Judge B	7	5	4	1	3	2	6	8

Find the Spearman's rank correlation coefficient.

Solution :

Prepare the following table and calculate d_i's and d_i^2's as given below

Judge A (X_i)	8	7	6	3	2	1	5	4	
Judge B (Y_i)	7	5	4	1	3	2	6	8	
$d_i = X_i - Y_i$	1	2	2	2	-1	-1	-1	-4	$\sum d_i = 0$
di^2	1	4	4	4	1	1	1	16	$\sum di^2 = 32$

Calculate rank correlation coefficient as follows

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6x32}{8(64-1)} = 1 - 0.3809 = 0.6191$$

24.4.2 When ranks are not given

When we are given the actual data and not the ranks, the following procedure is adopted to find out rank correlation coefficient.

1) First step is to convert the data into ranks the highest (smallest) observation is given rank 1. The next highest (smallest) observation is given rank 2 and so on. It is immaterial in such a way (descending or ascending) the ranks are assigned. However, the same approach should be followed for the entire variable under consideration.

2) Compute di the difference of ranks i.e., $d_i = X_i - Y_i$ difference between the ranks of i^{th} individual in two characters A and B.
3) Compute di^2 i.e., square of the rank difference.
4) Obtain the sum of squares of rank difference i.e., $\sum di^2$

5) Use the following formula to compute rank correlation

6) $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

The procedure is illustrated by the following example

Example 2: Calculate Spearman's rank Correlation for the following data on marks obtained.

X	80	73	85	36	54	93	65	96	58	88
Y	15	83	95	64	32	16	67	66	85	39

Solution :

Prepare the following table and calculate d_i's and d_i^2's as given below

X	Y	Ranks alloted to X x_i	Ranks alloted to Y y_i	$d_i = x_i - y_i$	di^2
80	15	6	1	5	25
73	83	5	8	-3	9
85	95	7	10	-3	9
36	64	1	5	-4	16
54	32	2	3	-1	1
93	16	9	2	7	49
65	67	4	7	-3	9
96	66	10	6	4	16
58	85	3	9	-6	36
88	39	8	4	4	16
					$\sum di^2 = 186$

Calculate rank correlation coefficient as follows

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6x186}{10(100-1)} = 1 - 1.1273 = -0.1273$$

24.4.3 When ranks are repeated (or Tie case)

In case of attributes where there is a tie i.e. if two or more individuals are placed at the same level in classification with respect to an attribute or if in case of variable data, there is more than one item with the same

value in either or both the series then Spearman's rank correlation formula breaks down, since in this case the variables X and Y don't take values from 1 to n and consequently $(\bar{X} \neq \bar{Y})$. In such cases, common ranks are assigned to repeated items. The common ranks are arithmetic mean of ranks which the items would have got if they were different from each other and next item will get the rank next to the rank used in computing the common rank. If there is large number of ranks with tie, it is advisable to apply correction factor or adjustment factor, (C.F)

$$\frac{m(m^2-1)}{12}$$

where m is the number of times an item is repeated. Then add this correction factor to $\sum di^2$.

$$r_s = 1 - \frac{6\left[\sum d_i^2 + \frac{m(m^2-1)}{12} + \frac{m(m^2-1)}{12} + ---\right]}{n(n^2-1)}$$

This correction factor is to be added for each repeated value in both the series.
The procedure is illustrated by the following example

Example 3: Calculate Spearman's rank Correlation for the following data

X	80	73	80	36	54	93	65	36	58	80
Y	15	83	15	64	32	16	67	64	85	64

Solution

Prepare the following table and calculate d_i's and d_i^2's as given below

X	Y	x_i	y_i	di	di^2
80	15	8	1.5	6.5	42.25
73	83	6	9	-3	9
80	15	8	1.5	6.5	42.25
36	64	1.5	6	-4.5	20.25
54	32	3	4	-1	1
93	16	10	3	7	49
65	67	5	8	-3	9
36	64	1.5	6	-4.5	20.25
58	85	4	10	-6	36
80	64	8	6	2	4
					$\sum di^2 = 233$

In X series, we see that the value 36 occurs twice. The common rank assigned to each of these values is 1.5, the arithmetic mean of 1 and 2 rank which these observations would have taken if they were different. The subsequent four values 54,58, 65 and 73 are allotted ranks as 3,4,5 and 6 respectively. Again, the value 80 occurs thrice .The common rank assigned to it is 8, the arithmetic mean of 7, 8 and 9 and the next value, viz., 93 gets the rank 10. Similarly in Y series the value 15 occurs twice and the common rank assigned to each is 1.5, the arithmetic mean of 1 and 2.The next value 16 gets the next rank viz., 3. Again the value 64 occurs thrice. The common rank assigned to it is 6, the arithmetic mean of 5, 6 and 7 and the next value, viz., 67 gets the rank 8 and so on. Hence we see that in the X-series the observation 36 occurs twice (m=2) and 80 occurs thrice (m=3) and in the Y series the observation 15 occurs twice (m=2) and 64 occurs thrice (m=3). Hence on applying the correction factor $\frac{m(m^2-1)}{12}$ for each repeated item, we get

Calculate rank correlation coefficient as follows

$$r_s = 1 - \frac{6\left[233 + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}\right]}{10(10^2-1)}$$

$$= 1 - \frac{6(233 + 0.5 + 2 + 0.5 + 2)}{10 \times 99} = 1 - 1.4424 = -0.4424$$

24.5 Test of Significance of Rank Correlation Coefficient

The significance of rank correlation coefficient is tested by t-test, as it is done in case of Karl Pearson's correlation coefficient. Here we test the null hypothesis Ho: $\rho_s=0$. vs H_1: $\rho_s \neq 0$. The test statistic

$$t = \frac{r_s}{\sqrt{1-r_s^2}} \; x\sqrt{n-2} \sim t_{n-2}$$

i.e., t follows student's t distribution with (n-2) d.f., where n is the number of paired observations and r_s is the rank correlation coefficient.

Lesson 25

LINEAR REGRESSION

25.1 Introduction

In Lesson 23 we have established the fact that if two variables are closely related we may be interested in estimating the value of one variable given the value of another. For example, if we know that in milk, the content of total solids and fat levels are correlated we want to find out expected total solids in milk for a given fat level. Similarly, if we know that spoilage of milk (in %) and the temperature (°C) of storage of milk in a dairy plant are closely related we may find out the level of temperature at which spoilage of milk starts. It is often of interest to determine how change of values of some variables influences the change of values of other variables. Regression analysis reveals average relationship between two variables and this makes possible estimation or prediction. The literal or dictionary meaning of the word 'Regression' is 'stepping back or returning to the average value'. The term was first used by British biometrician Sir Francis Galton in the later part of the 19^{th} century in connection with some studies made on estimating the extent to which the stature of the population. Actually regression means to regress i.e., to step back or to fall back or to return back to a former state. So regression means returning of retrogression. Falconer (1936) conducted an experiment in which he took two groups of parents, one group having more height while others having shorter than the normal height. It was found that the children of the first group of parents try to go back to normal height while the children of second group of parents try to reach the normal height. Regression analysis in the general sense means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the very important statistical tools which are extensively used in almost all branches of science-natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related casually and for estimation of demand and supply curves, cost functions, production and consumption functions, etc.

25.2 Definition of Regression

Regression analysis is one of the very scientific techniques for making predictions. In the words of M.M. Blair "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data".

According to Morris Hamburg the term 'regression analysis' "refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more values of other variables and to the measurement of the errors involved in this estimation process."

Ya-Lun Chou defined "Regression analysis attempts to establish the 'nature of the relationship' between variables-that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting".

It is clear from the above definitions that regression analysis is a statistical device with the help of which estimate (predict) the unknown values of one variable from known values of another variable. In regression analysis there are two types of variables. The variables whose value is influenced or is to be predicted is called dependent variable and the variable which influence the values or is used for prediction, is called independent variable. In regression analysis independent variable is also known as regressor or predictor or explanatory while the dependent variable is also known as regressed or explained variable.

25.3 Types of Regression Analysis

The main types of regression analysis are as follows:

a) Simple and Multiple
b) Linear and Non- Linear

25.3.1 Simple and multiple

The regression analysis confined to the study of only two variables at a time is termed as simple regression. In simple regression analysis one variable is dependent and another is independent. The functional relationship between total solids and fat content in milk samples is an example of simple regression. But quite often the values of a particular phenomenon may be affected by multiplicity of factors. The regression analysis where we study more than two variables at a time is known as multiple regression for

example, the study of effect of fat and SNF contents, on total solids in milk of samples; the study of Total Quality as affected by Methyl Blue Reduction Time (MBR) and Standard Plate Counts (SPC) etc.

25.3.2 Linear and non –linear

If the given bivariate data are plotted on a graph, the points so obtained on the scatter diagram will more or less concentrate round a curve, called the '***curve of regression***'. Often such a curve is not distinct and is quite confusing and sometimes complicated too. The mathematical equation of the regression curve, usually called the regression equation, enables us to study the average change in the value of the dependent variable for any given value of the independent variable. If the regression curve is a straight line, we say that there is linear regression between the variables under study. The equation of such a curve is the equation of a straight line, *i.e.,* a first degree equation in the variable X and Y. In case of linear regression, the values of the dependent variable increase by a constant absolute amount for a unit change in the value of the independent variable. However, if the curve of regression is not a straight line, the regression is termed as curved or non-linear regression. In that case the regression equation is a functional relation between X and Y involving transformed values of X and Y, i.e., involving terms of the type $X^2, Y^2, XY, \log X, \log Y$ etc. However, in this chapter we shall confine our discussion to linear regression between two variables only.

25.4 Simple Linear Regression

In practice, simple linear regression is often used and under this, regression lines, regression equations and regression coefficients are very important to be studied, which are discussed in the subsequent sections.

25.5 Regression Lines

The regression line shows the average relationship between two variables. It is the line which gives the best estimate of one variable for given value of other variable. The term best fit is interpreted in accordance with the Principle of Least Squares which consists in minimising the sum of the squares of the residuals or the errors of estimates, i.e., the deviations between the given observed values of the variable and their corresponding estimated values as given by the line of best fit. In case of two variables X and Y, we shall have two lines regression one for Y on X and the other for X on Y.

25.5.1 Regression line of Y on X

Line of regression of Yon X is the line which gives the best estimate for the value of Y for any specified value of X and is obtained by minimising the sum of squares of the errors parallel to Y-axis.

25.5.2 Regression line of X on Y

Line of regression of X on Y is the line which gives the best estimate for the value of X for any specified value of Y and is obtained by minimising the sum of squares of the errors parallel to X-axis.

25.6 Derivation of Line of Regression of Y on X

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n Y_n)$ be n pairs of observations on two variables under study. Let

$$Y = a + bX \qquad (25.1)$$

be the line of regression (best fit of Y on X). For a given point $P_i\ (X_i, Y_i)$ in the scatter diagram, the error estimate or residual as given by the line of best fit (eq. 25.1) is $P_i H_i$ as shown in figure 25.1. Now the X- coordinate of H_i is same as that of P_i, so X_i lies on the same line (25.1) the Y-coordinates of H_ii.e., H_iM is given by $(a + bX_i)$. Hence, the error of estimate for P_i is given by

$$P_i H_i = P_i M - H_i M = Y_i - (a + bX_i) \qquad (25.2)$$

This error is parallel to Y-axis for the i^{th} point and we compute such error for all points of scatter diagram. The $P_i H_i$which lie above the line be positive and below the line, the error will be negative. There will be several lines passing through these scatter of points and we have to find that particular line of best fit for which deviation or residual is minimum.

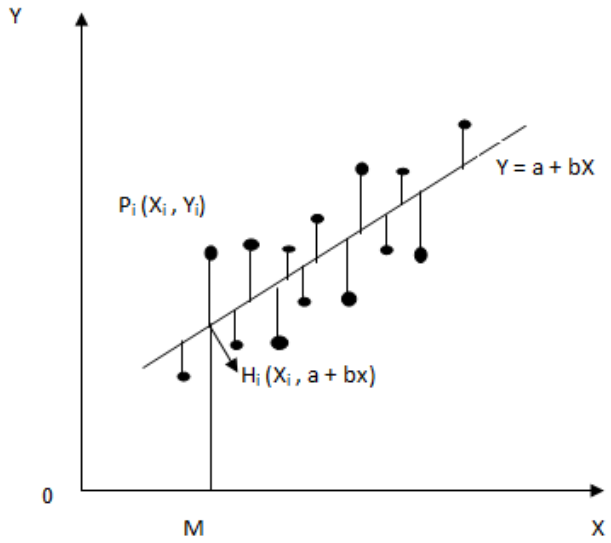


Fig. 25.1 Scatter diagram with an estimating line

According to principle of least squares, we have to determine the constants a and b in equation (25.1) such that the residual or deviation sum of squares of the errors is minimum. In other words we have to minimise the residual sum of squares due to error 'E'

$$E = \sum_{i=1}^{n} (P_i H_i)^2 = \sum_{i=1}^{n} \left(Y_i - (\hat{a} + \hat{b}X_i\right)^2 \qquad (25.3)$$

Differentiating E partially with respect to $\hat{a}$ and $\hat{b}$, we get

$$\frac{\partial E}{\partial \hat{a}} = \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)(-1) \Rightarrow \sum (Y_i - \hat{a} - \hat{b}X_i) = 0$$

$$\sum Y_i - n\hat{a} - \hat{b}\sum X_i = 0$$

$$\Rightarrow \sum Y_i = n\hat{a} + \hat{b}\sum X_i \qquad (25.4)$$

$$\frac{\partial E}{\partial \hat{b}} = \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)(-X_i)$$

$$= \sum_{i=1}^{n} (\hat{a} + \hat{b}X_i - Y_i)(X_i) = \sum_{i=1}^{n} (\hat{a} + \hat{b}X_i - Y_i)(X_i) = 0$$

$$\sum Y_i X_i = \hat{a}\sum X_i + \hat{b}\sum X_i^2$$

$$\sum Y_i X_i = \hat{a}\sum X_i + \hat{b}\sum X_i^2 \qquad (25.5)$$

Equation 25.4 and 25.5 are known as two normal equations. Solving these two normal equations, we get

$$\sum X_i \sum Y_i - n\sum X_i Y_i = \hat{b}\left[\left(\sum X_i\right)^2 n\sum X_i^2\right]$$

$$\Rightarrow \hat{b} = \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2} = \frac{Cov\,(X,Y)}{V(X)} = b_{YX} \qquad (25.6)$$

Putting the value of $\hat{b}$ in either of the normal equation we get

$$\hat{a} = \frac{(\sum X_i^2)(\sum Y_i) - (\sum X_i)(\sum X_i Y_i)}{n\sum X_i^2 - (\sum X_i)^2} \qquad (25.7)$$

Substituting these values of $\hat{a}$ & $\hat{b}$ from equation (25.7) and (25.6) respectively in equation (25.1) we get required equation of line regression of Y on X.

Dividing both the equation (25.4) by number of pairs of observation we get

$$\frac{\sum Y_i}{n} = \frac{\sum \hat{a}}{n} + \frac{b\sum X_i}{n} \Rightarrow \bar{Y} = \hat{a} + \hat{b}\bar{X} \qquad (25.8)$$

This implies that line of best fit passes through the point $\bar{X}, \bar{Y}$ or in other words points $\bar{X}, \bar{Y}$ lies on the line of regression of Y on X. The required equation of the line of regression of Y on X can be written as:

$$(Y - \bar{Y}) = b_{YX}(X - \bar{X}) = \frac{Cov(x,y)}{V(x)}(X - \bar{X}) \qquad (25.9)$$

But we know that $r = \frac{Cov.(X,Y)}{\sigma_X \sigma_Y} \Rightarrow Cov.(X,Y) = r\sigma_X\sigma_Y$

Substituting the value of Cov. (X,Y) in equation (25.9) we get

$$(Y - \bar{Y}) = \frac{r\sigma_Y}{\sigma_X}(X - \bar{X}) \qquad (25.10)$$

25.6.1 Line of Regression of X on Y

Similarly we can have a line of X on Y i.e., $X = a' + b'Y$

$$\hat{b}' = b_{XY} = \frac{Cov(X,Y)}{V(Y)} = \frac{r\sigma_Y}{\sigma_X}$$

The required equation of the line of regression of X on Y can be written as :

$$(X - \bar{X}) = b_{XY}(Y - \bar{Y}) = \frac{Cov(X,Y)}{V(X)}(Y - \bar{Y}) = \frac{r\sigma_X}{\sigma_Y}(Y - \bar{Y}) \qquad (25.11)$$

From equations (25.9) and (25.11) it is evident that both the lines of regression X on Y and Y on X pass through the point $(\bar{X}, \bar{Y})$. Hence $(\bar{X}, \bar{Y})$ is a point of intersection of Y on X and X on Y. The above procedure of fitting of regression equation is illustrated through the following example:

Example 1 : The following data pertains to spoilage of milk (in %)(X) and the temperature (^{0}C) (Y) of storage of milk in a dairy plant.

Spoilage of Milk (X)	27.3	29.5	26.8	29.5	30.5	29.7	25.6	25.4	24.6	23.6
Temperature(^{0}C) (Y)	33.9	34.6	34.5	36.9	37.1	37.3	28.8	29.6	31.2	30.7

Fit a linear regression for spoilage of milk (in %)(X) on the temperature (^{0}C) (Y)and vice versa . Also predict the spoilage of milk when temperature is 40^{0}C and value of temperature when spoilage in milk is 35 %.

Solution

$$\bar{X} = \frac{\sum X_i}{n} = \frac{272.5}{10} = 27.25, \bar{Y} = \frac{\sum Y_i}{n} = \frac{334.6}{10} = 33.46$$

Spoilage of Milk (X_i)	Temperature (^{0}C) (Y_i)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
27.3	33.9	0.05	0.44	0.0025	0.1936	0.022
29.5	34.6	2.25	1.14	5.0625	1.2996	2.565
26.8	34.5	-0.45	1.04	0.2025	1.0816	-0.468
29.5	36.9	2.25	3.44	5.0625	11.8336	7.7400
30.5	37.1	3.25	3.64	10.5625	13.2496	11.830

29.7	37.3	2.45	3.84	6.0025	14.7456	9.408
25.6	28.8	-1.65	-4.66	2.7225	21.7156	7.689
25.4	29.6	-1.85	-3.86	3.4225	14.8996	7.141
24.6	31.2	-2.65	-2.26	7.0225	5.1076	5.989
23.6	30.7	-3.65	-2.76	13.3225	7.6176	10.074
Total 272.5	334.6			53.3850	91.7440	61.990

Regression coefficient of Yon X (b_{YX}) :

$$b_{YX} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum(X_i - \bar{X})^2.} = \frac{61.990/10}{53.385/10} = 1.1612$$

Regression coefficient of X on Y (b_{XY}) :

$$b_{XY} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum(Y_i - \bar{Y})^2.} = \frac{61.990/10}{91.7440/10} = 0.6757$$

Regression equation of Y on X i.e., regression line of temperature on spoilage of milk is

$$(Y - \bar{Y}) = b_{YX}(X - \bar{X}) \Rightarrow (Y - 33.46) = 1.1612(X - 27.25)$$

The required equation of the line of regression of Y on X i.e. regression equation of temperature on spoilage in milk is Y=1.8173+1.1612X. To predict the value of temperature when spoilage in milk is 35%, we put X=35 in the above equation so we get Y=42.4593. It means when spoilage in milk is 35% the temperature will be 42.4593 ^{0}C.

Regression equation of X on Y i.e., regression line of spoilage of milk on temperature is

$$(X - \bar{X}) = b_{XY}(Y - \bar{Y}) \Rightarrow (X - 27.25) = 0.6757(X - 33.46)$$

The required equation of the line of regression of X on Y i.e. regression equation of spoilage of milk on temperature is X=4.6416+0.6757Y. To predict the value of spoilage in milk when temperature is 40^{0}C, we put Y=40 in the above equation so we get X=31.6693. It means when temperature is 40^{0}C then spoilage in milk will be 31.6693 %.

25.6.2 Why there are two regression lines

The line of regression of Y on X $(Y = a + b_{YX}X)$ is used to estimate/predict the value of Y for any given value of X i.e. Y is a dependent variable and X is an independent/explanatory variable. The estimates so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of the least squares. In order to predict or estimate X for any given value of Y we use the regression equation of X on Y$(X = a' + b_{XY}Y)$ which is obtained by minimizing sum of squares due to error of estimates in X. Here X is dependent variable and Y is independent/explanatory variable. Two regression equations are not reversible or interchangeable. Regression equation of Y on X is obtained by minimizing the sum of square of errors parallel to the Y-axis, while the regression equation of X on Y is obtained by minimizing the sum of squares of error parallel to X-axis. In a particular case of perfect correlation, positive or negative i.e., r=±1, the equation of line of regression of Y on X becomes:

$$(Y - \bar{Y}) = \pm\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \Rightarrow \frac{(Y - \bar{Y})}{\sigma_Y} = \pm\frac{(X - \bar{X})}{\sigma_X}$$

Similarly , the equation of the line of regression of X on Y becomes:

$$(X - \bar{X}) = \pm\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \Rightarrow \frac{(X - \bar{X})}{\sigma_X} = \pm\frac{(Y - \bar{Y})}{\sigma_Y}$$

Above two equations are same. Hence, in case of perfect correlation (r=±1) both the lines coincide. Therefore, in general we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

25.6.3 Angle Between Two Regression Lines :

The angle between two regression lines is given by

$$\theta = \tan^{-1}\left\{\frac{\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2}\left(\frac{r^2-1}{r}\right)\right\}$$

If r=±1, then $\theta = \tan^{-1}(0) \Rightarrow \theta = 0 \text{ or } \pi$ i.e., two lines are either coincident or they are parallel to each other. But since both the lines of regression intersect at the point $(\bar{X}, \bar{Y})$, they cannot be parallel. Hence in case of perfect correlation, positive or negative, the two lines of regression coincide. If r=0, then $\theta = \tan^{-1}(\infty) \Rightarrow \theta = \frac{\pi}{2}$ i.e., if the variables are uncorrelated, the two lines of regression become perpendicular to each other. Hence, for higher degree of correlation between the variables, the angle between the lines is smaller i.e., the two lines of regression are nearer to each other. On the other hand, the angle between the lines increases as the lines of regression move apart and the value of correlation decreases.

25.7 Coefficients of Regression

Just as there are two regression equations, similarity there are two regression coefficients. Regression coefficients measure the average change in the value of one variable for a unit change in the value of another variable. Regression coefficient, infact, represents the slope of a regression line. For two variables X and Y, there are two regression coefficients which are given as follows

25.7.1 Regression Coefficient of Y on X

This coefficient shows that with a unit change in the value of X variable, what will be the average change in the value of Y variable. This is represented by b_{YX} .

$$b_{YX} = \frac{Cov(x,y)}{V(x)} = \frac{r\sigma_Y}{\sigma_X}$$

25.7.2 Regression Coefficient of X on Y

This coefficient shows that with a unit change in the value of Y variable, what will be the average change in the value of Y variable. This is represented by b_{XY} .

$$b_{XY} = \frac{Cov(x,y)}{V(y)} = \frac{r\sigma_X}{\sigma_Y}$$

25.8 Properties of Regression Coefficient

The important properties of the regression coefficients are :

1) The correlation coefficient is the geometric mean between the regression coefficients i.e.,
$$r^2 = b_{YX}.b_{XY} \Rightarrow r = \sqrt{b_{YX}.b_{XY}}$$
2) Both the regression coefficients must have the same algebraic signs. This means that either both regression coefficients will be positive or negative i.e., when one regression coefficient is negative, the other would also be negative and if one regression coefficient is positive, the other would be also positive. It is never possible that one regression coefficient is negative while the other is positive.
3) The coefficient of correlation will have the same sign as that of regression coefficients.
4) If one of the regression coefficients is greater than unity (one), the other must be less than unity.
5) The arithmetic mean of the regression coefficients is greater than the correlation coefficient.
6) Regression coefficients are independent of change of origin but not of scale which is illustrated below.

If X_i and Y_i are the given variables and these variables are transformed to new variables U_i and V_i by the change of origin and scale

$$U_i = \frac{X_i - A}{h} \Rightarrow X_i = A + hU_i \quad \text{and} \quad V_i = \frac{Y_i - A}{K} \Rightarrow Y_i = B + kV_i$$

$$\bar{X} = A + h\bar{U} \quad \text{and} \quad \bar{Y} = B + k\bar{V}$$

$$X_i - \bar{X} = h(U_i - \bar{U}) \qquad Y_i - Y = k(V_i - \bar{V})$$

$$b_{YX} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum(X_i - \bar{X})^2} = \frac{hk\frac{1}{n}\sum(U_i - \bar{U}) \cdot (V_i - \bar{V})}{\frac{1}{n}\sum h^2(U_i - \bar{U})^2}$$

$$= \frac{hk\frac{1}{n}\sum(U_i - \bar{U}) \cdot (V_i - \bar{V})}{h^2\frac{1}{n}\sum(U_i - \bar{U})^2} = \frac{k}{h}\frac{\frac{1}{n}\sum(U_i - \bar{U})(V_i - \bar{V})}{\frac{1}{n}\sum(U_i - \bar{U})^2}$$

$$b_{YX} = \frac{k}{h}\frac{n\sum U_iV_i - (\sum U_i)(\sum V_i)}{n\sum {U_i}^2 - (\sum U_i)^2} = \frac{k}{h}\frac{Cov\,(U_i, V_i)}{\sigma_{U_i} \cdot \sigma_{V_i}} = \frac{k}{h}b_{U_iV_i}$$

Hence regression coefficient is independent of change of origin and but not of scale. The procedure is illustrated in the following example by taking the data from example 25.1.

Example 2 : Solve example 1 by change of origin and scale:

Solution : Let us define $U_i = X_i - 29.7$ and $V_i = Y_i - 34.5$

(X)	(Y)	$U_i = X_i - 29.7$	$V_i = Y_i - 34.5$	${U_i}^2$	${V_i}^2$	U_iV_i
27.3	33.9	-2.4	-0.7	5.76	0.49	1.68
29.5	34.6	-0.2	0	0.04	0	0
26.8	34.5	-2.9	-0.1	8.41	0.01	0.29
29.5	36.9	-0.2	2.3	0.04	5.29	-0.46
30.5	37.1	0.8	2.5	0.64	6.25	2
29.7	37.3	0	2.7	0	7.29	0
25.6	28.8	-4.1	-5.8	16.81	33.64	23.78
25.4	29.6	-4.3	-5	18.49	25	21.5
24.6	31.2	-5.1	-3.4	26.01	11.56	17.34
23.6	30.7	-6.1	-3.9	37.21	15.21	23.79
Total 272.5	334.6	-24.5	-11.4	113.41	104.74	89.92

$$b_{YX} = \frac{k}{h}\frac{n\sum U_iV_i - (\sum U_i)(\sum V_i)}{n\sum {U_i}^2 - (\sum U_i)^2} = \frac{10x89.92 - (-24.5)(-11.4)}{10x113.41 - (-24.5)^2} = \frac{619.9}{533.85} = 1.1612$$

$$b_{XY} = \frac{h}{k}\frac{n\sum U_iV_i - (\sum U_i)(\sum V_i)}{n\sum {V_i}^2 - (\sum V_i)^2} = \frac{10x89.92 - (-24.5)(-11.4)}{10x104.74 - (-11.4)^2} = \frac{619.9}{917.44} = 0.6757$$

These values are same as obtained in example 25.1, which shows that regression coefficients are independent of change of origin.

$r = \sqrt{b_{YX} \times b_{XY}} = \sqrt{1.1612 \times 0.6757} = 0.8858$

25.9 Coefficient of Determination

The total variation in the dependent variable Y can be split into two:

a) **Explained variation**: The variation in Y which is explained by the variation in X is known as explained variation in Y

b) **Unexplained variation**: The variation in Y which is unexplained by the variation in variable X and is due to some other factors (a variable) is called unexplained variation in Y.

Symbolically,

Total variation in Y = Explained variation in Y + Unexplained variation in Y

$$\sum(Y - \bar{Y})^2 = \sum(\hat{Y} - \bar{Y})^2 + \sum(Y - \hat{Y})^2$$

where $\hat{Y}$ = computed (or estimated) value of Y on the basis of regression equation

$\bar{Y}$ = Mean value of Y series

Y = Original value of Y series

A similar relationship we may have for X variable (Dependent) in terms of Y:

$$\sum(X-\bar{X})^2=\sum(\hat{X}-\bar{X})^2+\sum(X-\hat{X})^2$$

Coefficient of Determination: Based on above expression, the coefficient of determination (r^2) is defined as the ratio of the explained variation to total variation i.e.,

$$r^2=\frac{\text{Explained variation}}{\text{Total variation}}=\frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

It is clear that the objective of coefficient of determination is to determine the percentage variation in Y which is explained by variation in X. For example, let us suppose that the correlation coefficient between X and Y is +0.8, then coefficient of determination (r^2) = $(.8)^2$ = .64. It means that 64per cent variation in Y is due to variation in X and 36 per cent variation is due to other factors. Thus, explained variations and unexplained variation are 64 and 36 per cent respectively.

Coefficient of Non-Determination: The proportion of unexplained variation to total variation is termed as coefficient of non-determination. It is denoted by k^2, where $k^2=1-r^2$. It is also written as:

$$k^2=\frac{\text{Unexplained variation}}{\text{Total variation}}=1-r^2$$

The square root of k^2 is termed as coefficient of alientation i.e., $k=\sqrt{k^2}=\sqrt{1-r^2}$

Standard Error of Estimate: Standard error of estimates of Y on X and that of X on Y can also be calculated as:

$$S_{Y.X}=\sqrt{\frac{\sum(Y-\hat{Y})^2}{N}}=\sqrt{\frac{\text{Unexplained variation in Y}}{N}}$$

$$S_{X.Y}=\sqrt{\frac{\sum(X-\hat{X})^2}{N}}=\sqrt{\frac{\text{Unexplained variation in X}}{N}}$$

Example 3: Compute Coefficient of Determination in example 1

Solution

(X)	(Y)	$\hat{Y}$ =1.8173+1.1612X	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$	$(Y-\hat{Y})^2$
27.3	33.9	33.5181	0.1936	0.0034	0.1459
29.5	34.6	36.0727	1.2996	6.8262	2.1688
26.8	34.5	32.9375	1.0816	0.2730	2.4415
29.5	36.9	36.0727	11.8336	6.8262	0.6844
30.5	37.1	37.2339	13.2496	14.2423	0.0179
29.7	37.3	36.3049	14.7456	8.0937	0.9901
25.6	28.8	31.5440	21.7156	3.6710	7.5296
25.4	29.6	31.3118	14.8996	4.6148	2.9302
24.6	31.2	30.3828	5.1076	9.4690	0.6678
23.6	30.7	29.2216	7.6176	17.9639	2.1856
Total 72.5	334.6		91.7440	71.9836	19.7620

From above table the different variation are as

Total variation: $\sum(Y-\bar{Y})^2$ =91.7440

Explained variation in Y = $\sum(\hat{Y}-\bar{Y})^2$=71.9836

Unexplained variation in Y=$\sum(Y-\hat{Y})^2$=19.7620

Coefficient of determination (r^2) $= \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2} = \frac{71.9836}{91.7440} = 0.7846$

It means that 78.46 percent variations in Y is due to variation in X and 21.54 per cent variation is due to other factors. Moreover, coefficient of determination is square of correlation coefficient i.e., $(0.8858)^2$=0.7846 which is same as computed above.

Lesson 26

BASIC CONCEPTS OF STATSITICAL QUALITY CONTROL

26.1 Introduction

From the early days of industrial production, the emphasis had been on turning out products of uniform quality by ensuring use of similar raw materials, identical machines, and proper training of the operators. Inspite of these efforts, the causes of irregularity often crept in inadvertently. Besides, the men and machines are not infallible and give rise to the variation in the quality of the product. For keeping this variation within limits, in earlier days, the method used was 100 per cent inspection at various stages of manufacturing.

It was in 1924 that Dr. W.A. Shewhart of Bell Telephone Laboratories, USA developed a method based on statistical principles for controlling quality of products during the manufacturing and thus eliminating the need for 100 per cent inspection. This technique which is meant to be an integral part of any production process, does not provide an automatic corrective action but acts as sensor and signal for the variation in the quality. Therefore, the effectiveness of this method depends on the promptness with which a necessary corrective action is carried out on the process. This technique has since been developed by adding to its armory more and more charts, as a result of its extensive use in the industry during and after the Second World War. In this lesson various terms used in the context of Statistical Quality Control (SQC) have been illustrated.

26.2 Definitions of Various Terms Involved in Statistical Quality Control

The following terms are used to understand the concept of Statistical Quality Control

26.2.1 Quality

The most important word in the term ‘Statistical Quality Control’ is quality. By ‘Quality’ we mean an attribute of the product that determines its fitness for use. Quality can be further defined as “Composite product characteristics of engineering and manufacture that determine the degree to which the product in use will meet the expectations of the customer at reasonable cost.” Quality means conformity with certain prescribed standards in terms of size, weight, strength, colour, taste, package etc.

26.2.2 Quality characteristics

Quality of a product (or service) depends upon the various characteristics that a product possesses. For example, the Kulfi we buy should have the following characteristics.

(a) TS (b) Sugar (c) Flavour (d) Body & Texture.

All these individual characteristics constitute the quality of Kulfi. Of course, some of them are important (critical) without which the Kulfi is not acceptable. For example Minimum TS, Sugar, Body and Texture score is important. However, other characteristics such as Colour and Flavour may not be so important. The quality characteristics may be defined as the “distinguishing” factor of the product in the appearance, performance, length of life, dependability, reliability, durability, maintainability, taste, colour, usefulness etc. Control of these quality characteristics in turn means the control of the quality of product.

26.2.3 Types of characteristics

There are two types of characteristics viz., variable characteristics and attribute characteristics.

26.2.3.1 Variable characteristic

Whenever a record is made of an actual measured quality characteristic, such as dimension expressed in mm, cm etc. quality is said to be expressed by variables. This type of quality characteristics includes e.g., dimension (length, height, thickness etc.),hardness, temperature, tensile strength, weight, moisture percent, yield percent, fat percent etc.

26.2.3.2 Attribute characteristic

Whenever a record shows only the number of articles conforming and the number of articles failing to conform to any specified requirements, it is said to be a record of data by 'attributes'. These include:

- Things judged by visual examination
- Conformance judged by gauges
- Number of defects in a given surface area etc.

26.2.4 Control

Control means organizing the following steps:

- Setting up standards of performance.
- Comparing the actual observations against the standards.
- Taking corrective action whenever necessary.
- Modifying the standards if necessary.

26.2.5 Quality control

Quality control is a powerful productivity technique for effective diagnosis of lack of quality (or conformity to set standards) in any of the materials, processes, machines or end products. It is essential that the end products possess the qualities that the consumer expects of them, for the progress of the industry depends on the successful marketing of products. Quality control ensures this by insisting on quality specifications all along the line from the arrival of materials through each of their processing to the final delivery of goods.Quality control, therefore, covers all the factors and processes of production which may be broadly classified as follows:

- **Quality of materials**: Material of good quality will result in smooth processing there by reducing the waste and increasing the output. It will also give better finish to end products.
- **Quality of manpower**: Trained and qualified personnel will give increased efficiency due to the better quality production through the application of skill and also reduce production cost and waste.
- **Quality of machines**: Better quality equipment will result in efficient working due to lack or scarcity of break downs thus reducing the cost of defectives.
- **Quality of Management**: A good management is imperative for increase in efficiency, harmony in relations, growth of business and markets.

26.2.6 Chance and assignable causes of variation

Variation in the quality of the manufactured product in the repetitive process in the industry is inherent and inevitable. These variations are broadly classified as being due to two causes viz., (i) chance causes, and (ii)

assignable causes.

26.2.6.1 Chance causes

Some "Stable pattern of variation" or "a constant cause system" is inherent in any particular scheme of production and inspection. This pattern results from many minor causes that behave in a random manner. The variation due to these causes is beyond the control of human being and cannot be prevented or eliminated under any circumstance. Such type of variation has got to be allowed within the stable pattern, usually termed as Allowable Variation. The range of such variation is known as natural tolerance of the process.

26.2.6.2 Assignable causes

The second type of variation attributed to any production process is due to non-random or the so called assignable causes and is termed as Preventable Variation. The assignable causes may creep in at any stage of the process, right from the arrival of raw materials to the final delivery of the goods.

Some of the important factors of assignable causes of variation are substandard or defective raw material, new techniques or operations, negligence of the operators, wrong or improper handling of machines, faulty equipment, unskilled or inexperienced technical staff and so on. These causes can be identified and eliminated and are to be discovered in a production process before it goes wrong i.e., before the production becomes defective.

26.3 Statistical Quality Control

By Statistical Quality Control (SQC) we mean the various statistical methods used for the maintenance of quality in a continuous flow of manufactured goods. The main purpose of SQC is to devise statistical techniques which help us in separating the assignable causes from chance causes of variation thus enabling us to take remedial action wherever assignable causes are present. The elimination of assignable causes of erratic fluctuations is described as bringing a process under control. A production process is said to be in a state of statistical control if it is governed by chance causes alone, in the absence of assignable causes of variation.

In the above problem, the main aim is to control the manufacturing process so that the proportion of defective items is not excessively large. This is known as '**Process Control'**. In another type of problem we want to ensure that lots of manufactured goods do not contain an excessively large proportion of defective items. This is known as '**Product or Lot Control**'. The process control and product control are two distinct problems, because even when the process is in control, so that the proportion of defective products for the entire output over a long period will not be large, an individual lot of items may not be of satisfactory quality. Process Control is achieved mainly through the technique of '**Control Charts**' whereas Product Control is achieved through '**Sampling Inspection'**.

26.4 Stages of Production Process

Before production starts, a decision is necessary as to what is to be made. Next comes the actual manufacturing of the product. Finally it must be determined whether the product manufactured is what was intended. It is therefore necessary that quality of manufactured product may be looked at in terms of three functions of specification, production and inspection.

26.4.1 Specification

This tells us what is to be produced and of what specification. That is, it gives us dimension and limits within which dimension can vary. These specifications are laid down by the manufacturer.

26.4.2 Production

Here we should look into what we have manufactured and what was intended to.

26.4.3 Inspection

Here we examine with the help of SQC techniques whether the manufactured goods are within the specified limits or whether there is any necessity to widen the specifications or not. So SQC tells us as to what are the capabilities of the production process.

Therefore statistical quality control is considered as a kit of tools, which may influence decisions, related to the functions of specification, production or inspection. The effective use of SQC generally requires cooperation among those responsible for these three different functions or decisions at a higher level than any one of them. For this reason, the techniques should be understood at a management level that encompasses all the three functions.

Lesson 27

CONTROL CHARTS FOR VARIABLES

27.1 Introduction

In the previous lesson various terms used in the context of 'Statistical Quality Control' were described. As stated earlier the process control is carried out to control the manufacturing process so that the proportion of defective items is not excessively large. This is mainly achieved through the application of control charts which were developed by Dr. W. A. Shewart of Bell Telephone Laboratories, USA. These control charts provide a powerful tool of discovering and correcting the assignable causes of variation outside the "stable pattern "of chance causes, thus enabling us to stabilize and control our processes at desired performances and thus bring the process under statistical control. In industry, one can face with two kinds of problems: i) to check whether the process is conforming to standards laid down and ii) to improve the level of standards and reduce variability consistent with cost considerations. In this lesson we discuss various control charts for variables that are used in the industry.

27.2 Control Chart

To maintain the quality of the product manufactured, quality control charts are prepared to maintain the uniformity in the quality of the product manufactured. The power of the Shewhart technique of control chart lies in its ability to separate out the assignable causes of variability.

27.2.1 Merits of Control Chart

1. This makes possible the diagnosis and correction of many production troubles and often brings substantial improvements in product quality and reduction of spoilage and rework.
2. Moreover, by identifying certain of the quality variations as inevitable chance variations, the control chart tells when to leave to process alone and thus prevents unnecessary frequent adjustments that tend to increase the variability of the process rather than to decrease it.
3. By control charts, it is possible to detect assignable causes of variation and we can remove the factors responsible to bring back the production process within the stable system of variability.
4. It also permits better decisions on engineering tolerances and better comparisons between alternative designs and between alternative production methods.
5. Through improvement of conventional acceptance procedure it often provides better quality assurance at lower inspection cost.
6. It helps us in taking decision on matters relating to the quality. It provides us the basic variability of the quality characteristics. It estimates the capability of production process by estimating the inherent variation in the quality of the product manufactured.
7. It helps us in consistency of performance. The quality control charts tell us when to leave the production process alone and undisturbed and when to take remedial measures to bring back the quality under control.
8. It helps us to detect assignable causes of variation in the quality of product. This detection and correction of assignable causes of variation in quality of the product helps us in 3 ways viz.,
   (a) It ensures reliable quality level.

(b) It reduces the spoilage and rework.

(c) It builds us consumer confidence in the quality of the product which cannot be measured by any terms but which is of utmost importance.

Indirect benefits of SQC are:

(i) Introduction or improvement of the inspection department helps us to install an inspection department.

(j) It helps us to evaluate periodically the performance of the production process or department in terms of quality.

27.3 Objectives of Control Chart

a) To secure information to be used in establishing or changing specification or in determining whether a given process meet specifications.

b) To secure information to be used in establishing or changing production procedures. Such changes can be either elimination of assignable causes of variation or fundamental changes in production methods that may be called for whenever the production control charts makes it clear that specifications cannot be met with present methods.

c) To secure information to be used in establishing or changing inspection procedures or acceptance procedures or both.

d) To provide a basis for current decisions during production as to when to hunt for causes of variation and take action intended to correct them, and when to leave a process alone.

e) To provide a basis for current decisions on acceptance or rejection of the manufactured or purchased product.

27.4 Rational Subgroups

The Central idea in Shewhart's control chart technique is the division of observations into what are called rational subgroups. These are to be taken in such a way that variation within a subgroup may be attributed entirely to chance causes while systematic variation ; if at all exists, can occur only from one subgroup to another. That is subgroups should be selected in such a way that they are homogeneous as far as possible and that gives the maximum opportunity for variation from one subgroup to another so that different subgroups may indicate the presence of systematic variation.

The most obvious basis for the selection of subgroups is the order of production. As applied to control charts on production, this means that each subgroup should consist of the product of a machine or a homogeneous group of machines for a short period of time, so that there may not be any remarkable change in the cause system within that period. Therefore, if primary purpose of keeping the charts is to detect shifts in the process average, one subgroup should consist of items produced as nearly as possible at one time; The next subgroup should consist of items all produced at a single later time and so forth. The use of such subgroups would tend to reveal assignable causes of variation that come and go. However, there may be assignable causes that are not revealed merely by taking subgroups in the order of production e.g., two or more machines in a factory may have different patterns of variation. In this case it may be necessary to have different subgroups for different machines or for different operators or for different shifts. The problem of process control then boils down to the use of methods that would enable us to judge whether the distributions of the given quality characteristic for the

different subgroups are identical or not. In case the distributions are identical, the process may be supposed to be in control. Otherwise, the process will be considered to be out of control and one has to look for the source of trouble.

Shewhart suggested four as the ideal subgroup size. In the industrial use of the control chart, five seems to be the most common size. Because the essential idea of the control chart is to select subgroups in a way that gives minimum opportunity for variation within a group, it is desirable that subgroups be as small as possible. On the other hand, a size of four is better than three or two on statistical grounds because the distribution of $\bar{x}$ is nearly normal for subgroups of four or more even though the samples are taken from a non-normal universe; this fact is helpful in the interpretation of control chart limit. A reason for the use of five as the subgroup size is ease in computation of the average, which can be obtained by multiplying the sum by two and moving the decimal point one place to the left. Subgroups of two or three may often be used to good advantage, particularly where the cost of measurements is so high as to the use of larger subgroups. Larger subgroups such as 10 or 20 are sometimes advantageous where it is desired to make the control chart sensitive to small variations in the process average. The larger the subgroup size, the narrower the control limits on charts for $\bar{x}$ and the easier it is to detect small variations. Generally speaking, the larger the subgroup size, the more desirable it is to use standard deviation rather than range as a measure of subgroup dispersion. A practical working rule in this case is to use $\bar{x}$ and σ charts rather than $\bar{x}$ and R-charts whenever the subgroup size is greater than 15.

27.5 Techniques of Control Chart

Shewhart's control chart technique is a particular diagrammatic method of making this comparison and thus deciding whether the process is or is not affected by systematic variation. Let us first focus our attention on some parameter of the distribution say θ. Let T be the corresponding statistic. If the process is in control then θ must be same from subgroup to subgroup and consequently the fluctuations in the values of T from sample to sample should be due to random variation alone. Supposing in such a case

$E(T) = \mu_T$ and $V(T) = \sigma_T^2$

One may take any value of T lying outside the limits $(\mu_T - 3\sigma_T)$ and $(\mu_T + 3\sigma_T)$ as an indication of the presence of systematic variation reason being

$$P\{|T - \mu_T| \leq 3\sigma_T\} = 0.9973$$

Even when T is non-normal, we have from Chebyshev's inequality

$$P\{|T - \mu_T| \leq 3\sigma_T\} = 0.8889$$

The Central Limit Theorem also states that whatever be the distribution of parent population, when we draw samples from the population the distribution of sample mean $\bar{X}$ will follow Normal distribution. Thus if the observed T_i lies between the limits $(\mu_T - 3\sigma_T)$ and $(\mu_T + 3\sigma_T)$, it is taken to be a fairly good indication of non-

existence of assignable causes of variation at the same time when i^{th} sample was taken. If the observed T_i wanders outside the limits, one suspects the existence of assignable causes of variation and the process is supposed to be out of control. The obvious action is then to stop the process and to hunt for and remove the assignable causes. The testing is however, done by means of a graph where sample number is plotted on X-axis and the statistic T are plotted on Y-axis. The lower control limit (LCL) $\mu_T - 3\sigma_T$ and the upper control limit (UCL) $\mu_T + 3\sigma_T$ are shown on the chart by means of horizontal lines. The line corresponding to the mean value μ_T is called the Central line. The process is said to be out of control if any point falls below the LCL or above the UCL (see fig. 27.1). Even if all the points may be inside the control limits, indications of trouble or presence of assignable causes of variation in the process are sometimes evidenced from unusual patterns or arrangement of points e.g.,

(a) A series of points all falling close to one of the control limits.

(b) A long series predominantly on one side of the central line.

(c) A series of points exhibiting a trend.

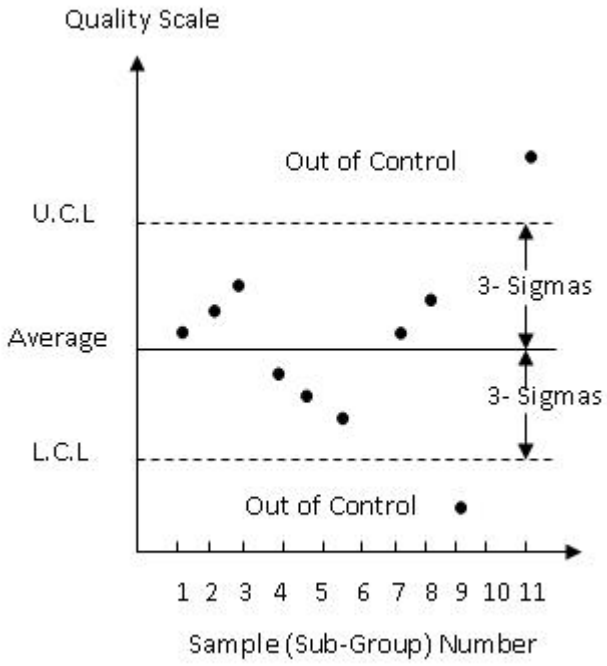


Fig. 27.1 Outline of a control chart

27.6 Types of Control Chart

Control charts may be of two types as stated below:

1. Control chart with respect to given standards: Here our purpose is to discover whether the observed values of $\bar{x}$, s, p etc. for samples of n items differ from standard values $\bar{x}, \sigma, p$ etc. by an amount greater than what should be attributed to chance. The standard values may be either established by authority as some desired values designated by specification or some economic standard levels provided by experience. These charts are used to maintain quality uniformly at the desired level.

2. Control chart with no standards given: Here we want to discover whether the observed values of $\bar{x}, \sigma, p$ etc. for samples of size n vary amongst themselves by amount greater than what should be attributed to chance. These charts are used to detect lack of constancy of the cause system. So far as the size of the samples for different subgroups are concerned, small samples at shorter intervals are always preferable to large samples at longer intervals.

27.7 Control Charts for Mean, Standard Deviation and Range

Suppose we are dealing with a quality characteristic like length, diameter or breaking strength, Moisture content in khoa, fat content in Ice-cream, filling of milk in bottles/pouches, filling of Ghee & condensed milk in tins, moisture in butter etc. For manufactured articles subject solely to random variation such a variable may be supposed to be normally distributed. So the different distributions of x for the different subgroups are then all supposed to be of the normal type, the i^{th} subgroup giving a distribution with mean μ_i and variance σ_i^2. To examine whether the process is in control, we need to see whether the μ'$_s$ and the σ's are the same. The four types of situations encountered are given below:-

(a) The process is in control.
(b) The mean is out of control but not the standard deviation (s.d)
(c) The s.d. is out of control but not the mean.
(d) Both mean and s.d. are out of control.

The appropriate statistics corresponding to μ and σ are $\bar{x}$ and s. Hence the whole judgment regarding control or lack of it is based on control charts for $\bar{x}$ and s. It is to be remembered, however that the range R inspite of its theoretical inferiority to s, is simpler and easier to compute. Hence in quality control, the range is often preferred to s.d. and one uses $\bar{x}$ and R charts instead of $\bar{x}$ and σ chart.

27.7.1 Control charts for mean

27.7.1.1 Standards given

For samples of size n per subgroup, we have $E(\bar{X}) = \mu$ and $\sigma_x = \sigma/\sqrt{n}$

Assuming that n observations in each subgroup are mutually independent .Hence if, the values, for μ and σ', are specified as $\bar{\bar{X}}'$ and σ', the control chart for $\bar{\bar{X}}$ will be given by

LCL= $\bar{\bar{X}}' - 3\frac{\sigma'}{\sqrt{n}} = \bar{\bar{X}}' - A\sigma'$

Central line = $\bar{\bar{X}}'$

UCL = $\bar{\bar{X}}' + 3\frac{\sigma'}{\sqrt{n}} = \bar{\bar{X}}' + A\sigma'$ where $A = 3/\sqrt{n}$

27.7.1.2 standards not given

Let there be m subgroups and let the successive sample means be $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots\ldots\ldots\ldots, \bar{X}_m$; the successive standard deviations be $s_1, s_2 \ldots . s_m$ and the successive ranges be $R_1, R_2, \ldots . R_m$. Since μ and σ are unspecified, these are estimated from sample themselves. Let

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{m} \bar{x}_i}{m}, \bar{s} = \frac{\sum_{i=1}^{m} s_i}{m} \text{ and } \bar{R} = \frac{\sum_{i=1}^{m} R_i}{m}$$

which are pooled mean, the mean of sample standard deviation and mean of sample ranges respectively.

The relations $E(\bar{\bar{X}}) = \mu$ and $E(s) = C_2\sigma$ (valid for a normal variable x)

where $C_2 = \frac{\Gamma(n/2)}{\Gamma(n-1/2)}\sqrt{\frac{2}{n}}$

and $E(R) = d_2\sigma$ (valid for a normal variable x) where d_2 is also a function of n but not as simple as C_2, provide us with an estimate of μ and two alternative estimates for σ, viz.

$$\hat{\mu} = \bar{\bar{x}} \qquad (27.1)$$

$$\hat{\sigma} = \bar{s}/C_2 \qquad (27.2)$$

$$\hat{\sigma} = \bar{R}/d_2 \qquad (27.3)$$

In case one uses the estimate (27.1)and (27.2), the control chart for mean will be based on

LCL = $\bar{\bar{X}} - 3\frac{\bar{s}}{C_2\sqrt{n}} = \bar{\bar{x}} - A_1\bar{s}$

Central line = $\bar{\bar{X}}$

UCL = $\bar{\bar{X}} + 3\frac{\bar{s}}{C_2\sqrt{n}} = \bar{\bar{x}} + A_1\bar{s}$

Where $A_1 = \frac{3}{C_2\sqrt{n}}$ and is tabulated together with C_2 for different values of n in tables provided at the end of the lesson.

On the other hand, if one uses the estimate (27.1) and (27.3), the control chart for mean is given by

LCL = $\bar{\bar{X}} - 3\frac{\bar{R}}{d_2\sqrt{n}} = \bar{\bar{X}} - A_2\bar{R}$

Central line = $\bar{\bar{X}}$

UCL = $\bar{\bar{X}} + 3\frac{\bar{R}}{d_2\sqrt{n}} = \bar{\bar{X}} + A_2\bar{R}$

Where $A_2 = \frac{3}{d_2\sqrt{n}}$ and is again tabulated for different values of n in tables provided at the end of the lesson.

27.7.2 Control chart for standard deviation

27.7.2.1 Standards given

For a normally distributed variable x, we have

E (s) = $C_2\sigma$

where σ_S = $\sigma\sqrt{\frac{n-1}{n} - C_2^{\,2}}$

If the standard value of σ is σ′, then the chart will be based on

LCL = $C_2\sigma' - 3\sigma'\sqrt{\frac{n-1}{n} - C_2^{\,2}}$

LCL = $\left[C_2 - 3\sqrt{\frac{n-1}{n} - C_2^{\,2}}\right]\sigma' = B_1\sigma'$

Central line = $C_2\sigma'$

UCL = $\left\{C_2\sigma' + 3\sigma'\sqrt{\frac{n-1}{n} - C_2^{\,2}}\right\} = \left\{C_2 + 3\sqrt{\frac{n-1}{n} - C_2^{\,2}}\right\}\sigma' = B_2\sigma'$

where, of course

$$B_1 = C_2 - 3\sqrt{\frac{n-1}{n} - C_2^{\,2}}$$

$$B_2 = C_2 + 3\sqrt{\frac{n-1}{n} - C_2^2}$$

where the values of B_1, B_2 and C_2 are provided in the tables for different values of n.

27.7.2.2 Standards not given

In this case we use the estimate of $\hat{\sigma} = \frac{\bar{s}}{C_2}$ for σ and get the control chart on replacing $C_2\sigma'$ by $\bar{s}$. Therefore,

LCL = $\bar{s} - 3\frac{\bar{s}}{C_2}\sqrt{\frac{n-1}{n} - C_2^2} = \left[1 - \frac{3}{C_2}\sqrt{\frac{n-1}{n} - C_2^2}\right]\bar{s} = B_3\bar{s}$

Central line = $\bar{s}$

UCL = $\bar{s} + 3\frac{\bar{s}}{C_2}\sqrt{\frac{n-1}{n} - C_2^2} = \left[1 + \frac{3}{C_2}\sqrt{\frac{n-1}{n} - C_2^2}\right]\bar{s} = B_4\bar{s}$

Where $B_3 = 1 - \frac{3}{C_2}\sqrt{\frac{n-1}{n} - C_2^2}$

$$B_4 = 1 + \frac{3}{C_2}\sqrt{\frac{n-1}{n} - C_2^2}$$

where the values of B_3 and B_4 are provided in the tables for different values of n.

In either case if LCL comes out to be negative, then it is to be taken as zero because in no case's' can be negative.

27.7.3 Control charts for range

27.7.3.1 Standards Given

For a normally distributed variable X, we have

E (R) =$d_2\sigma$; $\sigma_R = D\sigma$

If the standard value of σ is given to be σ′, then the R- chart will be given by

LCL = $d_2\sigma' - 3D\sigma' = (d_2 - 3D)\sigma' = D_1\sigma'$

CL = $d_2\sigma'$

UCL = $d_2\sigma' + 3D\sigma' = (d_2 + 3D)\sigma' = D_2\sigma'$

where $D_1 = d_2 - 3D$ and $D_2 = d_2 + 3D$

where values of d_2, D_1 and D_2 are provided in the tables for different values of n

27.7.3.2 Standards not Given

When no standard value of σ is specified, it is estimated by $\bar{R}/d_2$ i.e., $\hat{\sigma} = \bar{R}/d_2$

$$\text{LCL} = \bar{R} - 3\frac{D}{d_2}\bar{R} = \left(1 - 3\frac{D}{d_2}\right)\bar{R} = D_3\bar{R}$$

$$\text{CL} = \bar{R}$$

$$\text{UCL} = \bar{R} + 3\frac{D}{d_2}\bar{R} = \left(1 + 3\frac{D}{d_2}\right)\bar{R} = D_4\bar{R}$$

where the values of D, d_2, D_3 and D_4 are tabulated in the tables provided for different values of n.

In either case if LCL comes out to be negative, then it should be taken as zero as R can't be negative.

27.8 Criterion for Detecting Lack of Control in $\bar{X}$ and R Charts

The main objective of the control chart is to indicate when a process is not in control. The criteria for detecting lack of control are, therefore, of crucial importance. The following situations depict lack of control.

1. *A point outside the control limits:* The probabilistic considerations provide a basis for lack of control in such a situation. A point going beyond the control limits (below LCL and above UCL) is a clear indication of the presence of assignable causes of variation which must be identified and corrected. A point outside control limit may result from an increased dispersion or change in level. Lack of uniformity may be due to the variation in the quality of raw materials, deficiency in skill of the operators, loss of alignment among machines, change of working conditions etc. It may be indicated by a point (points) above the UCL for ranges.
2. *A run of seven of more points:* Although all the sample points are within control limits usually the pattern of points in the chart indicates assignable causes. A run of 7 or more points above or below the central line in the control chart indicate shift in the process level/process average. On the R-chart a run of points above the central line is indicative of increase in process spread and therefore represents a undesirable situation. A run of points below central line indicates an improvement in the sense that the variability has been reduced i.e., the process could hold to a closer tolerance.
3. One or more points in the vicinity of control limits or a run of points beyond some secondary limits e.g., A run of 2, 3 points beyond 2 σ limits or a run of 4, 5 points beyond 1σ limit.
4. The sample points on $\bar{X}$ and R charts, too close to the central line, exhibit another form of assignable causes. This situation represents systematic differences within samples or subgroups and results from improper selection of samples and biases in measurements
5. The upward or downward trends exhibited by sample points on the control chart are also an indication of assignable cause. This trend pattern is usually observed in engineering industry, indicating the gradual shift in the process level.
6. In some cases the cyclic pattern of points in the control chart indicates the presence of assignable causes of variation. Such pattern are due to material or/ and any mechanical reasons.

27.9 Interpretations of $\bar{X}$ and R-Chart

In order of judge whether the process is in a state of control, $\bar{X}$ and R charts should be examined together and process should be deemed in statistical control if both the charts show a state of control. Situations exist where R chart is in a state of control but $\bar{X}$ chart is not. Different situations have been summarized below in a tabular form along with the corresponding interpretation:

Situations in		Interpretation
$\bar{X}$ chart	R-chart	
1. Points beyond limits only on one side	In control	Level of process has shifted
2. Points beyond limit on both sides	In control	Level of process is changing in erratic manner needing frequent adjustments
3. Points beyond limit on both sides	Out of control	Variability has increased
4. Out of control on one side	Out of control	Both the level and variability have changed
5. Run of 7 or more points on one side of the central line	In control	Shift in the process level
6. Trend of 7 or more points no points out of control limits	In control	Process level is gradually changing
7. -	Runs of 7 or more points above central line	Variability has increased
8. -	Points too close to the central line	Systematic differences within subgroups
9. Points too close to the central line	-	Systematic differences within subgroups

No production process is perfect enough to produce all the items exactly alike. Some amount of variation in the produced items is inherent in any production process. This variation is the totaling of numerous characteristics of the production process viz., raw material, machine setting and handing operators etc. The control limits in $\bar{X}$ and R or $\bar{X}$ and σ charts are so placed that they reveal the presence or absence of assignable causes of variation (a) in the average-mostly related to machine setting (b) in the range-mostly related to the negligence on the part of operators.

The above procedure is illustrated through following example :

Example 27.1 : Khoa is manufactured in a continuous khoa making plant. The specifications for the moisture content are 30±2 percent. To keep control on the quality of khoa, it was decided to check the moisture content.

15 sub-groups of size 4 each were taken at an hourly interval and the moisture content noted. Set up control charts for mean, standard deviation and range.

Sub-group Numbers	Moisture content in khoa (in percentage)			
	1	2	3	4
1.	30.5	31.9	29.8	33.2
2.	28.0	28.3	29.6	30.7
3.	31.4	27.8	29.5	32.3
4.	28.4	30.0	27.9	29.3
5.	26.9	29.0	28.8	27.3
6.	28.5	27.9	30.4	31.3
7.	32.4	31.8	32.1	30.9
8.	31.5	32.1	30.6	28.5
9.	30.8	33.0	32.7	31.4
10.	27.8	27.9	30.0	29.1
11.	28.9	30.1	29.4	31.1
12.	27.6	30.4	28.2	29.6
13.	29.2	28.7	29.0	30.1
14.	31.1	31.8	30.6	28.5
15.	30.6	30.8	29.4	30.0

Solution :

Prepare the following table:

Sub-group Numbers	Moisture content in khoa (in percentage)				Total	Mean	Standard Deviation	Range
	1	2	3	4				
1	30.5	31.9	29.8	33.2	125.4	31.35	1.5111	**3.4**
2	28.0	28.3	29.6	30.7	116.6	29.15	1.2450	**2.7**
3	31.4	27.8	29.5	32.3	121	30.25	2.0075	**4.5**
4	28.4	30.0	27.9	29.3	115.6	28.90	0.9345	**2.1**
5	26.9	29.0	28.8	27.3	112	28.00	1.0551	**2.1**
6	28.5	27.9	30.4	31.3	118.1	29.53	1.5924	**3.4**
7	32.4	31.8	32.1	30.9	127.2	31.80	0.6481	**1.5**
8	31.5	32.1	30.6	28.5	122.7	30.68	1.5756	**3.6**
9	30.8	33.0	32.7	31.4	127.9	31.98	1.0468	**2.2**

10	27.8	27.9	30.0	29.1	114.8	28.70	1.0488	**2.2**
11	28.9	30.1	29.4	31.1	119.5	29.88	0.9535	**2.2**
12	27.6	30.4	28.2	29.6	115.8	28.95	1.2793	**2.8**
13	29.2	28.7	29.0	30.1	117	29.25	0.6028	**1.4**
14	31.1	31.8	30.6	28.5	122	30.50	1.4213	**3.3**
15	30.6	30.8	29.4	30.0	120.8	30.20	0.6325	**1.4**
Total					449.10	17.55	38.80	**449.10**

In order to calculate LCL and UCL and obtain the following

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{m} \bar{x}_i}{m} = \frac{449.10}{15} = 29.94,$$

$$\bar{s} = \frac{\sum_{i=1}^{m} s_i}{m} = \frac{17.55}{15} = 1.17 \text{ and } \bar{R} = \frac{\sum_{i=1}^{m} R_i}{m} = \frac{38.80}{15} = 2.5867$$

For sub-group of size 4

$A_1 = 1.880, B_3 = 0, B_4 = 2.266, C_2 = 0.7979 \quad D_3 = 0 \quad D_4 = 2.282$

Mean Chart ($\bar{\bar{X}}$ Chart)

LCL = $\bar{\bar{X}} - 3\frac{\bar{s}}{C_2\sqrt{n}} = \bar{\bar{x}} - A_1\bar{s} = 29.94 - 1.628 \times 1.17 = 28.03524$

Central line = $\bar{\bar{X}} = 29.94$

UCL = $\bar{\bar{X}} + 3\frac{\bar{s}}{C_2\sqrt{n}} = \bar{\bar{x}} + A_1\bar{s} = 29.94 + 1.628 \times 1.17 = 31.8447$

s- Chart (S. D. Chart)

LCL = $B_3\bar{s} = 0 \times 1.17 = 0$

Central line = $\bar{s} = 1.17$

UCL = $B_4\bar{s} = 2.266 \times 1.17 = 2.6519.$

The mean and standard deviation charts are shown in fig. 27.2

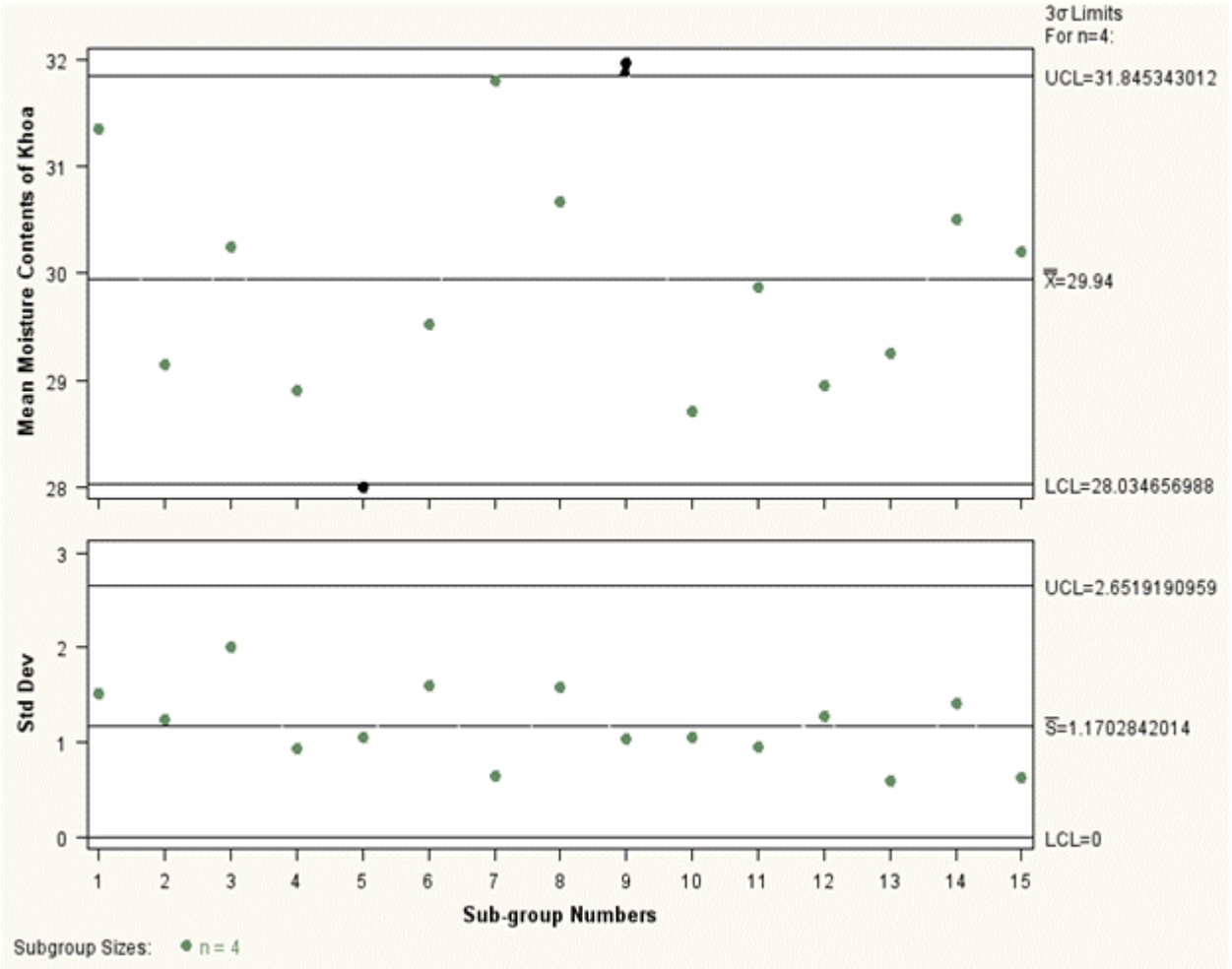


Fig 27.2 Mean and standard deviation chart for moisture contents of khoa

Range Chart

LCL $= D_3\bar{R} = 0$

CL $= \bar{R} = 2.5867$

UCL $= D_4\bar{R} = 2.282 \times 2.5867 = 5.9028$

The control chart is shown in fig. 27.3

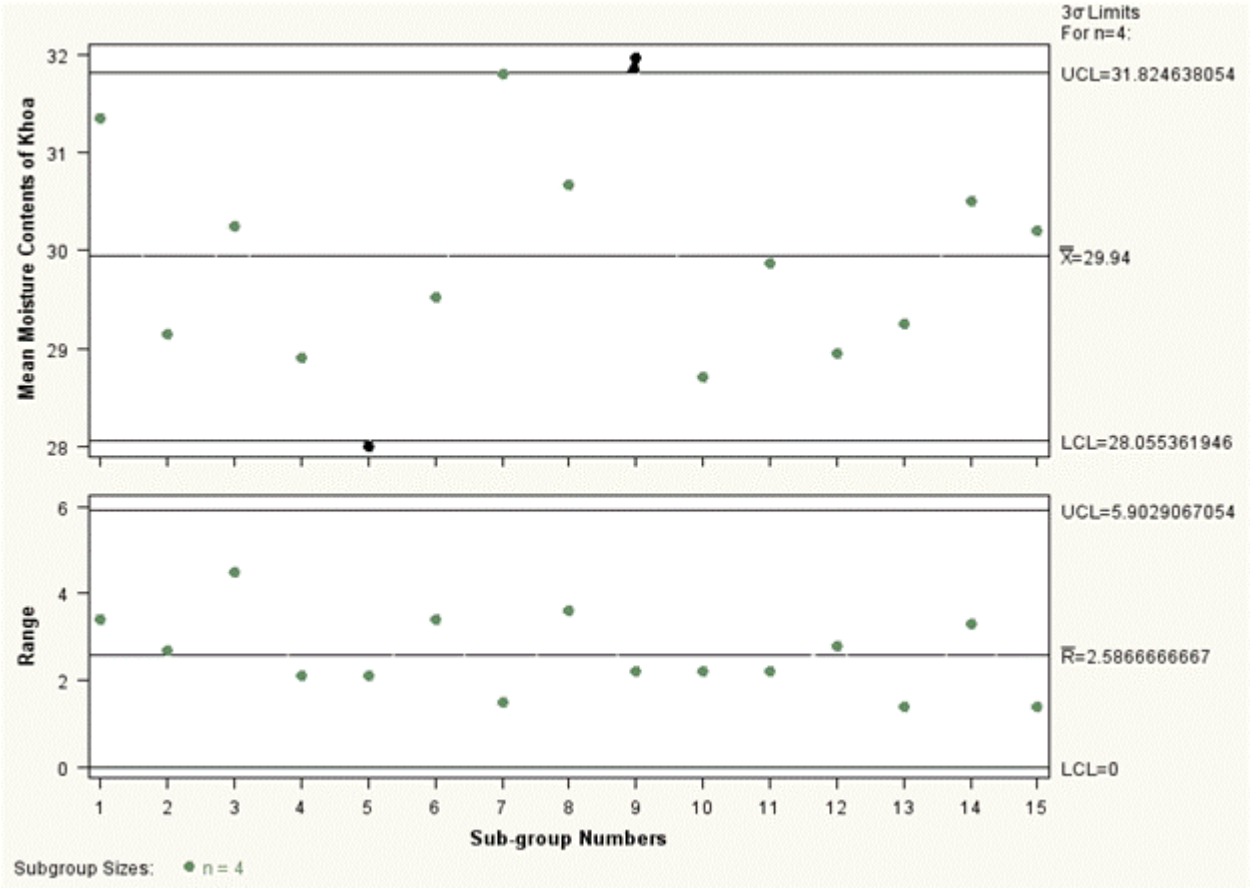


Fig 27.3 Mean and range chart for moisture contents of khoa

It may be seen from the above control charts for mean and standard deviation as well as mean and range , all the points are within the control limits except 5^{th} and 9^{th} sub-group in mean chart. Therefore, the process is out of control.

Lesson 28

CONTROLCHARTS FOR ATTRIBUTES

28.1 Introduction

In the previous lesson, we have discussed the control charts for variables. In spite of wide application of $\bar{X}$ and R (σ - charts) as a powerful tool of diagnosis of sources of trouble in a production process, their use is restricted because of the following limitations:

- These are the charts for variables i.e. for quality characteristics which can be measured and expressed in numbers.
- In certain situations they are impracticable and uneconomical e.g., if the number of measurable characteristics, each of which could be a possible candidate for $\bar{X}$ and R-chart, is too large, say 30000 or so then obviously there cannot be many control charts.

As an alternative to $\bar{X}$ and R charts, there are control charts for attributes which can be used for quality characteristics i.e., (i) which can be observed only as an attribute by classifying an item as defective or non defective that is conforming to specifications or not. (ii) which are actually observed as attributes even though they could be measured as variables. In this lesson, we will discuss the control charts for attributes.

28.2 Control Charts for Number of Defective and Fraction Defective

When the quality characteristic is an attribute, and each item is recorded as either defective or non defective, to judge whether the process is in control, one has to ascertain whether the population fraction defective P is same for all subgroups. This judgment may be made either on the number of defective say d in the sample or on the fraction defective $p = {}^{d}/_{n}$ in the sample, where n as before denotes the number of items inspected per subgroup.

28.2.1 Control charts for number of defective

28.2.1.1 Standards given

Assuming that each random sample is taken with replacements or even if taken without replacements, is taken from a practically infinite population, we may suppose that d = np is distributed in the binomial form with

$$E(np) = nP \qquad \sigma_{np} = \sqrt{np(1-p)}$$

P being the same for all subgroups if and only if the process is in control. Hence if þ′ be the specified standard value of P, The control limits for number of defective chart will be given as

$$LCL = np' - 3\sqrt{np'(1-p')}$$
$$CL = np'$$
$$UCL = np' + 3\sqrt{np'(1-p')}$$

28.2.1.2 Standards not given

If no standard value is specified for p, it will have to be estimated from the samples themselves. The appropriate estimate for the mean fraction defective is

$$\bar{p} = \frac{\sum_{i=1}^{m} p_i}{m}$$

The control limits for number defective chart will then be

$$\text{LCL} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$\text{CL} = n\bar{p}$$

$$\text{UCL} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

Note: Since $n\bar{p}$ can never be negative. Hence, if LCL, according to either of the formula comes out to be negative then it is to be taken as zero.

Example 1 : The following table gives the number of bottles broken in a sample of size 25:

Sample number	Number of defectives	Sample number	Number of defectives	Sample number	Number of defectives
1	3	10	3	19	3
2	3	11	2	20	3
3	2	12	3	21	2
4	7	13	1	22	2
5	1	14	9	23	1
6	8	15	1	24	0
7	2	16	0	25	1
8	1	17	4		
9	0	18	8		

Construct the control chart for number defective. State whether the process is in a state of control.

Solution : Here we have a fixed sample size m=25 for each lot .Prepare the following table :

Sample number	Number of defectives (d_i)	Fraction defective (p_i)	Sample number	Number of defectives (d_i)	Fraction defective (p_i)	Sample number	Number of defectives (d_i)	Fraction defective (p_i)
1	3	0.12	10	3	0.12	19	3	0.12

2	3	0.12	11	2	0.08	20	3	0.12
3	2	0.08	12	3	0.12	21	2	0.08
4	7	0.28	13	1	0.04	22	2	0.08
5	1	0.04	14	9	0.36	23	1	0.04
6	8	0.32	15	1	0.04	24	0	0
7	2	0.08	16	0	0	25	1	0.04
8	1	0.12	17	4	0.16	Total	70	2.8
9	0	0.04	18	8	0.32			

Calculate mean fraction defective as

$$\bar{p} = \frac{\sum_{i=1}^{m} p_i}{m} = \frac{70}{25 \times 25} = \frac{2.8}{25} = 0.112$$

(due to constant sample size)

The control limits for number defective chart will then be

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 25 \times 0.112 - 3\sqrt{25 \times 0.112 \times 0.888} = 2.8 - 4.73$$
$$= -1.93$$

(to be taken as 0 as LCL can't be negative as stated earlier)

$$CL = n\bar{p} = 25 \times 0.112 = 2.8$$

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 25 \times 0.112 + 3\sqrt{25 \times 0.112 \times 0.888} = 2.8 + 4.73$$
$$= 7.53$$

The control limits for np -chart and various points are shown in fig. 28.1

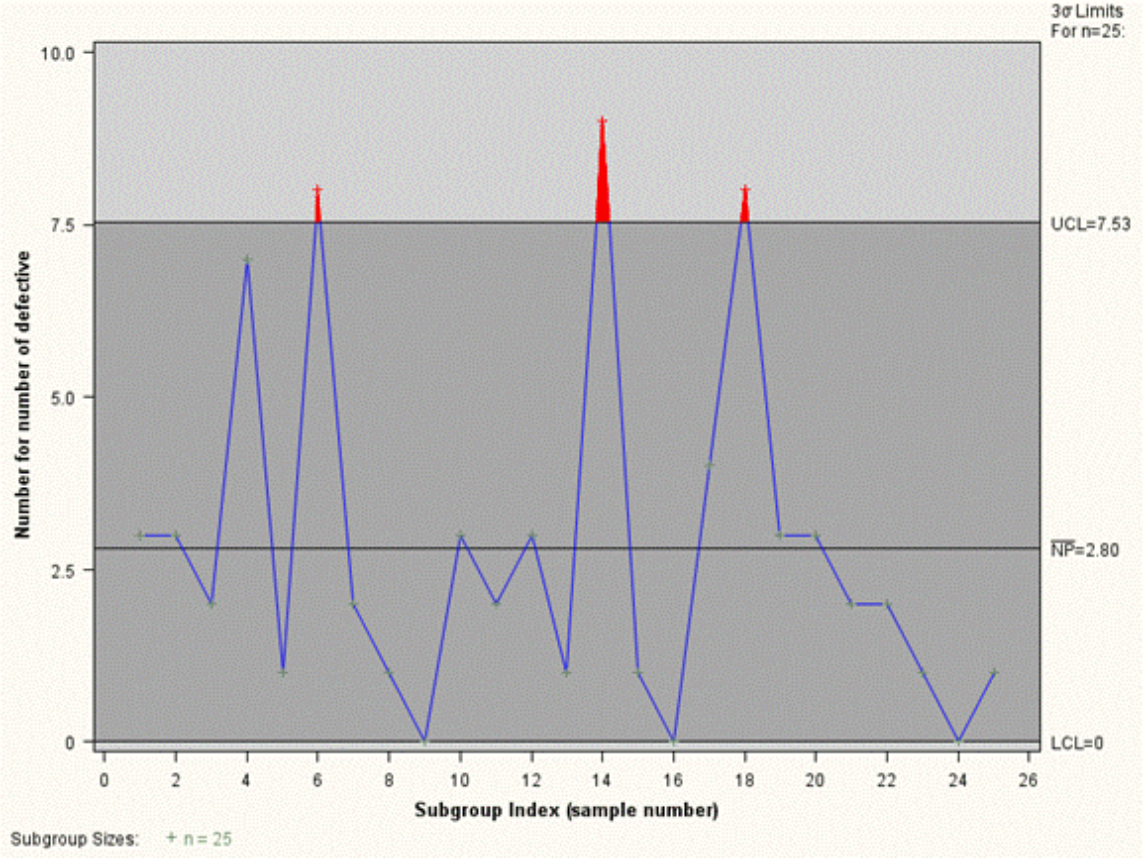


Fig. 28.1 Control chart for number defective (np- chart)

Since , some of the points i.e. 6^{th}, 14^{th} and 18^{th} are outside the control limits therefore the process is not in a state of control.

28.3.3 Control Charts for Fraction Defective (p - chart)

28.3.3.1 Standards given

In case one constructs a control chart for p instead of, np one uses the relations

Supposing p′ is the specified standard for P, the control limit for fraction defective chart will consist of

LCL = $p' - 3\sqrt{p'(1-p')/n} = p' - A\sqrt{p'(1-p')}$

CL = p'

UCL = $p' + 3\sqrt{p'(1-p')/n} = p' + A\sqrt{p'(1-p')}$ where $A = 3/\sqrt{n}$

28.3.2 Standards not given

Here the common value of P will be estimated by $\bar{p}$ and the control limits for fraction defective chart will be

LCL = $\bar{p} - 3\sqrt{\bar{p}(1-\bar{p})/n} = \bar{p} - A\sqrt{\bar{p}(1-\bar{p})}$

CL = $\bar{p}$

UCL = $\bar{p} + 3\sqrt{\bar{p}(1-\bar{p})/n} = \bar{p} + A\sqrt{\bar{p}(1-\bar{p})}$

Where $A = 3/\sqrt{n}$

Here also since 'þ′can never be negative. Hence if LCL comes out to be negative, then it is to be taken as zero.

A þ-chart or np-chart is advantageous because they may be used even for characters that are observed as variables. The cost of obtaining data on an attribute is usually less than that for data on variables. The cost of compiling a þ-chart may also be less since a þ-chart may be used for any number of characteristics and may replace pairs of $\bar{X}$, s, or $\bar{X}$, R charts. In case the sample size is constant, it is immaterial whether one uses the nþ-chart or the þ-chart. If however the sample size varies, then resulting chart will be highly confusing, whereas in the Þ-chart the central line will be invariant. It is, therefore, simpler and preferable to use þ-chart in case the sample size varies. Instead of computing control limits for each sample size separately, two sets of limits may be computed based on the minimum and maximum sample sizes. Action need not be taken for points lying within the inner set of limits, while action must be taken for points lying beyond the outer limits. For other points, action should be based on exact control limits. The confusion in þ-chart (or nþ-chart) with varying control limits can be avoided with some additional computation. For that, instead of plotting p in the control chart one should plot the standardized value viz.,

$$Z = \frac{p - p'}{\sqrt{p'(1-p')/n}} \text{ or } \frac{p - \bar{p}}{\sqrt{\bar{p}(1-\bar{p})/n}}$$

According as the standard value for þ is specified or not, $\bar{p}$ being the weighted mean of sample proportions with sample sizes as weight. The central line as well as the control limits becomes invariant with n, since obviously here

LCL = -3, CL = 0, UCL =3

28.3.3 Interpretations of p-chart

1. If all the sample points fall within the control limits without exhibiting any specific pattern, the process is said to be in control. In such a case, the observed variations in the fraction defective are attributed to the stable pattern of chance causes and the average fraction defective p is taken as the standard fraction defective P.
2. Points outside the UCL are termed as high spots. These suggest deterioration in the quality and should be regularly reported to the Production Engineers. The reason for such deterioration can be known and

removed if the details of conditions under which data were collected, it may be found, if there was any change of inspection or inspection standards.

3. Points below LCL are called low spots. Such points represent a situation showing improvement in the product quality. However, before taking this improvement for guarantee it should be investigated if there was any slackness in inspection or not.

4. When a number of points fall outside the control limits, a revised estimate of P should be obtained by eliminating all the points that fall above UCL (it is assumed that points that fall below LCL are not due to faulty inspection). The standard fraction defective P should be revised periodically in this way.

Example 2: Table below gives the results of inspection of nuts used in equipment. The nuts were packed in 20 lots of 100 each.

Lot number	Number defectives	Lot number	Number defectives
1	5	11	4
2	10	12	7
3	12	13	8
4	8	14	2
5	6	15	3
6	5	16	4
7	6	17	5
8	3	18	8
9	3	19	6
10	5	20	10

Construct the control chart for fraction defective. State whether the process is in a state of control.

Solution : Here we have a fixed lot size n=100 for each lot .Prepare the following table :

Lot size (n)	Number defective (np)	Fraction defective (p)	Lot no.	Lot size (n)	Number defective (np)	Fraction defective (p)
100	5	0.05	11	100	4	0.04
100	10	0.1	12	100	7	0.07
100	12	0.12	13	100	8	0.08
100	8	0.08	14	100	2	0.02
100	6	0.06	15	100	3	0.03
100	5	0.05	16	100	4	0.04
100	6	0.06	17	100	5	0.05

	100	3	0.03	18	100	8	0.08
	100	3	0.03	19	100	6	0.06
	100	5	0.05	20	100	10	0.1
			G. total		2000	120	1.2

$$\bar{p} = \frac{\sum_{i=1}^{m} p_i}{m} = \frac{120}{20 \times 100} = 0.06$$

The control limits for fraction defective chart will then be

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 - 3\sqrt{\frac{0.06 \times 0.94}{100}} = 0.06 - 0.0711 = -0.011$$

(to be taken as 0 as LCL can't be negative as stated earlier)

$$CL = \bar{p} = 0.06$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 + 3\sqrt{\frac{0.06 \times 0.94}{100}} = 0.06 + 0.0711 = 0.1311$$

The control limits and various points are shown in fig. 28.2

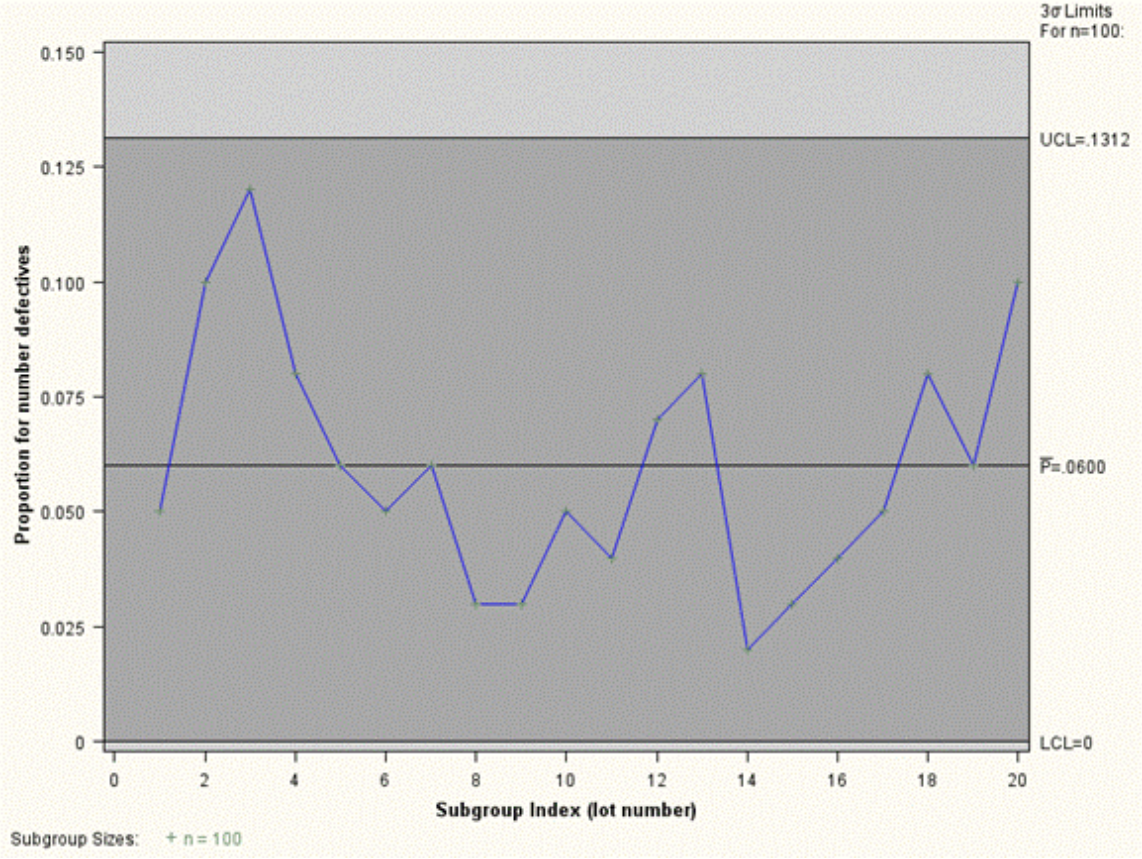


Fig. 28.2 Control chart for proportion number defective(fraction defective p-chart)

The control chart shown in fig. 28.2 shows that all the points are falling within control limits. Hence, the process is in a state of control.

28.4 Control Chart for Number of Defects Per Unit (C-Chart)

The $\bar{X}$ and R control charts may be applied to any quality characteristic that is measurable. The control chart for p may be applied to the results of any inspection that accepts or rejects individual items of a product. Thus, both these types of charts are broadly useful in any SQC programme. The control chart for number of defects per unit (c-chart) has a much more restricted field of usefulness. In many manufacturing plants there may be no opportunities for its economic use even though there are dozens of places where $\bar{X}$ & R charts and p charts can be used advantageously.

28.4.1 Distinction between a defect and a defective

A defective is an article that in some way fails to conform to one or more given specifications. Each instance of the article's lack of conformity to specifications is a defect. Every defective contains one or more defects.The n þ chart which was explained previously, applies to the number of defectives in subgroups of constant size. The c-chart which will now be explained, applies to the number of defects in subgroups of constant size. In most cases, each subgroup for the c-chart consists of a single article; the variable C consists of number of defects observed in one article. However, it is not necessary that the subgroup for C-chart be a single article; it is essential only that

the subgroup size be constant in the sense that the different subgroups have substantially equal opportunity for the occurrence of defects.

In many kinds of manufactured articles, the opportunities for defects are numerous, even though the chances of a defect occurring in any one spot are small. Whenever this is true, it is correct as a matter of statistical theory to base control limits on the assumption that the Poisson distribution is applicable. The limits on the control chart for C are based on this assumption. Some representative types of defects to which C chart may be applied are as follows

- C is the number of defective rivets in an aircraft wing or fuselage.
- C is the number of breakdowns at weak spots in insulation in a given length of insulated wire subjected to a specified test voltage.
- C is the number of surface defects observed in a galvanized sheet or painted, plated, or enameled surface of a given area or number of seeds on the surface of cheese or Paneer.
- C is the number of "seeds" (small air pockets) observed in a glass bottle.
- C is the number of imperfections observed in a bolt of cloth.
- C is the number of surface defects observed in a roll of coated paper or sheet of photographic film.

28.4.2 Control limits for C-chart

A control chart for C aims at detecting any differences, that may exist, among the Poisson distributions for the different subgroups or, in other words among the λ-values for the subgroups.

28.4.2.1 When standards specified

We know that for a Poisson Variables C with parameter λ

$$E(C) = \lambda \text{ and } \sigma_c = \sqrt{\lambda}$$

Hence if a standard value for λ say C′ is provided, then the control chart for C will be given by

$$LCL = C' - 3\sqrt{C'}$$

CL = C′

$$LCL = C' + 3\sqrt{C'}$$

28.4.2 Standards not specified

When no standards are specified then λ is estimated from observed C value. Supposing C_i (i=1, 2,---.m) is the C value for the sample taken from i^{th} subgroup, the appropriate estimate of λ will be $\bar{C} = \sum_{i=1}^{m} C_i / m$

In the above case we replace C′ by $\bar{C}$ the control limits for C-chart is given by

$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$

$$CL = \bar{C}$$

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$

Note: Since C can't be negative. Hence if LCL comes out to be negative, it should be taken as Zero. The above formulae relate to C-charts with samples of constant size from all subgroups. In most cases each subgroup sample will consist of a single article.

Example 3 : The following table gives the number of defects (C) noted at the final inspection of a dairy equipment.

Equipment No.	No. of defects	Equipment No.	No. of defects
1	7	9	20
2	15	10	11
3	13	11	22
4	18	12	15
5	10	13	8
6	14	14	24
7	7	15	14
8	10	16	8

Set up control limits for C chart and state whether the process is in a state of control for the process.

Solution : Here m=16

Let us calculate $\bar{C} = \sum_{i=1}^{m} C_i / m = 216/16 = 13.5$

The control limits for C Chart

$$LCL = \bar{C} - 3\sqrt{\bar{C}} = 13.5 - 3\left(\sqrt{13.5}\right) = 13.5 - 11.0227 = 2.4773$$

$$CL = \bar{C} = 13.5$$

$$UCL = \bar{C} + 3\sqrt{\bar{C}} = 13.5 + 3\left(\sqrt{13.5}\right) = 13.5 + 11.0227 = 24.5227$$

The control limits and various points are shown in fig. 28.3

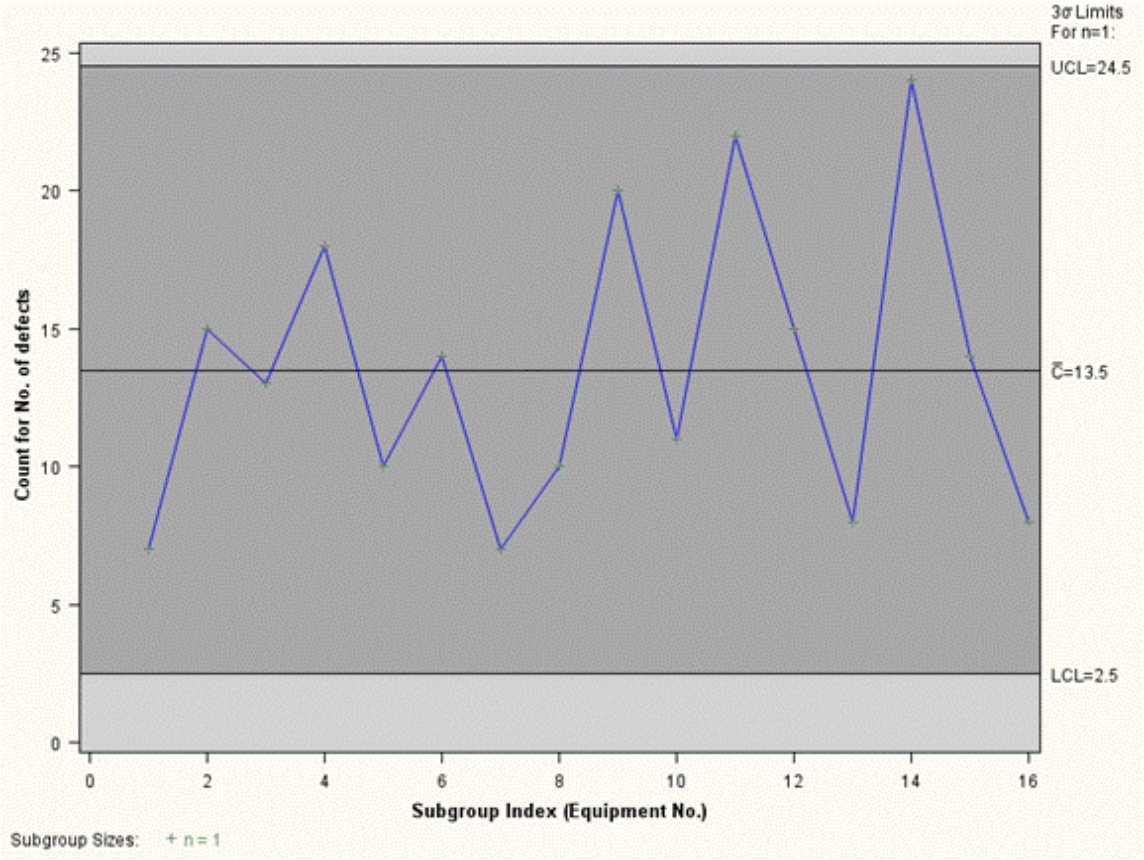


Fig. 28.3 Control chart for number of defects per unit (C-chart)

Since , all the points are within control limits therefore the process is in a state of control.

Lesson 29

FUNDAMENTAL CONCEPTS OF ACCEPTANCE SAMPLING PLAN BY ATTRIBUTES

29.1 Introduction

By SQC we mean the various statistical methods used for maintenance of quality in a continuous flow of manufactured products. In any manufacturing process it is not possible to produce goods of exactly the same quality. Variation is bound to be there or inevitable. There are two types of variation (1) Chance Causes of Variation—Allowable (2) Assignable Causes of Variation — Preventable. The main purpose of SQC is to device statistical methods to separate out allowable / chance Causes of Variation from Preventable / Assignable Causes of Variation. In this problem we try to control the manufacturing process so that the proportion of defective items is not excessively large. This is known as process control and is achieved mainly through control charts technique which has been discussed in the previous lesson.

On the other hand there is another type of problem where we may like to ensure that lots of manufactured goods do not contain an excessively large proportion of defective items. This is known as 'Product Control' and is achieved mainly through the technique of Acceptance Sampling Plans/ Sampling Inspection by Attributes. Product control or Lot control will be far more economical if the process is under statistical control because in that case the rejections of goods and the amount of sampling for arriving at a decision will be minimum. In this lesson we discuss various terms and definitions used in the context of acceptance sampling plan by attributes.

29.2 Basic Concepts

29.2.1 Producer

By Producer we mean a person or firm which manufacture goods or articles to be supplied to another person or firm or another section of the same firm.

29.2.2 Consumer

By Consumer we mean a person or firm which receive articles from the producer.

29.2.3 Acceptable quality level (AQL)

A lot with relatively small fraction defective (i.e. sufficiently of good quality) say p_1 that we do not wish to reject more often than a small proportion of time is sometimes referred to as a good lot. Usually

Pr. [Rejecting a lot of Quality p_1] =0.05

P_a = Pr. [Accepting a lot of Quality p_1] =0.95

$'p_1'$ is known as the Acceptable Quality Level and a lot of this Quality is considered as Satisfactory by the Consumer. If $\propto$ is the Producer's Risk then that level of quality which results in $100(1-\propto)\%$ acceptance of the good lots submitted for inspection.

29.2.4 Lot tolerance percent defective (LTPD)

The Lot Tolerance Fraction Defective P_t is the maximum Fraction Defective which the Consumer is prepared to tolerate in the accepted lot. In other words, this is the Quality Level which the Consumer regards as rejectable

(not acceptable) and is usually abbreviated as Rejected Quality Level (RQL). A lot of quality p_t stands to accept some arbitrary and small Fraction of time, usually 10%.

29.2.5 Process average fraction defective ($\bar{p}$)

The quality turned out by the Manufacturing process over a long period of time is known as Process Average Fraction defective denoted by $\bar{p}$. In Industry, the quality of any process tends to settle down to some level which is expected to be more or less the same every day for a particular machine. If this level could be maintained and if the process is working free from assignable causes of variation, the inspection could often be dispensed with. But in practice, as a result of the failure of machine and men, the quality of the product may suddenly deteriorate. The process average of any manufactured product is obtained by finding the percentage of defectives in the product over a fairly long time.

29.2.6 Consumer's risk

Any sampling scheme would involve certain risk on the part of the consumer in the sense that he/she has to accept certain percentage of undesirable bad lots. The probability of accepting a lot of unsatisfactory quality is known as consumer's risk. It p_t be the lot tolerance percent defective (LTPD) i.e. Maximum fraction defective in the lot which he/she will tolerate , then the probability of accepting a lot with fraction defective p_t is called consumer's risk and is written as P_c. Usually it is denoted by β.

$$P_c = P[\text{Accepting a lot of quality } p_t] = \beta$$

29.2.7 Producer's risk

The Producer has also to face the situation that some good lots will be rejected. He might demand adequate protection against such contingencies happening too frequently just as the Consumer can claim reasonable protection against accepting too many bad lots. The Probability of rejection of a lot with 100 $\bar{p}$ as the process average percent defective is called Producer's Risk and is usually denoted by $\propto$.

P_p = P [of rejecting a lot of Quality $\bar{p}$] = $\propto$

29.3 Sampling Inspection by Attributes

From economic consideration it is practically impossible to inspect the whole lot in product control. That is why we have to depend upon sampling inspection. By sampling inspection for attribute it is meant that the items are judged good or defective by inspection of the sampling and quality of the lot adjudged from sample fraction defective.

29.4 Benefits of Sampling Inspection

Sampling inspection is economical when compared to cent percent inspection and is the only recourse when the inspection procedure is destructive. The advantages of sampling inspection are that it:

- provides desired assurance of quality.
- minimizes damages from handling during inspection.

- exerts pressure on vendor to improve quality through rejection of entire lot (instead of returning only the defectives) thus placing the responsibility for sorting on him.
- needs fewer Inspectors who can be trained more effectively than a large number.
- increases inspector's importance (since his sample results lead to important decision) and creates a sense of responsibility.
- reduces errors due to Monotony since he has to inspect a few pieces from lots of different products.
- is easily adjustable to suit rate of incoming material without compromising on desired quality assurance.

29.5 Validity of Sampling

Inspection of sample tells

- Whether items in a sample are good or bad.
- Whether the process at the time when the samples were selected was satisfactory or not.
- Whether the uninspected Items made at the same time by the same process are acceptable or not.
- Whether the process is stable.
- Whether the items to be produced are going to be acceptable or not.

29.6 Acceptance Sampling by Attributes

In many a manufacturing process the producer in order to ensure that the manufactured goods are according to specifications of the consumer, gets his lot checked at strategic stages. On the other hand, the consumer is anxious to satisfy himself about the quality of goods he accepts. An ideal way of doing this seems to inspect each and every item presented for acceptance i.e., to resort to 100% inspection. Cent percent inspection should be resorted under the following conditions

- The occurrence of a defect may cause loss of life or serious casualty to personnel.
- A defect may cause serious malfunction of the equipment. 100% inspection may also be resorted when (i) the lot size is small, and (ii) the incoming quality is poor or unknown.

If testing is destructive, as for instance in the case of crackers, shells, bulbs etc., it is absolutely non-sense to talk of cent percent inspection. Even in these cases where 100% inspection is possible, it may not be desirable because

a) it is costly and time consuming and

b) due to fatigue, impossibility of proper check and variations in efficiencies of inspection in time, persons, and place, however careful one may be, the inspection lot is likely to contain a small number percentage of defective items.

So from practical point of view, sampling procedures are adopted i.e., a lot is accepted or rejected on the basis of the samples drown at random from the lot.

It has been found that if a scientifically designed sampling inspection plan is used, it provides adequate protection to producer as well as consumer. The main object of inspection is to control the quality of the product by critical examination at strategic points. Sampling inspection besides keeping down the cost of production also ensures that the quality of a lot accepted is according to the specifications of the consumer. The guidelines for a sampling procedure are that

a) It should give definite assurance again passing any unsatisfactory lot, and

b) Inspection expenses should be as low as possible subject to degree of protection afforded by (a).

29.6.1 An acceptance plan prescribes

- procedure of sampling from a specified lot.
- size of sample.
- statistic to be formed from sample observations.
- the decision criteria and
- the procedure for disposal of the rejected lot (to be returned to vendor or screened).

29.6.2 Characteristics of a good acceptance plan

- protect the producer against having any lot rejected when his product is in a state of control and satisfactory as to level and uniformity.
- protect the consumer against acceptance of the bad lot.
- give long run protection to consumer.
- encourage the producer to keep his process under control.
- minimize Cost of sampling, inspection and administration.
- provide information concerning the quality of the product.

29.7 Sampling Inspection by Attributes

From economic considerations, it is practically impossible to inspect the whole lot in lot control or product control. That is why we have to depend upon sampling inspection.

By sampling inspection for Attributes we mean that the items are judged good or defective by inspection and the quality of lot is adjudged from sample fraction defective.

In discussing acceptance sampling the following symbols will be of much use.

N = number of items or pieces in a given lot

n = number of items or pieces in a given sample

M = number of defective items or pieces in a given lot of size N

m = number of defective items or pieces in a given sample of size m.

C = acceptance number or maximum allowable number of defective pieces or articles in a sample of size n.

p = fraction defective. In a given submitted lot this is M/N and in a given sample it is m/n

P_a = probability of Acceptance.

p_1 = true process average fraction defective of a product submitted for inspection.

$\bar{p}$ = average fraction defective in observed samples.

P_a = probability of acceptance

α = Producer's Risk, the Probability of Rejection of product of some stated desirable quality.

$\alpha = 1 - P_a$ at the stated quality.

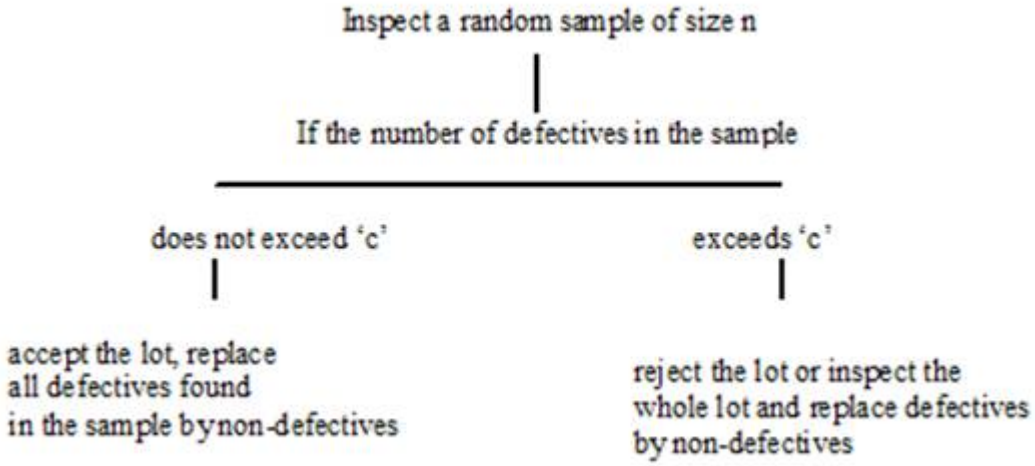
β = Consumer's Risk, the Probability of accepting a product of some stated undesirable quality. It is the value of P_a at the stated quality.

$\beta = P_a$

29.8 Single Sampling Plan

In acceptance sampling inspection, a defective article is one that fails to conform to specifications in one or more quality characteristics. A common procedure in acceptance sampling is to consider each submitted lot of product separately and to base the decision on acceptance or rejection of the lot on the evidence of one or more samples chosen at random from the lot. When the decision is made on the evidence only one sample, the acceptance plan is described as a Single Sampling Plan. Any systematic plan for single sampling requires three numbers to be specified. One is the number of articles N in the lot from which the sample is to be drawn. The second is the number of articles n in the random sample drawn from the lot. The third is the acceptance number C. The acceptance number is the maximum allowable number of defective articles in the sample. More than C defective will cause the rejected of the lot otherwise lot is said to be accepted c-g-, if we have a single sampling plan specified as N=50 , n=5, c=0. This can be interpreted as "Take a random sample of 5 from a lot of 50. If the sample contains more than zero defects, reject the lot; otherwise accept the lot."

A single Sampling plan by attributes can be described as follows:-



Remarks: Obviously in such a plan, the chance of cent-percent inspection increases as the percentage of defectives in the lot increases. Thus the amount of inspection automatically increases as the lot quality deteriorates.

29.9 Double Sampling Plan

Single Sampling Plan calls for decision on acceptance or rejection of a lot on the basis of evidence of one sample from the lot. Double Sampling Plan involves the possibility of putting off the decision on the lot until a second sample has been taken. A lot may be accepted at once if the first sample is good enough or rejected at once if the first sample is bad enough. If the first sample is neither good nor bad enough, the decision is based on the evidence of the first and second samples combined. In general Double Sampling Plan involves less inspection than single sampling for any given quality protection. They also have certain psychological advantages based on the idea of giving a second chance to doubtful lots. The symbols used in connection with the Double Sampling Plan are as follows:-

N = Number of pieces in the lot.

n_1 = Number of pieces in the first sample.

c_1= Acceptance number for the first sample i.e., the maximum number of defectives that permits the acceptance of the lot on the basis of the first sample.

n_2 = Number of pieces in the second sample

$(n_1 + n_2)$ = Number of pieces in the two samples Combined.

c_2 = Acceptance number for the two samples combined that is the maximum number of defective that will permit the acceptance of the lot on the basis of two samples.

The procedure of double sampling plan is illustrated through the following example:

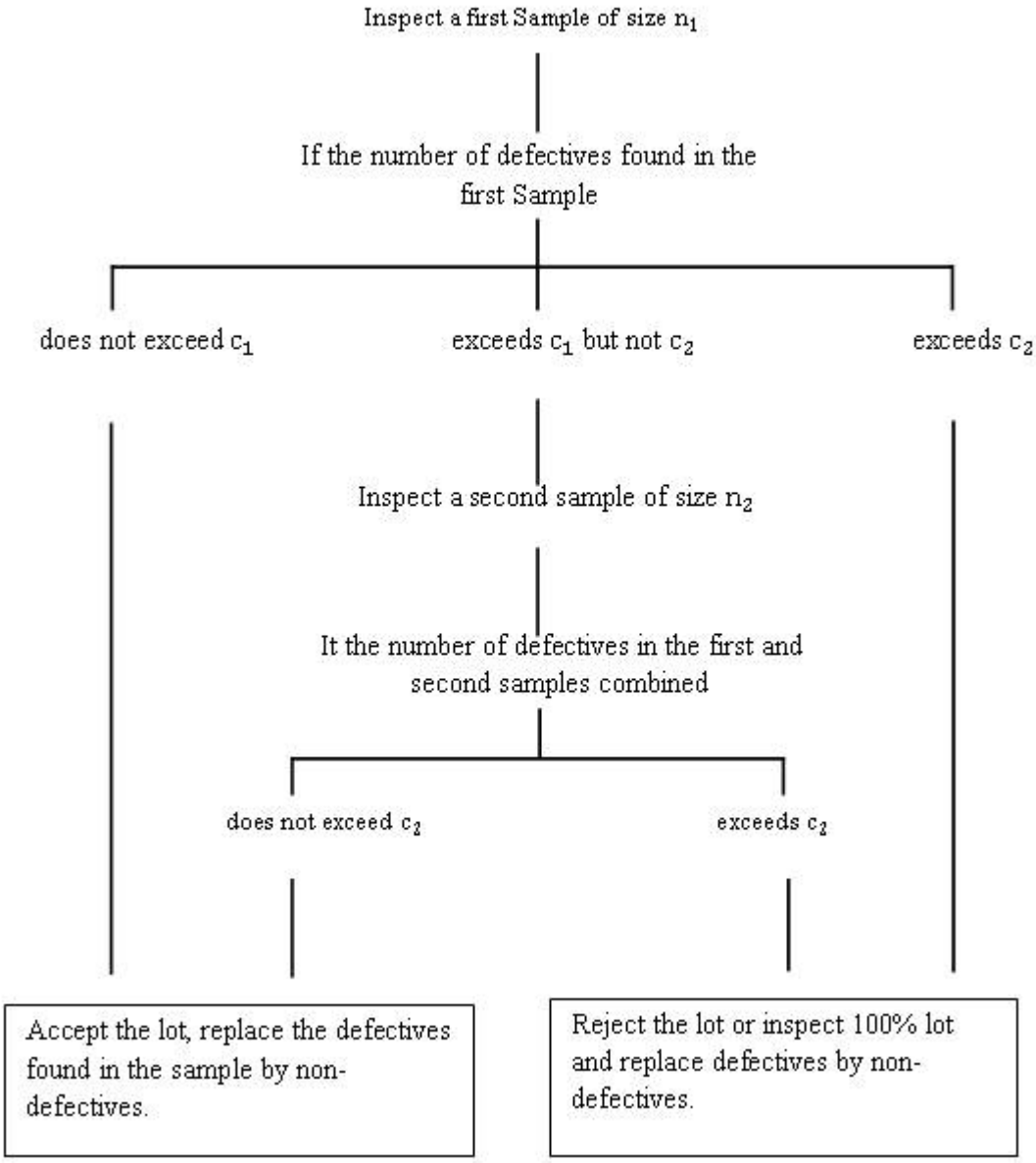
Example $N = 1000$, $n_1 = 36$, $c_1 = 0$, $n_2 = 59$, $c_2 = 3$

The above double sampling plan can be interpreted as follows through the following steps:

- Inspect a first sample of 36 items from a lot of 1000 items.
- Accept the lot on the basis of the first sample if the sample contains 0 defective.
- Reject the lot on the basis of the first sample if the sample contains more than 3 defectives.
- Inspect a second sample of 59 items if the first sample contains 1,2 or 3 defectives.
- Accept the lot on the basis of the combined sample of 95 items if the combined sample contains 3 or less defectives.
- Reject the lot on the basis of the combined sample if the combined sample contains more than 3 defectives.

There are four possibilities for acceptance or rejection of a lot submitted for double sampling plan namely

a) Acceptance after the First Sample.
b) Rejection after the First Sample.
c) Acceptance after the Second Sample.
d) Rejection after the Second Sample.



29.10 Advantages of Double Sampling Plan over Single Sampling Plan

- A psychological advantage of double sampling over single sampling scheme is that to a layman, it seems unfair to reject a lot on the basis of one sample alone and appears more convincing to say that the lot has been rejected after inspecting two samples. Further border line lots in a double sampling plan are given a second chance to be accepted and no lot is rejected because of a single defective article.
- Double sampling requires 25 to 33% less inspection on the average than the single sampling plan. This reduces total inspection compared to a Single Sampling Plan which gives the same quality protection. This reduces the lot of inspection. This requires 25 to 33 % inspection on the average than Single Sampling. This does not necessarily mean, however, that a double sampling plan should be costlier than single sampling plan. The double sampling plans being more complicated and the necessity of inspecting the second sample being unpredictable, the unit cost of inspection for a double sampling procedure may be higher than that for single sampling plan. For economy in overall inspection effort, Double sampling plan is preferred. However, if minimum variation in inspector's workload is desired, single sampling plan is preferred.

Lesson 30

OC AND AOQ CURVES

30.1 Introduction

The two main considerations on the basis of which sampling inspection plans may be compared are the Operating Characteristic (OC), Average Outgoing Quality (AOQ) and Average Sampling Number (ASN). Let us consider two equivalents sampling inspection plans viz., single and double sampling plan for which the OC curves are practically the same. The two plans are equivalent in the sense that they give the same amount of protection against rejection (or 100% inspection) of good lots or acceptance of bad lots. The average amount of inspection required per lot is maximum for single sampling plan and relatively less for double sampling plan. The exact amount of saving depends on the lot-quality and the particular plan under consideration. Generally speaking double sampling plan requires 25 to 33 percent less inspection on an average than single sampling plan. In this lesson, the OC and AOQ curves for single sampling plan have been discussed.

30.2 The Operating Curve (OC) Curve

The OC curve of an Acceptance Sampling Plan shows the ability of the plan to distinguish between good and bad lots. In judging various acceptance sampling plans it is desirable to compare their performance over a range of possible quality levels of submitted product. An excellent picture of this performance is given by the OC Curve. For any given fraction defective p in the submitted lot, the OC Curve shows the probability of acceptance P_a that such a lot will be accepted by the given sampling plan or it shows the long run percentage of submitted lots that would be accepted if a great many lots of any stated quality were submitted for Inspection. The OC is the Mathematical Expression L(p) or Pa, stating the probability of accepting a lot as a function of p, the fraction defective of the lot. The curve obtained by plotting the Operating Characteristic known as Probability of Acceptance Pa against p is called the OC curve. The steeper the OC Curve, the greater is the protection to the Consumer. An ideal plan, of course, would be one which rejects all lots which are of worse quality than some predetermined value of the fraction defective p and accepts all lots which are equal to or better than that quality. Such a plan, however, can never be attained. As ‘p’ the fraction defective increases, the probability of acceptance P_a decreases.

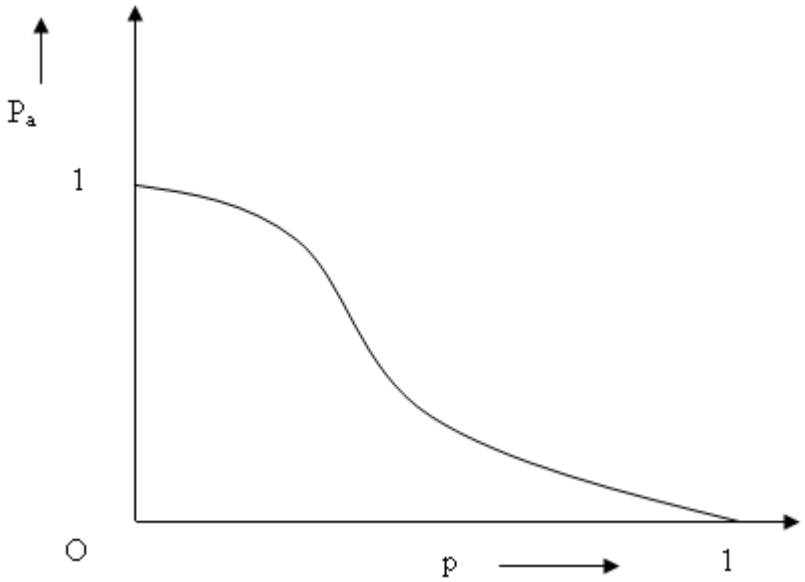


Fig 30.1 OC Curve

An ideal sampling plan would be one that rejected all lots that were worse than specified quality and accepted all lots of specified quality or better. The OC curve of such an ideal plan would be

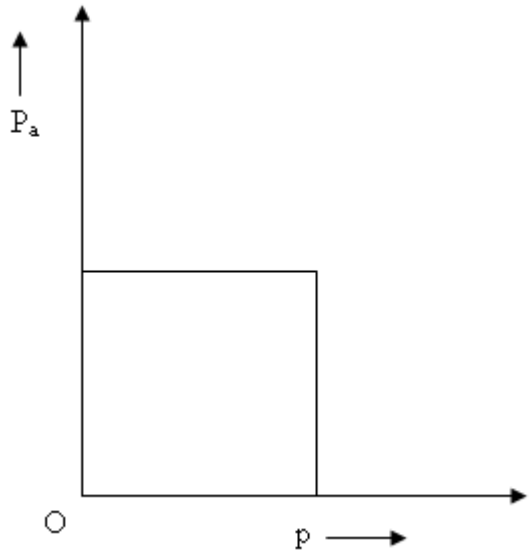


Fig. 30.2 OC Curve for ideal plan

The probability of acceptance (P_a) is calculated by using the following formula:

$$P_a = \sum_{r=0}^{c} n_{C_r} p^r q^{n-r} \quad when\ n\ is\ small\ n\ and\ < 50$$

The OC Curve for the incoming quality 'p' is given by

$$P_a = L(p) = \sum_{r=0}^{c} g(r,p) = \sum_{r=0}^{c} \binom{Np}{r}\binom{N-Np}{n-r}\binom{N}{n} \qquad (30.1)$$

When p<0.10. a good approximation to above equation (30.1) is given by the first (c+1) terms of the binomial expansion

$$\left[\left(1-\frac{n}{N}\right)+\frac{n}{N}\right]^{r} \qquad (30.2)$$

$$L(p)=\sum_{r=0}^{c}\binom{Np}{r}\left(\frac{n}{N}\right)^{r}\left(1-\frac{n}{N}\right)^{Np-r}$$

When p<0.10 and also n/N<0.10, the equation (30.1) is given by Poisson Distribution.

$$L(p)=\sum_{r=0}^{c}\frac{e^{-np}.(np)^{r}}{r!} \text{ (when n > 50)}$$

30.3 Average Outgoing Quality (AOQ) Curve

The expected fraction defective remaining in the lot after the application of the sampling plan is called the Average Outgoing Quality (AOQ). This is a function of 'p', the actual fraction defective in the lot. The maximum value of the average outgoing Quality, the maximum being taken with respect to p, is known as Average Outgoing Quality Limit (AOQL). Suppose a sample of n items is drawn from a lot of size N and P_a is the probability of acceptance of a lot of average quality level p then

$$AOQL=\frac{p(N-n)P_a}{N}$$

If the sampling fraction n/N is negligible, then AOQL = $p.P_a$

The AOQ Curve is obtained by plotting $p.P_a$ known as Average Outgoing Quality (AOQ) against 'p', the fraction defective of the lot.

30.4 Average Sample Number (ASN)

The expected value of the sample size required for coming to a decision i.e., for acceptance or rejection, under the sampling inspection plan of a lot is called the Average Sample Number (ASN). This is a function of p, the actual fraction defective of the lot. The curve obtained by plotting ASN against p is called the ASN curve. Obviously, other factors remaining the same, the lower the ASN Curve the better is the sampling inspection plan.

Statistical Tables

Table1: Random Numbers One-digit Numbers

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
3	3	2	6	1	6	8	0	4	5	6	0	8	0	0	7	3	5	4	5	5	4
2	7	0	7	3	6	0	7	5	1	2	4	6	1	9	9	6	7	3	8	1	8
1	3	5	5	3	8	5	8	5	9	8	8	6	7	0	0	3	7	2	6	7	1
5	7	1	2	1	0	1	4	2	1	8	8	6	5	2	2	4	2	8	6	7	3
0	6	1	8	4	4	3	2	5	3	2	3	0	9	2	7	8	5	0	0	9	4
8	7	3	5	2	0	9	6	4	3	8	4	6	7	4	2	9	6	3	7	2	7
2	1	7	6	3	3	5	0	2	5	8	3	3	0	8	7	8	6	1	0	4	8
1	2	8	6	7	3	5	8	0	7	4	4	1	7	5	5	7	3	9	5	2	5
1	5	5	1	0	0	1	3	4	2	9	9	3	2	0	0	8	2	8	0	0	4
9	0	5	2	8	4	7	7	2	7	0	8	1	3	1	3	7	5	6	9	2	9
0	6	7	6	5	0	0	3	1	0	5	5	2	6	7	2	5	1	9	2	7	5
2	0	1	4	8	5	8	8	4	5	1	0	2	3	8	1	7	0	5	0	0	3
3	2	9	8	9	4	0	7	7	2	9	3	2	8	3	4	5	0	3	4	1	4
8	0	2	2	0	2	5	3	5	3	8	6	4	0	6	6	8	9	2	0	8	7
5	4	4	2	0	6	8	7	9	8	3	5	1	7	3	1	8	8	9	2	2	9
1	7	7	6	3	7	1	3	0	4	0	7	0	0	8	8	3	8	7	9	3	1
7	0	3	3	2	4	0	3	5	4	9	7	2	4	3	3	9	3	4	8	6	6
0	4	4	3	1	8	6	6	7	9	9	4	9	7	7	2	7	8	0	5	7	3
1	2	7	2	0	7	3	4	4	5	9	9	6	1	9	9	1	3	9	5	0	9
5	2	8	5	6	6	6	0	4	4	3	8	5	7	0	0	4	5	0	5	0	9
0	4	3	3	4	6	0	9	5	2	6	8	5	0	4	2	4	2	4	6	7	7
1	3	5	8	1	8	2	4	7	6	1	5	5	8	2	7	7	4	4	7	6	2
9	6	4	6	9	2	4	2	4	5	9	7	3	2	8	5	3	8	2	7	5	7
1	0	4	5	6	5	0	4	2	6	1	1	1	2	1	8	7	7	3	4	9	6
3	4	2	5	2	0	5	7	2	7	4	0	5	1	7	2	7	5	9	5	4	1
6	0	4	7	2	1	2	9	6	8	0	2	5	4	6	7	0	2	6	5	5	4
7	6	7	0	9	0	3	0	8	6	3	8	3	9	4	7	4	8	2	5	8	7
1	6	9	2	5	3	5	6	1	6	0	2	7	6	3	9	9	2	2	5	0	4
4	0	0	1	7	4	9	1	6	2	4	8	8	6	7	4	8	3	3	1	2	7

0	0	5	2	4	3	4	8	8	5	2	7	8	4	6	4	5	9	7	9	4	9

Table2: Random Numbers Two-digit Numbers

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
51	51	00	83	63	22	55	39	65	36	63	70	77	45	85	50
68	97	87	64	81	07	83	73	71	98	16	04	29	18	94	51
30	79	20	69	22	40	98	72	20	56	20	11	72	65	71	08
81	69	40	23	72	51	39	75	17	26	99	76	89	37	20	70
90	60	73	96	53	97	86	37	48	60	82	29	81	30	15	39
46	15	38	26	61	70	04	68	08	02	80	72	83	75	46	30
99	05	48	67	26	43	18	14	23	98	61	67	70	52	85	01
98	35	55	03	36	67	68	49	08	96	21	44	25	27	99	41
11	53	44	10	13	5	57	78	37	06	08	43	63	61	62	42
06	71	95	06	79	88	54	37	21	34	17	68	86	96	83	23
83	45	19	90	70	99	00	14	29	09	34	04	87	83	07	55
49	90	65	97	38	20	46	68	43	28	06	36	49	52	83	51
39	84	51	67	11	52	49	10	43	67	29	70	80	62	80	03
16	17	17	95	70	45	80	44	38	88	39	54	86	97	37	44
13	74	63	52	52	01	41	90	59	59	19	51	85	39	52	85
68	93	60	61	97	22	61	41	47	10	25	62	97	05	31	03
01	07	98	99	46	50	47	91	94	14	63	19	75	89	11	47
74	97	76	38	03	29	63	80	06	54	18	66	09	18	94	06
19	33	53	05	70	53	30	67	72	77	63	48	84	08	31	55
43	70	02	87	40	41	45	59	40	24	13	27	79	26	88	86
95	80	35	14	97	35	33	05	90	35	89	95	01	61	16	96
82	15	94	51	33	41	67	44	43	80	69	98	46	68	05	14
65	31	91	51	80	32	44	61	81	31	96	82	00	57	25	60
85	23	65	09	29	75	63	42	88	07	10	05	24	98	65	63
65	79	20	71	53	20	25	77	94	30	05	39	28	10	99	00
81	06	01	82	77	45	12	78	83	19	76	16	94	11	68	84
00	52	53	43	37	15	26	87	76	59	61	81	43	63	64	61
50	28	11	39	03	34	25	91	43	05	96	47	55	78	99	95

53	32	40	36	40	96	76	84	97	77	72	73	09	62	06	65
69	84	99	63	22	32	98	87	41	60	76	83	44	88	96	07

Table3: (F-Variance ratio) at 5% level of significance

n_1 / n_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69

27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Source: Fisher and Yates Statistical Tables

Table4: (F-Variance ratio) at 1% level of significance

n_1 / n_2	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42

21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Source: Fisher and Yates Statistical Tables

Table5: Values of χ^2 and t at 1%, 5% and 10% level of significance.

d.f.	Significance level of t			Significance level of χ^2		
	.01	.05	.10	.01	.05	.10
1	63.66	12.71	6.31	6.63	3.84	2.71
2	9.92	4.30	2.92	9.21	5.99	4.61
3	5.84	3.18	2.35	11.34	7.81	6.25
4	4.60	2.78	2.13	13.28	9.49	7.78
5	4.03	2.57	2.02	15.09	11.07	9.24
6	3.71	2.45	1.94	16.81	12.59	10.64
7	3.50	2.36	1.90	18.50	14.07	12.02
8	3.36	2.31	1.86	20.09	15.51	13.36
9	3.25	2.26	1.83	21.67	16.92	14.68
10	3.17	2.23	1.81	23.21	18.31	16.00
11	3.11	2.20	1.80	24.72	19.68	17.28
12	3.06	2.18	1.78	26.22	21.03	18.55
13	3.01	2.16	1.77	27.69	22.36	19.81

14	2.98	2.14	1.76	29.14	23.68	21.06
15	2.95	2.13	1.75	30.58	25.00	22.31
16	2.92	2.12	1.75	32.00	26.30	23.54
17	2.90	2.11	1.74	33.41	27.59	24.77
18	2.88	2.10	1.73	34.80	28.87	25.99
19	2.86	2.09	1.73	36.19	30.14	27.20
20	2.84	2.09	1.72	37.57	31.41	28.41
21	2.83	2.08	1.72	38.93	32.67	29.62
22	2.82	2.07	1.72	40.29	33.92	30.81
23	2.81	2.07	1.71	41.64	35.17	32.01
24	2.80	2.06	1.71	42.98	36.42	33.20
25	2.79	2.06	1.71	44.31	37.65	34.38
26	2.78	2.06	1.71	45.64	38.88	35.56
27	2.77	2.05	1.70	46.96	40.11	36.74
28	2.76	2.05	1.70	48.28	41.34	37.92
29	2.76	2.04	1.70	49.59	42.56	39.09
30	2.75	2.04	1.70	50.89	43.77	40.26
70				100.42	90.53	85.53
∞	2.58	1.96	1.64			

Source: Fisher and Yate's Statistical Tables

TABLE 6. FACTORS USEFUL IN THE CONSTRUCTION OF CONTROL CHARTS *

Mean Chart				**Standard deviation chart**			**Range chart**		
Samples Size	**factors for control limits**			**Factor for central line**	**factors for control limits**		**Factor for central line**	**Factors for control limits.**	
n	**A**	**A1**	**A2**	**c2**	**B3**	**B4**	**d2**	**D3**	**D4**
2	2.121	3.760	1.880	0.5642	0	3.267	1.128	0	3.267
3	1.732	2.394	1.023	0.7236	0	2.568	1.693	0	2.575
4	1.500	1.880	0.729	0.7979	0	2.266	2.059	0	2.282
5	1.342	1.596	0.577	0.8407	0	2.089	2.326	0	2.115
6	1.225	1.410	0.483	0.8686	0.030	1.970	2.534	0	2.004
7	1.134	1.277	0.419	0.8882	0.118	1.882	2.704	0.076	1.924
8	1.031	1.175	0.373	0.9027	0.185	1.815	2.847	0.136	1.864
9	1.000	1.094	9.337	0.9139	0.239	1.761	2.970	0.184	1.816
10	0.949	1.028	0.308	0.9227	0.284	1.716	3.078	0.223	1.777

11	0.905	0.973	0.285	0.9330	0.321	1.679	3.173	0.256	1.744
12	0.866	0.925	0.266	0.9359	0.354	1.646	3.258	0.284	1.716
13	0.832	0.884	0.249	0.9410	0.382	1.618	3.336	0.308	1.692
14	0.802	0.848	0.235	0.9453	0.406	1.594	3.407	0.329	1.671
15	0.775	0.816	0.223	0.9490	0.428	1.572	3.472	0.348	1.652
16	0.750	0.788	0.212	0.9523	0.448	1.552	3.532	0.364	1.636
17	0.728	0.762	0.203	0.9551	0.466	1.534	3.588	0.379	1.621
18	0.707	0.738	0.194	0.9576	0.482	1.518	3.640	0.392	1.608
19	0.688	0.717	0.187	0.9599	0.497	1.503	3.689	0.404	1.596
20	0.671	0.697	0.180	0.9619	0.510	1.490	3.735	0.414	1.586
21	0.655	0.679	0.173	0.9638	0.523	1.477	3.778	0.425	1.575
22	0.640	0.662	0.167	0.9655	0.534	1.466	3.819	0.434	1.566
23	0.626	0.647	0.162	0.9670	0.545	1.455	3.858	0.443	1.557
24	0.612	0.632	0.157	0.9684	0.555	1.445	3.895	0.452	1.548
25	0.600	0.619	0.153	0.9696	0.565	1.435	3.931	0.459	1.541

* Reproduced from Table B2, ASTM SPT – 15C, Manual on Quality control of Materials, with the kind permission of the American Society for Testing and Materials.

REFERENCES

Agarwal, B.L. 1991. Basic Statistics. Wiley Eastern Ltd., New Delhi.

Amble,V.N. 1975. Statistical Methods in Animal Sciences. Indian Society of Agril. Statistics, New Delhi.

Goon, A.M., Gupta, M.K. and Gupta, B. D. 1979. Fundamental of Statistics. Vol. I and II. The World Press Pvt. Ltd., Kolkata.

Goulden, C.H. 1959. Methods of Statistical analysis. John Wiley and Sons, New York.

Gupta, S.C. 1987. Fundamental of Statistics. Himalaya Publishing House, New Delhi.

Gupta, S.C. and Kapoor, V.K. 1990. Fundamentals of Applied Statistics. Sultan Chand & Sons, New Delhi.

Gupta, S.P. 2010. Statistical Methods. Sultan Chand and Sons, New Delhi.

Handbook on Statistical Quality Control. 1986 . Indian Standards Institute, New Delhi.

Moroney, M.J. 1975. Facts from Figures. Penguin Books, England.

Snedecor, G.W. and Cochran, W.G. 1967. Statistical Methods. Oxford and IBH Publishing Co., New Delhi.