Module - I

1

CO-RELATION ANALYSIS AND REGRESSION ANALYSIS

Unit Structure :

1.1 INTRODUCTION

Statistics deals with quantitative phenomenon only. However, the quantitative character may arise in any of the following two ways:

1. In the first place, we measure the actual magnitude or size of the phenomenon. For example, we may measure the height of students of a class, their weight etc. Similarly, we may study the wage structure of the workers of a particular factory, the amount of a rainfall in a year. The characteristic of this type of phenomena is that they can be quantitatively measured. Data regarding such phenomena are known as statistics of variables. The various statistical techniques like measures of central tendency, dispersion, and correlation deal with such variables.
2. In the second place, there are certain phenomena like blindness, deafness etc. which are not capable of direct quantitative measurement. In such cases the quantitative

character arises only indirectly in the process of counting. For example, we can determine out of 1, 000 persons, how many are blind and how many are not blind but we cannot precisely measure blindness. Such phenomena, where direct quantitative measurement is not possible, i.e. where we can study only the presence or absence of a particular characteristic, are called statistics of attributes.

1.2 DIFFERENCE BETWEEN CORRELATION AND ASSOCIATION

The tool of correlation is used to measure the degree of relationship between two such phenomena as are capable of direct quantitative measurement. On the other hand, the method of association of attributes is employed to measure the degree of relationship between two phenomena whose size we cannot measure and where we cannot only determine the presence or absence of a particular attribute.

While dealing with statistics of attributes we have to classify the data. The classification is done on the basis of presence or absence of particular attribute or characteristic. When we are studying only one attribute, two classes are formed – one possessing that attribute and another not possessing it. For example, when are studying the attribute employment, two classes shall be formed; those who are employed and those who are not employed. When two attributes are studied, four classes shall be formed. If, besides employment, we study the sex-wise distribution, four classes shall be formed; number of males employed, number of females employed, number of males unemployed and number of females unemployed.

It should be noted that in some cases while classifying the attributes no clear-cut definition of an attribute and line of demarcation between classes can be drawn. For example, when the attribute ‘employment’ is being studied the data are classified into ‘Employed’ and ‘Unemployed’. But there can be further category of those people who are partially employed (i.e. part-time)/ Also there may be some persons who are employed before the survey but on the date of survey they are unemployed. So we cannot treat them as employed and also as unemployed because there is some difference between those persons who have not got any job, and those who have got some job but were retrenched after sometime. Hence, it is absolutely essential to lay down clear-cut definition of the various attributes under study. This is often a difficult task. Hence, this limitation must be kept in mind while studying association between attributes.

1.3 ASSOCIATION AND DISASSOCIATION

The word association as used in Statistics has a technical meaning, different from the one in ordinary speech. In common language one speaks of A and B as being 'associated' if they appear together in a number of cases. But in Statistics, A and B are associated only if they are independent. On the other hand, if this number (or proportion) is less than expected for independence, they are disassociated. Thus, in case of second order frequencies, A and B are:

i) Associated if (AB) (αβ) > (Aβ) (Bα)
ii) Associated if (AB) (αβ) < (Aβ) (Bα)

Hence, it should carefully be noted that association cannot be inferred from the mere fact that some A's and B's, however great the proportion.

1.4 METHODS OF STUDYING ASSOCIATION

In order to ascertain whether two attributes are associated or not the following methods may be used:

1. Comparison of Observed and Expected Frequencies Methods
2. Proportion Method
3. Yule's Coefficient of Association
4. Coefficient of Colligation
5. Coefficient of Contingency.

1. Comparison of Observed and Expected Frequencies Methods

When this method is applied, the actual observation is compared with the expectation. If the actual observation is equal to the expectation, the attributes are said to be independent; if actual observation is more than the expectation, the attributes are said to be positively associated and if the actual observation is less than the expectation, the attributes are said to be negatively associated. Symbolically, attributes A and B are:

i. Independent if (AB) = $\frac{(A)X\,(B)}{N}$ (expectation); (actual observation)

ii. Positively associated if (AB) > $\frac{(A)X\,(B)}{N}$ (expectation); and (actual observation)

iii. Negatively associated if (AB) < $\frac{(A)X\,(B)}{N}$ (expectation); and (actual observation)

The same is true for attributes α and B; α and β and A and β. Thus,

i. Independent, if (αβ) = (α) x (β)/N;
ii. Positively associated, if (αβ) > (α) x (β)/N;
iii. Negatively associated, if (αβ) < (α) x (β)/N;

i) EXAMPLE 1: As from the following data find out whether attributes (i) AB, (ii) (Aβ) (iii) (Bα) and (iv) (αB) are independent, associated or disassociated N = 100, (A) = 40, (B) = 80 and (AB) = 30.

Solution: (i) Apply the criterion of independence, i.e., attributes (AB) shall be called independent if (AB) = $\frac{(A)X\ (B)}{N}$; positively associated if (AB) > $\frac{(A)X\ (B)}{N}$ and negatively associated or disassociated if (AB) < $\frac{(A)X\ (B)}{N}$.

Expectation of (AB) = $\frac{(A)X\ (B)}{N}$, here (A) = 40, (B) = 80, N = 100.

Expectation of (AB) = $\frac{(40)X\ (80)}{100}$ = 32.

The actual observation (i.e., the given value of (AB) i.e., 30) is less than the expectation and hence the attributes are disassociated or negatively associated.

(ii) For finding out the nature of association between the attributes (Aβ), (αβ) and αβ we shall have to determine the unknown values. This can be done by preparing a nine-square table:

	A	**α**	**Total**
B	30	50	80
β	10	10	20
Total	40	60	100

From the table (Aβ) = 10, (Bα) = 50, (αβ) = 10, α = 60 and β = 20,

Attributes A and β shall be independent if (Aβ) = $\frac{(A)X\ (B)}{N}$

Expectation (Aβ) = (A) x (β)/N, where (A) = 40, β = 20 and N = 100.

∴ Expectation (Aβ) = $\frac{(40)X\ (80)}{100}$ = 8.

Thus the actual observation (i.e., (Aβ) = 10) is more than expectation and hence the attributes A and β are positively associated.

(iii) Attributes α and B shall be called independent if (αβ) = (α) x (β)/N where (α) = 60, B = 80 and N= 100.

∴ Expectation (Aβ) = $\frac{(60)X(80)}{100}$ = 48.

Thus actual observation (i.e., αβ = 50) is more than the expectation and hence the attributes are positively associated.

(iv) Attributes α and β shall be called independent if (αβ) (α) x (β)/N where α = 60; β = 20 and N = 100.

Expectation of (αβ) $\frac{(60)X(20)}{100}$ = 12.

Thus actual observation [(αβ) = 10] is less than the expectation (12). Hence, the attributes are disassociated.

Limitations

With the help of this method we can only determine the nature of association (i.e. whether there is positive or negative association or no association) and not the degree of association (i.e., where association is high or low). Yule's coefficient is superior because it provides information not only on the nature but also on the degree of association.

2. Proportion Method

If there is no relationship of any kind between two attributes A and B we expect to find the same proportion of A's amongst the B's as amongst the β's. Thus, if a coin is tossed we expect the same proportion of heads irrespective of whether the coin is tossed by the right hand or the left hand.

Symbolically, two attributes may be termed:

(i) Independent if $\frac{(AB)}{B} = \frac{(A\beta)}{\beta}$

(ii) Positively associated if $\frac{(AB)}{B} > \frac{(A\beta)}{\beta}$

(iii) Negatively associated if $\frac{(AB)}{B} < \frac{(A\beta)}{\beta}$

If the relation (i) holds good the corresponding relations

$\frac{\alpha B}{B} = \frac{\alpha\beta}{\beta}$; $\frac{(AB)}{A} = \frac{(\alpha B)}{\alpha}$; $\frac{(AB)}{A} = \frac{(\alpha B)}{\alpha}$

Must also hold true.

EXAMPLE 2

In a population of 500 students the number of married is 200. Out of 150 students who failed 60 belonged to the married group. It is required to find out whether the attributes marriage and failure are independent, positively associated or negatively associated.

Solution: Let A denote married students.
∴ α represents unmarried students.

Let B denote number of failures.
∴ β represents non-failures.

We are given the total number of students i.e. N = 500; (A) = 200, (B) = 150; and (AB) i.e., the number of married students who failed = 60.
Applying the proportion method

Attributes A and B shall be called independent if $\frac{(AB)}{A} = \frac{(\alpha B)}{\alpha}$

In other words, if the proportion of married students who failed is the same as the proportion of unmarried students who failed we say that the attributes, marriage and failure, are independent.

Proportion of unmarried students who failed : i.e. $\frac{(AB)}{A} = \frac{60}{200}$ = 30% or .3.

Proportion of unmarried students who failed i.e., $\frac{(\alpha B)}{\alpha} = \frac{90}{300}$ = 30% or .3

((αβ) = (B) – (AB) i.e. 150 – 60 = 90))
(α) = N – (A) i.e., 500 – 200 = 300)

Since the two proportions are the same we conclude that the attributes, marriage and failure are independent.

Limitations

Just like the previous method, in this method also we can only determine the nature of association not the degree of association.

3. Yule's Coefficient of Association

The most popular method of studying association is the Yule's Coefficient because here not only we can determine the nature of association, i.e. whether the attributes are positively associated, negatively associated or independent, but also the degree or extent to which the two attributes are associated. The

Yule’s Coefficient is denoted by the symbol Q = $\frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$.

The value of this coefficient lies between ± 1. When the value of Q + 1there is perfect positive association between the attributes. When Q -1 there is perfect negative association (or perfect disassociation) between the attributes and when the value of Q is zero the two attributes are independent.

The coefficient of association can be used to compare the intensity of association between two attributes with the intensity of association between two other attributes.

EXAMPLE 3

Investigate the association between eye colour of husbands and eye colour of wives from the data given below:

Husbands with light eyes and wives with light eyes = 309
Husbands with light eyes and wives with non-light eyes = 214
Husbands with non-light eyes and wives with light eyes = 132
Husbands with non-light eyes and wives with non-light eyes = 119

Solution:

Since we have to find out the association between eye colour of husband and that of wife, one attribute we would take as A and another as B.

Let A denote husbands with light eyes.

∴ α denote husbands with non-light eyes.

Let B denote wives with light eyes.

∴ β denote wives with non-light eyes.

The given data in terms of these symbols is

(AB) = 309; (Aβ) = 214; (αB) = 132; (αβ) = 119.

Applying Yule’s method: Q = $\frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

Substituting the above values in this formula

$$Q = \frac{(309\ X\ 119) - (214\ X\ 132)}{(309\ X\ 119) + (214\ X\ 132)} = \frac{8523}{65019} = 0.131$$

Thus, there is a very little association between the eye colour of husband and wife.

4. Coefficient of Colligation

Yule has computed another coefficient called the coefficient of ‘colligation’. It is denoted by the symbol γ and is obtained by applying the following formula:

$$\gamma = \frac{1 - \sqrt{\sqrt{\frac{(A\beta)x(\alpha B)}{(A\beta)x(\alpha B)}}}}{1 + \sqrt{\frac{(A\beta)x(\alpha B)}{(A\beta)x(\alpha B)}}}$$

From this coefficient we can obtain Yule’s Coefficient of Association i.e Q as follows:

$$Q = \frac{2\gamma}{1 + \gamma 2}$$

It should be noted that though γ and Q serve the same purpose. These coefficients are not directly comparable with each other. Further, in practice Q is more popularly used than γ as a measure of association.

1.5 INTRODUCTION - CORRELATION

Statistics that describe or make inferences about a single distribution are referred to as univariate Statistics. While univariate statistics form the basis for many other types of statistics, none of the issues concerning the relationships among variables can be answered by examining only a single variable.

1.6 CORRELATION ANALYSIS

Often an analysis of data concerning two or more quantitative variables is needed to look for any statistical relationship or association between them that can describe specific numerical features of the association. The knowledge of such a relationship is important to make inferences from the relationship between variables in a given situation. Few instances where the knowledge of an association or a relationship between two variables would be helpful to make decision are as follows:

- Family income and expenditure on luxury items.
- Yield of a crop and quantity of fertilizer used.
- Sales revenue and expenses incurred on advertising.
- Frequency of smoking and lung damage.
- Weight and height of individuals.

A statistical technique that is used to analyse the strength (magnitude) and direction of the relationship between two quantitative variables is called correlation analysis. A few definitions of correlation analysis are as follows:

- An analysis of the relationship of two or more variables is usually called correlation.

  - A. M. Tuttle
- 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

The coefficient of correlation is a number that indicates the strength and direction of statistical relationship between two variables.

- The strength of the relationship is determined by the closeness of the points to a straight line when a pair of values of two variables are plotted on a graph. A straight line is used as the frame of reference for evaluating the relationship.
- The direction is determined by whether one variable generally increases or decreases when the other variable increases.

The following questions determine the importance of examining the statistical relationship between two or more variables and accordingly requires the statistical methods to answer these questions:

i. Is there an association between two or more variables? If yes, what is the form and degree of that relationship?
ii. Is the relationship strong or significant enough to be useful to arrive at a desirable conclusion?
iii. Can the relationship be used to predict the most likely value of a dependent variable for the given value of independent variable or variables?

The first two questions will be answered in this chapter. For correlation analysis, the data on values of two variables must come from sampling in pairs, one for each of the two variables.

1.7 SIGNIFICANCE OF MEASURING CORRELATION

The objective of any scientific research is to establish relationships between two or more sets of observations or variables to arrive at some valid conclusion. Few advantages of measuring an association (or correlation) between two or more variables are as under:

1. Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important

variables on which others depend, may reveal to the economist the connection by which disturbances spread and suggest to him the paths through which stabilising forces may become effective. - W. A. Neiswagner

2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality. - Tippett
3. In economic theory, an association (or correlation) between two or more variables, such as price, supply and quantity demanded; customers retention is related to convenience, amenities and service standards; yield of a crop is related to quantity of fertilizer applied, type of soil, quality of seeds, rainfall and so on is established.
4. In healthcare, an association (or correlation) between two or more variables such as validity and reliability of clinical measures; effect on health due to certain biological or environmental factors, blood pressure and age of a person; inter-observer reliability for two doctors who are assessing a patient's disease, and so on is established.

1.8 CORRELATION AND CAUSATION

Correlation is one of the three criteria for establishing a causal relationship between two or more variables. While correlation coefficient only measures the strength of a linear relationship but it does not necessarily imply a causal relationship. The following factors should be examined to interpret the nature and extent of relationship between two or more variables:

i. Chance Coincidence: The inferences drawn from the value of correlation coefficient may not be of any statistical significance because variables might be entirely different and unrelated. Any association between them may be only by a chance. For example, (i) a positive correlation between growth in population and wheat production in the country has no statistical significance, and (ii) the correlation in sales revenue and expenditure on advertisements over a period of time should be statistically significant and not just due to biased sampling or sampling error.

ii. Influence of Third Variable: Clinically, it has been proved that smoking causes lung damage. However, there are often multiple reasons such as stress, quality of food and air pollution, of health problems. Similarly, the yield of rice and tea is positively correlated because both the crops are influenced by the amount of rainfall. But the yield of any one is not influenced by other.

iii. Mutual Influence: Although two variables might be highly correlated, still it is difficult to say as to which variable is influencing the other. For example, variables like price supply, and demand of a commodity are mutually correlated. As price of a commodity increases, its demand decreases, so price influences the demand level. But when demand of a commodity increases, its price also increases so demand influences the price.

TYPES OF CORRELATIONS

There are three broad types of correlations:

i. Positive and negative
ii. Linear and non-linear
iii. Simple, partial, and multiple

1. POSITIVE AND NEGATIVE CORRELATION

The positive (or direct) correlation refers to an association between two variables where their values change (i.e., increasing and decreasing) in the same direction. The negative (or inverse) correlation refers to an association between two variables where their values change (i.e., increasing or decreasing) in the opposite direction.

ILLUSTRATION :

Positive Correlation

Increasing	→	x : 5	8	10	15	17
Increasing	→	y : 10	12	16	18	20
Decreasing	→	x : 17	15	10	8	5
Decreasing	→	y : 20	18	16	12	10

Negative Correlation

Increasing	→	x : 5	8	10	15	17
Decreasing	→	y : 20	18	16	12	10
Decreasing	→	x : 17	15	10	8	5
Increasing	→	y : 10	12	16	18	20

Remarks: the change (increasing or decreasing) in values of both the variables may not be proportional or fixed.

2. LINEAR AND NON-LINEAR CORRELATION

A linear correlation refers to an association between two variables where variation in their values is either proportional or fixed. The following pattern of variation in the values of two variables x and y reveals linear correlation.

x	: 10	20	30	40	50
y	: 40	60	80	100	120

When these pairs of values of x and y are plotted on a graph paper, the line joining these points would be a straight line.

A non-linear (or curvy linear) correlation refers to an association between two variables where variation in their values is neither proportional nor fixed. The following pattern of variation in the values of two variables x and y reveals non-linear correlation.

x	:	8	9	9	10	10	28	29	30
y	:	80	130	170	150	230	560	460	600

When these pairs of values of x and y are plotted on a graph paper, the line joining these points would be a straight line, rather it would be curvy linear.

3. SIMPLE, PARTIAL AND MULTIPLE CORRELATIONS

The distinction between simple, partial and multiple correlations is based upon the number of variables involved in the correlation analysis.

If only two variables are chosen to study correlation between them, then such a correlation is referred to as **simple correlation.** A study on the yield of a crop with respect to only amount of fertilizer used, or sales revenue with respect to amount of money spent on advertisement, are a few examples of simple correlation.

In **partial correlation,** two variables are chosen to study the correlation between them but the effect of other influencing variables is kept constant. For example (i) yield of a crop is influenced by the amount of fertilizer applied, whereas effect of other influencing variables such as rainfall, quality of seed, type of soil and pesticides is kept constant, and (ii) sales revenue from a product is influenced by the level of advertising expenditure, whereas effect of other influencing variables such as quality of the product, price, competitors, distribution and so on is kept constant.

In **multiple correlation**, more than two variables are chosen to study the correlation among them. For example, (i) employer-employee relationship in any organisation may be examined with reference to, training and development facilities; medical, housing and education to children facilities; salary structure; grievances handling system; and so on. and (ii) sales revenue from a product may be examined in relation with the level of advertising expenditure, quality of the product, price, competitors, distribution and so on.

1.9 METHODS OF CORRELATION ANALYSIS

The correlation between two ratio-scaled (numeric) variables is represented by the letter, r, which takes on values between -1 and +1 only and is referred to as **'Pearson Product Moment Correction or correlation coefficient."** The correlation coefficient is a relative (scale free) number and its interpretation is independent of the units of measurement of two variables x and y.

In this chapter, the following methods of calculating a correlation coefficient between two variables x and y are discussed:

1. Scatter Diagram
2. Karl Pearson's Coefficient of Correlation Method
3. Spearman's Rank Correlation Method
4. Method of Least Squares

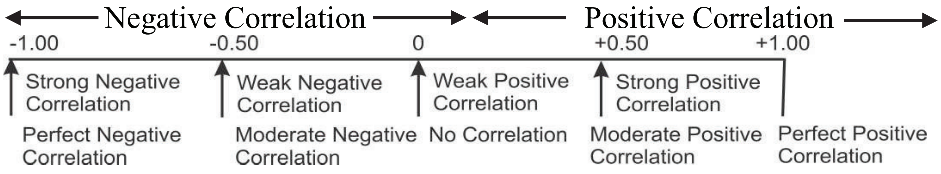

Negative Correlation
Positive Correlation
-1.00
-0.50
0
+0.50
+1.00
Strong Negative Correlation
Weak Negative Correlation
Weak Positive Correlation
Strong Positive Correlation
Perfect Negative Correlation
Moderate Negative Correlation
No Correlation
Moderate Positive Correlation
Perfect Positive Correlation

Figure 1: Interpretation of Correlation Coefficient

1.10 SCATTER DIAGRAM METHOD

The Scatter Diagram method is an at-a-glance method to understand an apparent relationship (if any) between two variables. A scatter diagram (or a graph) can be traced on a graph paper by plotting pairs of values of variables, x and y, taking values of variable, x on the x-axis and values of variable y on the y-axis. The horizontal and vertical axes are scaled in units corresponding to the variables x and y respectively. A straight line drawn through these pair of values describes different types of relationships between the two variables.

Figure 2 shows examples of different types of relationships based on pairs of values of x and y in a sample data. The patterns shown in figure 2 (a) and (b) represent linear relationships since the patterns are described by straight lines. The pattern in figure 2(a) shows a positive relationship since the value of y tends to increase as the value of x increases, whereas pattern in figure 2(b) shows a negative relationship since the value of y tends to decrease as the value of x increases.

The pattern shown in figure 2(c) illustrates very low or no relationship between the values of x and y, whereas figure 2(d) represents a curvilinear relationship since it is described by a curve rather than a straight line. The wider scattering indicates that there is a lower degree of association between the two variables x and y tan there is in figure 2(a).

Interpretation of Correlation Coefficients

While interpreting correlation coefficient r, the following points should be taken into account:

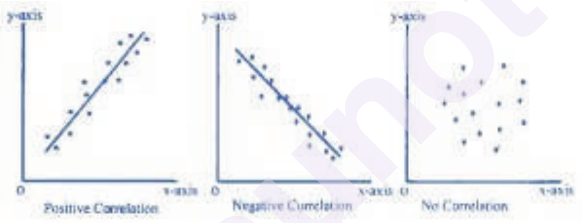
- A low positive or negative value of correlation coefficient, r, indicates that the relationship is poorly described by a straight line. A non-linear relationship may also exist.
- A correlation is an observed association and does not indicate any cause-and-effect relationship.

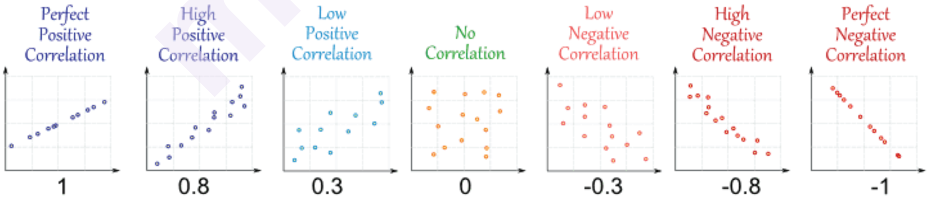
Types of Correlation Coefficients

Table 1 shows several types of correlation coefficients used in statistics along with the conditions of their use. All of them are appropriate for quantifying linear relationship between two variables x and y.

Coefficient	Conditions Applied for Use
φ	Both x and y variables are measured on a nominal scale
ρ	Both x and y variables are measured on, or changed to, ordinal scales (rank data)
r	Both x and y variables are measured on an interval or ration scale scales (numeric data)

The correlation coefficient, denoted by η (eta), is used for quantifying non-linear relationships.





Features of the Correlation Coefficient

The following are the common features among all correlation coefficient:

i. The value of correlation coefficient, r, depends on the slope of the line passing through the data points and the scattering of the pair of values of variables x and y about this line.

ii. The sign of the correlation coefficient indicates the direction of the relationship. The positive correlation denoted by + (positive sign) indicates that the direction of increase (or decrease) in the value of two variables is same. While negative correlation is denoted by - (minus sign) indicates that direction of increase (or decrease) in the value of two variables is opposite.

iii. The values of the correlation coefficient range from +1 to - 1 regardless of the units of measurements of x and y. That is, correlation coefficient is a pure number independent of the unit of measurement.

iv. The value of correlation coefficient r = +1 or - 1 indicates perfect linear association (relationship) between two variables, x and y. A perfect correlation implies that every observed pair of values of x and y falls on the straight line.

v. The value of correlation coefficient indicates the strength of association (relationship) between two variables, i,e,, a closeness of the observed pair of values of x and y to the straight line. The sign of the correlation coefficient indicates the strength of the linear relationship.

vi. The value of correlation coefficient remains unchanged when a constant value is subtracted from every pair of values of variables x and y (also referred to as change of origin), also when a pair of values of variables x and y are divided or multiplied by a constant (also referred to as change of scale).

vii. The value of correlation coefficient, r = 0, indicates that the straight line through the data points is horizontal, and therefore no association (relationship) between two variables x and y.

viii. The square, r^2, of correlation coefficient, r, value is referred to as coefficient of determination.

EXAMPLE 1

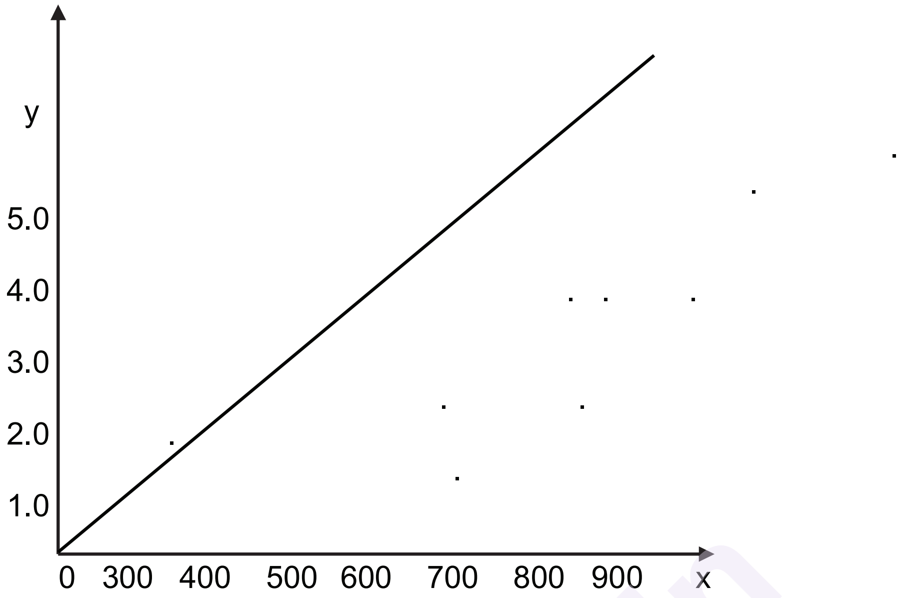
Given the following data:

Student :	1	2	3	4	5	6	7	8	9	10
Management Aptitude Score:	400	675	475	350	425	600	550	325	675	450
Grade Point Average:	1.8	3.8	2.8	1.7	2.8	3.1	2.6	1.9	3.2	2.3

a. Draw this data on a graph paper.

b. Is there any correlation between per capita national income and per capita consumer expenditure? If yes, what is your opinion?

Solution: By taking an appropriate scale on the x and y axes, the pairs of observation are plotted on a graph paper as shown in figure 1. The scatter diagram in figure 1 with straight line represents the relationship between x and y 'fitted' through it.



Interpretation: Since pairs of values of two variables are very close to a straight line passing through them, therefore it appears that there is a high degree of association between two variable values. The pattern of dotted points also indicates a high degree of linear positive correlation.

1.11 KARL PEARSON'S CORRELATION COEFFICIENT

Karl Pearson's Correlation Coefficient quantitatively measures the degree of association (relationship) between two variables x and y. For a set of n pairs of values of x and y, Pearson's Correlation Coefficient, r, is given by

$$r = \frac{Covariance\ (x, y)}{\sqrt{var\ x}\sqrt{var y}} = \frac{Cov\ (x, y)}{\sigma x \sigma y}$$

where Cov (x, y) = 1/n Σ (x - $\bar{x}$) (y - $\bar{y}$)

$\sigma_x = \frac{\sqrt{\Sigma(x-x)2}}{n}$ Standard deviation of sample data on variable x

$\sigma_y = \frac{\sqrt{\Sigma(y-y)2}}{n}$ Standard deviation of sample data on variable y

Substituting values of Cov (x, y), σ_x and σ_y, we have r = 1/n Σ (x - $\bar{x}$) (y - $\bar{y}$)/ $\frac{\sqrt{\Sigma(x-x)2}}{n}$ $\frac{\sqrt{\Sigma(y-y)2}}{n}$ = n Σxy - (Σx)(Σy)/$\sqrt{n\Sigma x2} - (\Sigma x)2$ $\sqrt{n\Sigma y2} - (\Sigma y)2$ ----------------------------(1)

Step Deviation Method for Ungrouped Data

If actual mean values of variables x and y are in fraction, then calculation of Pearson's correlation coefficient can be simplified by taking deviations, $d_x = x - A$ and $d_y = y - B$, of x and y values from their assumed means A and B, respectively. The formula (1) becomes

r = n Σ $d_x d_y$ - (Σ d_x)(Σ d_y)/$\sqrt{n\Sigma\, dx2} - (\Sigma dx)2$ $\sqrt{n\Sigma dy2} - (\Sigma dy)2$ ----------------------- (2)

Step Deviation Method for Grouped Data

If values of variables x and y values are classified into a frequency distribution, then formula (2) is modified as

r = n Σ $fd_x d_y$ - (Σ f d_x)(Σ fd_y)/$\sqrt{n\Sigma\, fdx2} - (\Sigma fdx)2$ $\sqrt{n\Sigma fdy2} - (\Sigma fdy)2$ ----------------------- (2)

Assumptions for Using Pearson's Correlation Coefficient

1. Pearson's correlation coefficient is used only when both variables x and y are measured on an interval or a ratio scale.
2. Pearson's correlation coefficient is used only when two variables x and y are linearly related.

Merits and Demerits of Pearson’s method of studying correlation:

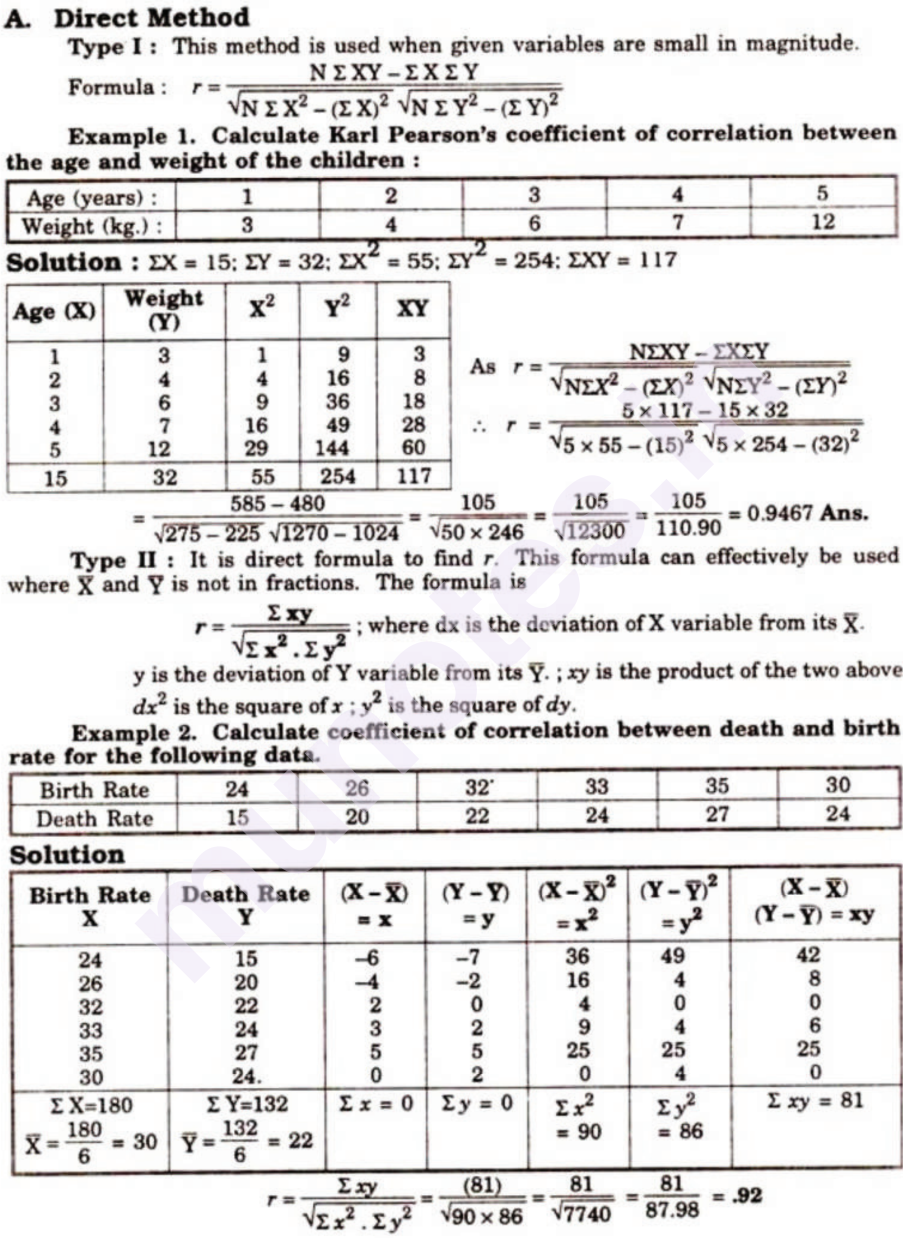
Merits:

1. This method indicates the presence or absence of correlation between two variables and gives the exact degree of their correlation.
2. In this method, we can also ascertain the direction of the correlation; positive, or negative.
3. This method has many algebraic properties for which the calculation of co-efficient of correlation, and other related factors, are made easy.

Demerits:

1. It is more difficult to calculate than other methods of calculations.
   ADVERTISEMENTS:
2. It is much affected by the values of the extreme items.

3. It is based on a many assumptions, such as: linear relationship, cause and effect relationship etc. which may not always hold good.
4. It is very much likely to be misinterpreted in case of homogeneous data.



A. Direct Method

Type I : This method is used when given variables are small in magnitude.

Formula : $r = \dfrac{N\,\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{N\,\Sigma X^2 - (\Sigma X)^2}\,\sqrt{N\,\Sigma Y^2 - (\Sigma Y)^2}}$

Example 1. Calculate Karl Pearson's coefficient of correlation between the age and weight of the children :

Age (years) :	1	2	3	4	5
Weight (kg.) :	3	4	6	7	12

Solution : $\Sigma X = 15$; $\Sigma Y = 32$; $\Sigma X^2 = 55$; $\Sigma Y^2 = 254$; $\Sigma XY = 117$

Age (X)	Weight (Y)	X^2	Y^2	XY
1 2 3 4 5	3 4 6 7 12	1 4 9 16 29	9 16 36 49 144	3 8 18 28 60
15	32	55	254	117

As $r = \dfrac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\,\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$

$\therefore\ r = \dfrac{5 \times 117 - 15 \times 32}{\sqrt{5 \times 55 - (15)^2}\,\sqrt{5 \times 254 - (32)^2}}$

$= \dfrac{585 - 480}{\sqrt{275 - 225}\,\sqrt{1270 - 1024}} = \dfrac{105}{\sqrt{50 \times 246}} = \dfrac{105}{\sqrt{12300}} = \dfrac{105}{110.90} = 0.9467$ **Ans.**

Type II : It is direct formula to find r. This formula can effectively be used where $\overline{X}$ and $\overline{Y}$ is not in fractions. The formula is

$r = \dfrac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$; where dx is the deviation of X variable from its $\overline{X}$.

y is the deviation of Y variable from its $\overline{Y}$. ; xy is the product of the two above dx^2 is the square of x ; y^2 is the square of dy.

Example 2. Calculate coefficient of correlation between death and birth rate for the following data.

Birth Rate	24	26	32˙	33	35	30
Death Rate	15	20	22	24	27	24

Solution

Birth Rate X	Death Rate Y	$(X - \overline{X}) = x$	$(Y - \overline{Y}) = y$	$(X - \overline{X})^2 = x^2$	$(Y - \overline{Y})^2 = y^2$	$(X - \overline{X})(Y - \overline{Y}) = xy$
24 26 32 33 35 30	15 20 22 24 27 24.	–6 –4 2 3 5 0	–7 –2 0 2 5 2	36 16 4 9 25 0	49 4 0 4 25 4	42 8 0 6 25 0
$\Sigma X=180$ $\overline{X} = \frac{180}{6} = 30$	$\Sigma Y=132$ $\overline{Y} = \frac{132}{6} = 22$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 90$	$\Sigma y^2 = 86$	$\Sigma xy = 81$

$r = \dfrac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \dfrac{(81)}{\sqrt{90 \times 86}} = \dfrac{81}{\sqrt{7740}} = \dfrac{81}{87.98} = .92$

1.12 SPEARMAN'S RANK CORRELATION (COEFFICIENT OF RANK CORRELATION)

In certain types of characteristics it is not possible to get numerical measurements; but we can rank the individuals in order according to our own judgement; e.g., smartness, beauty. If two persons rank a given group of individuals and we have to find how far the two judges agree with each other, the technique of rank correlation can be used. In some cases though actual measurements are available we may be interested in only the ranks, that is the relative position of an individual in the group. Here also rank correlation is used.

The formula for Spearman's rank correlation is R = $1 - 6\Sigma d^2/N(N^2 - 1)$

where d = difference between the ranks of the same individual

N = number of individuals in the groups.

The value of this coefficient also lies between -1 and 1.

When all the ranks are the same, that is two ranks are in complete agreement with each other, R = 1, because in that case all the values of d^2 will be zero. The correlation is perfect positive in this case. When the ranks are exactly opposite, that is the ranks in one are in the reverse order of the other, R = -1. Here the correlation is perfect negative.

Example 3: The ranks according to judges in a beauty contest are given below:

Rank I	1	2	3	4	5	6
Rank II	4	1	2	3	6	5

Solution:

Rank I - R_1	Rank II - R_2	$\lvert d \rvert = \lvert R1 - R2 \rvert$	d^2
1	4	3	9
2	1	1	1
3	2	1	1
4	3	1	1
5	6	1	1
6	5	1	1
N = 6			14

$R = 1 - 6\Sigma d^2/N(N^2 - 1) = 1 - 6(14)/6(6^2 - 1) = 0.6$

In some cases ranks are not given and we have to rank the individuals first, considering the given numerical measurements. We consider below an example where ranks are to be determined by us.

Example 4: Calculate coefficient of rank correlation for the following data giving the number of hours of daily practice and the number of minutes taken to run a track by 5 runners.

No. of hours	2.00	1.50	2.50	1.75	2.75	3.00	1.60
No. of Minutes	9.00	9.25	8.30	8.10	8.20	7.00	8.80

Solution:

No. of hours	No. of Minutes	R_1	R_2	$\lvert d \rvert$ = $\lvert R1 - R2 \rvert$	d^2
2.00	9.00	4	2	2	4
1.50	9.25	7	1	6	36
2.50	8.30	3	4	1	1
1.75	8.10	5	6	1	1
2.75	8.20	2	5	3	9
3.00	7.00	1	7	6	36
1.60	8.80	6	3	3	9
					96

$R = 1 - 6\Sigma d^2/N(N^2 - 1) = 1 - 6(96)/7(7^2 - 1) = 1 - 576/336 = 1 - 1.71 = 0.71$

In the above example, the ranks were different for all the individuals but in some cases, two or more items may have the same numerical measurements and ranks should be the same for these individuals. Suppose we give the ranks 1, 2, 3 and then the next two persons have to be given the same rank. In this case the next two ranks are 4 and 5. These are to be distributed equally. Therefore both the individuals will get the rank 4 + 5/2 and the next one will get the rank 6.

When there are groups getting the same rank, there is some adjustment in the formula also. If m_1 denotes the number of persons having the same rank in the first group, m_2 denotes the number of persons having the same rank in the second group and so on, the coefficient of Rank Correlation will be

$$R = 1 - \frac{6\left\{\Sigma d2 + \frac{1}{12\left[\left(m1^3 - m1\right) + \left(m2^3 - m2\right) + \ldots\ldots\right]}\right\}}{N\,(N2 - 1)}$$

The following example will illustrate this method.

Example 5: Find the coefficient of rank correlation from the following data giving the number of hours of daily practice and the number of minutes taken to run a track by 8 runners.

No. of Hours	No. of Minutes
2	7
1.5	9.5
2	8
2.5	8
1.5	10
3	6
2	7.5
2.25	6.5

Solution:

No. of hours	No. of Minutes	R_1	R_2	$\lvert d \rvert = \lvert R1 - R2 \rvert$	d^2
2	7	5	6	1	1
1.5	9.5	7.5	2	5.5	30.25
2	8	5	3.5	1.5	2.25
2.5	8	2	3.5	1.5	2.25
1.5	10	7.5	1	6.5	42.25
3	6	6	1	8	49.00
2	7.5	7.5	5	5	0
2.25	6.5	6.5	3	7	16
Total					143.00

Here N = 8,

We have in all three groups. 2 in R_1 and 1 in $R_{2.}$

$m_1 = 3$, $m_2 = 2$, $m_3 = 2$

$$R = 1 - \frac{6\left\{\Sigma d2 + \frac{1}{12\left[(m1^3 - m1) + (m2^3 - m2) + \ldots\ldots\right]}\right\}}{N\,(N2 - 1)}$$

$$R = 1 - (6\{143 + 1/12\ (24 + 6 + 6))/(8\ (63))$$

$$R = 1 - (6\{143 + 1/12\ (36))/(8\ (63))$$

$$R = 1 - \frac{6\{143 + 3)}{8\ (63)}$$

$$R = 1 - \frac{6\{146)}{8\ (63)}$$

$= 1 - 1.74$

$= - 0.74$

1.13 CORRELATION IS NOT NECESSARILY CAUSATION

We have seen how to ascertain the correlation between two series. If we find a certain amount of correlation between two variables, does it mean that there is a cause and effect relation between the two variables? That is, is the change in one variable, the effect of the change in the other variable? This may not be so. There are various situations, which give us high degree of correlation even though the cause and effect relationship does not hold well between the two variables under consideration. It may be just an index of co-variability.

1. If we consider the heights of brothers and sisters, we will find that tall brothers have tall sisters. But it is clear that there is no cause and effect relationship between the two. Both are the children of tall parents i.e., we may have two series which are affected by the third factor either in the same way or in opposite ways and therefore they show correlation. If we consider the number of telephones subscribers and the number of car owners, the two may show high positive correlation. Here the third factor is high income of these people who can afford to have both, the car and the telephone. The presence of third factor is particularly seen in time series analysis. Two time series may show a high positive correlation because both are affected in the same manner by the business cycle.
2. Each variable may be sometimes the cause and sometimes the effect. If we consider production and price, for some commodity price will come down because of increased production, while in some cases production will increase with increase in price.
3. Correlation may not be due to chance also. If we find that there is positive correlation between the heights of father and the marks that their sons get in the examination, we cannot say that there is any relation between the two variables. It is clear that the values for the two variables in that particular sample varied together only by chance.
4. The observed correlation may be just the opposite to the one that really exists. If we consider the marks of the students and time spent in studies, we may find that there is negative correlation. The students who get more than average marks study less than those who get less than average marks. Here logically it is evident that the student dost not get good marks because he studies less. The reason for this negative correlation is that the students who get good marks are intelligent and therefore get good marks with limited amount of time spent in studies. But those who are dull have to study more even to pass. Here intelligence is the factor affecting the two given variables in opposite directions. Intelligence is positively correlated to marks and negatively correlated to the time spent in studies.

When we see high correlation between two variables, which of the above possibilities exists, whether there is a cause and effect relationship, is a matter of interpretation in the light of our knowledge about the two variables. The conclusion should be based on logical considerations. The observed correlation coefficient may only suggest the existence of cause and effect relationship but cannot prove it.

1.14 EXERCISES

Questions

1. What is a correlation coefficient intended to measure?
2. Explain with suitable examples the difference between positive and negative correlation.
3. Explain briefly the idea of correlation. Define the coefficient of correlation r. State the limiting values of r and comment on the same.
4. Mention methods used for ascertaining correlation between two variables and explain the method of rank correlation. Discuss whether coefficient of correlation is a measure of co-variation or does it indicate particular cause and effect relationship.
5. Explain the concept of correlation with the help of examples.
6. What is meant by (i) positive correlation (ii) perfect correlation.

Solve the following

1. Following data give the age of husband & wife for 8 couples. Find Pearson's coefficient of correlation.

Age of Husband in Years	35	38	40	45	48	48	50	53
Age of wife in years	30	30	33	35	40	42	45	48

2. Average prices of rice and wheat for the year 1992 to 1996 are given below. Find the coefficient of correlation between price of wheat & price of rice.

Year	**Price per quintal in Rs.**	
	Rice	**Wheat**
1992	150	82
1993	175	72
1994	200	90
1995	220	99
1996	198	81

3. Find the coefficient of correlation between the marks in Mathematics & Physics from the following data:

Roll No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Marks in Mathematics	40	37	90	85	67	75	80	52	80
Marks in Physics	50	40	80	85	75	80	85	65	85

4.

No. of new houses (100)	Appliances Sales (1000 Rs.)
2.0	5.0
2.5	5.5
3.2	6.0
3.6	7.0
3.3	7.2
4.0	7.7

The above data gives the number of new houses and sales in domestic appliances. Find the coefficient of correlation.

5. Marks of 6 students in class work and annual examination are given below. Find coefficient of correlation.

Class work	12	14	23	18	10	19
Annual Examination	68	78	85	75	70	74

❖❖❖❖

Module - II

2

REGRESSION ANALYSIS

Unit Structure :

2.1 INTRODUCTION

The statistical technique that expresses a functional (or algebraic) relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable, is called regression analysis. The variable whose value is to be estimated is called dependent (or response) variable and the variable whose value is used to estimate this value is called independent (regression or predictor) variable. The linear algebraic equations that express a dependent variable in terms of an independent variable are called linear regression equation.

Sir Francis Galton in 1877, while studying the relationship between the height of father and sons found that though 'tall father has tall sons', the average height of tall father is x above the general height and the average height of sons is 2x/3 above the general height. He described such a fall in the average height as 'regression to mdeiocrity'. The term regression in the literary sense is all as referred to as 'moving backward.'

2.2 DIFFERENCES BETWEEN CORRELATION AND REGRESSION ANALYSIS

1. Developing an algebraic equation between two variables based on the given data and estimating the value of a dependent variable given the value of an independent variable is referred to as regression analysis.
2. Measuring the strength (or degree) and direction of the relationship between two variables is referred as correlation analysis. The direction (direct or inverse) of the relationship is indicated by the correlation coefficient, and the absolute value of correlation coefficient indicates the extent (strength or degree) of the relationship.
3. Correlation analysis determines the strength (or degree) of association between two variables x and y but does not establish a cause-and-effect relationship. Regression analysis establishes the cause-and-effect relationship between x and y, that is, a change in the value of independent variable x causes a change (effect) in the value of dependent variable, y assuming that all other factors that may affect y remain unchanged.
4. In linear regression analysis one variable is considered as dependent variable and other as independent variable, while in correlation analysis both variables are considered to be independent.
5. The coefficient of determination r^2 indicates the proportion of total variance in the dependent variable that is explained or accounted for due to variation in the independent variable. Since value of r^2 is determined from a sample, its value is subject to sampling error.

2.3 ADVANTAGES OF REGRESSION ANALYSIS

The following are few advantages of regression analysis:

1. Regression analysis helps in developing an algebraic equation between two variables based on the given data and estimating the value of a dependent variable given the value of an independent variable.
2. Regression analysis helps to determine standard error of estimate to measure the variability or spread of values of a dependent variable around the regression line. Closer the pair

of values (x, y) fall around the regression line, better the line fits the data and hence smaller the variance and error of estimate. Thus, a good estimate can be made of the value of variable y when all the points fall on the line, i.e.. standard error of estimate equals zero.

3. If the sample size is large ($n \geq 30$), then interval estimation for predicting the value of a dependent variable based on standard error of estimate is considered to be acceptable by changing the values of either x or y. The magnitude of r^2 remains the same regardless of the values of the two variables.

2.4 TYPES OF REGRESSION MODELS

A regression model is an algebraic equation between two variables based on the given data and estimating the value of a dependent variable based on the known values of one or more independent variables. A particular form of regression model depends upon the nature of the problem under study and the type of data available.

Simple and Multiple Regression Models

If a regression model represents the relationship between a dependent, y, and only one independent variable, x, then such a regression model is called simple regression model. But if more than one independent variable is associated with a dependent variable, then such a regression model is called a multiple regression model. For example, sales turnover of a product (a dependent variable) is associated with more than one independent variables such as price of the product, expenditure on advertisement, quality of the product, competitors and so on. Thus, estimation of possible sales turnover with respect to only one of these independent variables is an example of a simple regression model, otherwise multiple regression model.

LINEAR AND NON-LINEAR REGRESSION MODELS

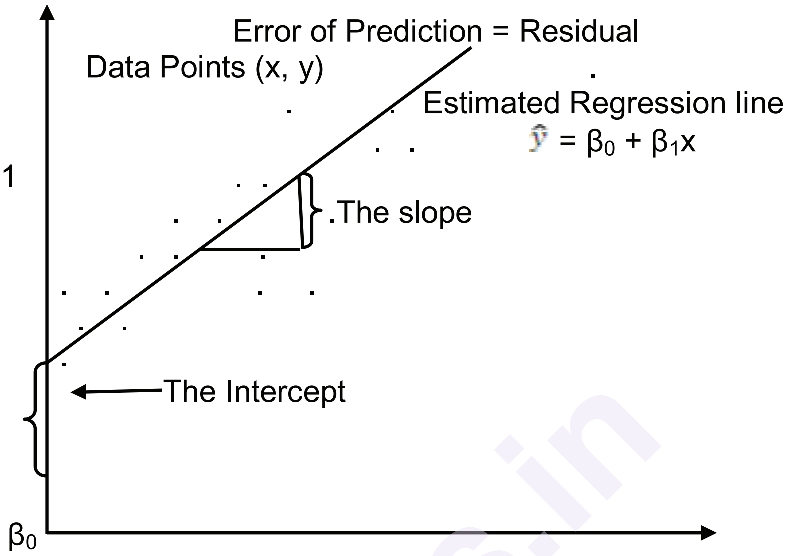
If the change (increase or decrease) in the values of a dependent (response) variable y in a regression model is directly proportional to a unit change (increase or decrease) in the values of independent (predictor) variable x, then such a model is called a linear regression model. Thus, the relationship between these two variables can be represented by a straight-line relationship in terms of population parameters β_0 and β_1 as follows:

$$E(y) = \beta_0 + \beta_1 x................(1)$$

where β_0 = y intercept that represents mean (or average) value of the dependent variable y when x = 0.

β_1 = slope of the regression line that represents the expected change (positive or negative) in the value of dependent variable, y for a unit change in the value of independent variable, x.

Figure 1


Error of Prediction = Residual
Data Points (x, y)
Estimated Regression line
$\hat{y} = \beta_0 + \beta_1 x$
.The slope
The Intercept
β_0

The intercept β_0 and the slope β_1 are unknown regression coefficients. The value of both β_0 and β_1 is to be calculated to predict average value of y for a given value of x by substituting these values in equation (1).

Figure 1 presents a scatter diagram of each pair of values of (x_i, y_i) around the regression line. Although, mean (or average) value of dependent variable, y, is a linear function of independent variable, x, but not all values of y fall exactly on the straight line. Since few points do not fall on the regression line, therefore values of y are not exactly equal to the values obtained by equation (1). Thus, such a straight line is also called line of mean deviation of observed y value from the regression line. This situation arises due to random error (also called the residual variation or residual error) in the prediction of the x, is not alone responsible for all variability in the value of variable, y. For example, sales volume is related to the level of expenditure on advertisement, but if other factors related to sales such as price of the product, quality of the product, competitors etc., are ignored, then a regression equation to predict the sales volume (y) based on budget of advertising (x) only may cause an error. Thus for a fixed value of independent variable, x, the actual value of dependent variable, y, is determined by the mean value function plus a random error term, e, as follows:

$$y = \text{Mean value function} + \text{Deviation} = \beta_0 + \beta_1 x + e \quad \text{.........................(2)}$$

The equation (2) is referred to as simple probabilistic linear regression model. The error term, e, in equation (2) is called random error because its value associated with each value of variable, y, is assumed to vary unpredictably. The extent of random error associated with each value of variable, y, for a given value of x is measured by the error variance. Lower the value of, e, better the regression model fit to a sample data.

The random errors corresponding to different observations (x_i, y_i) for all i are assumed to follow a normal distribution with mean zero and (unknown) constant standard deviation.

If the line passing through the pair of values of variables x and y is not linear, then the relationship between variables x and y is non-linear. A non-linear relationship implies that expected change (positive or negative) in the value of dependent variable, y, is not directly proportional to a unit change in the value of independent variable, x. A non-linear relationship is not very useful for predictions.

2.5 ESTIMATION: THE METHOD OF LEAST SQUARES

A sample of n pairs of observations (x_1, y_1), (x_2, y_2) ....(x_n, y_n) is drawn from the population under study to estimate the values of regression coefficients β_0 and β_1. The method that provides the best linear unbiased estimate of the values of β_0 and β_1 is called the method of least squares. The estimated values of β_0 and β_1 should result in a straight line where most pairs of observations (x_1, y_1), (x_2, y_2) ....(x_n, y_n) fall very close (best fit) to it. Such a straight line is referred to as 'best fitted' (least squares or estimated) regression line because the sum of the squares of the vertical deviations (difference between the actual values of y and the estimated values predicted from the fitted line) is as small as possible.

Rewrite equation (2) as follows:
$y_i = \beta_0 + \beta_1 x_i + e_i$ or $e_i = y_i - (\beta_0 + \beta_1 x_i)$ for all i

Mathematically, we intend to minimize

$$L = \sum_{i=1}^{n} ei\ ^{2} = \Sigma_{\downarrow}(i = 1)^{\uparrow} n \equiv \{\ y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Let b_0 and b_1 be the best least squares estimators of β_0 and β_1 respectively. The least squares estimators b_0 and b_1 must satisfy the following equations:

$$\frac{\partial L}{\partial \beta 0}\Bigg|\; b_0b_1 = -2\ \Sigma_{\downarrow}(i = 1)^{\uparrow} n \equiv (\ y_i - b_0 - b_1\ x_i) = 0$$

$$\frac{\partial L}{\partial \beta 1}\Bigg|\; b_0b_1 = -2\ \Sigma_{\downarrow}(i = 1)^{\uparrow} n \equiv (\ y_i - b_0 - b_1\ x_i)\ x_i = 0$$

After simplifying these two equations, we get

$$\sum_{i=1}^{n} yi = n\ b_0 + b_1 \sum_{i=1}^{n} xi \quad \text{.. (3)}$$

$$\sum_{i=1}^{n} xi\ yi = b_0 \sum_{i=1}^{n} xi + b_1 \sum_{i=1}^{n} xi^{2}$$

Equation (3) is called the least squares normal equation. The values of least squares estimators b_0 and b_1 can be obtained by solving equation (3). Hence, the fitted or estimated regression line is given by

$$\hat{y} = b_0 + b_1x$$

where $\hat{y}$ (called y hat) is the predicted value of y falling on the fitted regression line for a given value of x and $e_i = yi - \hat{y}i$ is called that describes the amount of error in fitting of the regression line to the values of yi.

Remark:

$\sum ei$ **= 0 i.e., sum of the residuals is zero for any least-squares regression line.**

ASSUMPTIONS FOR THE SIMPLE LINEAR REGRESSION MODEL

To form a basis for application of simple linear regression models, certain assumptions about the population from which a sample of observations is drawn.

Assumptions

1. There is a linear relationship between the dependent variable y and independent variable x. This relationship can be described by a linear regression equation y = a + bx + e, where e represents the deviation in the value of dependent variable, y, from its expected value for a given value of independent variable, x.
2. The set of expected (or mean) values of the dependent variable, y, for given values of independent variable, x are normally distributed. The mean of these normally distributed values falls on the line of regression.

3. The dependent variable y is a continuous random variable, whereas values of the independent variable x are fixed and not random.
4. The sampling error associated with the expected value of the dependent variable y is assumed to be an independent random variable distributed normally with mean zero and constant standard deviation. The amount of deviation (error) in the value of dependent variable, y, may be different in successive observations.
5. The standard deviation and variance of expected values of the dependent variable about the regression line are constant for all values of the independent variable x for the set of observations in a sample.
6. The expected value of the dependent variable cannot be obtained for a value of an independent variable falling outside the range of values in the sample.

2.6 PARAMETERS OF SIMPLE LINEAR REGRESSION MODEL

The objective regression analysis is to ascertain that a regression equation (line) should provide the best fit of sample data to the population data so that the error of variance is as small as possible. J. R. Stockton stated that the device used for estimating the values of one variable from the value of the other consists of a line through the points, drawn in such a manner as to represent the average relationship between the two variables.

Such a line is called line of regression.

The two variables x and y which are correlated can be expressed in terms of each other in the form of straight line equations called regression equations as follows:

- The regression equation of y on x
  $y = a + bx$
  is used for estimating the value of y for given values of x.
- Regression equation of x on y
  $x = c + dy$
  is used for estimating the value of x for given values of y.

Remarks:

1. Regression line coincide (overlap) when variables x and y are perfectly correlated (either positive or negative).
2. Higher the degree of correlation, the two regression lines come closer to each other. Lesser the degree of correlation, the two regression lines are away from each other. Also, when variables x and y are not correlated, i.e. $r = 0$, the two regression lines are at right angle to each other.

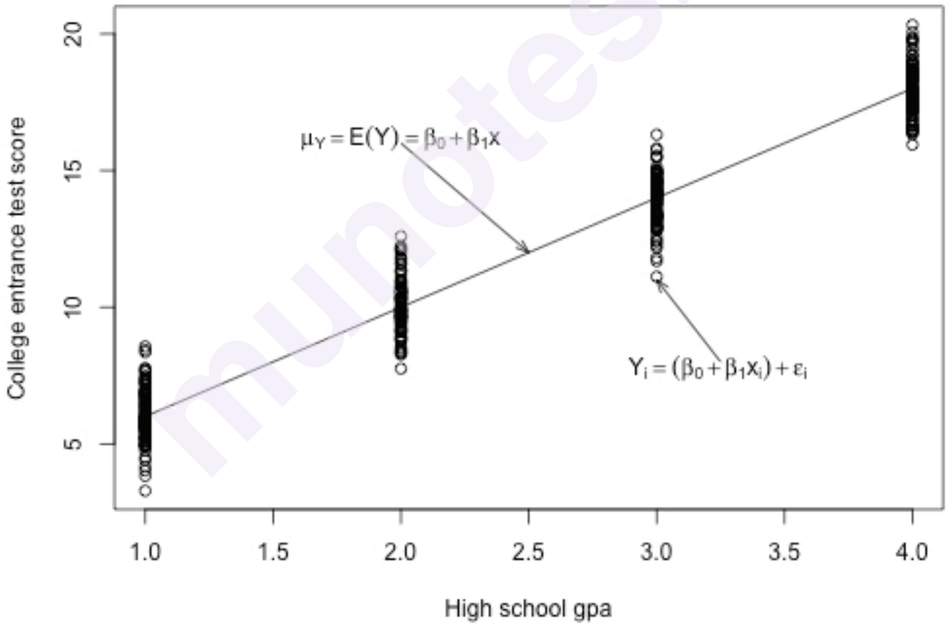
3. The point of interaction of two linear regression lines represents the average value of variables x and y.

2.7 THE SIMPLE LINEAR REGRESSION MODEL

What do *b_0* and *b_1* estimate?

Let's investigate this question with another example. Below is a plot illustrating a potential relationship between the predictor "high school grade point average (gpa)" and the response "college entrance test score." Only five groups ("subpopulations") of students are considered — those with a gpa of 1, those with a gpa of 2, ..., and those with a gpa of 4.

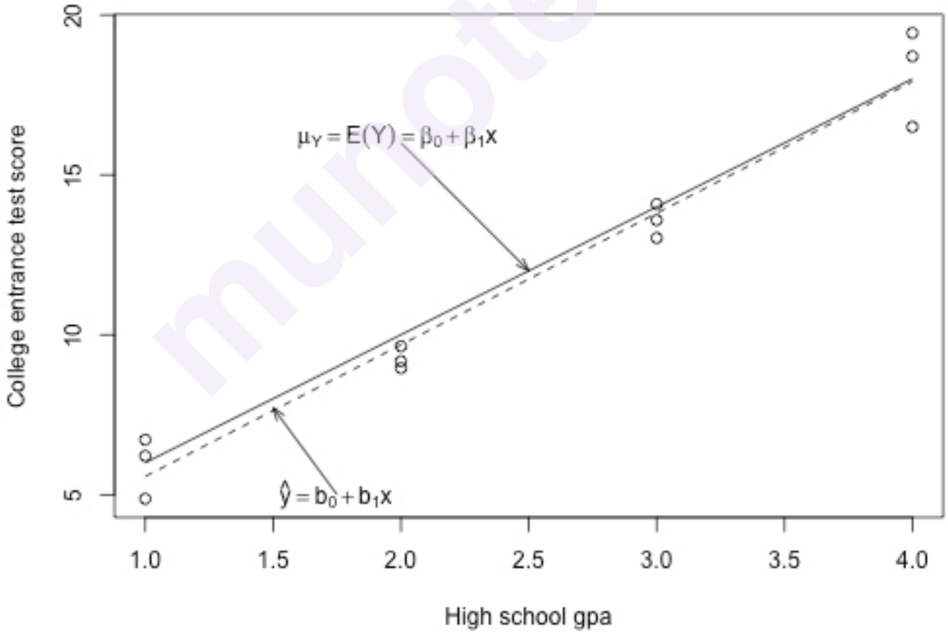
Let's focus for now just on those students who have a gpa of 1. As you can see, there are so many data points — each representing one student — that the data points run together. That is, the data on the entire subpopulation of students with a gpa of 1 are plotted. And, similarly, the data on the entire subpopulation of students with gpas of 2, 3, and 4 are plotted.



Now, take the average college entrance test score for students with a gpa of 1. And, similarly, take the average college entrance test score for students with a gpa of 2, 3, and 4. Connecting the dots — that is, the averages — you get a line, which we summarize by the formula $\mu Y=E(Y)=\beta 0+\beta 1x\mu Y=E(Y)=\beta 0+\beta 1x$. The line — which is called the "**population regression line**" — summarizes the trend *in the population* between the predictor *x* and the mean of the responses *μ_Y*. We can also express the average college entrance test score for the *i*-th student, $E(Yi)=\beta 0+\beta 1xiE(Yi)=\beta 0+\beta 1xi$. Of course, not every student's college entrance test score will equal the average E(Yi)E(Yi). There will be some error. That is, any student's response y_i will be the linear trend β0+β1xiβ0+β1xi plus some error ϵiϵi. So, another way to write the simple linear regression model is yi=E(Yi)+ϵi=β0+β1xi+ϵiyi=E(Yi)+ϵi=β0+β1xi+ϵi.

When looking to summarize the relationship between a predictor *x* and a response *y*, we are interested in knowing the population regression line μY=E(Y)=β0+β1xμY=E(Y)=β0+β1x. The only way we could ever know it, though, is to be able to collect data on everybody in the population — most often an impossible task. We have to rely on taking and using a sample of data from the population to estimate the population regression line.

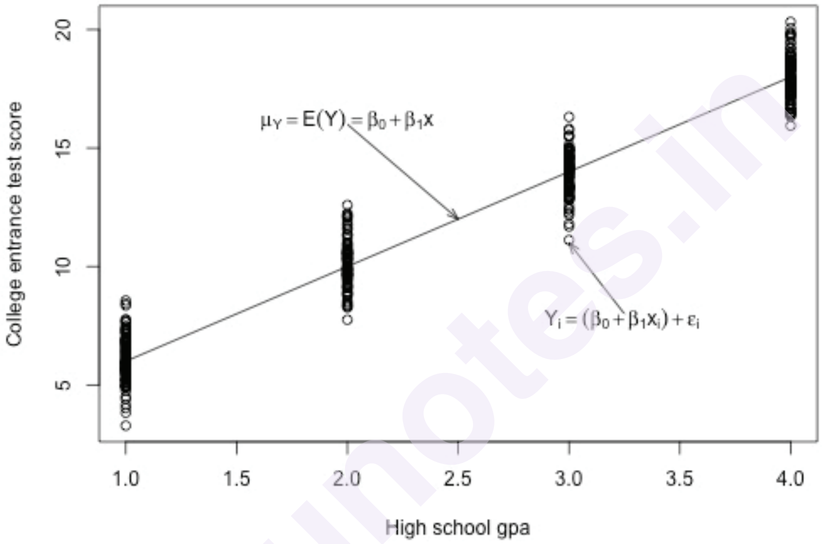
Let's take a sample of three students from each of the subpopulations — that is, three students with a gpa of 1, three students with a gpa of 2, ..., and three students with a gpa of 4 — for a total of 12 students. As the plot below suggests, the least squares regression line y^=b0+b1xy^=b0+b1x through the sample of 12 data points estimates the population regression line μY=E(Y)=β0+β1xμY=E(Y)=β0+β1x. That is, the sample intercept b_0 estimates the population intercept β_0 and the sample slope b_1 estimates the population slope β_1.



The least squares regression line doesn't match the population regression line perfectly, but it is a pretty good estimate. And, of course, we'd get a different least squares regression line if we took another (different) sample of 12 such students. Ultimately, we are going to want to use the sample slope b_1 to learn about the parameter we care about, the population slope β_1. And, we will use the sample intercept b_0 to learn about the population intercept β_0.

In order to draw any conclusions about the population parameters β_0 and β_1, we have to make a few more assumptions about the behavior of the data in a regression setting. We can get a pretty good feel for the assumptions by looking at our plot of gpa against college entrance test scores.

First, notice that when we connected the averages of the college entrance test scores for each of the subpopulations, it formed a line. Most often, we will not have the population of data at our disposal as we pretend to do here. If we didn't, do you think it would be reasonable to assume that the mean college entrance test scores are **linearly related** to high school grade point averages?



Again, let's focus on just one subpopulation, those students who have a gpa of 1, say. Notice that most of the college entrance scores for these students are clustered near the mean of 6, but a few students did much better than the subpopulation's average scoring around a 9, and a few students did a bit worse scoring about a 3. Do you get the picture? Thinking instead about the errors, ϵiϵi, most of the errors for these students are clustered near the mean of 0, but a few are as high as 3 and a few are as low as -3. If you could draw a probability curve for the errors above this subpopulation of data, what kind of a curve do you think it would be? Does it seem reasonable to assume that the errors for each subpopulation are **normally distributed**?

Looking at the plot again, notice that the spread of the college entrance test scores for students whose gpa is 1 is similar to the spread of the college entrance test scores for students whose gpa is 2, 3, and 4. Similarly, the spread of the errors is similar, no matter the gpa. Does it seem reasonable to assume that the errors for each subpopulation have **equal variance**?

Does it also seem reasonable to assume that the error for one student's college entrance test score is **independent** of the error for another student's college entrance test score? I'm sure you can come up with some scenarios — cheating students, for example — for which this assumption would not hold, but if you take a random sample from the population, it should be an assumption that is easily met.

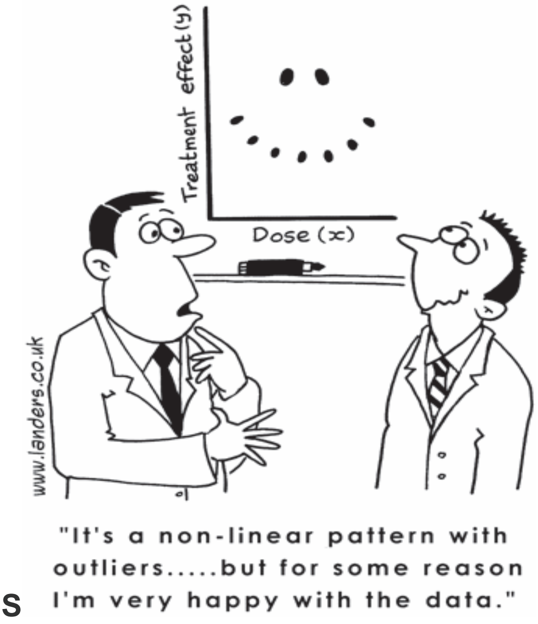
We are now ready to summarize the four conditions that comprise "**the simple linear regression model**:"

- The mean of the response, E(Yi)E(Yi), at each value of the predictor, xixi, is a **Linear function** of the xixi.
- The errors, ε_i, are **Independent**.
- The errors, ε_i, at each value of the predictor, xixi, are **Normally distributed**.
- The errors, ε_i, at each value of the predictor, xixi, have **Equal variances** (denoted σ^2).

Do you notice what the first letters that are colored in blue spell? "**LINE**." And, what are we studying in this course? Lines! Get it? You might find this mnemonic a useful way to remember the four conditions that make up what we call the "simple linear regression model." Whenever you hear "simple linear regression model," think of these four conditions!

An equivalent way to think of the first (linearity) condition is that the mean of the error, E(ϵi)E(ϵi), at each value of the predictor, xixi, is **zero**. An alternative way to describe all four assumptions is that the errors, ϵiϵi, are independent normal random variables with mean zero and constant variance, σ2σ2.

POPULAR APPLICATIONS OF LINEAR REGRESSION FOR BUSINESSES



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

In Statistics, Linear regression refers to a model that can show relationship between two variables and how one can impact the other. In essence, it involves showing how the variation in the “dependent variable” can be captured by change in the “independent variables”.

In Business, this dependent variable can also be called the predictor or the factor of interest for eg., sales of a product, pricing, performance, risk etc. Independent variables are also called explanatory variables as they can explain the factors that influence the dependent variable along with the degree of the impact which can be calculated using “parameter estimates” or “coefficients”. These coefficients are tested for statistical significance by building confidence intervals around them so that the model that we are building is statistically robust and based on objective data. The elasticity based on the coefficient can tell us the extent to which a certain factor explains the dependent. Further, a negative coefficient can be interpreted to have a negative or an inverse relation with the dependent variable and positive coefficient can be said to have a positive influence. The key factor in any statistical models is the right understanding of the domain and its business application.

Linear Regression is a very powerful statistical technique and can be used to generate insights on consumer behaviour, understanding business and factors influencing profitability. Linear regressions can be used in business to evaluate trends and make estimates or forecasts. For example, if a company’s sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months.

Linear regression can also be used to analyze the marketing effectiveness, pricing and promotions on sales of a product. For instance, if company XYZ, wants to know if the funds that they have invested in marketing a particular brand has given them substantial return on investment, they can use linear regression. The beauty of linear regression is that it enables us to capture the isolated impacts of each of the marketing campaigns along with controlling the factors that could influence the sales. In real life scenarios there are multiple advertising campaigns that run during the same time period. Supposing two campaigns are run on TV and Radio in parallel, a linear regression can capture the isolated as well as the combined impact of running this ads together.

Linear Regression can be also used to assess risk in financial services or insurance domain. For example, a car insurance company might conduct a linear regression to come up with a suggested premium table using predicted claims to Insured Declared Value ratio. The risk can be assessed based on the attributes of the car, driver information or demographics. The results of such an analysis might guide important business decisions.

In the credit card industry, a financial company maybe interested in minimizing the risk portfolio and wants to understand the top five factors that cause a customer to default. Based on the results the company could implement specific EMI options so as to minimize default among risky customers.

While Linear regression has limited applicability in business situations because it can work only when the dependent variable is of continuous nature, it still is a very well known technique in the situations it can be used. It assumes a linear relation between the independent and dependent variables. It must be noted that sometimes transformations can also be applied to non linear relationships to make them applicable in a linear regression model.

General Purpose

The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating). You may also detect "outliers," that is, houses that should really sell for more, given their location and characteristics.

Personnel professionals customarily use multiple regression procedures to determine equitable compensation. You can determine a number of factors or dimensions such as "amount of responsibility" (*Resp*) or "number of people to supervise" (*No_Super*) that you believe to contribute to the value of a job. The personnel analyst then usually conducts a salary survey among comparable companies in the market, recording the salaries and respective characteristics (i.e., values on dimensions) for different positions. This information can be used in a multiple regression analysis to build a regression equation of the form:

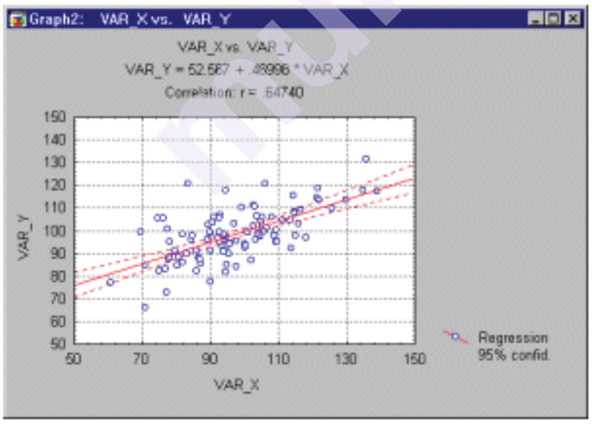
Salary = .5*Resp + .8*No_Super

Once this so-called regression line has been determined, the analyst can now easily construct a graph of the expected (predicted) salaries and the actual salaries of job incumbents in his or her company. Thus, the analyst is able to determine which position is underpaid (below the regression line) or overpaid (above the regression line), or paid equitably.

In the social and natural sciences multiple regression procedures are very widely used in research. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of ...". For example, educational researchers might want to learn what are the best predictors of success in high-school. Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt and be absorbed into society.

Computational Approach

The general computational problem that needs to be solved in multiple regression analysis is to fit a straight line to a number of points.



In the simplest case - one dependent and one independent variable - you can visualize this in a scatterplot.

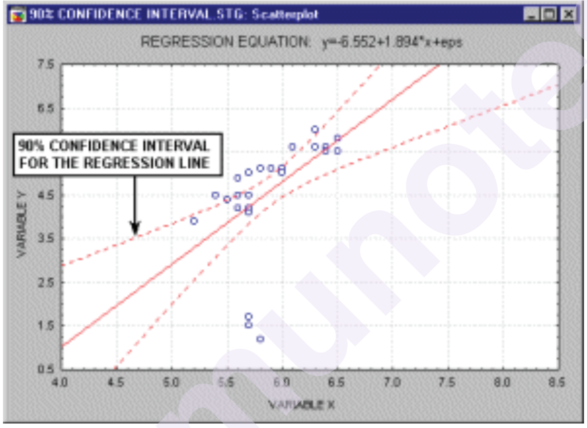
LEAST SQUARES

In the scatterplot, we have an independent or *X* variable, and a dependent or *Y* variable. These variables may, for example, represent IQ (intelligence as measured by a test) and school achievement (grade point average; GPA), respectively. Each point in the plot represents one student, that is, the respective student's IQ and GPA. The goal of linear regression procedures is to fit a line through the points. Specifically, the program will compute a line so that the squared deviations of the observed points from that line are minimized. Thus, this general procedure is sometimes also referred to as least squares estimation.

THE REGRESSION EQUATION

A line in a two dimensional or two-variable space is defined by the equation *Y=a+b*X*; in full text: the *Y* variable can be expressed in terms of a constant (*a*) and a slope (*b*) times the *X* variable. The constant is also referred to as the *intercept*, and the slope as the *regression coefficient* or *B coefficient*. For example, GPA may best be predicted as *1+.02*IQ*. Thus, knowing that a student has an *IQ* of 130 would lead us to predict that her GPA would be 3.6 (since, 1+.02*130=3.6).

For example, the animation below shows a two dimensional regression equation plotted with three different confidence intervals (90%, 95% and 99%).



In the multivariate case, when there is more than one independent variable, the regression line cannot be visualized in the two dimensional space, but can be computed just as easily. For example, if in addition to *IQ* we had additional predictors of achievement (e.g., *Motivation, Self- discipline*) we could construct a linear equation containing all those variables. In general then, multiple regression procedures will estimate a linear equation of the form:

$$Y = a + b_1{*}X_1 + b_2{*}X_2 + ... + b_p{*}X_p$$

2.8 UNIQUE PREDICTION AND PARTIAL CORRELATION

Note that in this equation, the regression coefficients (or *B* coefficients) represent the *independent* contributions of each independent variable to the prediction of the dependent variable.

Another way to express this fact is to say that, for example, variable X_1 is correlated with the *Y* variable, after controlling for all other independent variables. This type of correlation is also referred to as a *partial correlation* (this term was first used by Yule, 1907). Perhaps the following example will clarify this issue. You would probably find a significant negative correlation between hair length and height in the population (i.e., short people have longer hair). At first this may seem odd; however, if we were to add the variable *Gender* into the multiple regression equation, this correlation would probably disappear. This is because women, on the average, have longer hair than men; they also are shorter on the average than men. Thus, after we remove this gender difference by entering *Gender* into the equation, the relationship between hair length and height disappears because hair length does *not* make any unique contribution to the prediction of height, above and beyond what it shares in the prediction with variable *Gender*. Put another way, after controlling for the variable *Gender*, the partial correlation between hair length and height is zero.

2.9 PREDICTED AND RESIDUAL SCORES

The regression line expresses the best prediction of the dependent variable (*Y*), given the independent variables (*X*). However, nature is rarely (if ever) perfectly predictable, and usually there is substantial variation of the observed points around the fitted regression line (as in the scatter plot shown earlier). The deviation of a particular point from the regression line (its predicted value) is called the *residual* value.

2.10 RESIDUAL VARIANCE AND R-SQUARE

R-Square, also known as the *Coefficient of determination* is a commonly used statistic to evaluate model fit. *R-square* is 1 minus the *ratio of residual variability*. When the variability of the residual values around the regression line relative to the overall variability is small, the predictions from the regression equation are good. For example, if there is no relationship between the X and Y variables, then the *ratio of the residual variability* of the Y variable to the original variance is equal to 1.0. Then R-square would be 0. If X and Y are perfectly related then there is no residual variance and the ratio of variance would be 0.0, making R-square = 1. In most cases, the ratio and *R-square* will fall somewhere between these extremes, that is, between 0.0 and 1.0. This ratio value is immediately interpretable in the following manner. If we have an *R-square* of 0.4 then we know that the variability of the Y values around the regression line is 1-0.4 times the original variance; in other words we have explained 40% of the original variability, and are left with 60% residual variability. Ideally, we would like to explain most if not all of the original variability. The *R-square* value is an indicator of how well the model fits the data (e.g., an *R-square* close to 1.0 indicates that we have accounted for almost all of the variability with the variables specified in the model).

2.11 INTERPRETING THE CORRELATION COEFFICIENT R

Customarily, the degree to which two or more predictors (independent or *X* variables) are related to the dependent (*Y*) variable is expressed in the correlation coefficient *R*, which is the square root of *R-square*. In multiple regression, *R* can assume values between 0 and 1. To interpret the direction of the relationship between variables, look at the signs (plus or minus) of the regression or *B* coefficients. If a *B* coefficient is positive, then the relationship of this variable with the dependent variable is positive (e.g., the greater the IQ the better the grade point average); if the *B* coefficient is negative then the relationship is negative (e.g., the lower the class size the better the average test scores). Of course, if the *B* coefficient is equal to 0 then there is no relationship between the variables.

2.12 ASSUMPTIONS, LIMITATIONS, PRACTICAL CONSIDERATIONS ASSUMPTION OF LINEARITY

First of all, as is evident in the name multiple *linear* regression, it is assumed that the relationship between variables is linear. In practice this assumption can virtually never be confirmed; fortunately, multiple regression procedures are not greatly affected by minor deviations from this assumption. However, as a rule it is prudent to *always* look at bivariate scatterplot of the variables of interest. If curvature in the relationships is evident, you may consider either transforming the variables, or explicitly allowing for nonlinear components.

NORMALITY ASSUMPTION

It is assumed in multiple regression that the residuals (predicted minus observed values) are distributed normally (i.e., follow the normal distribution). Again, even though most tests (specifically the *F*-test) are quite robust with regard to violations of this assumption, it is *always* a good idea, before drawing final conclusions, to review the distributions of the major variables of interest. You can produce histograms for the residuals as well as normal probability plots, in order to inspect the distribution of the residual values.

LIMITATIONS

The major conceptual limitation of all regression techniques is that you can only ascertain *relationships*, but never be sure about underlying *causal* mechanism. For example, you would find a strong positive relationship (correlation) between the damage that a fire does and the number of firemen involved in fighting the blaze. Do we conclude that the firemen cause the damage? Of course, the most likely explanation of this correlation is that the size of the fire (an external variable that we forgot to include in our study) caused the damage as well as the involvement of a certain number of firemen (i.e., the bigger the fire, the more firemen are called to fight the blaze). Even though this example is fairly obvious, in real correlation research, alternative causal explanations are often not considered.

2.13 CHOICE OF THE NUMBER OF VARIABLES

Multiple regression is a seductive technique: "plug in" as many predictor variables as you can think of and usually at least a few of them will come out significant. This is because you are capitalizing on chance when simply including as many variables as you can think of as predictors of some other variable of interest. This problem is compounded when, in addition, the number of observations is relatively low. Intuitively, it is clear that you can hardly draw conclusions from an analysis of 100 questionnaire items based on 10 respondents. Most authors recommend that you should have at least 10 to 20 times as many observations (cases, respondents) as you have variables; otherwise the estimates of the regression line are probably very unstable and unlikely to replicate if you were to conduct the study again.

2.14 MULTICOLLINEARITY AND MATRIX ILL-CONDITIONING

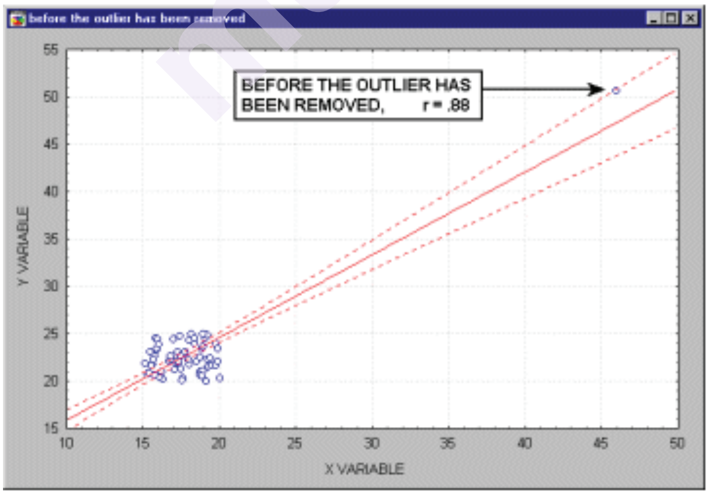
This is a common problem in many correlation analyses. Imagine that you have two predictors (*X* variables) of a person's height: (1) weight in pounds and (2) weight in ounces. Obviously, our two predictors are completely redundant; weight is one and the same variable, regardless of whether it is measured in pounds or ounces. Trying to decide which one of the two measures is a better predictor of height would be rather silly; however, this is exactly what you would try to do if you were to perform a multiple regression analysis with height as the dependent (*Y*) variable and the two measures of weight as the independent (X) variables. When there are very many variables involved, it is often not immediately apparent that this problem exists, and it may only manifest itself after several variables have already been entered into the regression equation. Nevertheless, when this problem occurs it means that at least one of the predictor variables is (practically) completely redundant with other predictors. There are many statistical indicators of this type of redundancy (tolerances, semi-partial *R*, etc., as well as some remedies (e.g., *Ridge regression*).

2.15 FITTING CENTERED POLYNOMIAL MODELS

The fitting of higher-order polynomials of an independent variable with a mean not equal to zero can create difficult multi-collinearity problems. Specifically, the polynomials will be highly correlated due to the mean of the primary independent variable. With large numbers (e.g., Julian dates), this problem is very serious, and if proper protections are not put in place, can cause wrong results. The solution is to "center" the independent variable (sometimes, this procedures is referred to as "centered polynomials"), i.e., to subtract the mean, and then to compute the polynomials. See, for example, the classic text by Neter, Wasserman, & Kutner (1985, Chapter 9), for a detailed discussion of this issue (and analyses with polynomial models in general).

2.16 THE IMPORTANCE OF RESIDUAL ANALYSIS

Even though most assumptions of multiple regression cannot be tested explicitly, gross violations can be detected and should be dealt with appropriately. In particular outliers (i.e., extreme cases) can seriously bias the results by "pulling" or "pushing" the regression line in a particular direction (see the animation below), thereby leading to biased regression coefficients. Often, excluding just a single extreme case can yield a completely different set of results.


before the outlier has been removed
BEFORE THE OUTLIER HAS BEEN REMOVED, r = .88
55
50
45
40
35
30
25
20
15
Y VARIABLE
10
15
20
25
30
35
40
45
50
X VARIABLE

REGRESSION COEFFICIENTS

Example 1

Use least squares regression line to estimate the increase in sales revenue expected from an increase of 7.5 percent in advertising expenditure.

Firm	**Annual Percentage Increase in Advertising Expenditure**	**Annual Percentage Increase in Sales Revenue**
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

Solution: Assume sales revenue (y) is dependent on advertising expenditure (x). Calculations for regression line using following normal equations are shown in Table 1.

$\sum y$ = na + b$\sum x$ and $\sum xy$ = a$\sum x$ + b$\sum x$ ^{2}

Calculations for normal equations

Sales Revenue	Advertising Expenditure x	X^2	Xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
40	56	524	373

Normal Equation Approach

$\sum y$ = na + b$\sum x$ or 40 = 8a + 56b

$\sum xy$ = a$\sum x$ + b$\sum x$ ^{2} or 373 = 56a + 524b

Solving these equations, we get a = 0.072 and b = 0.704

Substituting these values in the regression equation

$$y = a + bx = 0.072 + 0.704x$$

For x = 7.5 % or 0.075 an increase in advertising expenditure, the estimated increase in sales revenue will be

$$y = 0.072 + 0.704(0.075) = 0.1248 \text{ or } 12.48\%$$

Short-cut method

$$b = S_{xy}/ S_{xx} = 93/132 = 0.704,$$

$$\text{where } S_{xy} = \sum xy - \sum xy /n = 373 - 40 \times 56/8 = 93$$

$$S_{xx} = \sum x^2 - \sum x^2/n = 524 - (56)^2/8 = 132$$

The intercept 'a' on the y-axis is calculated as

$$a = \bar{y} - b\bar{x} = 40/8 - 0.704 \times 56/8 = 5 - 0.704 \times 7 = 0.072$$

Substituting the values of a = 0.072 and b = 0.704 in the regression equation, we get

$$y = a + bx = 0.072 + 0.704x$$

For x = 0.075, we have y = 0.072 + 0.704 (0.075) = 0.1248 or 12.48%.

Example 2

The owner of a small garment shop is hopeful that his sales are rising significantly week by week. Treating the sales for the previous six weeks as a typical example of this rising trend, he recorded them in Rs. 1000's and analysed the results.

Week	**1**	**2**	**3**	**4**	**5**	**6**
Sales	2.69	2.62	2.80	2.70	2.75	2.81

Fit a linear regression equation to suggest to him the weekly rate at which his sales are rising and use this equation to estimate expected sales for the 7th week.

Solution:

Assume sales (y) are dependent on weeks (x). Then the normal equations for regression equation y = a + bx are written as

$$\sum y = na + b\sum x \text{ and } \sum xy = a\sum x + b\sum x^2$$

Calculations for sales during various weeks are shown in table 1.

Calculation of Normal Equations

Week (x)	Sales (y)	X^2	Xy
1	2.69	1	2.69
2	2.62	4	5.24
3	2.80	9	8.40
4	2.70	16	10.80
5	2.75	25	13.75
6	2.81	36	16.86
21	16.37	91	57.74

The gradient 'b' is calculated as

$b = S_{xy}/ S_{xx} = 0.445/17.5 = 0.025$; $S_{xy} = \sum xy - \sum xy /n = 57.74 - 21 \times 16.37/6 = 0.445$

$$S_{xx} = \sum x\ ^2 - \sum x\ ^2/n = 91 - (21)^2/6 = 17.5$$

The intercept 'a' on the y-axis is calculated as

$$a = \bar{y} - b\bar{x} = 16.37/6 - 0.025 \times 21/6 = 2.728 - 0.25 \times 3.5 = 2.64$$

Substituting the values a = 2.64 and b = 0.025 in the regression equation, we have

$$y = a + bx = 2.64 + 0.025x$$

For x = 7, we have y = 2.64 + 0.025(7) = 2.815

Hence, the expected sales during the 7th week are likely to be Rs. 2.815 (in Rs. 1000's).

2.17 MULTIPLE CORRELATION ANALYSIS

The multiple correlation coefficient is denoted by $R_{1.23} = \sqrt{R}\ ^2{}_{1.23}$ for a dependent variable x_1 and two independent variables x_2 and x_3 measures the extent of the association between a dependent variable and several independent variables taken together.

The coefficient of multiple correlation can be expressed in terms of simple linear correlation coefficients r_{12}, r_{13} and r_{23} as
$R_{1.23} = \sqrt{\square}\, r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23} /1 - r_{23}^2$

By summary, we may also write
$R_{2.13} = \sqrt{\square}\, r_{21}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23} /1 - r_{13}^2$ and $R_{3.12} = \sqrt{\square}\, r_{31}^2 + r_{32}^2 - 2\, r_{12}\, r_{13}\, r_{23} /1 - r_{12}^2$

The values of the multiple correlation coefficient always lies between 0 and 1. When $R_{y.12} = 0$, implies no linear relationship between the variables. Further, when $R_{y.12} = 1$, the correlation is called perfect.

Example 3

In a trivariate distribution it is found that $r_{12} = 0.70$, $r_{13} = 0.61$, and $r_{23} = 0.40$. Find the values of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$.

Solution:

The partial correlation between variables 2 and 3 keeping the influence of variable 1 constant is given by

$r_{23.1} = r_{23} - r_{12}\, r_{13}/\sqrt{1} -\ r_{12}^2 \sqrt{\ }\ 1 - r_{13}^2$

Substituting the given values, we get

$r_{23.1} = 0.40 - 0.70 \times 0.61/\sqrt{\ }1 - (0.70)^2\sqrt{\ }1 - (0.61)^2 = 0.40 - 0.427/\sqrt{\ }0.51\ \sqrt{\ }0.6279$

$= 0.027/0.714 \times 0.7924 = 0.027/0.5657 = 0.0477.$

Similarly, we get

$r_{13.2} = r_{13} - r_{12}\ r_{23}/\sqrt{1} - r_{12}^2\sqrt{\ }\ 1 - r_{23}^2 = 0.61 - 0.70 \times 0.40/\sqrt{1} - (0.70)^2\ \sqrt{1} - (0.40)^2 = 0.504$

$r_{12.3} = r_{12} - r_{13}\ r_{23}/\sqrt{1} - r_{13}^2\sqrt{\ }\ 1 - r_{23}^2 = 0.70 - 0.61 \times 0.40/ \sqrt{1} - (0.61)^2\ \sqrt{1} - (0.40)^2 = 0.633.$

Example 4

Based on the following data, calculate $R_{1.23}$, $R_{3.12}$ and $R_{2.13}$.

$\overline{x1}$ = 6.8	**$\overline{x2}$ = 7.0**	**$\overline{x3}$ = 74**
S_1 = 1.0	S_2 = 0.8	S_3 = 9.0
r_{12} = 0.6	r_{13} = 0.7	r_{23} = 0.65

Solution:

The coefficient of multiple determination of two independent variables coded as 2 and 3 is given by

$R_{1.23} = \sqrt{\ }r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23} /1 - r_{23}^2$

Substituting the given values, we get

$R_{1.23} = \sqrt{\ }(0.6)^2 + (0.7)^2 - 2 \times 0.6 \times 0.7 \times 0.65 /1 - (0.65)^2 = \sqrt{\ }0.36 + 0.49 - 0.546/0.5775 = \sqrt{\ }0.526 = 0.725.$

Similarly,

$R_{3.12} = \sqrt{\ }(0.7)^2 + (0.65)^2 - 2 \times 0.6 \times 0.7 \times 0.65 /1 - (0.6)^2 = \sqrt{\ }0.49 + 0.4225 - 0.546/1 - 0.36 = \sqrt{\ }0.573 = 0.757.$

$R_{2.13} = \sqrt{\ } r_{21}^2 + r_{23}^2 - 2\ r_{12}\ r_{13}\ r_{23}\ /1 - r_{13}^2 = \sqrt{\ }(0.6)^2 + (0.7)^2 - 2 \times 0.6 \times 0.7 \times 0.65\ /1 - (0.7)^2 = \sqrt{\ }0.36 + 0.4225 - 0.546/0.51 = \sqrt{\ }0.464 = 0.681.$

2.18 STANDARD ERROR OF ESTIMATE FOR MULTIPLE REGRESSION

When we measure variation in the dependent variable y in multiple linear regression analysis we calculate three types of variations exactly in the same way as in simple linear regression. These variations are:

- Total variation or total sum of squares deviation

SST = $\sum_{i=1}^{n} y - \bar{y}_{2}$

- Explained variation resulting from regression relationship

between x and y SSR = $\sum_{i=1}^{n} y - \bar{y}_{2}$

Unexplained variation resulting from sampling error

SSE = $\sum_{i=1}^{n} y - \bar{y}_{2}$

The calculations of each of these sum of squares is associated with a certain number of degrees of freedom. SST has n - 1 degrees of freedom (n sample observations minus one due to fixed sample mean), SSR has k degrees of freedom (k independent variables involved in estimating value of y), SSE has n - (k + 1) degrees of freedom (n sample observations minus k + 1 constants a, b_1, b_2 ..... b_k), because in multiple regression we estimate k slope parameters b_1, b_2 ..... b_k and an intercept a from a data set containing n observations.

The three types of variations are related by the following equation:

$$\sum_{i=1}^{n} y - \bar{y}_{2} = \sum_{i=1}^{n} y - \bar{y}_{2} + \sum_{i=1}^{n} y - \bar{y}_{2}$$

The significant of regression effect is tested by computing the F-test statistic as shown in table 1.

ANOVA table for Multiple Regression

Source of Variation	**Sum of squares**	**Degrees of Freedom**	**Mean Square**	**F-Ratio**
Regression	SSR	k	MSR = SSR/k	F = MSR/MSE
Residual (or Error)	SSE	n - (k + 1)	MSE = SSE/ n - (k + 1)	
Total	SST	n-1		

Decision Rule: If the calculated value of F_{cal} is more that its table value at a given level of significance and degrees of freedom k for numerator and n - k - 1 for denominator, then H_0 is rejected.

If two variables are involved in the least-squares regression equation to predict the value of the dependent variable, then the standard error of estimate denoted by $S_{y.12}$ is given by

$S_{y.12} = \sqrt{\square}SSE/n - 3 = \sqrt{\Sigma \equiv [(y - y^\wedge]}^2)/n - 3 = \sqrt{\Sigma [(y2 - b1\Sigma x1y - b2]\ \Sigma x2y}/n - 3$

The subscript (y.12) list the dependent variable y for which the prediction is being made with respect to the values of two independent variables coded as 1 and 2. The denominator of this equation indicates that in a multiple regression with 2 independent variables the standard error has n - (2 + 1) = n - 3 degrees of freedom (number of unrestricted chances for variation in the measurement being made). This occurs because the degrees of freedom is reduced from n to 2 + 1 = 3 numerical constants a, b_1 and b_2 that have all been estimated from the sample.

The standard error of estimate y (say x_1) on x_2 and x_3 is defined as

$S_{1.23} = \sqrt{\Sigma \equiv [(x - x^\wedge]}^2)/n - 3.$

An alternative method of computing $S_{1.23}$ in terms of correlation coefficient r_{12} r_{13} and r_{23} is

$S_{1.23} = s_1 \sqrt{1 -}\ r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12}\, r_{13}\, r_{23} /1 - r_{23}^2$

By summary, we may write

$S_{2.13} = s_2 \sqrt{1 -}\ r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12}\, r_{13}\, r_{23} /1 - r_{13}^2$

$S_{3.12} = s_3 \sqrt{1 -}\ r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12}\, r_{13}\, r_{23} /1 - r_{12}^2$

2.19 COEFFICIENT OF MULTIPLE DETERMINATION

The coefficient of determination in multiple regression denoted $R^2_{y.12}$ or $R^2_{1.23}$ represents the proportion (fraction) of the total variation in the multiple values of dependent variable y, accounted for or explained by the independent variables in the multiple regression model. The value of R^2 varies from zero to one.
$R^2_{y.12}$ = SSR/SST = SST - SSE/SST = 1 - SSE/SST = 1 - $S^2_{y.12}/S^2y$.
Adjusted R^2: The value of SST (= S^2y) remains the same even if additional independent variables are added to the regression model because it represents the sum of squares of the dependent variable. However, additional independent variables are likely to increase value of SSR, so value of R^2 is also likely to increase for additional independent variables.

In particular, the coefficient of multiple determination as a measure of the proportion of total variation in the dependent variable x_1 which is explained by the combined influence of the variations in the independent variables x_2 and x_3 can be defined as
$R^2_{1.23} = 1 - S^2_{1.23}/ S^2_1$

where S^2_1 is the variance of the dependent variable x_1.

By summary, we may write
$R^2_{2.13} = 1 - S^2_{2.13}/ S^2_2$ and $R^2_{3.13} = 1 - S^2_{3.12}/ S^2_{3.}$

An adjusted (or corrected) coefficient of determination is defined as
Adjusted R^2_a = 1 - SSE/(n - k - 1)/SST/(n -1) = 1 - (1 - R^2) n - 1/n - (k + 1).

Since MSE = SSE/(n - k -1) , adjusted R^2_a may be considered as the mixture of the two measures of the performance of a regression model: R^2 and MSE. The decision to add an additional independent variable in a regression model should weigh the increase in R^2 against the loss of one degree of freedom for error resulting from the addition of the variable.

2.20 EXERCISES

Questions

1. a. Explain the concept of regression and point out its usefulness in dealing with business problems

Distinguish between correlation and regression. Also point out the properties of regression coefficients.

2. Explain the concept of regression and point out its importance in business forecasting.

3.. Under what conditions there can be a regression line? Explain.

4. What are the purpose and meaning of the error terms in regression?

5. Give examples of business situations where you believe a straight line relationship exists between two variables. What would be the uses of regression model in each of these situations?

Solve the following:

1. Given the following bivariate data:

x:	-1	5	3	2	1	1	7	3
y:	-6	1	0	0	1	2	1	5

a. Fit a regression line of y on x and predict y if x = 10.
b. Fit a regression line of x on y and predict x if y = 2.5.

2. The coefficient of correlation between the ages of husbands and wives in a community was found to be +0.8, the average of husbands age was 25 years and that of wives age 22 years. Their standard deviations were 4 and 5 years respectively. Find with the help of regression equations:

a. the expected age of husband when wife's age is 16 years and

b. the expected age of wife when husband's age is 33 years.

3. The following table gives the age of cars of a certain make and their annual maintenance costs. Obtain the regression equation for costs related to age.

Age of cars: (in Years)	2	4	6	8
Maintenance Costs (Rs. in 100's):	10	20	25	30

4. The following zero order correlation coefficient are given:
r_{12} = 0.98, r_{13} = 0.44, and r_{23} = 0.54.
Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independents.

5. For a given set of values x_1, x_2 and x_3, the computer has found that r_{12} = 0.96, r_{13} = 0.36, and r_{23} = 0.78. Examine whether these computations may be said to be free from errors.

❖❖❖❖

Module - III

3

FORECASTING TECHNIQUES

Unit Structure:

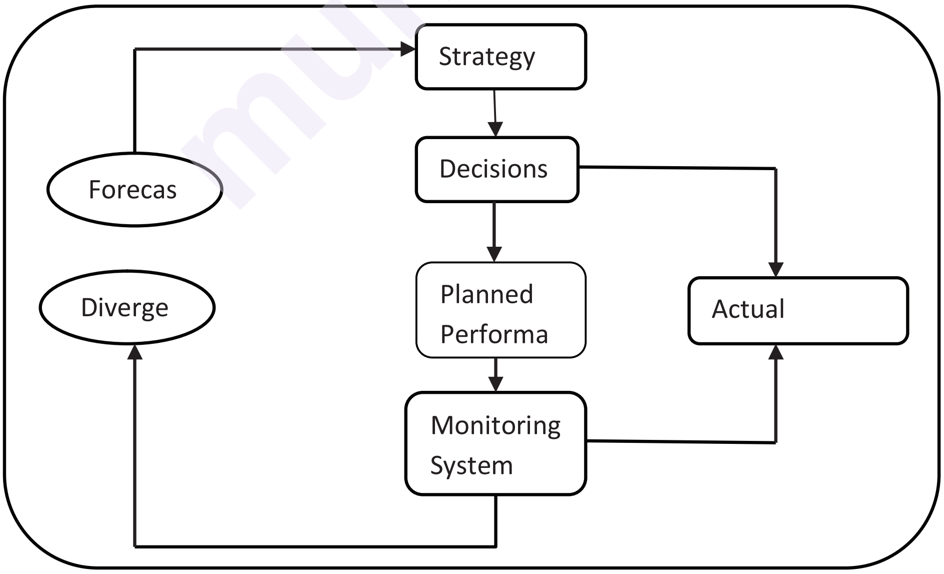
3.1 INTRODUCTION

The increasing complexity of the business environment together with changing demands and expectations are compelling every organization to understand consequences of its decisions (actions or strategies) on the future businesses or services. The knowledge of forecasting methods is essential for decision makers to make reliable and accurate estimates and assess or evaluate the future consequences of decisions in the face of uncertainty.

A flow chart of forecasts and the decision-making process is shown in figure 1. In general, the decisions are influenced by the chosen strategy with regard to an organisation's future priorities and activities. Once decisions are taken, the consequences are measured in terms of expectation to achieve the desired products/services levels.

Decisions also get influenced by the additional information obtained from the forecasting method used. Such information and the accuracy of the forecasts may also affect the strategy formulation of an organization. Thus, an organization needs to establish a monitoring system to compare planned performance with the actual. Divergence, if any, and no matter what is the cause of such divergence, should be fed back into the forecasting process, to generate new forecasts. A few objectives of forecasting are as follows:

- The creation of plans of action because it is not possible to evolve a system of business control without an acceptable system of forecasting.
- Monitoring of the progress of action plans continuously based on forecasts.
- Developing a warning system of the critical factors because they might have drastically after the performance of the plan.



3.2 TYPES OF FORECASTS

The objectives of any organization are facilitated by a number of different types of forecasts. These may be related to

cash flows, operating budgets, personnel requirement, inventory levels, and so on. However, a broad classification of the types of forecasts is as follows:
Demand Forecasts: These are concerned with predictions of demand for products and/or services based on sales and marketing information. These forecasts facilitate in formulating material and capacity plans, and also serves as inputs to financial, marketing and manpower planning.

Environmental Forecasts: These are concerned with the social, political and economic environments of the state and/or the country. Economic forecasts are helpful in predicting inflation rates, money supplies, operating budget and so on.

Technological Forecasts: These are concerned with new developments in existing technologies. The technological forecast is important for technologically advanced companies dealing with computers, aerospace, nuclear and so on.

3.3 TIMING OF FORECASTS

Forecasts are usually classified according to time period and use. The three broad categories of forecasts are as follows:

Short-range Forecast: The short-range forecast has a time-span of up to one year but usually is less than three months. It is normally used for job scheduling, work force levels, job assignments and production levels.

Medium-range Forecast: The medium-range forecast has a time-span from one to three years. It is normally used for sales planning, production planning, budgeting and so on.

Long-range Forecast: The long-range forecast has a time-span of three or more years. It is used for designing and installing new plants, facility location, capital expenditures, research and development and so on.

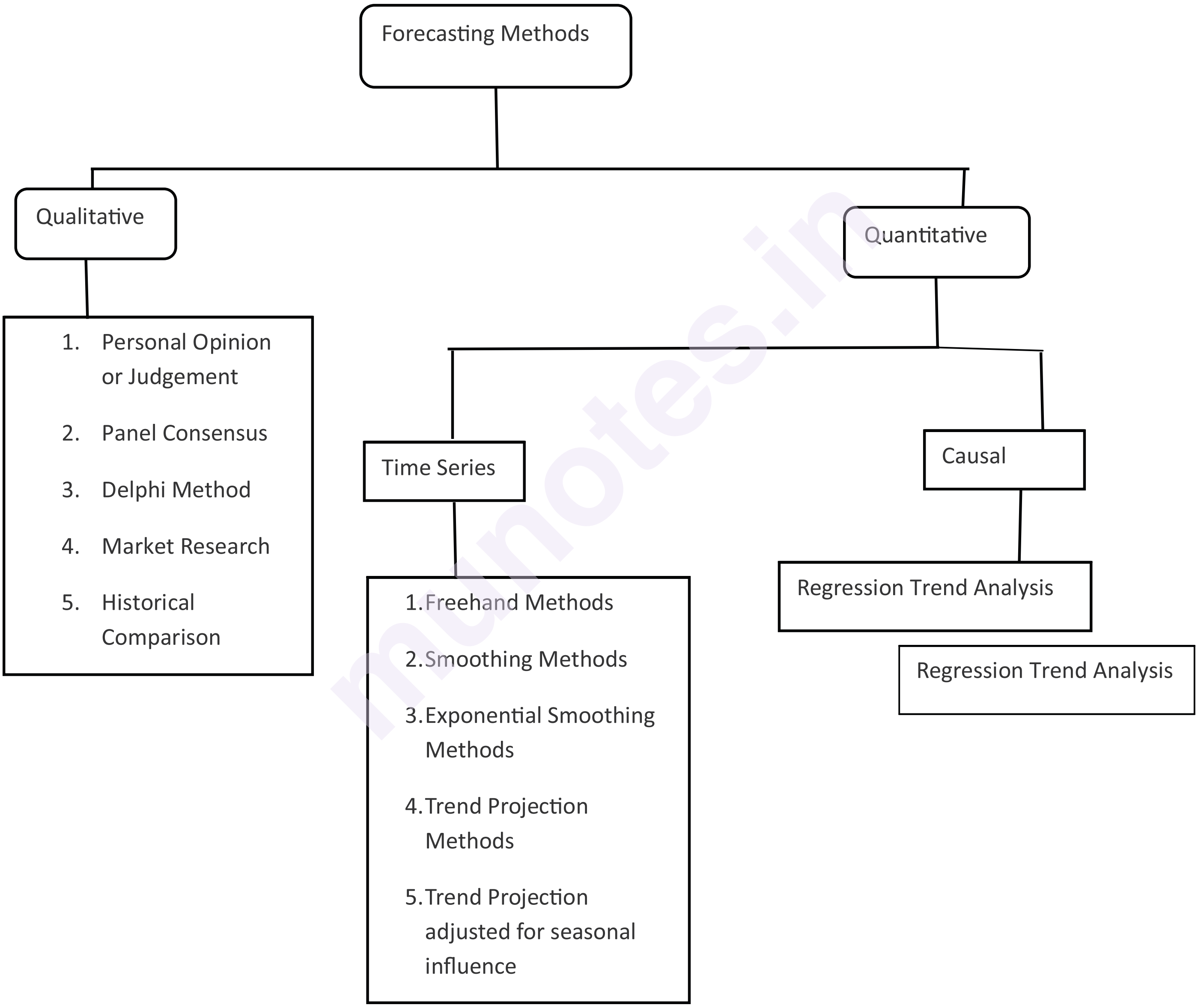
The medium and long-range forecasts differ from short-range forecast on account of following three factors:

i. Medium and long-range forecasts deal with more comprehensive issues and support decisions regarding design and development of new products, plants and processes.
ii. Mathematical techniques such as moving averages, exponential smoothing and trend extrapolation are used for short-range forecasts.
iii. The short-range forecasts tend to be more accurate than long-range forecasts. For example, sales forecasts need to be

updated regularly in order to maintain proper inventory level of products. After each sales period, the forecast should be reviewed and revised.

3.4 FORECASTING METHODS

Forecasting methods may be classified as either quantitative or qualitative (opinion or judgmental) as shown in figure 2.


Forecasting Methods
Qualitative
Quantitative
1. Personal Opinion or Judgement
2. Panel Consensus
3. Delphi Method
4. Market Research
5. Historical Comparison
Time Series
Causal
1. Freehand Methods
2. Smoothing Methods
3. Exponential Smoothing Methods
4. Trend Projection Methods
5. Trend Projection adjusted for seasonal influence
Regression Trend Analysis
Regression Trend Analysis

3.5 QUANTITATIVE FORECASTING METHODS

These methods are applied when

i. Past data about the variable being forecast is available.
ii. Information can be quantified and
iii. Pattern of the past will continue into the future.

The quantitative methods of forecasting are further classified into two categories:

1. **Time-series Forecasting Methods:** A time-series is a set of measurements of a variable that changes through time. The time variable fluctuates uniformly in the same direction from past to future rather than arbitrarily. Thus, there is a freedom to choose the time periods at which observations can be made. The time-series data are gathered on a variable characteristic over a period of time at regular intervals.
   The time-series methods attempt to predict the outcome for a future time period by analyzing patterns, cycles or trends over a period of time.

2. **Causal Forecasting Methods:** The causal forecasting methods are based on the assumptions that the variable value to be forecasted has a cause-effect relationship with one or more other variables. A linear regression analysis is one of the causal forecasting methods.

3.6 QUALITATIVE FORECASTING METHODS

The qualitative forecasting methods are used for collecting opinions and judgements of individuals who are expected to have the best knowledge of current activities or future plans of the organization. For example, marketing professionals through regular contact with customers are presumably familiar with retail market segment, trends by product line, demand trend and so on.

In qualitative forecasting methods, decision makers can incorporate subjective experience as inputs along with objective data. Since each human being has different knowledge, experience and perspective of reality, intuitive forecasts are likely to differ from one individual to another individual. The quantification of data gives decision makers a more precise meaning than words which are inexact and capable of being misunderstood. The following are few qualitative forecasting methods:

i. Personal Opinion: In this approach, an individual does some forecast about a variable of interest based on his/her own judgment or opinion without using a formal quantitative model. Such forecast can be relatively reliable and accurate.

This approach is usually recommended when conditions in the present are not likely to hold in the future. For example, an assessment whether inventory levels are likely to last until the next replenishment a machine will require repair in the next month and so on.

ii. Panel Consensus: In this approach, it is possible to develop consensus among group of individuals to reduce the prejudices and ignorance that may arise in the individual judgment. Such a panel of individuals is encouraged to share information, opinions and assumptions (if any) to predict the future value of a variable of interest.

The disadvantage of this method is that it is dependent on group dynamics and frequently requires a facilitator or convener to coordinate the process of developing a consensus.

iii. Delphi Method: In this method, a panel of experts uses the collective experience and judgment. The panel members may be located in different places, never meet and do not know each other. Each member is given a questionnaire to complete relating to the area under investigation. Based on the responses in questionnaire form from members, a summary is prepared and a copy of it is sent to each member for revision of responses, if any, based on the summary report. This process of updating the summary report is repeated until the desirable consensus is reached among members. This method produces a narrow range of forecasts rather than a single view of the future.

iv. Market Research: This method is used to collect data based on well-defined objectives and assumptions about the future value of a variable. For market research, a questionnaire related to the subject of interest is distributed among respondents. A summary report based on the responses in questionnaire form from respondents is prepared to develop survey results.

v. Historical Comparison: In this method, the data are arranged chronologically and the time-series approach is used to facilitate comparison between one time period and to the next. It provides a basis for making comparisons by isolating the effects of various influencing factors on the patterns of variable values.

3.7 STEPS OF FORECASTING

The following are the general steps to present a systematic procedure of initiating, designing and implementing a forecasting system:

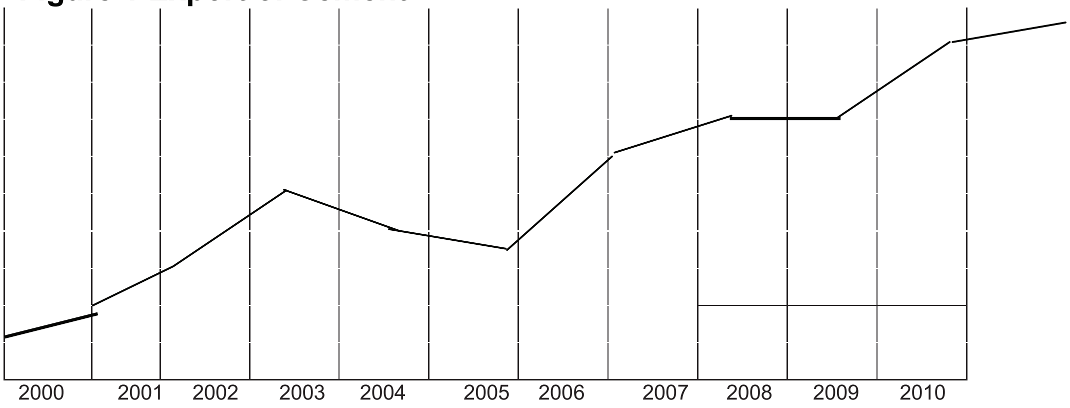
1. Define organisation's objectives of forecasting in order to make use of the best available information to guide future activities and policies to be achieved.
2. Select the variables to be forecasted such as capital investment, employment level, inventory level and purchase of new equipment.
3. Determine the time horizon - short, medium or long term - of the forecast in order to predict changes which may follow the present level of activities.
4. Select an appropriate forecasting method to make projections of the future keeping in view the reasons of changes in the past.
5. Collect the relevant data required for forecasting.
6. Make the forecast and implement its results.

3.8 TIME-SERIES ANALYSIS

A time-series is a set of numerical values of some variable obtained at regular period over time. These numerical values are usually tabulated or graphed to understand the behaviour of the variable. Figure 1 presents the export of cement (in tons) by a cement company between 2000 and 2010. The graph suggests that the series is time dependent. Through such a graph, the management of the company may determine time dependence of the series and develop a procedure to predict the future levels with some degree of reliability. The nature of the time dependence is often analysed by decomposing the time-series into its components.

Year	Export (Tonnes)
2000	2
2001	3
2002	6
2003	10
2004	8
2005	7
2006	12
2007	14
2008	14
2009	18
2010	19

Figure 1 Export of Cement



Year
X-Axis = Tonnes of Cement (4, 8, 12,16, 20)

3.9 OBJECTIVES OF TIME-SERIES ANALYSIS

1. In time-series analysis, it is assumed that the various factors which have already influenced the patterns of change in the value of the variable under study will continue to do so almost in the same manner in future also. Thus, one of the objectives of time-series analysis is to identify the patterns and isolate the influencing factors (or effects) for prediction, planning and control of future values of the variable.

2. The review and evaluation of progress in any phenomenon are made based on time-series data. For example, evaluation of the policy of controlling inflation and price rise is done based on various price indices that are based on the analysis of time-series.

3.10 TIME-SERIES PATTERNS

In time-series, it is assumed that the data consist of a patter along with random fluctuations. This may be expressed in the following form:

Actual value of the variable at time t = Mean value of the variable at time t + Random deviation
from mean value of
the variable at time t

$$\hat{y} = \text{pattern} + e$$

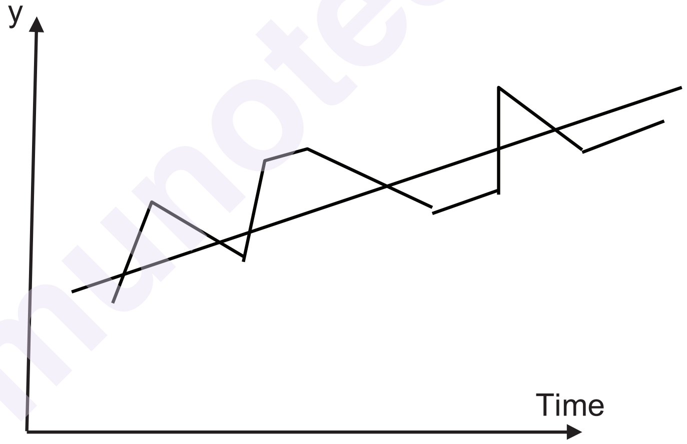
where $\hat{y}$ is the forecast variable at period t, pattern is the mean value of the forecast variable at period t, and e is the random fluctuation from the pattern that occurs of the forecast variable at period t.

3.11 COMPONENTS OF A TIME-SERIES

The time-series data contain four components trend, cyclicality, seasonality and irregularity. Not all time-series have all these components. Figure 2 shows the effects of these time-series components over a period of time.

TREND: Sometimes a time-series displays a steady tendency of either upward or downward movement in the average (or mean) value of the forecast variable y over time. Such a tendency is called a trend. When observations are plotted against time, a straight line describes the increase or decrease in the time-series over a period of time.

CYCLES: Upward and downward movements in the variable value about the trend time over a time period are called cycles. A business cycle may vary in length, usually more than a year but less than 5 to 7 years. The movement is through four phases from peak (prosperity) to contradiction (recession) to trough (depression) to expansion (recovery or growth) as shown in figure 2.



SEASONAL: It is a special case of a cycle component of time-series in which fluctuations are repeated usually within a year (e.g. daily, weekly, monthly and quarterly) with a high degree of regularity. For example, average sales for a retail store may increase greatly during festival seasons.

IRREGULAR: Variations are rapid charges or bleeps in the data caused by short-term unanticipated and non-recurring factors. Irregular fluctuations can happen as often as day to day.

3.12 TIME-SERIES DECOMPOSITION MODELS

The analysis of time-series consists of two major steps:

i. Identifying the various factors or influences which produce the variations in the time-series.

ii. Isolating, analysing and measuring the effect of these factors independently holding other things constant.

The purpose of decomposition is to break a time-series into its components: trend (T), cyclical (C), seasonality (S) and irregularity (I). Such decomposition helps to isolate influence of each of the four components on the actual series and becomes the basis for forecasting. Two commonly used models for decomposition of a time-series are discussed below:

3.13 MULTIPLICATIVE MODEL

The actual values of a time-series, Y can be found by multiplying its four components at a particular time period. The effect of four components on the time-series is interdependent. The multiplicative time-series model is defined as

$Y = T X C X S X I$.

This model is useful in situations where the effect of C, S and I is measured in relative sense instead of absolute sense. The geometric mean of C, S and I is assumed to be less than one. For example, suppose sales of a product for a period of 20 months is Y_{20} = 423.36. Decomposing sales into its components are trend component (mean sales) 400; effect of current cycle (0.90) which decreases sales by 10 per cent, and seasonality of the series (1.20) that increases sales by 20 per cent. In the absence of random fluctuation, the expected value of sales for the given period is 400 X 0.90 X 1.20 = 432. If the random factor decreases sales by 2 per cent in this period, then the actual sales volume will be 432 X 0.98 = 423.36.

3.14 ADDITIVE MODEL

In this model, it is assumed that the effect of various components on a time-series can be estimated by adding these components. The additive time-series model is defined as

$Y = C + S + T + I$

where C, S, and I are absolute quantities and can have positive or negative values. It is assumed that these four components are independent of each other.

3.15 QUANTITATIVE FORECASTING METHODS

The quantitative forecasting methods are classified into two general categories:

1. Time-Series Methods: This method takes into consideration an observed historical pattern for any variable and projects the same into the future using a mathematical formula. These methods do not suggest that a variable will take some future value.

2. Causal methods: In this method, regression analysis and correlation analysis identify factors that influence or cause variation in the value of any variable in some predictable manner.

3.16 FREEHAND METHOD

A freehand curve drawn through the data values is an easy and adequate representation of time-series. In figure 1, a straight line connecting the 1997 and 2004 exports volumes is fairly a good representation of the given data.

The forecast is done by extending the trend line. A trend line fitted by the freehand method should confirm to the following conditions:

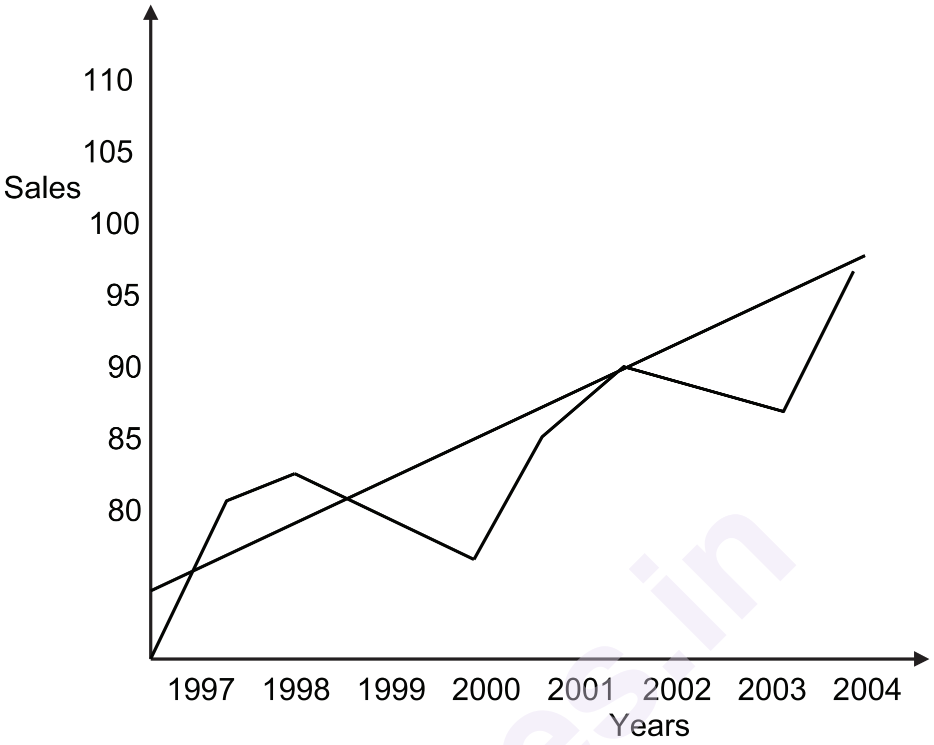
i. The trend line should be smooth.
ii. The sum of the vertical deviations of the observations above the trend line should equal the sum of the vertical deviations of the observations below the trend line.
iii. The sum of squares of the vertical deviations of the observations from the trend line should be as small as possible.
iv. The trend line should bisect the cycles so that area above and below the trend line for each full cycle.

EXAMPLE 1:
Fit a trend line to the following data by using the freehand method:

Year:	1997	1998	1999	2000	2001	2002	2003	2004
Sales turnover: (in Rs. Lakh)	80	90	92	83	94	99	92	104

Solution: Freehand graph of sales turnover (Rs. in Lakhs) from 1997 to 2004 is shown in figure 3. Forecast can be obtained simply by extending the trend line.

Figure 3: Graph of Sales of Turnover



3.17 LIMITATIONS OF FREEHAND METHOD

i. This method is highly subjective because the trend line depends on personal judgment and therefore good-fit for one individual may not be same for another.

ii. The trend line used as a basis for predictions cannot have much value.

iii. The construction of a freehand trend is very time-consuming provided a careful job is to be done.

3.18 SMOOTHING METHODS

The smoothing methods provide pattern of movement in the data over time by eliminating random variation due to irregular components of the time-series. The following three smoothing methods are discussed in this section:

1. Moving Averages
2. Weighted Moving Averages
3. Semi-averages

MOVING AVERAGES: To observe the movement of some variable values over a period of time and then to project this movement into the future, it is essential to smooth out first the irregular pattern in the historical values of the variable, and later use this as the basis for a future projection. This can be done by using the method of **moving averages.**

This is a subjective method and depends on the length of the period for calculating moving averages. To remove the effect of irregular pattern, the period of chosen for calculating moving averages should be an integer value. Preferably the period chosen should be a multiple of the estimated average length of cycle in the series.

The moving averages as an estimator of the value of a variable in the next period given a period of length *n* is expressed as

Moving average, MA_{i+1}= $(\Sigma\{Di + Di - 1 + Di - 2 + \cdots ... + Di - n + 1)/n$

where i is the current time period; D is the actual data which is exchanged each period and n is the length of time period.

In this method, the term 'moving' is used because each time an average is computed by summing the values from a given number of periods through deleting the oldest value and adding a new value.

ADVANTAGE The major advantage of moving average method is the opportunity to focus on the long-term trend (and cyclical) movements in a time-series without the obscuring effect of short-term influences.

Limitations

i. As the length of period chosen for computing the averages increases, it smoothens the variations better but it also makes the method less sensitive to real changes in the data.

ii. It is difficult to choose the optimal length of time for computing the moving average but it cannot be computed for the first (n-1)/2 years or the last (n-1)/2 year of the series.

iii. Moving averages cannot predict trends very well because averages always stay within past levels and do not predict a change to either a higher or a lower level.

iv. Moving averages do not usually adjust for such time-series effects as trend, cycle or seasonality.

EXAMPLE 2

Use following data to compute a three-year moving average for all available years. Also determine the trend and short-term error.

Year	Production (in '000 tonnes)	Year	Production (in '000 tonnes)
2000	21	2005	22
2001	22	2006	25
2002	23	2007	26
2003	25	2008	27
2004	24	2009	26

Solution: Computing moving average for the first 3 years as follows:

Moving average = $\frac{21 + 22 + 23}{3}$ = 22.

This average value of first three-year can be used to forecast the production volume in fourth year, 2008. Since 25, 000 tonnes production was made in 2008; therefore, error of the forecast becomes 25, 000 - 22, 000 = 3000 tonnes.

Similarly, moving average for the year 2006 to 2008 is

Moving average = $\frac{22 + 23 + 25}{3}$ = 23.33

Calculations of three-year moving average are shown in Table 1.

Year	Production y	3-year moving total	3-yearly moving Average (Trend Values) $\hat{y}$		Forecast Error (y-$\hat{y}$)
2000	21	-	-		-
2001	22	(21 + 22 + 23) = 66	66/3 22.00	=	0
2002	23	(22 + 23 + 25) = 70	70/3 23.33	=	-0.33
2003	25	(23 + 25 + 24) = 72	72/3 24.00	=	1.00
2004	24	71	23.67		
2005	22	71	23.67		
2006	25	73	24.33		
2007	26	78	26.00		
2008	27	(26 + 27 + 26) = 79	79/3 26.33	=	0.67
2009	26	-	-		-

3.19 ODD ANS EVEN NUMBERS OF YEARS

If period of length n is an odd number, then moving average period is centered on middle period in the consecutive sequence of n periods. For example, if n = 5, then the first five year moving average, MA_3 (5) is centered on the third year, MA_4 (5) is centered on the fourth year and MA_9 (5) is centered on the ninth year.

Since moving average cannot be obtained for the first (n-1)/2 years or the last (n-1)/2 year of the series; therefore for a 5-year moving average, it cannot be computed for the just two years or the last two years of the series.

If period of length n is an even number, then moving average can be computed by taking an average of each part. For example, if n = 4, then the first four year moving average, MA_3 (4) is an average of the first four year data, and the second moving average MA_4 (4) is the average of data values 2 through 5.

3.20 WEIGHTED MOVING AVERAGES

A moving average where some time periods are weighted differently than others is called a weighted moving average. Choice of weights is arbitrary because there is no set rule to do so. Generally, the most recent observation is assigned larger weightage and the weightage decreases for older data values. A weighted average is computed as:

Weighted Moving average =

$$\frac{\Sigma(Weight\ for\ period\ n)(Data\ value\ in\ period\ n)}{\Sigma\ weights}$$

EXAMPLE 1: Vacuum cleaner sales for 12 months are given below. The owner of the supermarket decides to forecast sales by weighing the past three months as follows:

Weight Applied	Month
3	
2	
1	

6	

Months:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Actual Sales: (in units)	**10**	**12**	**13**	**16**	**19**	**23**	**26**	**30**	**28**	**18**	**16**	**14**

Solution: The results of 3-month weighted average are shown in Table 2.

$\bar{x}_{\text{weighted}} = 3M_{t-1} + 2M_{t-2} + 1M_{t-3}$

= 1/6 [3 X sales last month + 2 x Sales two months ago + 1 x Sales three months ago)

TABLE 2 Weighted Moving Average

Month	Actual Sales	Three-month weighted moving average
1	10	----
2	12	----
3	13	----
4	16	[(13 x 13) + (2 x 12) + (1 x 10)] = 121/6
5	19	[(3 x 16) + (2 x 13) + (1 x 12)] = 141/3
6	23	[(3 x 19) + (2 x 19) + (1 x 13)] = 17
7	26	[(3 x 23) + (2 x 19) + (1 x 16)] = 201/2
8	30	[(3 x 26) + (2 x 23) + (1 x 19)] = 235/6
9	28	[(3 x 30) + (2 x 26) + (1 x 23)] = 271/2
10	18	[(3 x 28) + (2 x 30) + (1 x 26)] = 289/3
11	16	[(3 x 18) + (2x 28) + (1 x 30)] = 231/3
12	14	[(3 x 16) + (2 x 18) + (1 x 28)] = 182/3

3.21 SEMI-AVERAGE METHOD

The semi average method is used to estimate the slope and intercept of the trend line provided time-series is represented by a linear function. In this method, the data are divided into two parts and their respective arithmetic means are computed. The two arithmetic mean points are plotted corresponding to the midpoint of the class interval covered by the respective part and ten these points are joined by a straight line to get the required trend line. The arithmetic mean of the first part is the intercept value, and the slope (change per unit time) is determined by the ratio of the difference in the arithmetic means of the number of years between them to get a time-series of the form $\hat{y}$ = a + bx. The $\hat{y}$ equation should always be stated with reference to the year where x = 0 and a description of the units of x and y.

3.22 EXPONENTIAL SMOOTHING METHOD

Exponential smoothing method compares past data from previous time periods with exponentially decreasing importance in the forecast so that the most recent data carries more weight in the moving average. Simple exponential smoothing makes no explicit adjustment for trend effects whereas adjusted exponential smoothing does take trend effects into account.

3.23 SIMPLE EXPONENTIAL SMOOTHING

Simple exponential smoothing method is helpful in forecasting the value for the present time period X_t multiplied by an exponential smoothing constant α(not the same as used for Type I error) falling between 0 and 1 plus the product of the present time period forecast F_t and (1 - α). Mathematically, it is written as

$F_{t-1} = \alpha X_t + (1 - \alpha) F_t = F_t + \alpha (X_t - F_t)$

where F_{t+1} is the forecast for the next period (t + 1); F_t is the forecast for the present time period (t); α is a weight called exponentially smoothing constant ($0 \leq \alpha \leq 1$) and X_t is the actual value for the present time period (t).

If exponential smoothing is used over long period of time, then forecast for F_t is expressed as

$F_t = \alpha X_{t-1} + (1 - \alpha) F_{t-1}.$

More weight is given to past data when smoothing constant α is low, and more weight is given to recent data when it is high. For example, if α = 0.9 then 99.99 per cent of the forecast value is determined by the four most recent periods. But if α = 0.1, then only 34.30 per cent of the forecast value is determined by these last 4 periods and the smoothing effect is equivalent to a 19-period arithmetic moving average.

If α = 1, then each forecast would reflect total adjustments to the recent period value and the forecast would simply be last period's actual value, i.e., $F_t = 1.0D_{t-1}$. Since fluctuations are random, the value of α is generally kept in the range of 0.005 to 0.30 in order to 'smooth' the forecast.

The following table illustrates forecast value in different time periods. For example, when α = 0.5, the new forecast is based on the values in the last three or four periods. When α = 0.1, a very small weight is given to recent values and takes a 19-period arithmetic moving average.

Weight Assigned to Smoothing Constant	Most recent period (α)	2nd Most recent Period α (1 - α)	3rd most recent period α (1 - α)^2	4th most recent period α (1 - α)^3	5th most recent period A (1 - α)^4
α = 0.1	0.1	0.09	0.081	0.073	0.066
α = 0.5	0.5	0.25	0.125	0.063	0.031

3.24 SELECTING SMOOTHING CONSTANT

To get an accurate forecast, it is important to assign an appropriate value to the exponential smoothing constant, α.
The most appropriate value of α which is equal to an arithmetic moving average, in terms of degree of smoothing, can be estimated as α = 2/(n + 1). Lower the difference between forecasted values and the actual or observed values, the accuracy of a forecasting model is judged more.

Error: The error of an individual forecasting model is defined as
Forecast error = Actual values - Forecasted Values
$e_t = X_t - F_t$

The mean absolute deviation (MAD) method is used to measure the forecast error for a forecasting model. The MAD is computed by taking the sum of the absolute values of the individual forecast errors and then dividing by number of periods n as follows:

$$\text{MAD} = \frac{\Sigma|\textit{Forecast Errors}|}{n}$$

where standard deviation σ= 1.25 MAD.

The exponential smoothing method facilitates continuous updating of the estimate of MAD. The current MADt is given by
MADt = α $|\textit{Actual Values} - \textit{Forecasted Values}|$ + (1 - α)$\text{MAD}_{t-1.}$
Higher values of smoothing constant, α make the current MAD more responsive to current forecast errors.

EXAMPLE 1: A firm uses simple exponential smoothing with α = 0.1 to forecast demand. The forecast for the week of February 01 was 500 units whereas actual demand turned out to be 450 units.
a. Forecast the demand for the week of February 08.
b. Assume the actual demand during the week of February 08 turned out to be 505 units. Forecast the demand for the week of February 15. Continue forecasting through March 15, assuming the subsequent demands were actually 516, 488, 467, 554 and 510 units.

Solution: Given $F_{t-1} = 500$, $D_{t-1} = 450$ and α = 0.1

a. $F_t = F_{t-1} + \alpha (D_{t-1} - F_{t-1}) = 500 + 0.1 (450 - 500) = 495$ units.

b. Forecast of demand for the week of February 15 is shown in table 2.

TABLE 2

Week	**Demand** **D_{t-1}**	**Old Forecast** **F_{t-1}**	**Forecast Error** **$(D_{t-1} - F_{t-1})$**	**Correction** **α $(D_{t-1} - F_{t-1})$**	**New Forecast** **F_{t-1} + α $(D_{t-1} - F_{t-1})$**
Feb 1	450	500	-50	-5	495
8	505	495	10	1	496
15	516	496	20	2	498
22	488	498	-10	-1	497
Mar 1	467	497	-30	-3	404
8	554	494	60	6	500
15	510	500	10	1	501

If no previous forecast value is known, the old forecast starting point may be estimated or taken to be an average of some preceding periods.

3.25 EXERCISES

1. What is forecasting? Discuss in brief the various theories and methods of business forecasting.
2. Briefly describe the steps that are used to develop a forecasting system.
3. For what purpose do we apply time-series analysis to data collected over a period of time?
4. What is the difference between a causal model and a time-series model?
5. How can one benefit from determining past patterns?

SOLVE THE FOLLOWING

Q1. The owner of a small company manufactures a product. Since he started the company, the number of units of the product he has sold is represented by the following time series:

Year	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Units Sold	100	120	95	105	108	102	112

Find the trend line that describes the trend by using the method of semi-averages.

2. Fit a trend line to the following data by the freehand method:

Year	Production of Steel (miilion tonnes)	Year	Production of Steel (million tonnes)
2000	20	2005	25
2001	22	2006	23
2002	24	2007	26
2003	21	2008	25
2004	23		

3. The owner of small manufacturing company has been concerned about the increase in manufacturing costs over the past 10 years, The following data provide a time-series of the cost per unit for the company's leading product over the past 10 years.

Year	Cost per Unit	Year	Cost per Unit
2000	332	2005	405
2001	317	2006	410
2002	357	2007	427
2003	392	2008	405
2004	402	2009	438

Calculate a 5-year moving average for the unit cost of the product.

4. In January, a city hotel predicted a February demand for 142-room occupancy. Actual February demand was 153 rooms. Using a smoothing constant of $\alpha = 0.20$, forecast the March demand using the exponential smoothing model.

5.A shoe manufacturer, using exponential smoothing with $\alpha = 0.1$, has developed a January trend forecast of 400 units for a ladies' shoe. This brand has seasonal indexes of 0.80, 0.90 and 1.20 respectively for the first three months of the year. Assuming actual sales were 344 units in January and 414 units in February, what would be the seasonalised March forecast?

❖❖❖❖

Module IV

DECISION THEORY AND DECISION TREES

4

DECISION MAKING UNDER CONDITIONS OF CERTAINTY AND UNCERTAINTY

Decision making under conditions of Certainty, Decision Making under conditions of Uncertainty

Unit Structure :

4.1 INTRODUCTION

In this Unit-IV - Chapter 4.1, we shall discuss the concepts of decision, condition of certainty, uncertainty and decision making under these conditions

4.2 OBJECTIVES

At the end of this unit the learners will be able to understand

- What is a decision?
- What are conditions of certainty?
- What are conditions of uncertainty?
- What is the decision making?

4.3 WHAT IS A DECISION?

Decision of any managerial decision includes any decision regarding the operations of a firm, such as setting target growth rates, hiring or firing employees, and deciding what products to sell.

Decision making is crucial for running a business enterprise which faces a large number of problems requiring decisions.

Which product to be produced, what price to be charged, what quantity of the product to be produced, what and how much advertisement expenditure to be made to promote the sales, how much investment expenditure to be incurred are some of the problems which require decisions to be made by managers.

The five steps involved in managerial decision making process are explained below:

1. Establishing the Objective:

The first step in the decision making process is to establish the objective of the business enterprise. The important objective of a private business enterprise is to maximize profits. However, a business firm may have some other objectives such as maximization of sales or growth of the firm.

But the objective of a public enterprise is normally not of maximization of profits but to follow benefit-cost criterion. According to this criterion, a public enterprise should evaluate all social costs and benefits when making a decision whether to build an airport, a power plant, a steel plant, etc.

2. Defining the Problem:

The second step in decision making process is one of defining or identifying the problem. Defining the nature of the problem is important because decision making is after all meant for solution of the problem. For instance, a cotton textile firm may find that its profits are declining.

It needs to be investigated what are the causes of the problem of decreasing profits; whether it is the wrong pricing policy, bad labour-management relations or the use of outdated technology which is causing the problem of declining profits. Once the source or reason for falling profits has been found, the problem has been identified and defined.

3. Identifying Possible Alternative Solutions (i.e. Alternative Courses of Action):

Once the problem has been identified, the next step is to find out alternative solutions to the problem. This will require considering the variables that have an impact on the problem. In this way, relationship among the variables and with the problems has to be established.

In regard to this, various hypotheses can be developed which will become alternative courses for the solution of the problem. For example, in case of the problem mentioned above, it is identified that the problem of declining profits is due to be use of technologically inefficient and outdated machinery in production.

4. Evaluating Alternative Courses of Action:

The next step in business decision making is to evaluate the alternative courses of action. This requires the collection and analysis of the relevant data. Some data will be available within the various departments of the firm itself, the other may be obtained from the industry and government.

The data and information so obtained can be used to evaluate the outcome or results expected from each possible course of action. Methods such as regression analysis, differential calculus, linear programming, and cost- benefit analysis are used to arrive at the optimal course. The optimum solution will be one that helps to achieve the established objective of the firm. The course of action which is optimum will be actually chosen. It may be further noted that for the choice of an optimal solution to the problem, a manager works under certain constraints.

The constraints may be legal such as laws regarding pollution and disposal of harmful wastes; they may be financial (i.e. limited financial resources); they may relate to the availability of physical infrastructure and raw materials, and they may be technological in nature which set limits to the possible output to be produced per unit of time. The crucial role of a business manager is to determine optimal course of action and he has to make a decision under these constraints.

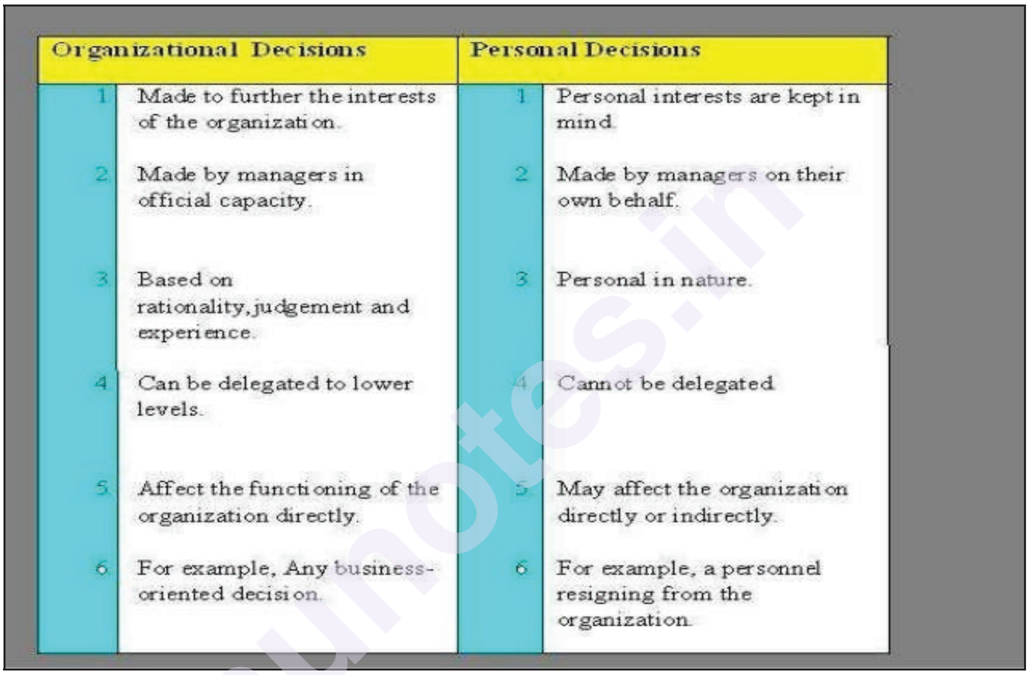
5. Implementing the Decision:

After the alternative courses of action have been evaluated and optimal course of action selected, the final step is to implement the decision. The implementation of the decision requires constant monitoring so that expected results from the optimal course of action are obtained. Thus, if it is found that expected results are not forthcoming due to the wrong implementation of the decision, then corrective measures should be taken.

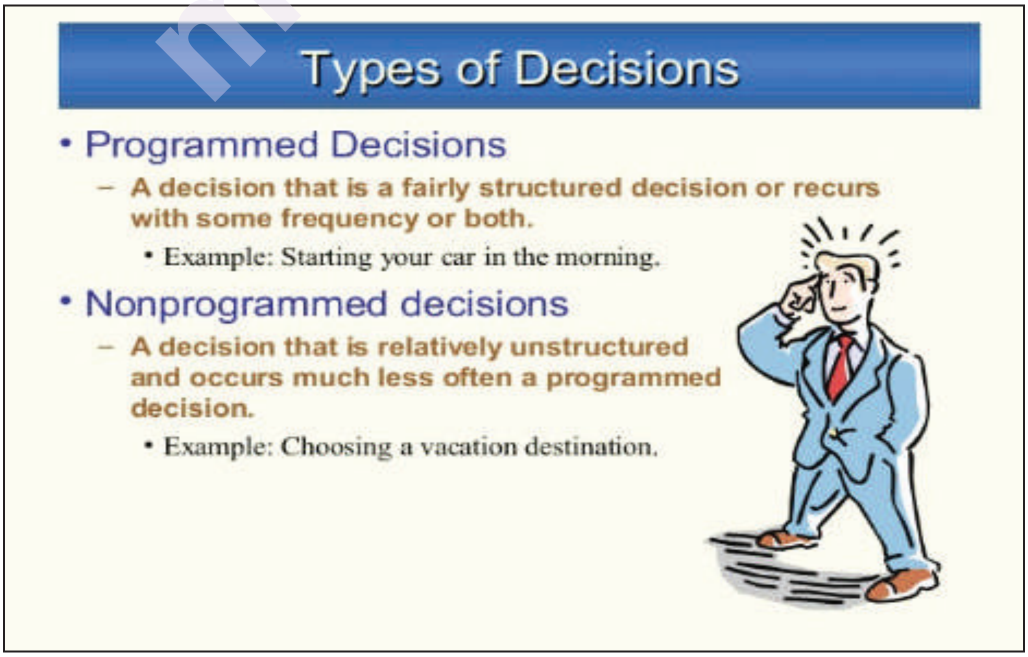
However, it should be noted that once a course of action is implemented to achieve the established objective, changes in it may become necessary from time to time in response in changes in conditions or firm's operating environment on the basis of which decisions were taken.

There are two situations of decision making:

- Programmed decisions are those made in routine, repetitive, well-structured situations through the use of pre-determined decision rules.
- Non-programmed decisions are those for which pre-determined decision rules are impractical because the situations are novel or ill-structured.



Organizational Decisions		Personal Decisions	
1	Made to further the interests of the organization.	1	Personal interests are kept in mind.
2	Made by managers in official capacity.	2	Made by managers on their own behalf.
3	Based on rationality,judgement and experience.	3	Personal in nature.
4	Can be delegated to lower levels.	4	Cannot be delegated.
5	Affect the functioning of the organization directly.	5	May affect the organization directly or indirectly.
6	For example, Any business-oriented decision.	6	For example, a personnel resigning from the organization.



Types of Decisions

- Programmed Decisions
  - A decision that is a fairly structured decision or recurs with some frequency or both.
    - Example: Starting your car in the morning.
- Nonprogrammed decisions
  - A decision that is relatively unstructured and occurs much less often a programmed decision.
    - Example: Choosing a vacation destination.

4.4 DECISION MAKING UNDER CONDITIONS OF CERTAINTY

An outcome defines what will happen if a particular alternative or course of action is chosen. Knowledge of outcomes is important when there are multiple alternatives. In the analysis of decision making, three types of knowledge with respect to outcomes are usually distinguished:

- Certainty: Complete and accurate knowledge of outcome of each alternative. There is only one outcome for each alternative.

- Risk: Multiple possible outcomes of each alternative can be identified and a probability of occurrence can be attached to each.

- Uncertainty: Multiple outcomes for each alternative can be identified but there is no knowledge of the probability to be attached to each.

If the outcomes are known and the values of the outcomes are certain, the task of the decision maker is to compute the optimal alternative or outcome with some optimization criterion in mind.

As an example: if the optimization criterion is least cost and you are considering two different brands of a product, which appear to be equal in value to you, one costing 20% less than the other, then, all other things being equal, you will choose the less expensive brand.

However, decision making under certainty is rare because all other things are rarely equal.

Linear programming is one of the techniques for finding an optimal solution under certainty. Linear programming problems normally need computations with the help of a computer.

In this type of decision making under certainty environment, there is only one type of event that can take place. It is very difficult to find complete certainty in most of the business decisions. However, in much routine type of decisions, almost complete certainty can be noticed. These decisions, generally, are of very little significance to the success of business.

4.5 DECISION MAKING UNDER CONDITIONS OF UNCERTAINTY AND RISK

In the environment of uncertainty, more than one type of event can take place and the decision maker is completely in dark regarding the event that is likely to take place. The decision maker is not in a position, even to assign the probabilities of happening of the events. Such situations generally arise in cases where happening of the event is determined by external factors. For example, demand for the product, moves of competitors, etc. are the factors that involve uncertainty.

Decisions under uncertainty (outcomes known but not the probabilities) must be handled differently because, without probabilities, the optimization criteria cannot be applied.Some estimated probabilities are assigned to the outcomes and the decision making is done as if it is decision making under risk.

Under the condition of risk, there are more than one possible events that can take place. However, the decision maker has adequate information to assign probability to the happening or non-happening of each possible event. Such information is generally based on the past experience.

Virtually, every decision in a modern business enterprise is based on interplay of a number of factors. New tools of analysis of such decision making situations are being developed. These tools include risk analysis, decision trees and preference theory.

The making of decisions under risk, when only the probabilities of various outcomes are known, is similar to certainty.Instead of optimizing the outcomes, the general rule is to optimize the expected outcome.

As an example: if you are faced with a choice between two actions one offering a 1% probability of a gain of $10000 and the other a 50% probability of a gain of $400, you as a rational decision maker will choose the second alternative because it has the higher expected value of $200 as against $100 from the first alternative.

4.6 EXAMPLES

Risk

- Must make a decision for which the outcome is not known with certainty
- Can list all possible outcomes & assign probabilities to the outcomes

Uncertainty

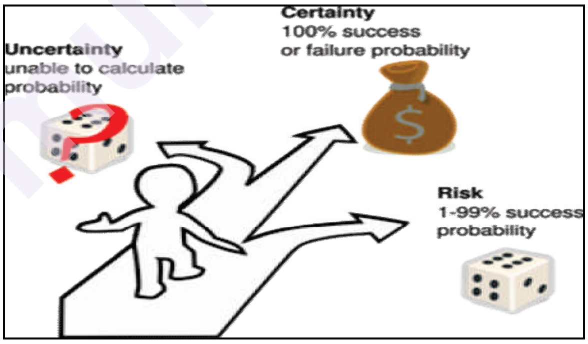
- Cannot list all possible outcomes
- Cannot assign probabilities to the outcomes

Virtually all decisions are made in an environment of at least some uncertainty. However, the degree will vary from relative certainty to great uncertainty. There are certain risks involved in making decisions.

In a situation involving certainty, people are reasonably sure about what will happen when they make a decision. The information is available and is considered to be reliable, and the cause and effect relationships are known.

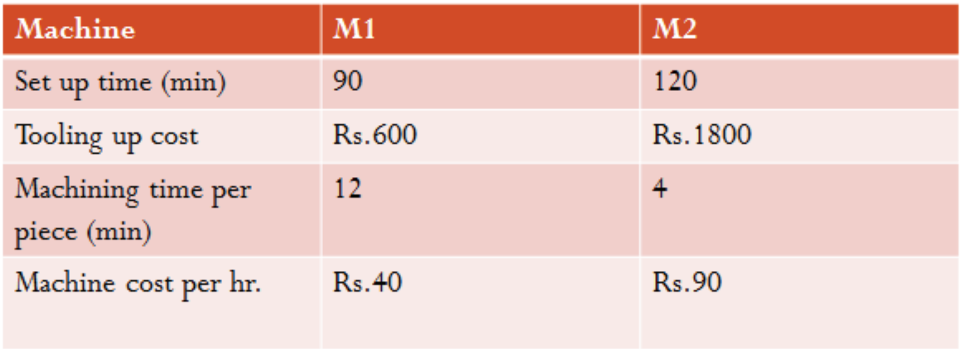
In a situation of uncertainty, on the other hand, people have only a meager data base, they do not know whether or not the data are reliable, and they are very unsure about whether or not situation may change. Moreover, they cannot evaluate the interactions of the different variables. For examples, a corporation that decides to expand its operation in a strange country may know little about the country and its culture, laws, economic environment, and politics. The political situation may be so volatile that even the experts cannot predict a possible change in government.

In a risk situation, factual information may exist, but it may be incomplete. To improve decision making, one may estimate the objective probabilities of an outcome by using, for example, mathematical models. On the other hand, subjective probability, based on judgment and experience, may be used. Fortunately, there are a number of tools available that helps make more effective decisions.



Example of Certain decision :

A manufacturer has two different kinds of machines – M1 and M2. He has received an order which can be processed either on M1 or on M2. The following data is available:



Machine	M1	M2
Set up time (min)	90	120
Tooling up cost	Rs.600	Rs.1800
Machining time per piece (min)	12	4
Machine cost per hr.	Rs.40	Rs.90

Which of the two machines will you choose to do the job if the order quantity is 1000 numbers?

- Step 1 : Find the Total Cost for both machine

Total Cost = Set up Cost + Tooling up Cost + Machining Cost

Set up Cost

for 60 min the cost is 40 Rs.

∴ for 90 min the cost is (?) Rs. → $x = \frac{90 * 40}{60} = 60\ Rs$

Tooling up Cost = 600 Rs.

Machining Cost

for producing 1 piece required time is 12 min.

∴ for producing 1000 piece time required =12000 min

(Same as set up time)

$$\therefore x = \frac{40 * 12000}{60} = 8000\ Rs.$$

∴ Total Cost = Set up Cost + Tooling up Cost + Machining Cost

∴ Total Cost = 60 + 600 + 8000

For machine M1,

∴ Total Cost = 8660 Rs.

Same method is apply for Machine M2 and find the Total Cost For machine M2,

∴ Total Cost = 7980 Rs.

∴ Machine M2 is more suitable machine for doing this job.

4.7 LET US SUM UP

In this unit you have learnt the concept of decision making under certainty, risk and uncertainty.

4.8 EXERCISES

1. Explain the term certainty with your own example.
2. What do you understand by programmed and non-programmed decision? Give Examples.
3. What does the concept risk denotes?
4. What do you understand by uncertainty? Give examples.

4.9 SUGGESTED READINGS

Decision making section in any of the reference / text books

❖❖❖❖

Module - IV
DECISION THEORY AND DECISION TREES

5

DECISION MAKING UNDER CONDITIONS OF CERTAINTY AND UNCERTAINTY

Decision making under conditions of Risk

Unit Structure :

5.1 INTRODUCTION

In this Unit-IV - Chapter 4.2, we shall discuss decision making under conditions of risk.

5.2 OBJECTIVES

At the end of this unit the learners will be able to understand

- What is a risk?
- What is decision making under conditions of risk?

4.2.3 DECISION MAKING UNDER CONDITIONS OF RISK

Definition of Risk :

A risk can be defined as a probability or threat of damage, injury, liability, loss, or any other negative occurrence that is caused by external or internal vulnerabilities, and that may be avoided through preemptive action.

A risk is not an uncertainty (where neither the probability nor the mode of occurrence is known), a peril (cause of loss), or a hazard (something that makes the occurrence of a peril more likely or more severe).

Examples of Risk :

Finance: The probability that an actual return on an investment will be lower than the expected return. Financial risk is divided into the following categories: Basic risk, Capital risk, Country risk, Default risk, Delivery risk, Economic risk, Exchange rate risk, Interest rate risk, Liquidity risk, Operations risk, Payment system risk, Political risk, Refinancing risk, Reinvestment risk, Settlement risk, Sovereign risk, and Under-writing risk.

Food industry: The possibility that due to a certain hazard in food there will be a negative effect to a certain magnitude on the health of a consumer.

Insurance: A situation where the probability of a variable (such as burning down of a building) is known but when a mode of occurrence or the actual value of the occurrence (whether the fire will occur at a particular property) is not.

Securities trading: The probability of a loss or drop in value of the security. Trading risk is divided into two general categories: (1) Systemic risk that affects all securities in the same class and is linked to the overall capital-market system and therefore cannot be eliminated by diversification, also is called as market risk. (2) Nonsystematic risk is any risk that isn't market-related or is not systemic. Also it is called as non-market risk, extra-market risk, or un-systemic risk.

Workplace: it is the consequence and probability of a hazardous event or phenomenon occurring at the workplace. For example, the risk of developing cancer is estimated as the incremental probability of developing cancer over a lifetime as a result of exposure to potential carcinogens (cancer-causing substances).

All things in the world vary - one from another, over time, and with different environments. The variation will occur in the economy due to its emphasis on decision making for the future.

For example, the estimate that cash flow next year will be Rs. 4500/- is one of the certainty. Decision making under certainty is, of course, not present in the real world now and surely not in the future. We can observe outcomes with a high degree of certainty, but even this depends upon the accuracy and precision of the scale or measuring instrument.

To allow a parameter of an economy study to vary implies that *risk,* and possibly *uncertainty,* is introduced.

When there may be two or more observable values for a parameter *and* it is possible to estimate the chance that each value may occur, **risk** is present. Virtually all decision making is performed *under risk.*

As an illustration, decision making under risk is introduced when an annual cash flow estimate has a 50-50 chance of being either Rs. 1000/- or Rs. 500/-.

Decision making under **uncertainty** means there are two or more values observable, but the chances of their occurring cannot be estimated or no one is willing to assign the chances. The observable values in uncertainty analysis are often referred to as *states of nature.*

For example, consider the states of nature to be the rate of national inflation in a particular country during the next 2 to 4 years: remain low, increase 2% to 6% annually, or increase 6%to 8% annually. If there is absolutely no indication that the three values are equally likely, or that one is more likely than the others, ***this is a statement that indicates decision making under uncertainty***.

Illustrative Example - 1

CMS in Fairfield, Virginia received three bids each from vendors for two different pieces of large equipment, A and B. One of each piece of equipment must be purchased. Tom, an engineer at CMS, performed an evaluation of each bid and assigned it a rating between 0 and 100, with100 points being the best of the three. The total for each piece of equipment is 100%. The bid amounts and ratings are shown at the top of the figure below:

a) Consider the ratings as the chance out of 100 that the bid will be chosen, and plot cost versus chance for each vendor.

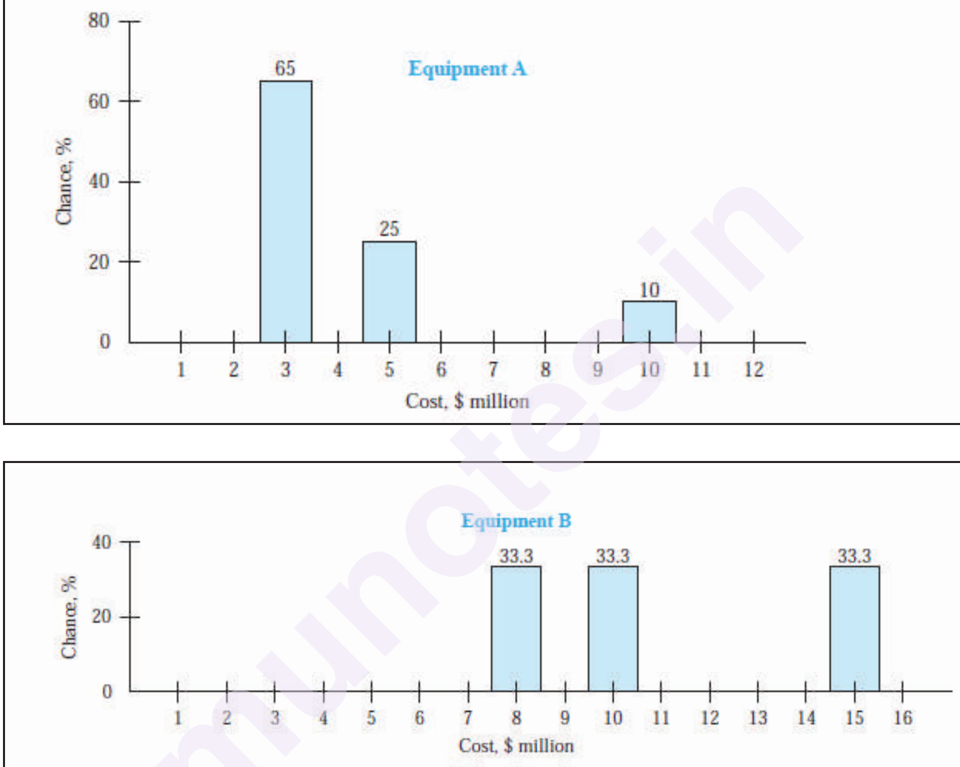
b) Since one each of A and B must be purchased, the total cost will vary somewhere between the sum of the lowest bids ($11 million) and the sum of the highest bids ($25 million). Plot this range with an equal chance of 1 in 14 that any amount in between these limits is possible.

c) Discuss the significant difference between the values of the cost (*x* axis values) in the graphs in (*a*) and (*b*) above and how the chances are stated (*y* axis values).

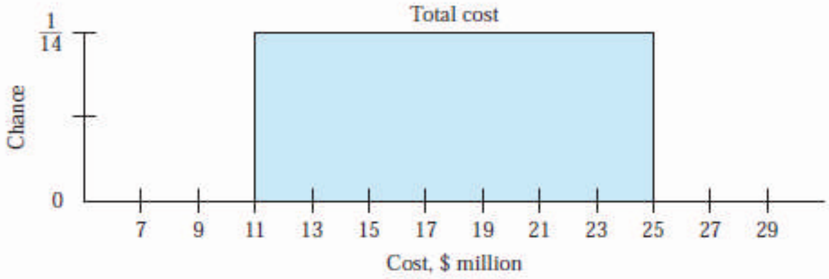
Plot of cost estimates versus chance for (a) each piece of equipment And (b) total cost range

Equipment A	
Bid, $1000	Rating, %
3,000	65
5,000	25
10,000	10

Equipment B	
Bid, $1000	Rating, %
8,000	33.3
10,000	33.3
15,000	33.3



(*a*) Specific values



(*b*) Continuous range

Solution

(a) Figure above plots the specific bids for equipment A and B. The chances (ratings) for A and for B add to 100%. No values

between the specific bids have any chance of occurring, according to the single-estimate bids from the three vendors.

(b) The range of total cost is between $11 million and $25 million, as shown in Figure. Tom decided to make his estimate of total cost continuous between these two extremes. This means that the discrete sums of bids ($11 million, $15 million, and $25 million) are no longer used. Rather the entire range from $11 million to $25 million with a chance for every total cost in between is included. Every value has a chance of 1 in 14 of being observed. Now, the sum is a continuous value.

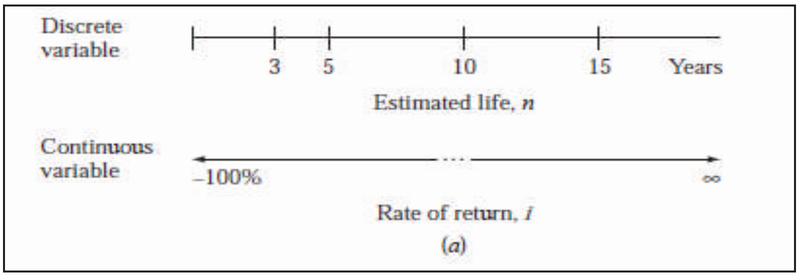
(c) In the graph for bid values, only specific or discrete estimates are included on the *x* axis. In the graph for the sum of the cost for equipment A and B, the *y* axis values are continuous over a specific range.

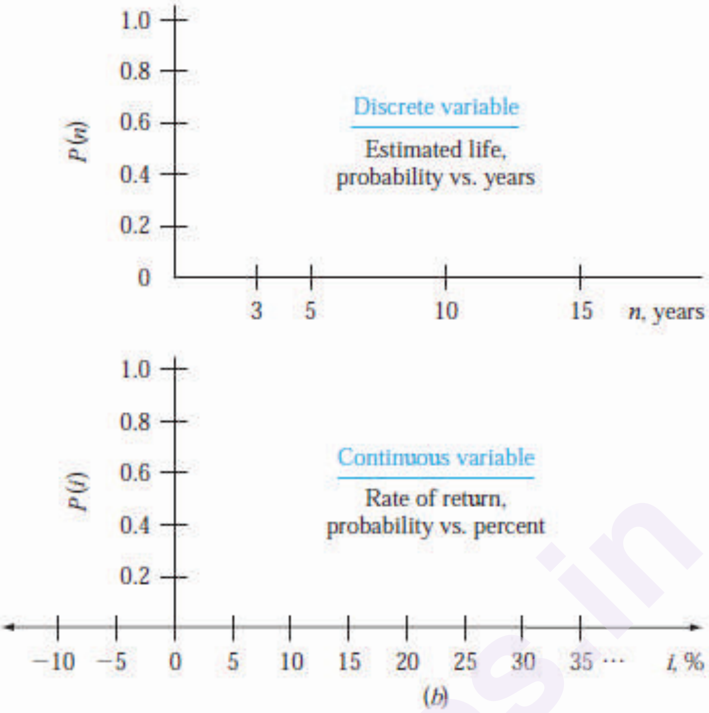
Important Concepts :

When chances are not known for the identified states of nature (or values) of the uncertain parameters, the use of expected value–based decision making under risk is not an option. In fact, it is difficult to determine what criterion to use to even make the decision. If it is possible to agree that each state is equally likely, then all states have the same chance, and the situation reduces to one of decision making under risk, because expected values can be determined. Because of the relatively inconclusive approaches necessary to incorporate decision making under uncertainty into an economy study, the techniques can be quite useful but are beyond the intended scope of this text.

In an economy study, observed parameter values will vary from the value estimated at the time of the study. However, when performing the analysis, not all parameters should be considered as probabilistic (or at risk). Those are estimable with a relatively high degree of certainty should be fixed for the study.

Accordingly, the methods of sampling, simulation, and statistical data analysis are selectively used on parameters deemed important to the decision-making process.





Elements important to decision making under risk :

Some basics of probability and statistics are essential to correctly perform decision making under risk via expected value or simulation analysis. They are the *random variable, probability, probability distribution*, and *cumulative distribution*, as defined here.

A **random variable** or **variable** is a characteristic or parameter that can take on any one of several values. Variables are classified as *discrete* or *continuous.* Discrete variables have several specific, isolated values, while continuous variables can assume any value between two stated limits, called the *range* of the variable.

The estimated life of an asset is a discrete variable. For example, *n* may be expected to have values of *n* = 3, 5, 10, or 15 years, and no others.

The rate of return is an example of a continuous variable; *i* can vary from −100% to ∞, that is, −100% <= *i* <= ∞. The ranges of possible values for *n* (discrete) and *i* (continuous) are shown as the *x* axes in the previous figure. (In probability texts, capital letters symbolize a variable, say *X*, and small letters *x* identify a specific value of the variable.

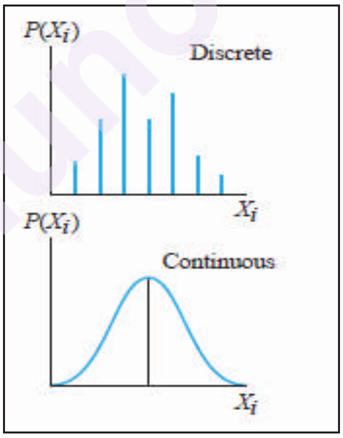
Probability is a number between 0 and 1.0 that expresses the chance in decimal form that a random variable (discrete or continuous) will take on any value from those identified for it. Probability is simply the amount of chance, divided by 100.

Probabilities are commonly identified by *P* (*Xi*) or *P* (*X*=*Xi*), which is read as the probability that the variable *X* takes on the value *Xi*. (Actually, for a continuous variable, the probability at a single value is zero, as shown in a later example.) The sum of all *P* (*Xi*) for a variable must be 1.0, a requirement already discussed. The probability scale, like the percentage scale for chance in previous figure of the illustrative example-1, is indicated on the ordinate (*y* axis) of a graph. This figure shows that the 0 to 1.0range of probability for the variables *n* and *i.*

A **probability distribution** describes how probability is distributed over the different values of a variable.

Discrete variable distributions look significantly different from continuous variable distributions, as indicated by the inset at the right.

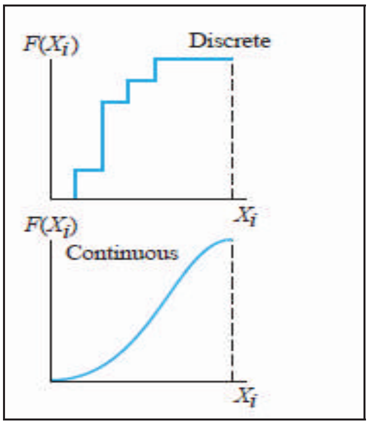
The individual probability values are stated as *P* **(***Xi* **) =**probability that *X* equals *Xi.*



The distribution may be developed in one of two ways: by listing each probability value for each possible variable value or by a mathematical description or expression that states probability in terms of the possible variable values.

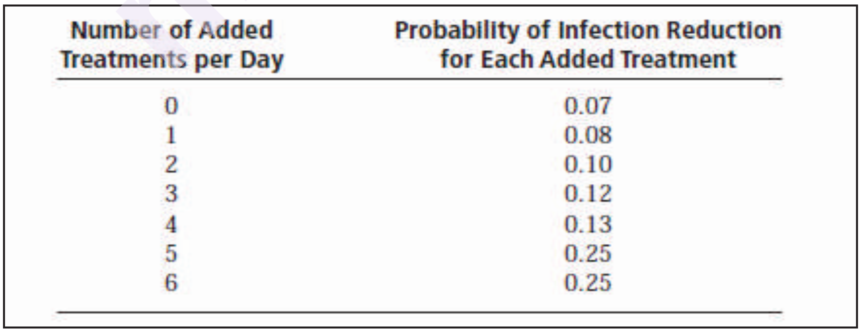
Cumulative distribution, also called the **cumulative probability distribution,** is the accumulation of probability over all values of a variable up to and including a specified value and is identified by F (Xi), each cumulative value is calculated as F (Xi)= sum of all probabilities through the value Xi = P (X<= Xi)

As with a probability distribution, cumulative distributions appear differently for discrete (stair stepped) and continuous variables (smooth curve).



Illustrative Example – 2

Alvin is a medical doctor and biomedical engineering graduate who practices at Medical Center Hospital. He is planning to start prescribing an antibiotic that may reduce infection in patients with flesh wounds. Tests indicate the drug has been applied up to 6 times per day without harmful side effects. If no drug is used, there is always a positive probability that the infection will be reduced by a person's own immune system. Published drug test results provide good probability estimates of positive reaction (i.e., reduction in the infection count) within 48 hours for increased treatments per day. Use the probabilities listed below to construct a probability distribution and a cumulative distribution for the total number of treatments per day.



Number of Added Treatments per Day	Probability of Infection Reduction for Each Added Treatment
0	0.07
1	0.08
2	0.10
3	0.12
4	0.13
5	0.25
6	0.25

Define the random variable *T* as the number of added treatments per day. Since *T* can take on only seven different values, it is a **discrete variable.** The probability of infection reduction is listed for each value in column 2 of table. The cumulative probability *F* (*Ti*) is determined by adding all *P* (*Ti*) values through *Ti*, as indicated in column 3 of the table and the plots of the probability distribution and cumulative distribution respectively are shown below. The summing of probabilities to obtain *F* (*Ti*) gives the cumulative distribution the stair-stepped appearance, and in all cases the final *F* (*Ti*) = 1.0, since the total of all *P* (*Ti*) values must equal 1.0.

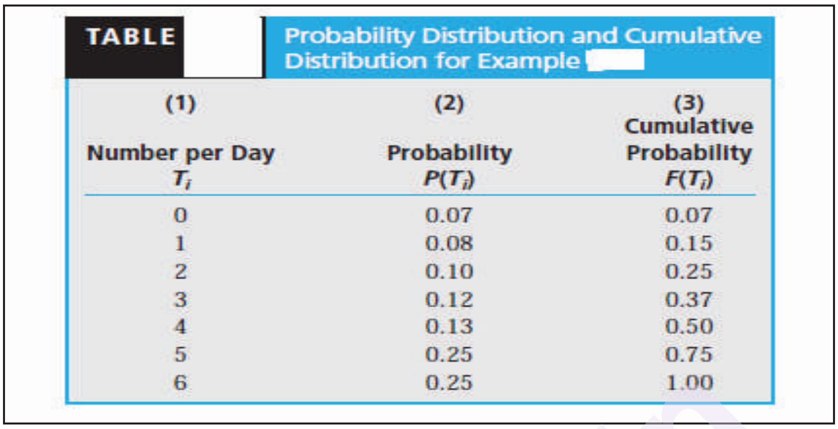
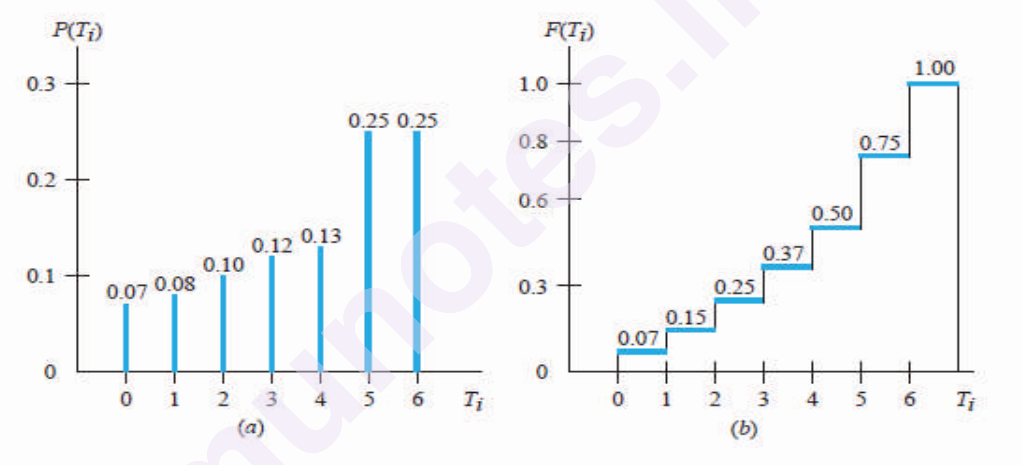


TABLE Probability Distribution and Cumulative Distribution for Example

(1) Number per Day T_i	(2) Probability $P(T_i)$	(3) Cumulative Probability $F(T_i)$
0	0.07	0.07
1	0.08	0.15
2	0.10	0.25
3	0.12	0.37
4	0.13	0.50
5	0.25	0.75
6	0.25	1.00



Rather than use a tabular form as in Table to state *P* (*Ti*) and *F* (*Ti*) values, it is possible to express them for each value of the variable.

$$P(T_i) = \begin{cases} 0.07 & T_1 = 0 \\ 0.08 & T_2 = 1 \\ 0.10 & T_3 = 2 \\ 0.12 & T_4 = 3 \\ 0.13 & T_5 = 4 \\ 0.25 & T_6 = 5 \\ 0.25 & T_7 = 6 \end{cases} \qquad F(T_i) = \begin{cases} 0.07 & T_1 = 0 \\ 0.15 & T_2 = 1 \\ 0.25 & T_3 = 2 \\ 0.37 & T_4 = 3 \\ 0.50 & T_5 = 4 \\ 0.75 & T_6 = 5 \\ 1.00 & T_7 = 6 \end{cases}$$

In basic economy situations, the probability distribution for a **continuous variable** is commonly expressed as a mathematical function, such as a *uniform distribution, a triangular distribution*, or the more complex, but commonly used, *normal distribution.* For continuous variable distributions, the symbol *f* (*X*) is routinely used instead of *P* (*Xi*), and *F* (*X*) is used instead of *F* (*Xi*), simply because the point probability for a continuous variable is zero. Thus, *f* (*X*) and *F*(*X*) are continuous lines and curves.

Illustrative Example – 3 - Random Samples

Estimating a parameter with a single value in previous chapters is the equivalent of taking a *random sample of size 1 from an entire population* of possible values. As an illustration, assume that estimates of first cost, annual operating cost, interest rate, and other parameters are used to compute one present worth (PW) value in order to accept or reject an alternative. Each estimate is a sample of size 1from an entire population of possible values for each parameter. Now, if a second estimate is made for each parameter and a second PW value is determined, a sample of size 2 has been taken.

If all values in the population were known, the probability distribution and cumulative distribution would be known. Then a sample would not be necessary. When we perform an engineering economy study and utilize decision making under certainty, we use one estimate for each parameter to calculate a measure of worth (i.e., a sample of size 1for each parameter). The estimate is the most likely value, that is, one estimate of the expected value. We know that all parameters will vary somewhat; yet some are important enough, or will vary enough, that a probability distribution should be determined or assumed for it and the parameter treated as a random variable. This is using risk, and a sample from the parameter's probability distribution— *P* (*X*) for discrete or *f* (*X*) for continuous—helps formulate probability statements about the estimates. This approach complicates the analysis somewhat; however, it also provides a sense of confidence (or possibly a lack of confidence in some cases) about the decision made concerning the economic viability of the alternative based on the varying parameter.

Illustrative Example – 4 - Random Samples

A **random sample** of size *n* is the selection in a random fashion of *n* values from a population with an assumed or known probability distribution, such that the values of the variable have the **same chance of occurring** in the sample as they are expected to occur in the population.

Suppose Yvon is an engineer with 20 years of experience working for the Aircraft Safety Commission. For a two-crew aircraft, there are three parachutes on board. The safety standard states that 99% of the time, all three chutes must be "fully ready for emergency deployment."

Yvon is relatively sure that nationwide the probability distribution of *N*, the specific number of chutes fully ready, may be described by the probability distribution

$$P(N = N_i) = \begin{cases} 0.005 & N = 0 \text{ chutes ready} \\ 0.015 & N = 1 \text{ chute ready} \\ 0.060 & N = 2 \text{ chutes ready} \\ 0.920 & N = 3 \text{ chutes ready} \end{cases}$$

This means that the safety standard is clearly not met nationwide. Yvon is in the process of sampling 200 (randomly selected) corporate and private aircraft across the nation to determine how many chutes are classified as fully ready. If the sample is truly random and Yvon's probability distribution is a correct representation of actual parachute readiness, the observed *N* values in the 200 aircraft will approximate the same proportions as the population probabilities, that is, 1 aircraft with 0 chutes ready, etc. Since this is a sample, it is likely that the results won't track the population exactly. However, if the results are relatively close, the study indicates that the sample results may be useful in predicting parachute safety across the nation.

To develop a random sample, use **random numbers** (**RN**) generated from a uniform probability distribution for the discrete numbers 0 through 9, that is,

$P\ (\ X_i) = 0.1$ for $X_i = 0, 1, 2, \ldots , 9$

In tabular form, the random digits so generated are commonly clustered in groups of two digits, three digits, or more. Table below a sample of 264 random digits clustered into two-digit numbers.

This format is very useful because the numbers 00 to 99 conveniently relate to the cumulative distribution values 0.01 to 1.00. This makes it easy to select a two-digit RN and enter F (X)to determine a value of the variable with the same proportions as it occurs in the probability distribution.

To apply this logic manually and develop a random sample of size n from a known discrete probability distribution P (X) or a continuous variable distribution f (X), the following procedure may be used.

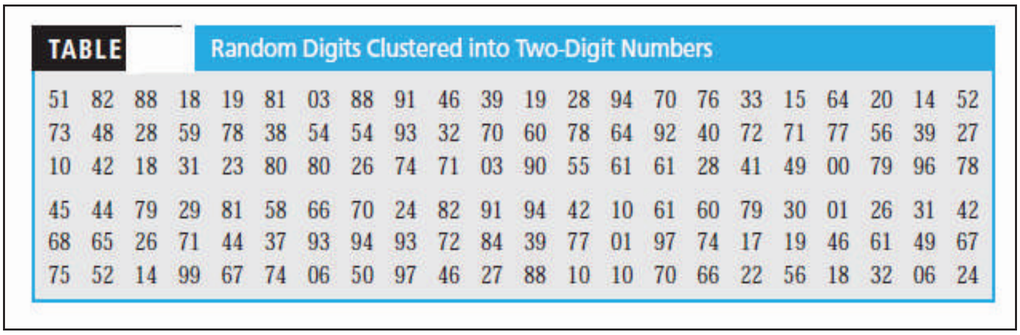


TABLE Random Digits Clustered into Two-Digit Numbers

51	82	88	18	19	81	03	88	91	46	39	19	28	94	70	76	33	15	64	20	14	52
73	48	28	59	78	38	54	54	93	32	70	60	78	64	92	40	72	71	77	56	39	27
10	42	18	31	23	80	80	26	74	71	03	90	55	61	61	28	41	49	00	79	96	78
45	44	79	29	81	58	66	70	24	82	91	94	42	10	61	60	79	30	01	26	31	42
68	65	26	71	44	37	93	94	93	72	84	39	77	01	97	74	17	19	46	61	49	67
75	52	14	99	67	74	06	50	97	46	27	88	10	10	70	66	22	56	18	32	06	24

1. Develop the cumulative distribution *F* (*X*) from the probability distribution. Plot *F* (*X*).

1. Assign the RN values from 00 to 99 to the *F* (*X*) scale (the *y* axis) in the same proportion as the probabilities. For the parachute safety example, the probabilities from 0.0 to 0.15 are represented by the random numbers 00 to 14. Indicate the RNs on the graph.

2. To use a table of random numbers, determine the scheme or sequence of selecting RN values—down, up, across, diagonally. Any direction and pattern is acceptable, but the scheme should be used consistently for one entire sample.
3. Select the first number from the RN table, enter the *F* (*X*) scale, and observe and record the corresponding variable value. Repeat this step until there are *n* values of the variable that constitute the random sample.

4. Use the *n* sample values for analysis and decision making under risk. These may include
   - Plotting the sample probability distribution.
   - Developing probability statements about the parameter.
   - Comparing sample results with the assumed population distribution.
   - Determining sample statistics
   - Performing a simulation analysis.

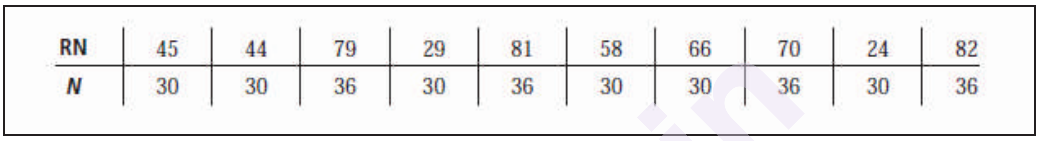
Illustrative Example – 5

Develop a random sample of size 10 for the variable N, number of months, as described by the probability distribution

$$P(N=N_i) = \begin{cases} 0.20 & N=24 \\ 0.50 & N=30 \\ 0.30 & N=36 \end{cases}$$

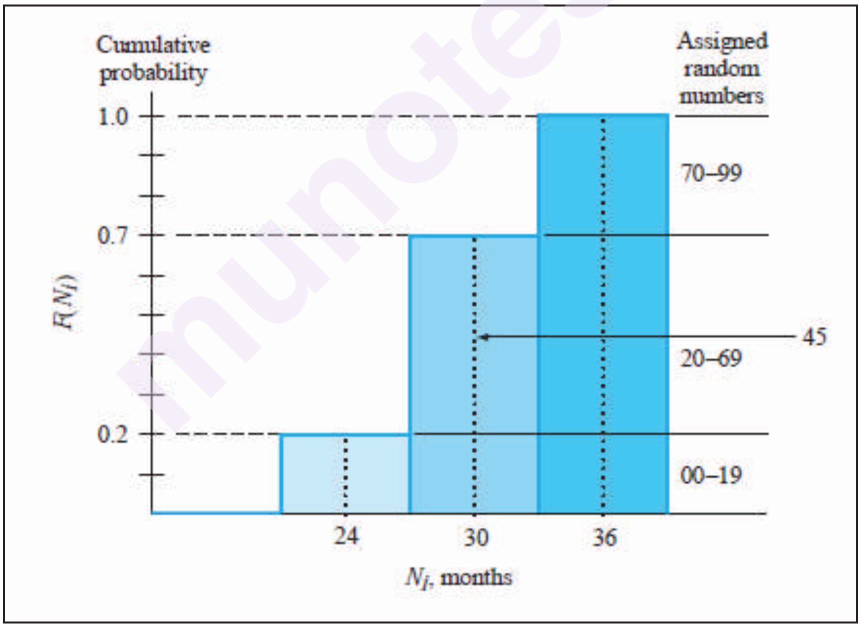
Apply the procedure above, using the *P* (*N=N i*) values in above equation.

1. The cumulative distribution is given below for the discrete variable *N*, which can assume three different values.
2. Assign 20 numbers (00 through 19) to *N*1 = 24 months, where *P* (*N* = 24) = 0.2; 50 numbers to *N*2 = 30; and 30 numbers to *N*3 = 36.
3. Initially select any position in the table of random numbers shown above, and go across the row to the right and onto the row below toward the left. (Any routine can be developed, and a different sequence for each random sample may be used.)
4. Select the initial number 45 (4th row, 1st column), and enter in the RN rangeof 20 to 69 to obtain *N*= 30 months
5. Select and record the remaining nine values from the table of random numbers as shown below:

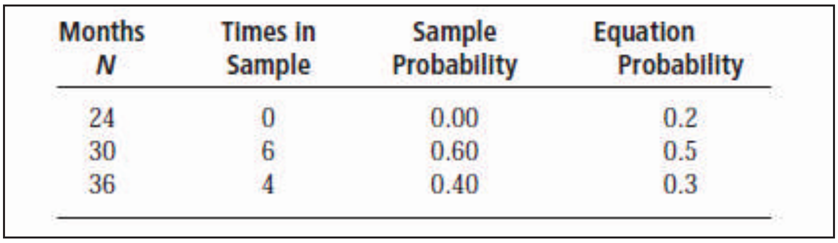


RN	45	44	79	29	81	58	66	70	24	82
N	30	30	36	30	36	30	30	36	30	36

Cumulative distribution with random number values assigned in proportion to probabilities



Now, using the 10 values, develop the sample probabilities



Months *N*	Times in Sample	Sample Probability	Equation Probability
24	0	0.00	0.2
30	6	0.60	0.5
36	4	0.40	0.3

With only 10 values, we can expect the sample probability estimates to be different from the values in question equation. Only the value *N*=24 months is significantly different, since no RN of 19 or less occurred. A larger sample will definitely make the probabilities closer to the original data.

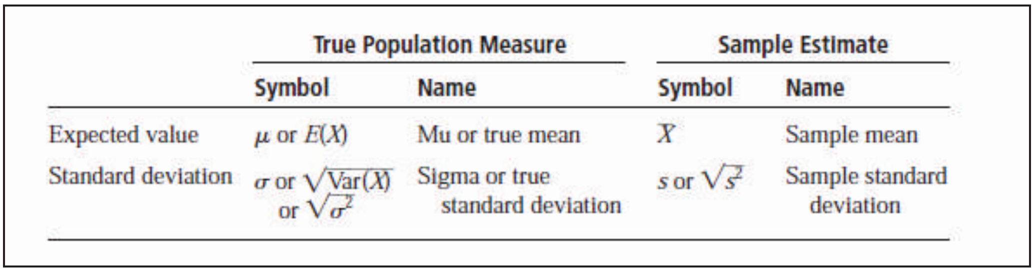
An initial question in random sampling usually concerns the **minimum size of *n*** required to ensure confidence in the results. Without detailing the mathematical logic, sampling theory, which is based upon the law of large numbers and the central limit theorem (check a basic statistics book to learn about these), indicates that an *n* of 30 is sufficient. However, since reality does not follow theory exactly, and since engineering economy often deals with sketchy estimates, samples in the *range of 100 to 200* are the common practice. But samples as small as 10 to 25provide a much better foundation for decision making under risk than the single-point estimate for a parameter that is known to vary widely.

Illustrative Example – 6

Expected Value and Standard Deviation :

Two very important measures or properties of a random variable are the expected value and standard deviation. If the entire population for a variable were known, these properties would be calculated directly. Since they are usually not known, random samples are commonly used to estimate them via the sample mean and the sample standard deviation, respectively. The following is a brief introduction to the interpretation and calculation of these properties using a random sample of size n from the population.

The usual symbols are Greek letters for the true population measures and English letters for the sample estimates.

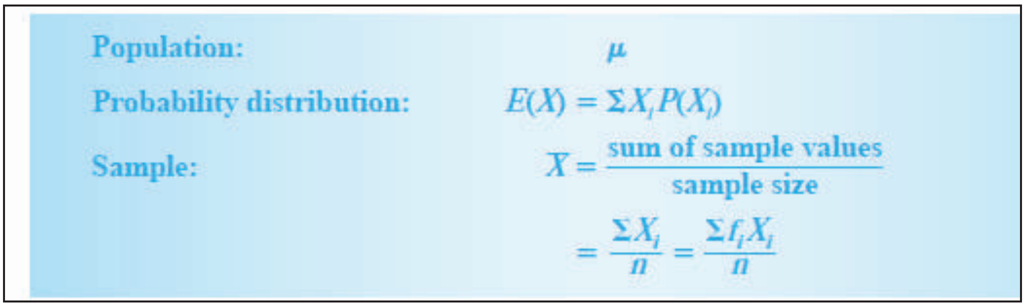


	True Population Measure		Sample Estimate	
	Symbol	Name	Symbol	Name
Expected value	μ or $E(X)$	Mu or true mean	$\bar{X}$	Sample mean
Standard deviation	σ or $\sqrt{\text{Var}(X)}$ or $\sqrt{\sigma^2}$	Sigma or true standard deviation	s or $\sqrt{s^2}$	Sample standard deviation

The **expected value** *E(X)* is the long-run expected average if the variable is sampled many times.

The population expected value is not known exactly, since the population itself is not known completely, so *μ* is estimated either by *E* (*X*) from a distribution or by

$\bar{X}$, the sample mean. The following equation, is used to compute the *E* (*X*) of a probability distribution, and the sample mean, also called the *sample average*.



Population: μ

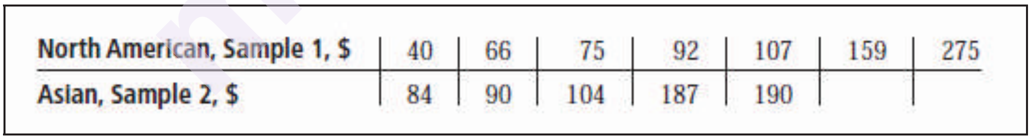
Probability distribution: $E(X) = \Sigma X_i P(X_i)$

Sample:

$$\bar{X} = \frac{\text{sum of sample values}}{\text{sample size}} = \frac{\Sigma X_i}{n} = \frac{\Sigma f_i X_i}{n}$$

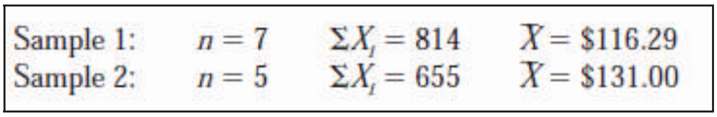
The *f i* in the equation is the frequency of *Xi*, that is, the number of times each value occurs in the sample. The resulting $\bar{X}$, is not necessarily an observed value of the variable; it is the long-run average value and can take on any value within the range of the variable (We omit the subscript *i* on *X* and *f* when there is no confusion introduced).

Kayeu, an engineer with Pacific NW Utilities, is planning to test several hypotheses about residential electricity bills in North American and Asian countries. The variable of interest is *X*, the monthly residential bill in U.S. dollars (rounded to the nearest dollar). Two small samples have been collected from different countries of North America and Asia. Estimate the population expected value. Do the samples (from a non-statistical viewpoint) appear to be drawn from one population of electricity bills or from two different populations?



North American, Sample 1, $	40	66	75	92	107	159	275
Asian, Sample 2, $	84	90	104	187	190		

Use the above equation for the sample mean.



Sample 1:	$n = 7$	$\Sigma X_i = 814$	$\bar{X} = \$116.29$
Sample 2:	$n = 5$	$\Sigma X_i = 655$	$\bar{X} = \$131.00$

Based solely on the small sample averages and the approximate $15 difference, which is only 11% of the larger average bill, does not seem sufficiently large to conclude that the two populations are different. There are several statistical tests available to determine if samples come from the same or different populations. (Check a basic statistics text to learn about them)

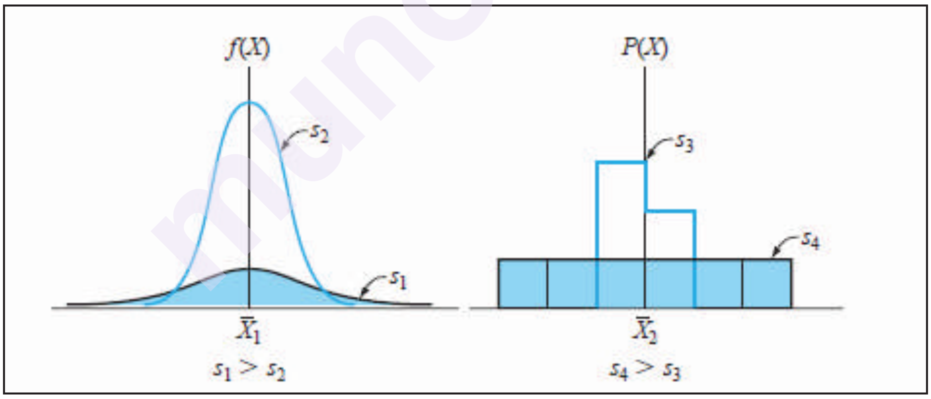
There are three commonly used measures of central tendency for data. The sample average is the most popular, but the *mode* and the *median* are also good measures. The mode, which is the most frequently observed value. The *median is the middle value* of the sample. It is not biased by extreme sample values, as is the mean. The two medians in the samples are $92 and $104. Based solely on the medians, the conclusion is still that the samples do not necessarily come from two different populations of electricity bills.

Illustrative Example – 7 Standard Deviation :

The **standard deviation** *s* or *s* (*X*) is the dispersion or spread of values **about the expected value *E* (*X*)** or sample average $\bar{X}$,

The sample standard deviation *s* estimates the property σ, which is the population measure of dispersion about the expected value of the variable. A probability distribution for data with strong central tendency is more closely clustered about the center of the data, and has a smaller *s*, than a wider, more dispersed distribution.

In the figure below, the samples with larger *s* values— *s*1 and*s*4 —have a flatter, wider probability distribution.



Sketches of distributions with different separate lines standard deviation values.

Actually, the variance *s*2 is often quoted as the measure of dispersion. The standard deviation is simply the square root of the variance, so either measure can be used. The *s* value is what we use routinely in making computations about risk and probability. Mathematically, the formulas and symbols for variance and standard deviation of a discrete variable and a random sample of size *n* are as follows:

Population: $\sigma^2 = \mathrm{Var}(X)$ **and** $\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{Var}(X)}$

Probability distribution: $\mathrm{Var}(X) = \Sigma[X_i - E(X)]^2 P(X_i)$

Sample:

$$s^2 = \frac{\text{sum of (sample value} - \text{sample average)}^2}{\text{sample size} - 1}$$

$$= \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

$$s = \sqrt{s^2}$$

$$s^2 = \frac{\Sigma X_i^2}{n-1} - \frac{n}{n-1}\bar{X}^2 = \frac{\Sigma f_i X_i^2}{n-1} - \frac{n}{n-1}\bar{X}^2$$

The standard deviation uses the sample average as a basis about which to measure the spread or dispersion of data via the calculation $(X - \bar{X})$, which can have a minus or plus sign. To accurately measure the dispersion in both directions from the average, the quantity $(X - \bar{X})$ is squared. To return to the dimension of the variable itself, the square root of s2 equation is extracted. The term $(X - \bar{X})^2$ is called the *mean-squared deviation,* and *s* has historically also been referred to as the *root-mean-square deviation.* The *fi* in the second form of above equation uses the frequency of each *X i* value to calculate *s*2.

One simple way to combine the average and standard deviation is to determine the percentage or fraction of the sample that is within $\pm 1, \pm 2, \text{or } \pm 3$ standard deviations of the average, that is, $\bar{X} \pm ts \quad \text{for } t = 1, 2, \text{or } 3$

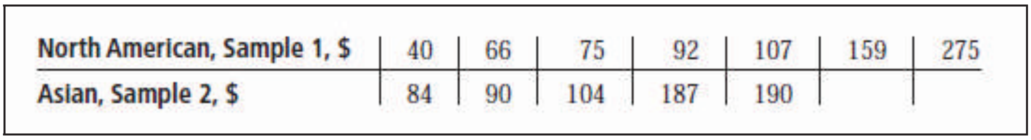
In probability terms, this is stated as

$$P(\bar{X} - ts \le X \le \bar{X} + ts)$$

Virtually all the sample values will always be within the $\pm 3s$ range of $\bar{X}$, but the percent within $\pm 1s$ will vary depending on how the data points are distributed about $\bar{X}$.

a) Use the two samples to estimate population variance and standard deviation for electricity bills.

b) Determine the percentages of each sample that are inside the ranges of 1 and 2 standard deviations from the mean.



North American, Sample 1, $	40	66	75	92	107	159	275
Asian, Sample 2, $	84	90	104	187	190		

a) For illustration purposes only, apply the two different relations to calculate s for the two samples. For sample 1 (North American) with n = 7, use X to identify the values. Table below presents the computation of $\Sigma(X-\bar{X})^2$ with $\bar{X} = \$116.29.$
The resulting s2 and s values are:

$$s^2 = \frac{37{,}743.40}{6} = 6290.57$$

$$s = \$79.31$$

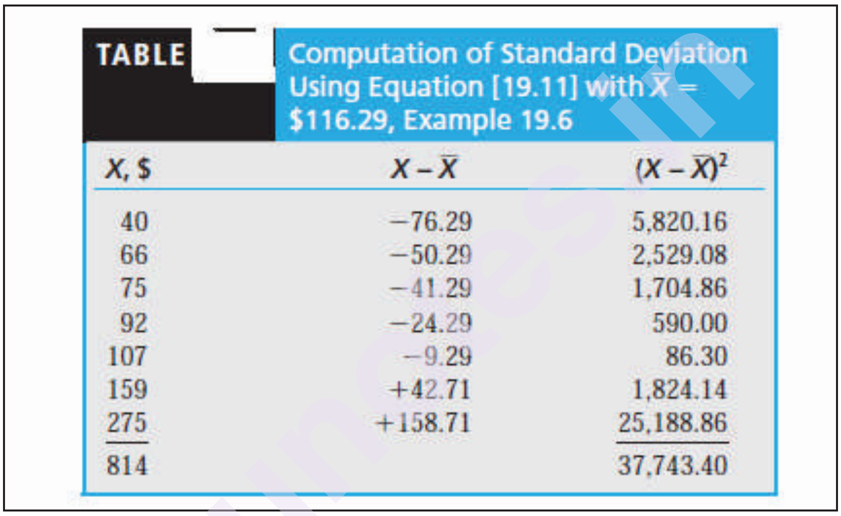


TABLE Computation of Standard Deviation Using Equation [19.11] with $\bar{X}$ = $116.29, Example 19.6

X, $	$X - \bar{X}$	$(X - \bar{X})^2$
40	−76.29	5,820.16
66	−50.29	2,529.08
75	−41.29	1,704.86
92	−24.29	590.00
107	−9.29	86.30
159	+42.71	1,824.14
275	+158.71	25,188.86
814		37,743.40

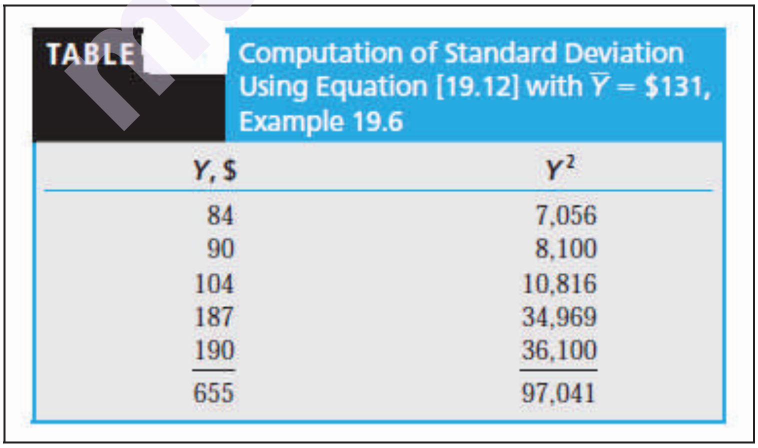


TABLE Computation of Standard Deviation Using Equation [19.12] with $\bar{Y}$ = $131, Example 19.6

Y, $	Y^2
84	7,056
90	8,100
104	10,816
187	34,969
190	36,100
655	97,041

For sample 2 (Asian), use Y to identify the values. With n = 5, and $\bar{Y} = 131$, the table above shows ΣY^2 for the equation given earlier and reproduced below:

$$s^2 = \frac{\Sigma X_i^2}{n-1} - \frac{n}{n-1}X^2 = \frac{\Sigma f_i X_i^2}{n-1} - \frac{n}{n-1}X^2$$

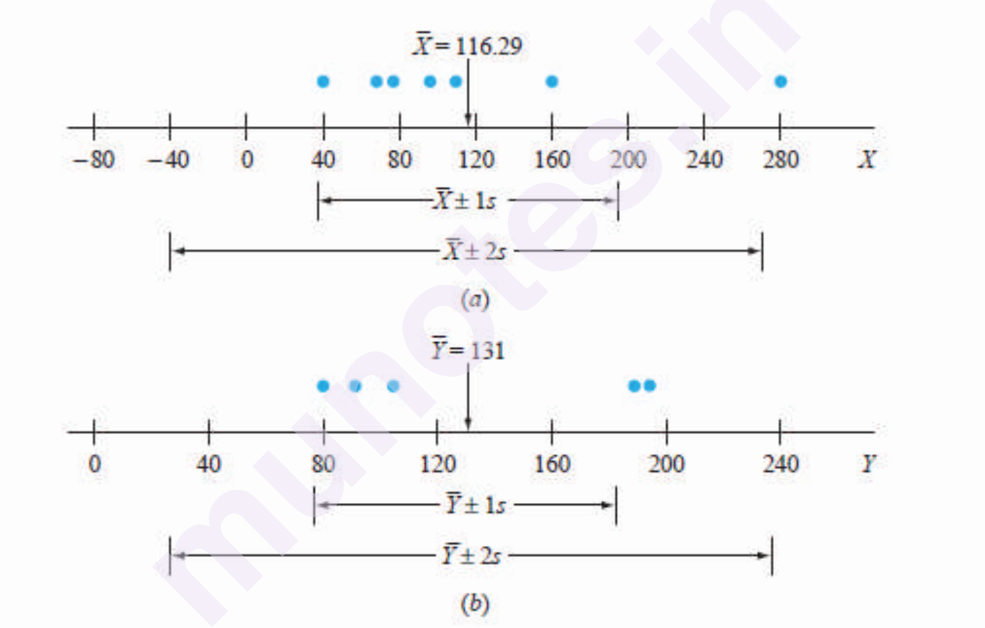
Then

$$s^2 = \frac{97{,}041}{4} - \frac{5}{4}(131)^2 = 42{,}260.25 - 1.25(17{,}161) = 2809$$

$$s = \$53$$

The dispersion is smaller for the Asian sample ($53) than for the North American sample ($79.31).

b) Count the number of sample data points between the limits, and calculate the corresponding percentage. See Figure for a plot of the data and the standard deviation ranges.



Values, averages, and standard deviation ranges for (*a*) North American and(*b*) Asian samples

North American sample

Six out of seven values are within this range, so the percentage is 85.7%.

$X \pm 1s = 116.29 \pm 79.31$ for a range of \$36.98 to \$195.60

There are still six of the seven values within the $X \pm 2s$ range. The limit $ -42.33 is meaningful only from the probabilistic perspective; from the practical viewpoint, use zero, that is, no amount billed.

Asian sample

$$\bar{Y} \pm 1s = 131 \pm 53 \qquad \text{for a range of \$78 to \$184}$$

There are three of five values, or 60%, within the range.

$$\bar{Y} \pm 2s = 131 \pm 106 \qquad \text{for a range of \$25 to \$237}$$

All five of the values are within the $\bar{Y} \pm 2s$ range.

A second common measure of dispersion is the *range,* which is simply the largest minus the smallest sample values. In the two samples here, the range estimates are $235 and $106.

Summary about Normal Distribution and Sampling :

The normal distribution is also referred to as the *bell-shaped curve,* the *Gaussian distribution*, or the *error distribution*. It is, by far, the most commonly used probability distribution in all applications. It places exactly one-half of the probability on either side of the mean or expected value. It is used for continuous variables over the entire range of numbers. The normal distribution is found to accurately predict many types of outcomes, such as IQ values; manufacturing errors about a specified size, volume, weight, etc.; and the distribution of sales revenues, costs, and many other business parameters around a specified mean, which is why it may apply in this situation.

The normal distribution, identified by the symbol $N(\mu, \sigma^2)$, where μ is the expected value or mean and σ^2 is the variance, or measure of spread, can be described as follows:

The mean μ locates the probability distribution, and the spread of the distribution varies with variance, growing wider and flatter for larger variance values.

- When a sample is taken, the estimates are identified as sample mean $\bar{X}$ for μ and sample standard deviation s for σ.
- The normal probability distribution $f(X)$ for a variable X is quite complicated, because its formula is:

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

Where exp represents the number $e = 2.71828$.

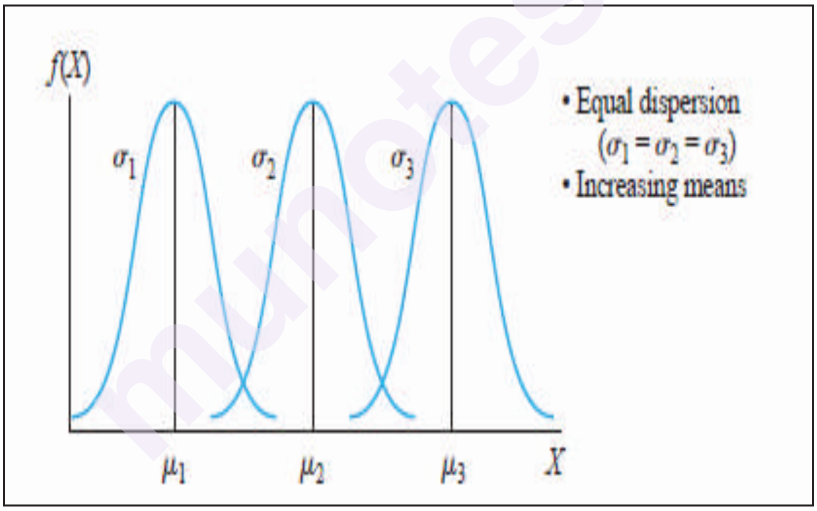
Since $f(X)$ is so unwieldy, random samples and probability statements are developed using a transformation, called the *standard normal distribution (SND),* which uses μ and σ (population) or X and s (sample) to compute values of the variable Z .

Population : $$Z = \frac{\text{deviation from mean}}{\text{standard deviation}} = \frac{X - \mu}{\sigma}$$

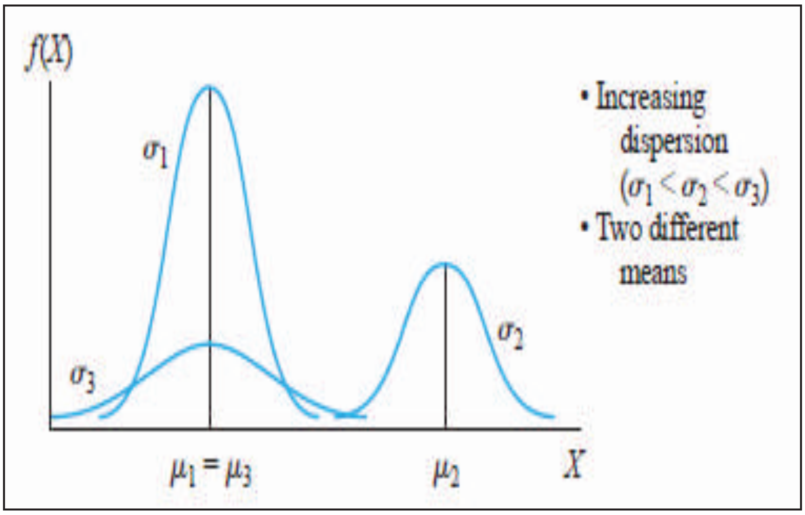
Sample: $$Z = \frac{X - X}{s}$$

The SND for *Z* is the same as for *X*, except that it always has a mean of 0 and a standard deviation of 1, and it is identified by the symbol *N* (0, 1). Therefore, the probability values under the SND curve can be stated exactly. It is always possible to transfer back to the original values from sample data by solving Equation Z for *X* as $X = Z\sigma + \mu$.

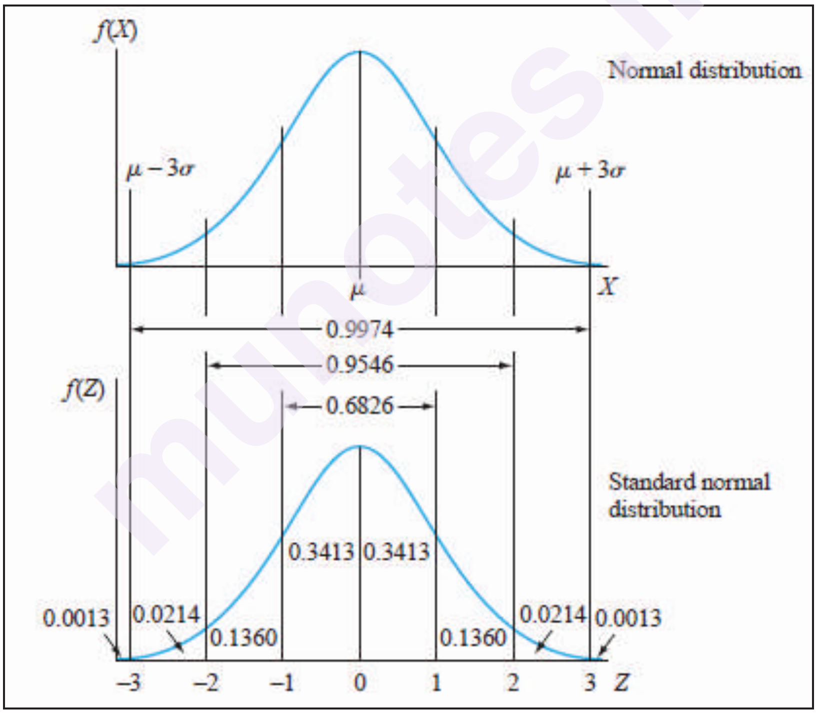
Normal distribution showing different means values μ



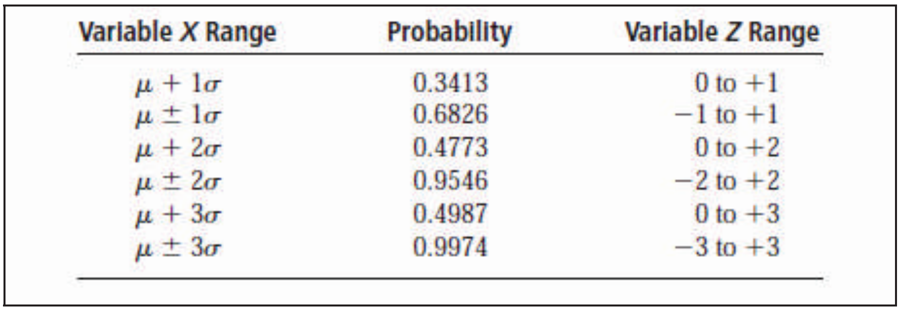
Normal distribution showing different standard deviation σ values



Normal distribution showing relation of normal X to standard normal Z.

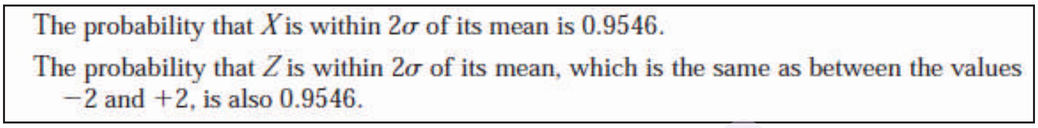


Several probability statements for *Z* and *X* are summarized in the following table and are shown on the distribution curve for *Z* given above.



Variable *X* Range	Probability	Variable *Z* Range
$\mu + 1\sigma$	0.3413	0 to +1
$\mu \pm 1\sigma$	0.6826	−1 to +1
$\mu + 2\sigma$	0.4773	0 to +2
$\mu \pm 2\sigma$	0.9546	−2 to +2
$\mu + 3\sigma$	0.4987	0 to +3
$\mu \pm 3\sigma$	0.9974	−3 to +3

As an illustration, probability statements from this tabulation and the above figure for *X* and *Z* are as follows:



The probability that X is within 2σ of its mean is 0.9546.

The probability that Z is within 2σ of its mean, which is the same as between the values −2 and +2, is also 0.9546.

5.4 LET US SUM UP

In this unit you have learnt the concept of decision making under certainty, risk and uncertainty.

5.5. EXERCISES

Q.1. Given that X is a random variable that is normally distributed with μ = 30 and σ = 4. Determine the following:

a. P (30 < x < 35)
b. P (x > 21)
c. P (x < 40)

Ans: Here we are simply finding the area under the standard type of normal curve under given conditions.

a. Now, Z = (30−30)4(30−30)4 = 0. Also, Z = (35−30)4(35−30)4 = 1.25
b. Thus P (30 < x < 35) = P (0 < z < 1.25) = 0.3944
c. Z = (21−30)4(21−30)4 = -2.25
d. Thus P (x > 21) = P (z > -2.25) = 0.9878
e. Z = (40−30)4(40−30)4 = 2.5
f. Thus P (x < 40) = P (z < 2.5) = 0.9938

Q.2. The ages of all students in a class is normally distributed. The 95 percent of total data is between the age 15.6 and 18.4. Find the mean and standard deviation of the given data.

Ans: From 68-95-99.7 rule, we know that in normal distribution 95 percent of data comes under 2-standard deviation.
Mean of the data = 15.6+18.4215.6+18.42 = 1717

From the mean, 17, to one end, 18.4, there are two standard deviations.

Standard deviation = 18.4−17218.4−172 = 0.7

Q.3. If mean of a given data for a random value is 81.1 and standard deviation is 4.7, then find the probability of getting a value more than 83.

Ans: Standard deviation, σ = 4.74.7
Mean μ = 81.1
Expected value, X = 83
Z-score, $z = (X-\mu) / \sigma$
$z = (83-81.1)/4.7= 0.404255$
Looking up the z-score in the z-table, 0.6700.
Hence, probability (1−0.6700) = 0.33.

Q.4 The average speed of a car is 65 kmph with a standard deviation of 4. Find the probability that the speed is less than 60 kmph.

Ans.: Mean μ =65
Standard deviation, σ = 4
Expected value, X = 4
Z-score, $z = (X-\mu)/\sigma$
$z = (60-65)/4 = -1.25$

Looking up the z-score in the z-table, we get 0.1056.
Hence, probability is 0.1056.

Q.5. The average score of a statistics test for a class is 85 and standard deviation is 10. Find the probability of a random score falling between 75 and 95.

Ans: The probability of score falling between 75 and 95 can be found after finding the respective z-scores.

For X = 75, z = 75−851075−8510 = -1.001.00

For X = 95, z = 95−851095−8510 = 2.002.00

Probability is, $P(-1.00 < z < 2.00) = P(z < 2.00) - P(z < -1.00) = 0.9772 - 0.1587 = 0.8185$.

5.6 SUGGESTED READINGS

Decision making section in any of the reference / text books.

❖❖❖❖

Module - IV
DECISION THEORY AND DECISION TREES

6

CONCEPT OF DECISION TREE, DECISION TREE ANALYSIS

Concept of Decision Tree, Decision Tree Analysis

Unit Structure :

6.1 INTRODUCTION

In this Unit-IV - Chapter 4.3, we shall discuss concept of decision tree and decision tree analysis

6.2 OBJECTIVES

At the end of this unit the learners will be able to understand

- What is a decision tree?
- What is the decision tree analysis?

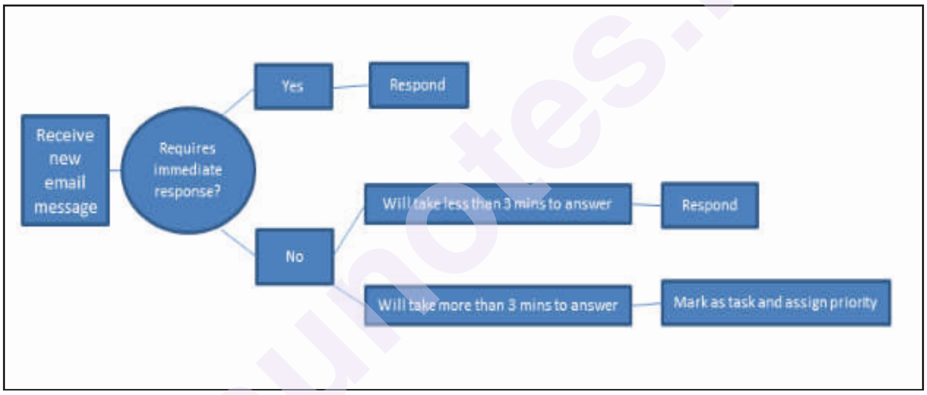
6.3 CONCEPT OF DECISION TREE, DECISION TREE ANALYSIS

What is a decision Tree?

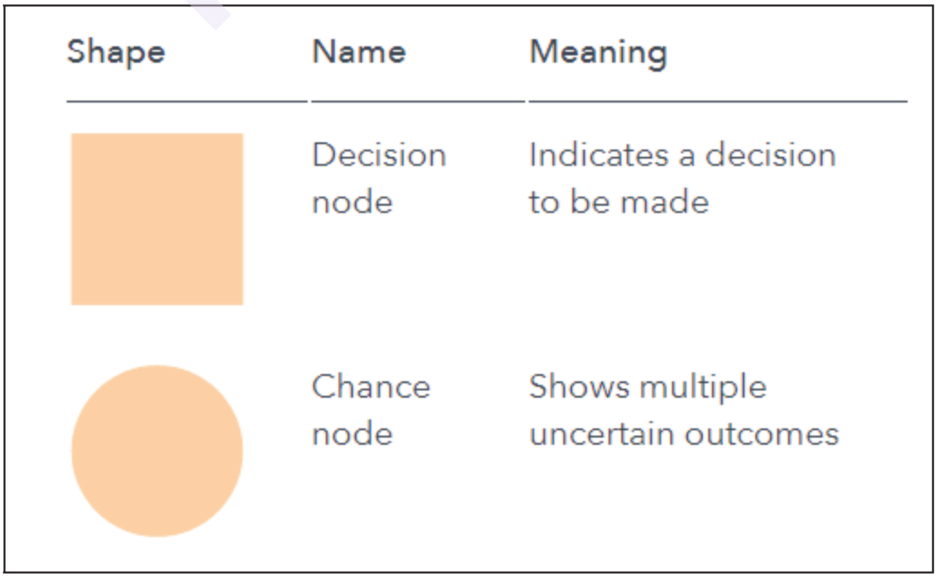
A **decision tree** is a graphical representation of possible solutions to a decision based on certain conditions. It's called a decision tree because it starts with a single box (or root), which then branches off into a number of solutions, just like a tree.

Decision trees are helpful, not only because they are graphics that help you 'see' what you are thinking, but also because making a decision tree requires a systematic, documented thought process. Often, the biggest limitation of our decision making is that we can only select from the known alternatives. Decision trees help formalize the brainstorming process so we can identify more potential solutions.

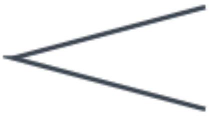
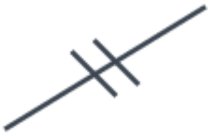
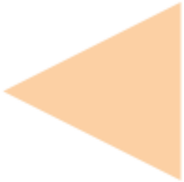
Here’s a simple example: An email management decision tree might begin with a box labeled “Receive new message.” From that, one branch leading off might lead to “Requires immediate response.” From there, a “Yes” box leads to a single decision: “Respond.” A “No” box leads to “Will take less than three minutes to answer” or “Will take more than three minutes to answer.” From the first box, a box leads to “Respond” and from the second box, a branch leads to “Mark as task and assign priority.” The branches might converge after that to “Email responded to? File or delete message”.


Receive new email message
Requires immediate response?
Yes
Respond
No
Will take less than 3 mins to answer
Respond
Will take more than 3 mins to answer
Mark as task and assign priority

Decision Tree Symbols

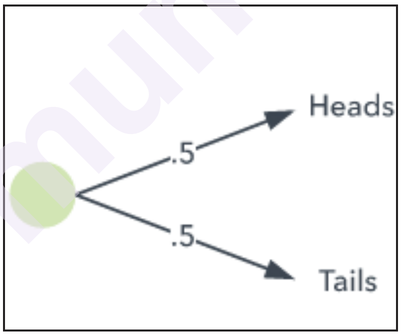


Shape	Name	Meaning
■	Decision node	Indicates a decision to be made
●	Chance node	Shows multiple uncertain outcomes

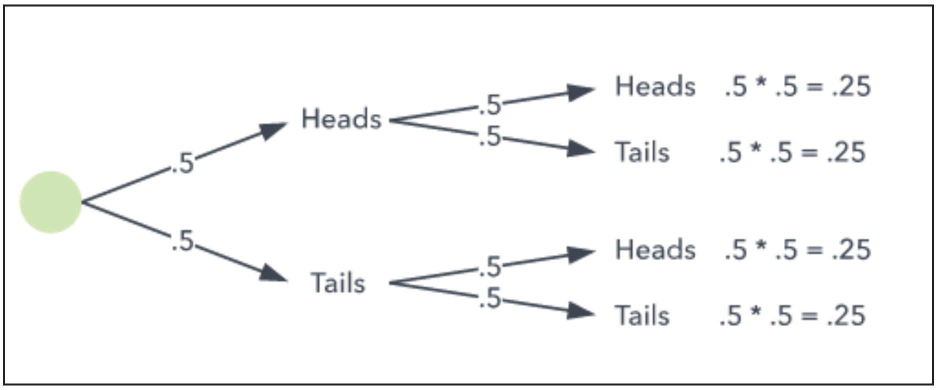
	Alternative branches	Each branch indicates a possible outcome or action
	Rejected alternative	Shows a choice that was not selected
	Endpoint node	Indicates a final outcome

How to draw a decision tree?

- **Start with the main decision**. Draw a small box to represent this point, and then draw a line from the box to the right for each possible solution or action. Label them accordingly. Consider the example of tossing a coin. When an unbiased coin is tossed a head or a Tail appears with each having a probability of 0.5.



- **Add chance and decision nodes to expand the tree as follows:**



- **Continue to expand until every line reaches an endpoint**, meaning that there are no more choices to be made or chance outcomes to consider. Then, assign a value to each possible outcome. It could be an abstract score or a financial value. Add triangles to signify endpoints.

The triangle values are P(HH) = 0.25, P(HT)=0.25, P(TH)=0.25 and P(TT)=0.25

Advantages and Disadvantages of Decision Trees :

Decision trees remain popular for reasons like these:

- How easy they are to understand
- They can be useful with or without hard data, and any data requires minimal preparation
- New options can be added to existing trees
- Their value in picking out the best of several options
- How easily they combine with other decision making tools.

However, decision trees can become excessively complex.

6.4 LET US SUM UP

In this unit you have learnt the concept of decision tree and it analysis.

6.5 EXERCISES

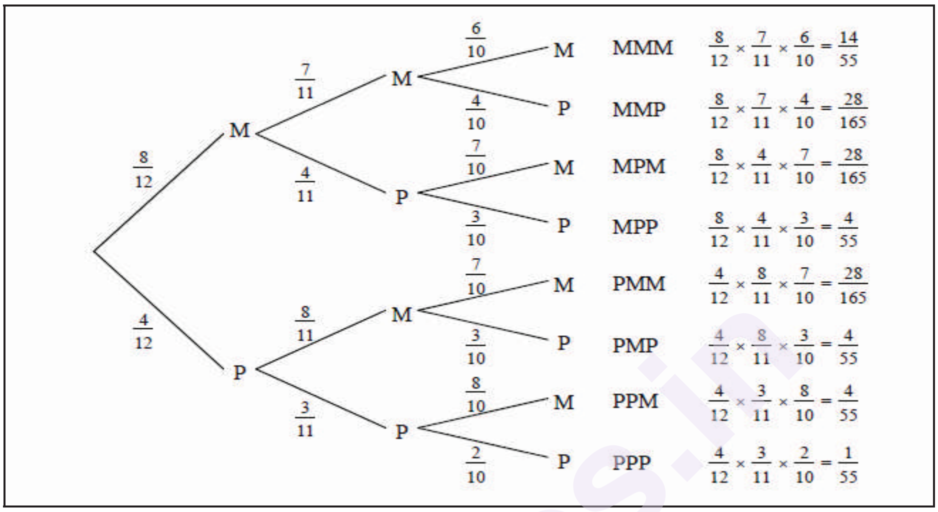
Q.1. A box of chocolates contains 8 milk chocolates and 4 plain chocolates. A chocoholic eats three chocolates.

Calculate the probability that:

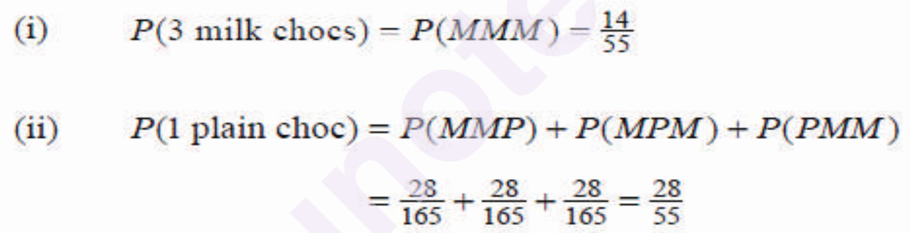
(i) all three are milk chocolates

(ii) exactly one is a plain chocolate

Ans: If *M* is the event "eats a milk chocolate" and *P* is the event "eats a plain chocolate", then the tree diagram is as follows:



M
P
M
P
M
P
M
P
M
P
M
P
M
P
$\frac{8}{12}$
$\frac{4}{12}$
$\frac{7}{11}$
$\frac{4}{11}$
$\frac{8}{11}$
$\frac{3}{11}$
$\frac{6}{10}$
$\frac{4}{10}$
$\frac{7}{10}$
$\frac{3}{10}$
$\frac{7}{10}$
$\frac{3}{10}$
$\frac{8}{10}$
$\frac{2}{10}$
MMM $\frac{8}{12}\times\frac{7}{11}\times\frac{6}{10}=\frac{14}{55}$
MMP $\frac{8}{12}\times\frac{7}{11}\times\frac{4}{10}=\frac{28}{165}$
MPM $\frac{8}{12}\times\frac{4}{11}\times\frac{7}{10}=\frac{28}{165}$
MPP $\frac{8}{12}\times\frac{4}{11}\times\frac{3}{10}=\frac{4}{55}$
PMM $\frac{4}{12}\times\frac{8}{11}\times\frac{7}{10}=\frac{28}{165}$
PMP $\frac{4}{12}\times\frac{8}{11}\times\frac{3}{10}=\frac{4}{55}$
PPM $\frac{4}{12}\times\frac{3}{11}\times\frac{8}{10}=\frac{4}{55}$
PPP $\frac{4}{12}\times\frac{3}{11}\times\frac{2}{10}=\frac{1}{55}$



(i) $P(\text{3 milk chocs}) = P(MMM) = \frac{14}{55}$

(ii) $P(\text{1 plain choc}) = P(MMP) + P(MPM) + P(PMM)$

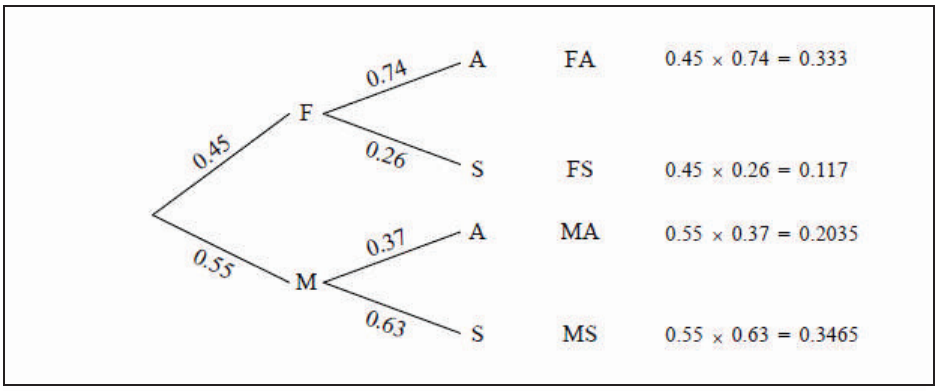
$$= \frac{28}{165} + \frac{28}{165} + \frac{28}{165} = \frac{28}{55}$$

Q.2. In a restaurant, 45% of the customers are female. 74% of females choose from the *à lacarte* menu, whilst only 37% of males do. The rest choose from the set menu. What is the probability that:

(i) a customer orders from the set menu

(ii) a customer ordering from the *à la carte* menu is female?

Ans: If F is the event "is female", M is the event "is male", A is the event "chooses from "*à la carte*", and S is the event "chooses from the set menu", then the tree diagram is as follows:



(i) $P(S) = P(FS) + P(MS)$

$= 0.117 + 0.3465$

$= 0.4635$

(ii) This is a conditional probability as we are told that the customer *has* chosen from the *à la carte* menu – so this is already *given*.

$$P(F \mid A) = \frac{P(F \cap A)}{P(A)} = \frac{0.333}{0.333 + 0.2035} = 0.621$$

6.6 SUGGESTED READINGS

Decision making / probability theory section in any of the reference / text books

❖❖❖❖

Module V
STATISTICAL QUALITY CONTROL

7

INTRODUCTION TO STATISTICAL QUALITY CONTROL

Introduction to Statistical Quality Control
Causes of Variation in Quality, Techniques of SQC

Unit Structure :

7.1.1 INTRODUCTION

In this Unit-V - Chapter 5.1, we shall discuss the concepts of Statistical Quality Control, Causes of Variation in Quality and Techniques of SQC

7.2 OBJECTIVES

At the end of this unit the learners will be able to understand

- What is Statistical Quality Control?
- What are Causes of Variation in Quality?
- What are the Techniques of SQC?

7.3 STATISTICAL QUALITY CONTROL

The main concept of ***product quality*** can be primarily described as the collection of features and characteristics of a product that contribute to its ability to meet given requirements and satisfy the customer's wants and needs in exchange for monetary considerations.

Those features of a saleable good, which determine its desirability, can be controlled by a manufacturer to meet certain basic requirements.

So, product quality is worked on the product's ability to fulfill the expectations and needs set by the end user, and the product must work reliably and perform all of its functions. Product quality can be the sum of two viewpoints, and more.

The control of ***Product Quality*** was defined over decades of creating standards for producing acceptable products: From evolving mature methods to control quality by statistical approaches and sampling techniques by the mid-1950s, and extended their use to the service industry during the1960s.

While during1960–1980, consumers became more conscious of the cost and quality of products and services, and Firms began to focus on total production systems for achieving quality at minimum cost.

Till now, that trend of1980s has continued, and today the goals of quality control are largely driven by consumer concerns and preferences.

Most businesses that produce goods for sale have a product quality or assurance department that monitors outgoing products for consumer acceptability.

To describe the overall quality of a product, we have to consider THREE points of view: The manufacturer, The Consumer and The Quality itself.

1. The view of Manufacturer concerns with the design, engineering, and manufacturing processes involved in fabricating the product.

Quality is measured by the degree of conformance to pre-determined specifications and standards, and deviations from these standards can lead to poor quality and low reliability. Eliminating defects as well as the need for scrap and rework, and hence overall reductions in production costs, are essentials in quality improvement.

2. The view of Customers: that defines the high-quality product is one that well satisfies their preferences and expectations. This consideration can include a number of characteristics, some of

which contribute little or nothing to the functionality of the product but are significant in providing customer satisfaction.

3. While the view of Quality is to consider the product itself as a system and to incorporate those characteristics that pertain directly to the operation and functionality of the product. This approach should include overlap of the manufacturer and customer views.

When we think about the basic elements of product quality, EIGHT dimensions can be identified as Product Quality Framework:

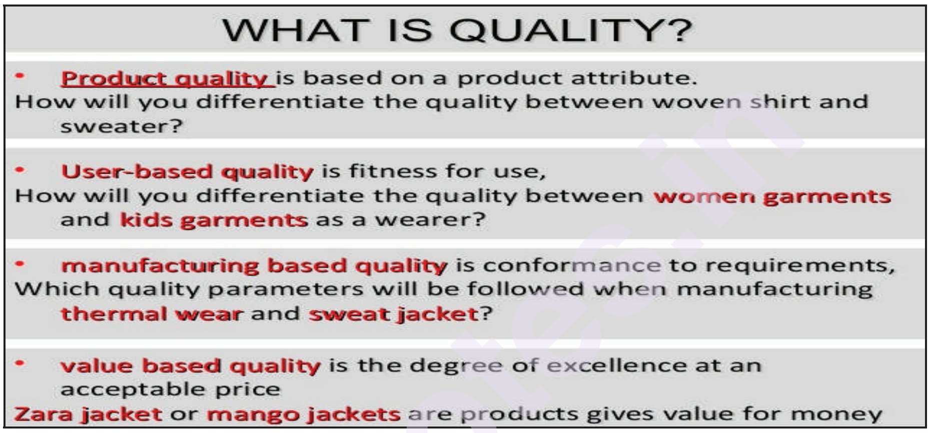
- Performance,
- Features (product characteristics),
- Reliability (the probability of a product's failing within a specified period of time),
- Conformance (the degree to which a product's design and operating characteristics match pre-established standards),
- Durability (a measure of product life, has both economic and technical dimensions),
- Serviceability (the speed, courtesy, and competence of repair),
- Aesthetics (how a product looks, feels, sounds, tastes, or smells — is clearly matters of personal judgment, and reflections of individual preferences),
- Perceived Quality (Evaluation of product's objective characteristics on their images, advertising, or brand names).

Product quality is rapidly becoming an important competitive issue. Companies need to actively shift one's approach to quality as products move from design to market, and to cultivate such differing perspectives, for they are essential to the successful introduction of high-quality products. Reliance on a single definition of quality is a frequent source of problems like: When such manufacturer discovers that its newly released product failed to satisfy customers even though it met the Industrial Standard. Conformance was excellent, reflecting a manufacturing-based approach to quality, but acceptance was poor. While other newly released products generated no customer complaints even though they failed to meet the standard. Other manufacturers' products may be well received by customers, and highly rated by Consumer Reports. Besides, reject, scrap, and warranty costs are so high, however, that large losses are incurred. While the product's design matched customers' needs, the failure to follow through with tight conformance in manufacturing cost the company dearly.

All three views are necessary and must be consciously cultivated.

A process that ignores anyone of the following steps in characterizing quality will not result in a quality product:

- Identified through market research (a user-based approach to quality), next
- Translated into identifiable product attributes (a product-based approach to quality), and followed by
- Organization of manufacturing process to ensure that products are made precisely to these specifications (a manufacturing-based approach to quality).



WHAT IS QUALITY?

- Product quality is based on a product attribute.

How will you differentiate the quality between woven shirt and sweater?

- User-based quality is fitness for use,

How will you differentiate the quality between women garments and kids garments as a wearer?

- manufacturing based quality is conformance to requirements,

Which quality parameters will be followed when manufacturing thermal wear and sweat jacket?

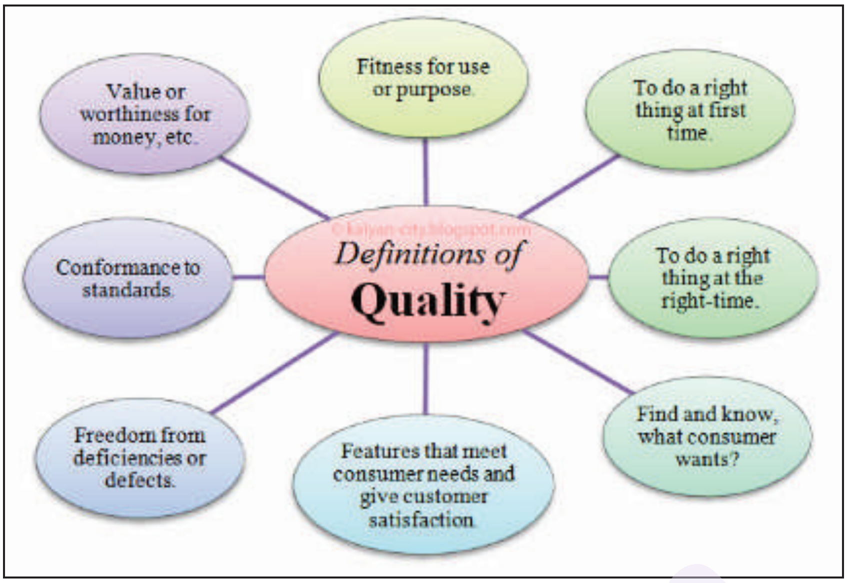
- value based quality is the degree of excellence at an acceptable price

Zara jacket or mango jackets are products gives value for money

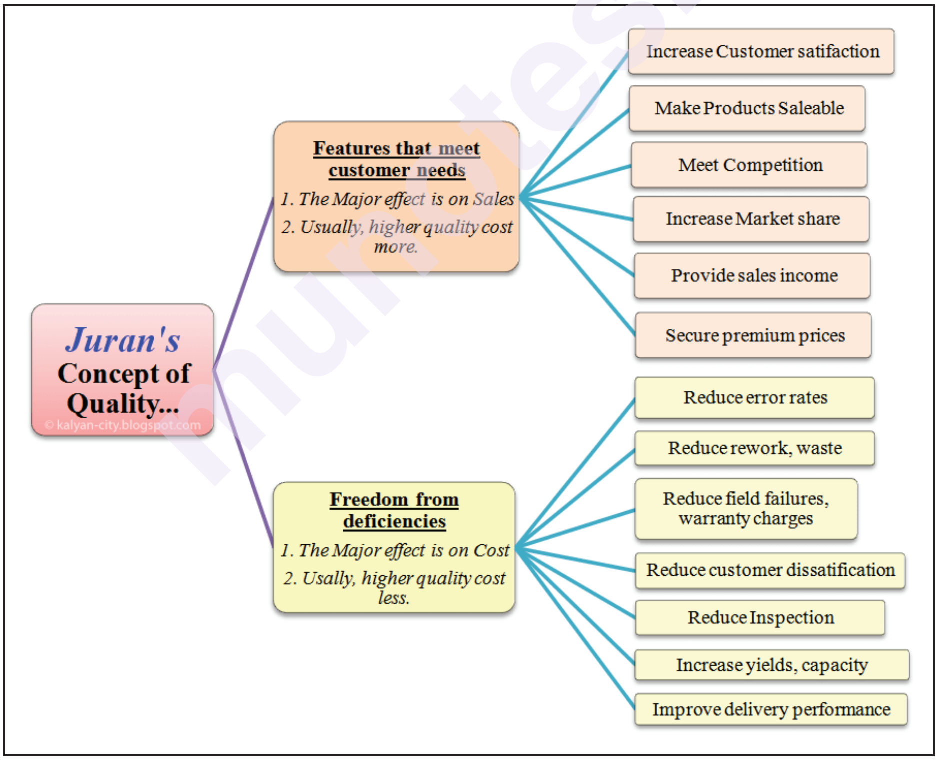
It is not easy to define the word **Quality** since it is perceived differently by the different set of individuals. If experts are asked to define quality, they may give varied responses depending on their individual preferences. These may be similar to following listed phrases.

According to experts, the word quality can be defined either as;

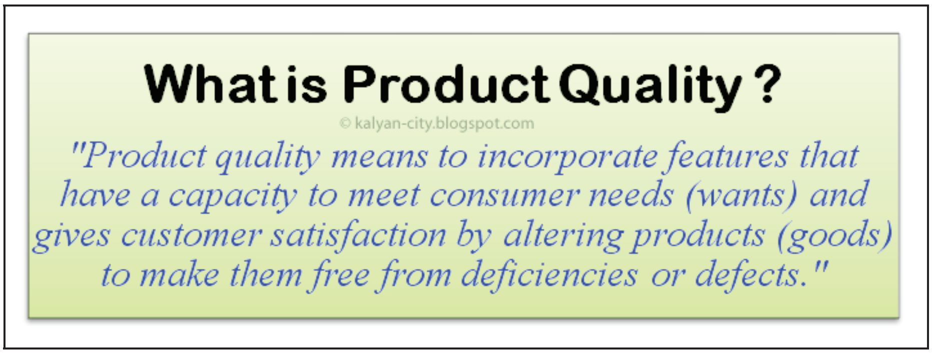
- Fitness for use or purpose.
- To do a right thing at first time.
- To do a right thing at the right-time.
- Find and know what consumer wants?
- Features that meet consumer needs and give customer satisfaction.
- Freedom from deficiencies or defects.
- Conformance to standards.
- Value or worthiness for money, etc.


Value or worthiness for money, etc.
Fitness for use or purpose
To do a right thing at first time
Conformance to standards
© kalyan-city.blogspot.com
Definitions of Quality
To do a right thing at the right-time.
Freedom from deficiencies or defects.
Features that meet consumer needs and give customer satisfaction
Find and know, what consumer wants?

Dr. **Joseph Juran** coined a short definition of quality as, “Product's fitness for use.”


Juran's Concept of Quality...
© kalyan-city.blogspot.com
Features that meet customer needs
1. The Major effect is on Sales
2. Usually, higher quality cost more.
Freedom from deficiencies
1. The Major effect is on Cost
2. Usally, higher quality cost less.
Increase Customer satifaction
Make Products Saleable
Meet Competition
Increase Market share
Provide sales income
Secure premium prices
Reduce error rates
Reduce rework, waste
Reduce field failures, warranty charges
Reduce customer dissatification
Reduce Inspection
Increase yields, capacity
Improve delivery performance

Hence, based on the above discussion, definition of product quality can be stated as follows:



What is Product Quality ?

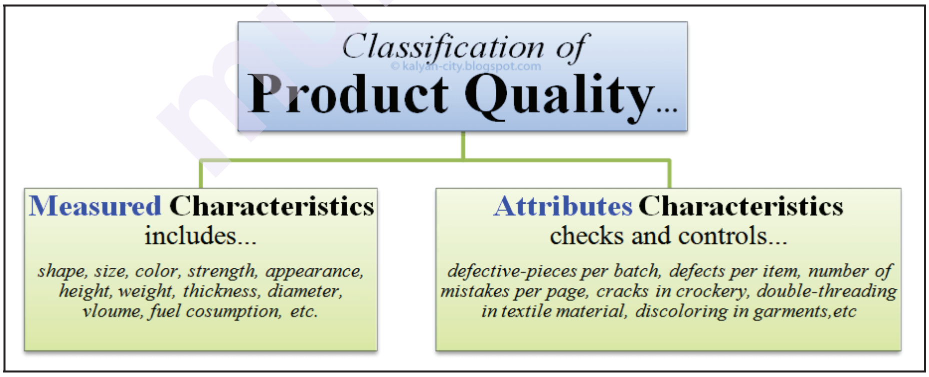
"Product quality means to incorporate features that have a capacity to meet consumer needs (wants) and gives customer satisfaction by altering products (goods) to make them free from deficiencies or defects."

"Product quality means to incorporate features that have a capacity to meet consumer needs (wants) and gives customer satisfaction by improving products (goods) and making them free from any deficiencies or defects."

Meaning of Product Quality

Product quality mainly depends on important factors like:

- The type of raw materials used for making a product.
- How well are various production-technologies implemented?
- Skill and experience of manpower that is involved in the production process.
- Availability of production-related overheads like power and water supply, transport, etc.
- Product quality has two main characteristics viz; measured and attributes.



Classification of
Product Quality...

Measured Characteristics
includes...
shape, size, color, strength, appearance, height, weight, thickness, diameter, vloume, fuel cosumption, etc.

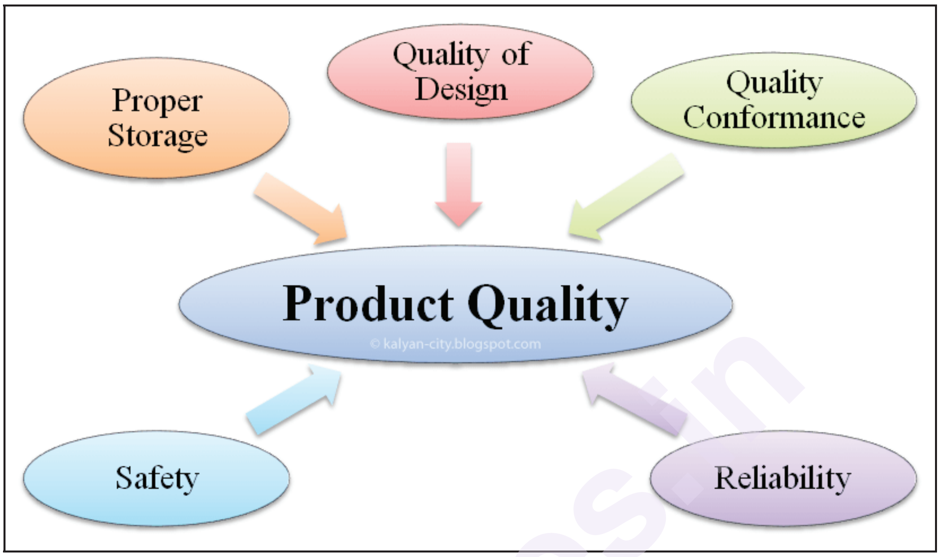
Attributes Characteristics
checks and controls...
defective-pieces per batch, defects per item, number of mistakes per page, cracks in crockery, double-threading in textile material, discoloring in garments,etc

- Measured characteristics include features like shape, size, color, strength, appearance, height, weight, thickness, diameter, volume, fuel consumption, etc. of a product.

- Attributes characteristics checks and controls defective-pieces per batch, defects per item, number of mistakes per page, cracks in crockery, double-threading in textile material, discoloring in garments, etc.

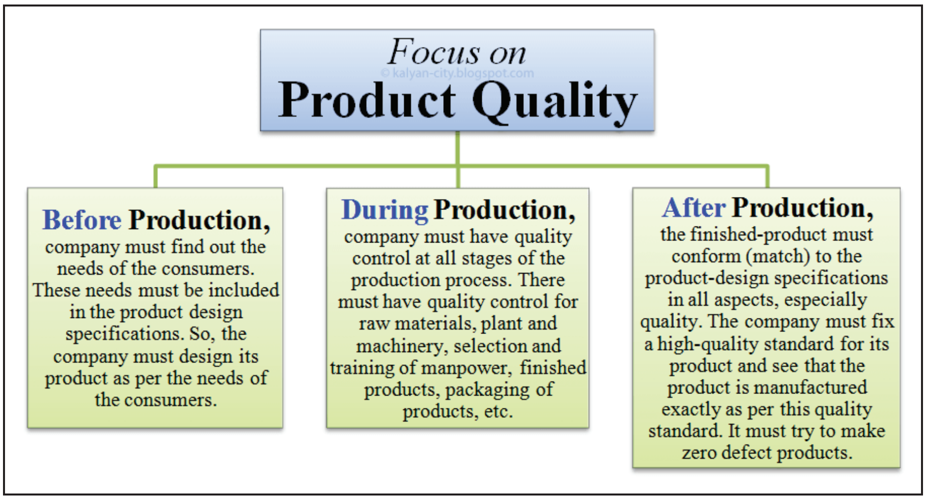
Based on this classification, we can divide products into good and bad. So, product quality refers to the total of the goodness of a product.

The five main aspects of product quality are depicted and listed below:



- Quality of design: The product must be designed as per the consumers' needs and high-quality standards.
- Quality conformance: The finished products must conform (match) to the product design specifications.
- Reliability: The products must be reliable or dependable. They must not easily breakdown or become non-functional. They must also not require frequent repairs. They must remain operational for a satisfactory longer-time to be called as a reliable one.
- Safety: The finished product must be safe for use and/or handling. It must not harm consumers in any way.
- Proper storage: The product must be packed and stored properly. Its quality must be maintained until its expiry date.

Company must focus on product quality, before, during and after production:



Focus on Product Quality

© kalyan-city.blogspot.com

Before Production, company must find out the needs of the consumers. These needs must be included in the product design specifications. So, the company must design its product as per the needs of the consumers.

During Production, company must have quality control at all stages of the production process. There must have quality control for raw materials, plant and machinery, selection and training of manpower, finished products, packaging of products, etc.

After Production, the finished-product must conform (match) to the product-design specifications in all aspects, especially quality. The company must fix a high-quality standard for its product and see that the product is manufactured exactly as per this quality standard. It must try to make zero defect products.

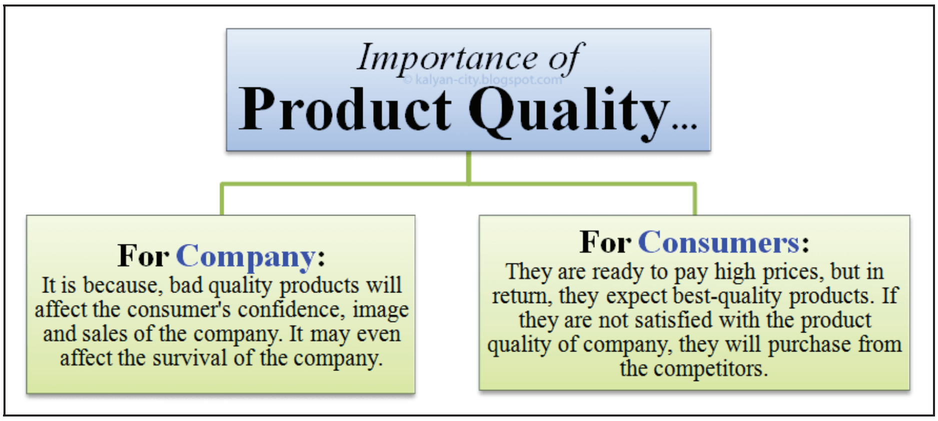
Before production, company must find out the needs of the consumers. These needs must be included in the product design specifications. So, the company must design its product as per the needs of the consumers.

During production, company must have quality control at all stages of the production process. There must have quality control for raw materials, plant and machinery, selection and training of manpower, finished products, packaging of products, etc.

After production, the finished-product must conform (match) to the product-design specifications in all aspects, especially quality. The company must fix a high-quality standard for its product and see that the product is manufactured exactly as per this quality standard. It must try to make zero defect products.

Importance of Product Quality :

Image depicts importance of product quality for company and consumers.



Importance of Product Quality...

© kalyan-city.blogspot.com

For Company: It is because, bad quality products will affect the consumer's confidence, image and sales of the company. It may even affect the survival of the company.

For Consumers: They are ready to pay high prices, but in return, they expect best-quality products. If they are not satisfied with the product quality of company, they will purchase from the competitors.

For company: Product quality is very important for the company. This is because, bad quality products will affect the consumer's confidence, image and sales of the company. It may even affect the survival of the company. So, it is very important for every company to make better quality products.

For consumers: Product quality is also very important for consumers. They are ready to pay high prices, but in return, they expect best-quality products. If they are not satisfied with the quality of product of company, they will purchase from the competitors. Nowadays, very good quality international products are available in the local market. So, if the domestic companies don't improve their products' quality, they will struggle to survive in the market.

7.4 CAUSES OF VARIATION IN QUALITY

Variation in quality characteristic is inevitable, however well the process may be controlled and tests carried out. Thus, for example, measurements taken on some components are bound to vary from piece to piece even though the process may be well under control.

This is a common shop experience. This variation is attributable to two types of causes: Natural or Chance Causes and Assignable Causes.

Variations due to natural or chance causes are inherent in a process. It is due to multitude of causes which are difficult to identify and uneconomical to eliminate. Further, variations due to chance causes follow statistical laws. Assignable causes of variation are generally due to few individual causes which can be identified and eliminated. These two types of causes can be distinguished as follows:

Chance Causes	Assignable Causes
Consists of many individual causes	Consists of a few individual causes
Any one chance cause results in only a very minute amount of variation.	Any one assignable cause can result in a large amount of variation.
Some typical chance causes of variation are small; variation in raw material; small vibration of a machine; lack of perfection in reading instruments.	Some typical assignable causes of variations are: A batch of defective raw material; Faulty set-up; New operator.
As a practical matter chance variation cannot economically be eliminated from a process.	The presence of assignable variation can be detected and action to eliminate the causes is usually economically justified.
Follow statistical laws of variation.	Do not follow any statistical law.

A process is said to be in a state of statistical control if the variations in the quality characteristics of a product follow the appropriate statistical law of variation. Thus, the objectives of controlling a process can be stated as:

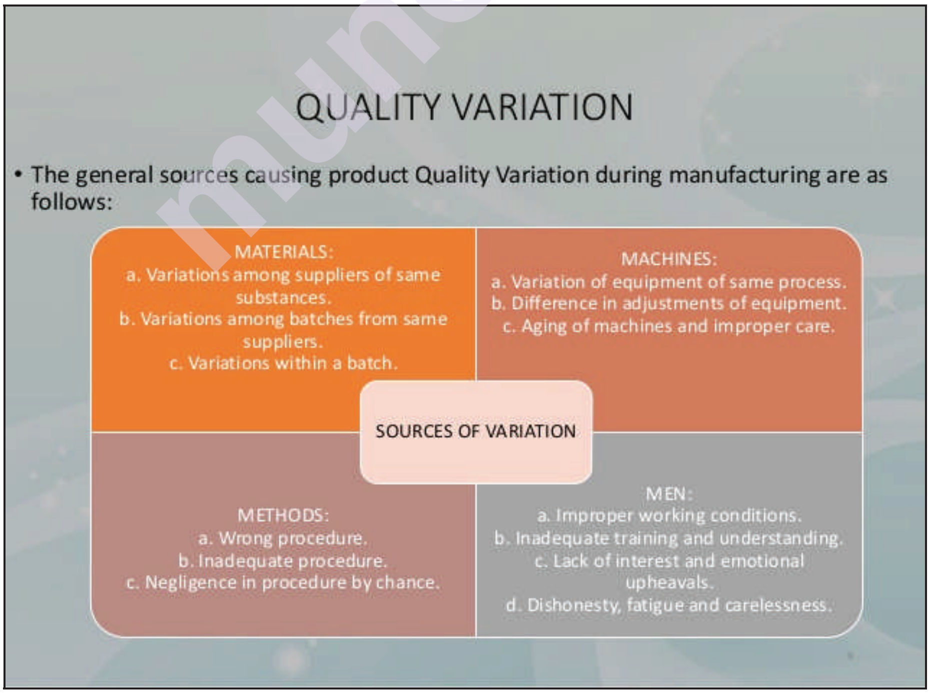
a) to restrict the causes of variation in the quality characteristic to chance causes;

b) to detect and eliminate the assignable causes of variation.

Analysis of pattern of variation exhibited by data on quality characteristics gives a clue to the behaviour of the process and helps in bringing the process under control.

Frequency Distribution - A Simple Method for Analysis of Data:

A first step in examining a set of data is to summarize them in a form such that the main features viz., the extent of variation and the pattern of the observed readings in that range are available in necessary details. This is accomplished by forming a frequency distribution of the data. An understanding of frequency distribution will provide information on ways to measure variability in products, processes and other operations and permits one to develop the ability of recognizing and explaining types and significance of variations.



QUALITY VARIATION

- The general sources causing product Quality Variation during manufacturing are as follows:

MATERIALS:
a. Variations among suppliers of same substances.
b. Variations among batches from same suppliers.
c. Variations within a batch.

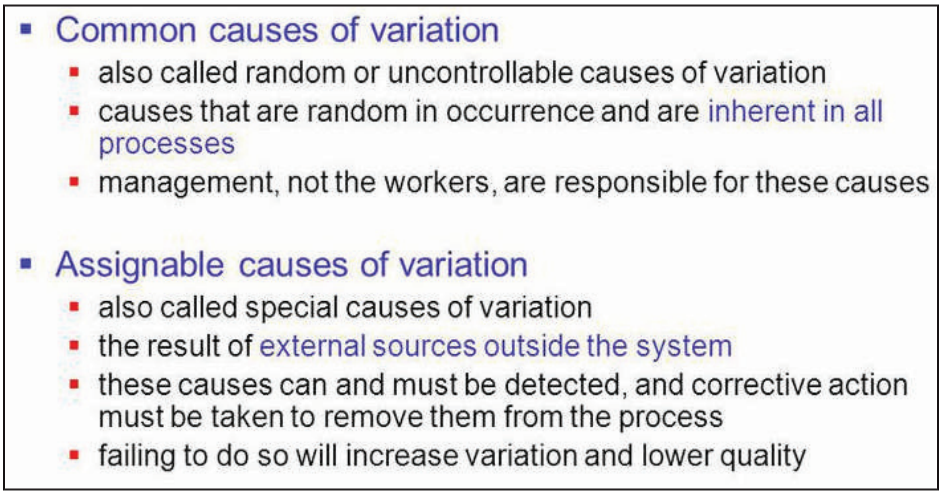
MACHINES:
a. Variation of equipment of same process.
b. Difference in adjustments of equipment.
c. Aging of machines and improper care.

SOURCES OF VARIATION

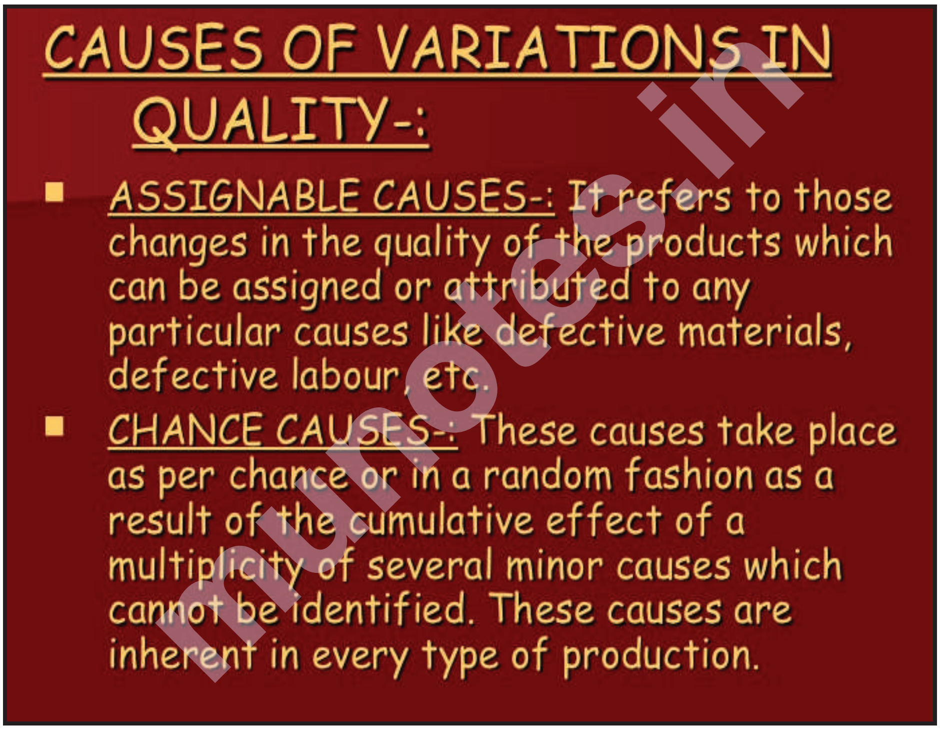
METHODS:
a. Wrong procedure.
b. Inadequate procedure.
c. Negligence in procedure by chance.

MEN:
a. Improper working conditions.
b. Inadequate training and understanding.
c. Lack of interest and emotional upheavals.
d. Dishonesty, fatigue and carelessness.

Summary of Causes of Variation in Quality



- Common causes of variation
  - also called random or uncontrollable causes of variation
  - causes that are random in occurrence and are inherent in all processes
  - management, not the workers, are responsible for these causes
- Assignable causes of variation
  - also called special causes of variation
  - the result of external sources outside the system
  - these causes can and must be detected, and corrective action must be taken to remove them from the process
  - failing to do so will increase variation and lower quality



CAUSES OF VARIATIONS IN QUALITY-:

- ASSIGNABLE CAUSES-: It refers to those changes in the quality of the products which can be assigned or attributed to any particular causes like defective materials, defective labour, etc.
- CHANCE CAUSES-: These causes take place as per chance or in a random fashion as a result of the cumulative effect of a multiplicity of several minor causes which cannot be identified. These causes are inherent in every type of production.

7.5 TECHNIQUES OF SQC

There are many ways to implement statistical quality or statistical process control. Key monitoring and investigating tools include:

- Histograms
- Check Sheets
- Pareto Charts
- Cause and Effect Diagrams
- Defect Concentration Diagrams
- Scatter Diagrams
- Control Charts

Histogram

A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) and was first introduced by Karl Pearson. It is a kind of bar graph.

To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency — the number of cases in each bin. A histogram may also be normalized to display "relative" frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.

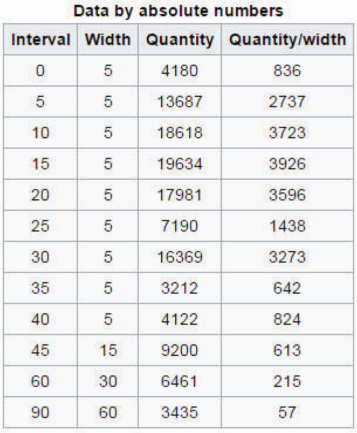
However, bins need not be of equal width; in that case, the erected rectangle is defined to have its area proportional to the frequency of cases in the bin. The vertical axis is then not the frequency but frequency density — the number of cases per unit of the variable on the horizontal axis.

As the adjacent bins leave no gaps, the rectangles of a histogram touch each other to indicate that the original variable is continuous.

Histograms give a rough sense of the density of the underlying distribution of the data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the x-axis is all 1, then a histogram is identical to a relative frequency plot.

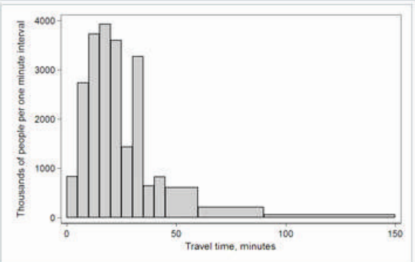
A histogram can be thought of as a simplistic kernel density estimation, which uses a kernel to smooth frequencies over the bins. This yields a smoother probability density function, which will in general more accurately reflect distribution of the underlying variable. The density estimate could be plotted as an alternative to the histogram, and is usually drawn as a curve rather than a set of boxes. The histogram is one of the seven basic tools of quality control.

The U.S. Census Bureau found that there were 124 million people who work outside of their homes. Using their data on the time occupied by travel to work, the table below shows the absolute number of people who responded with travel times "at least 30 but less than 35 minutes" is higher than the numbers for the categories above and below it. This is likely due to people rounding their reported journey time. The problem of reporting values as somewhat arbitrarily rounded numbers is a common phenomenon when collecting data from people.



Data by absolute numbers

Interval	Width	Quantity	Quantity/width
0	5	4180	836
5	5	13687	2737
10	5	18618	3723
15	5	19634	3926
20	5	17981	3596
25	5	7190	1438
30	5	16369	3273
35	5	3212	642
40	5	4122	824
45	15	9200	613
60	30	6461	215
90	60	3435	57

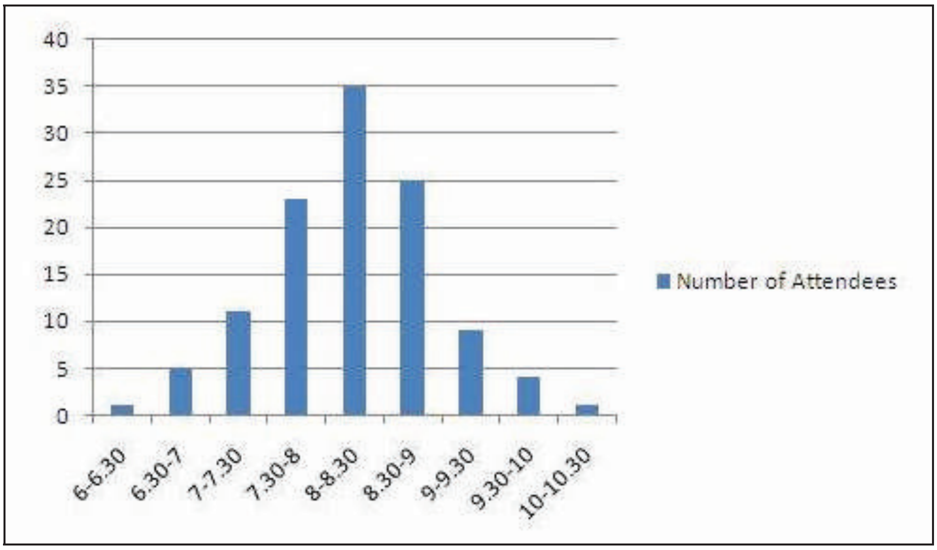


Histogram of travel time (to work), US 2000 census. Area under the curve equals the total number of cases. This diagram uses Q/width from the table.

This histogram shows the number of cases per unit interval as the height of each block, so that the area of each block is equal to the number of people in the survey who fall into its category. The area under the curve represents the total number of cases (124 million). This type of histogram shows absolute numbers, with Q in thousands.

A histogram represents a frequency distribution by means of rectangles whose widths represent class intervals and whose areas are proportional to the corresponding frequencies: the height of each is the average frequency density for the interval. The intervals are placed together in order to show that the data represented by the histogram, while exclusive, is also contiguous. E.g., in a histogram it is possible to have two connecting intervals of 10.5–20.5 and 20.5–33.5, but not two connecting intervals of 10.5–20.5 and 22.5–32.5. Empty intervals are represented as empty and not skipped.

The following histogram shows morning attendance of a class. The X-axis is the number of students and the Y-axis the time of the day.

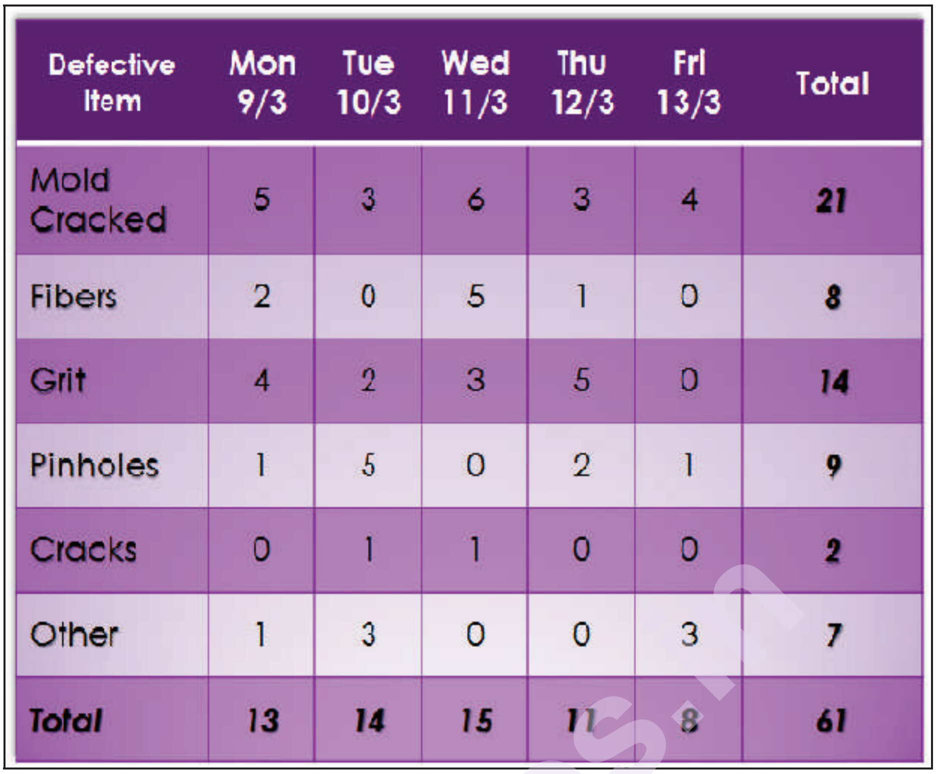


Check Sheets

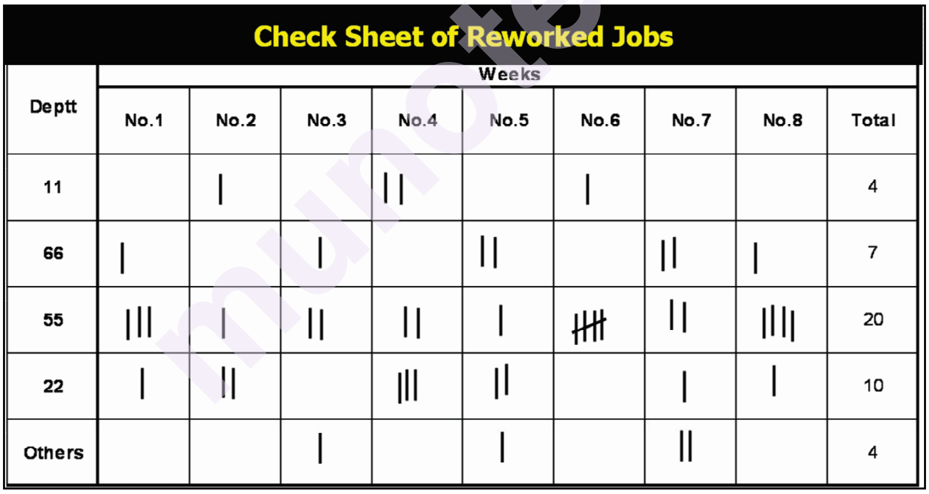
Check sheets are the paper forms for collecting data in real time easily and concisely.

According to the data on check sheets, frequency or pattern of the events, problems, defects, etc. can be observed. These data are then used as input data for other seven quality tools such as histograms and Pareto diagrams.

Furthermore, collected data on check sheets can be used as input to understand the real situation, analyze occurring problem, control the process, make the decision, and develop planning.



Defective Item	Mon 9/3	Tue 10/3	Wed 11/3	Thu 12/3	Fri 13/3	Total
Mold Cracked	5	3	6	3	4	**21**
Fibers	2	0	5	1	0	**8**
Grit	4	2	3	5	0	**14**
Pinholes	1	5	0	2	1	**9**
Cracks	0	1	1	0	0	**2**
Other	1	3	0	0	3	**7**
Total	***13***	***14***	***15***	***11***	***8***	***61***



Check Sheet of Reworked Jobs

Deptt	Weeks								
	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8	Total
11		ǀ		ǀǀ		ǀ			4
66	ǀ		ǀ		ǀǀ		ǀǀ	ǀ	7
55	ǀǀǀ	ǀ	ǀǀ	ǀǀ	ǀ	𝍸	ǀǀ	ǀǀǀǀ	20
22	ǀ	ǀǀ		ǀǀǀ	ǀǀ		ǀ	ǀ	10
Others			ǀ		ǀ		ǀǀ		4

Pareto Charts

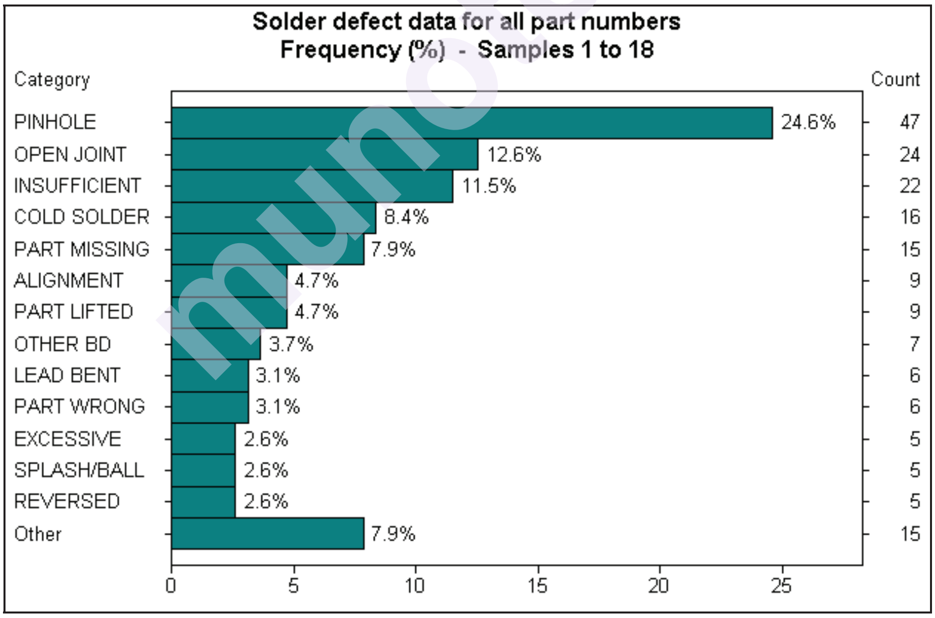
Pareto diagrams consist of bar graph and line graph that describe the proportion of each problem cause toward the overall cause. According to these quality tools, the major factor of the problem can be identified. Thus, it helps process owner in prioritizing the problem to be solved first.

As Dr. J. M. Juran said, 80% of the problems are caused by 20% of the potential sources.

Not only to identify the major problem, pare to diagram can also be used to analyze the improvement after solving some problems. Process owner can evaluate the process before and after solving the problems.

A Pareto diagram is a simple bar chart that ranks related measures in decreasing order of occurrence. The principle was developed by Vilfre do Pareto, an Italian economist and sociologist who conducted a study in Europe in the early 1900s on wealth and poverty. He found that wealth was concentrated in the hands of the few and poverty in the hands of the many. The principle is based on the unequal distribution of things in the universe. It is the law of the "significant few versus the trivial many." The significant few things will generally make up 80% of the whole, while the trivial many will make up about 20%.

The purpose of a Pareto diagram is to separate the significant aspects of a problem from the trivial ones. By graphically separating the aspects of a problem, a team will know where to direct its improvement efforts. Reducing the largest bars identified in the diagram will do more for overall improvement than reducing the smaller ones.



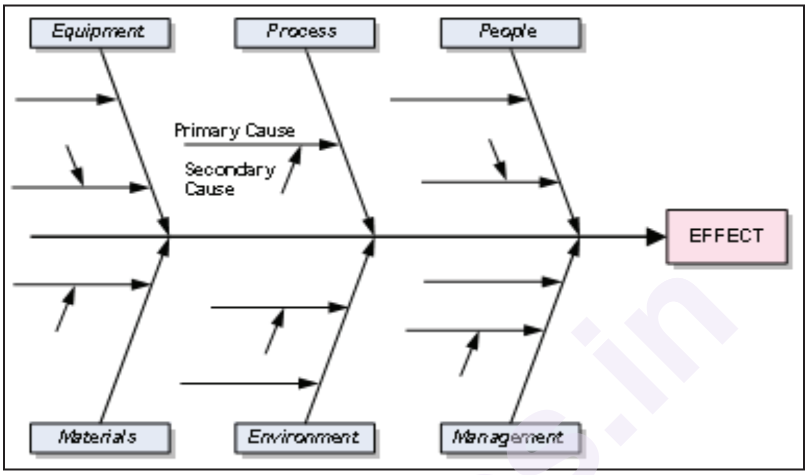
Cause and Effect Diagrams

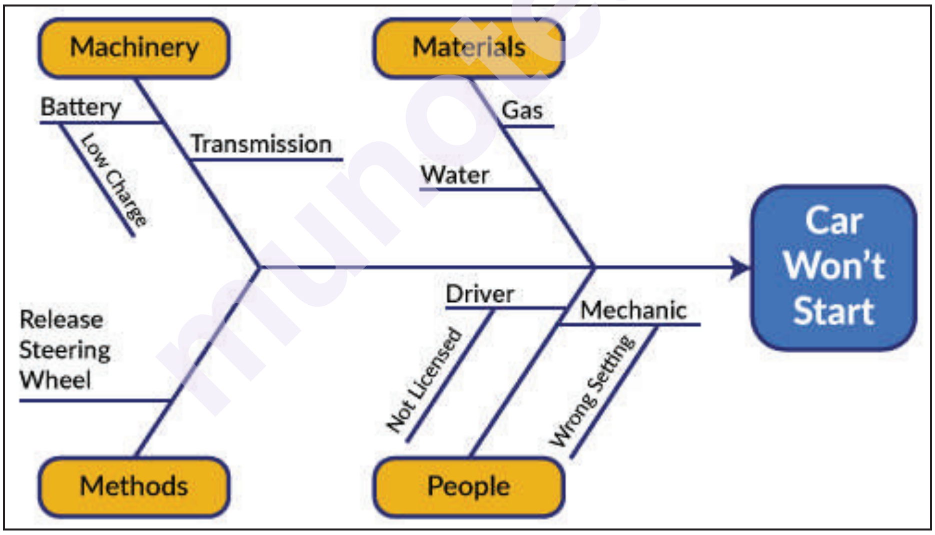
Cause and effect diagrams are also called Ishikawa diagrams or fishbone diagrams. These diagrams show a relationship between the effect of the problem and the causes.

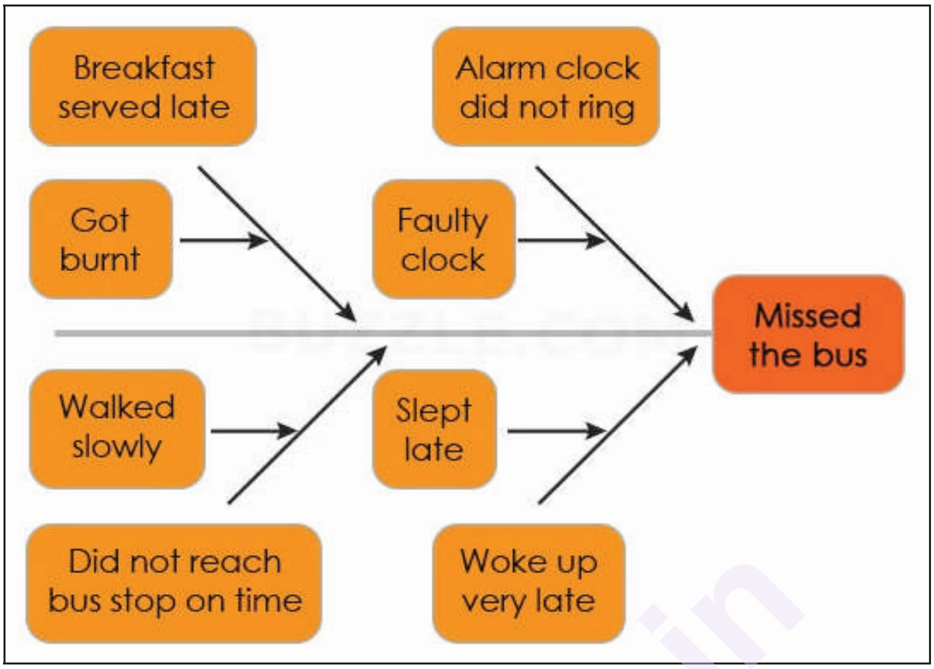
Although cause and effect diagrams can be developed by individual, but it is better to be brainstormed by a team.

To help in identifying the causes, people used to take 5 potential factors which are man, machine, material, method, and environment as the guidance.

Cause and effect diagrams consist of 2 sides. The right side lists the effect or the problem, while the left side lists the causes of the problem.





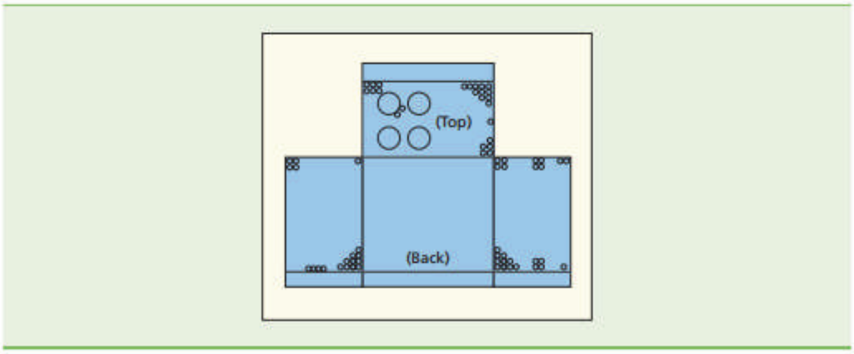


Defect Concentration Diagrams

The defect concentration diagram (also problem concentration diagram) is a graphical tool that is useful in analyzing the causes of the product or part defects. It is a drawing of the product (or other item of interest), with all relevant views displayed, onto which the locations and frequencies of various defects are shown.

A defect concentration diagram is a picture of the product. It depicts all views—for example, front, back, sides, bottom, top and so on. The various kinds of defects are then illustrated on the diagram. Often, by examining the locations of the defects, we can discern information concerning the causes of the defects. For example, in the October 1990 issue of Quality Progress, The Juran Institute presents a defect concentration diagram that plots the locations of chips in the enamel finish of a kitchen range. If the manufacturer of this range plans to use protective packaging to prevent chipping, it appears that the protective packaging should be placed on the corners, edges, and burners of the range.

Defect Concentration Diagram Showing the Locations of Enamel Chips on Kitchen Ranges



Source: "The Tools of Quality Part V: Check Sheets," from *QI Tools: Data Collection Workbook*, p. 11. Copyright © 1989. Juran Institute, Inc. Reprinted with permission from Juran Institute, Inc.

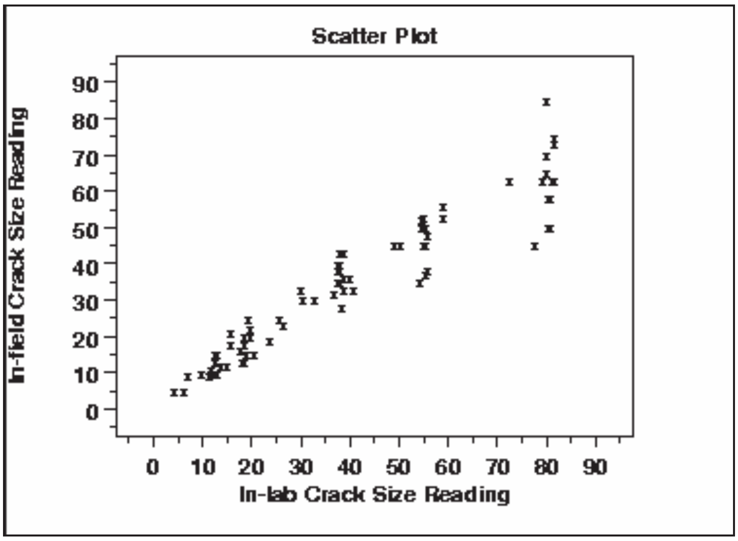
Scatter Diagrams

Scatter diagrams describe the correlation between 2 variables, so it can be identified whether those 2 variables are related.

After developing cause and affect diagrams, sometimes people use scatter diagrams to determine objectively whether the cause and the effect are related.

By looking at a glance to the scatter diagrams, process owner can analyze the positive/negative correlation of 2 variables. When the trend line goes from the bottom left to the upright, it means those 2 variables have positive correlation. Otherwise, if the trend line goes from the up left to the bottom right, and then it means those 2 variables have negative correlation.

Hence the scatter diagram graphs pairs of numerical data, with one variable on each axis, to look for a relationship between them. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the tighter the points will hug the line.



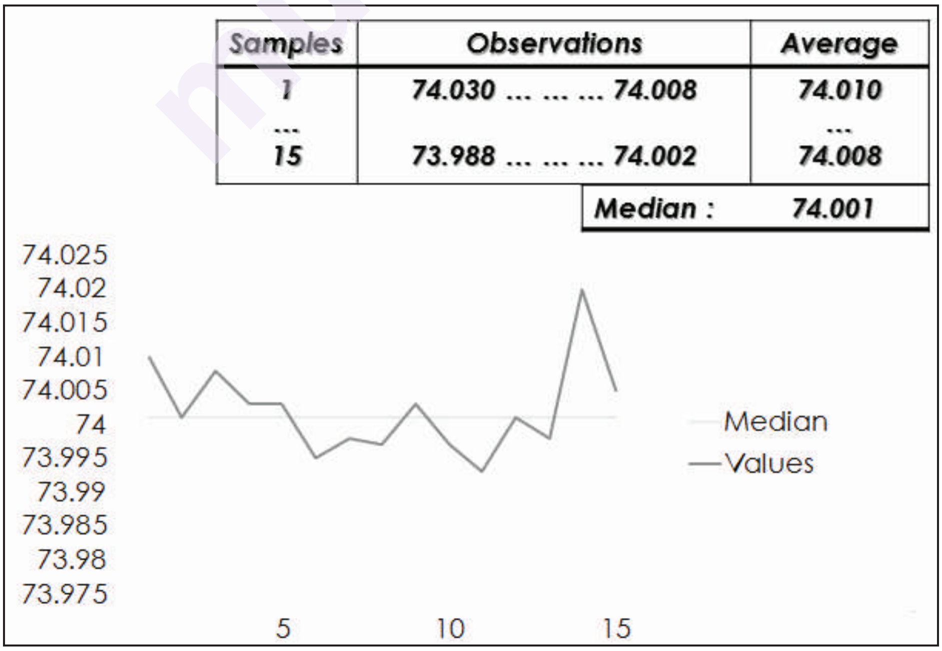
Control Charts

Control charts were invented by Walter A. Shewhart in 1920s. The maximum limit often called upper control limit and the minimum limit often called lower control limit, while the central line represents an estimate of the process mean.

If the value is within control area, the process can be stated as a controllable process. The value outside control area indicates that the process is no longer stable because of some variation causes. Thus, this process needs proper corrective actions to eliminate the sources of variation.

Hence, a control chart is a graph used to study how a process changes over time. Data are plotted in time order. A control chart always has a central line for the average, an upper line for the upper control limit and a lower line for the lower control limit. These lines are determined from historical data. By comparing current data to these lines, you can draw conclusions about whether the process variation is consistent (in control) or is unpredictable (out of control, affected by special causes of variation).

Control charts for variable data are used in pairs. The top chart monitors the average, or the centering of the distribution of data from the process. The bottom chart monitors the range, or the width of the distribution. If your data were shots in target practice, the average is where the shots are clustering, and the range is how tightly they are clustered. Control charts for attribute data are used singly.

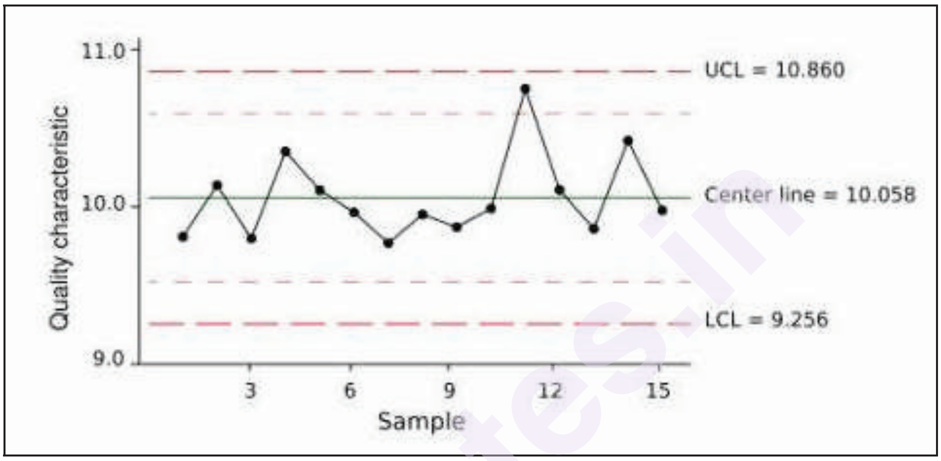


Samples	Observations	Average
1	74.030 … … … 74.008	74.010
…		…
15	73.988 … … … 74.002	74.008
	Median :	74.001

Control chart is the best tool for monitoring the performance of a process. These types of charts can be used for monitoring any processes related to function of the organization.

These charts allow you to identify the following conditions related to the process that has been monitored.

- Stability of the process
- Predictability of the process
- Identification of common cause of variation
- Special conditions where the monitoring party needs to react.



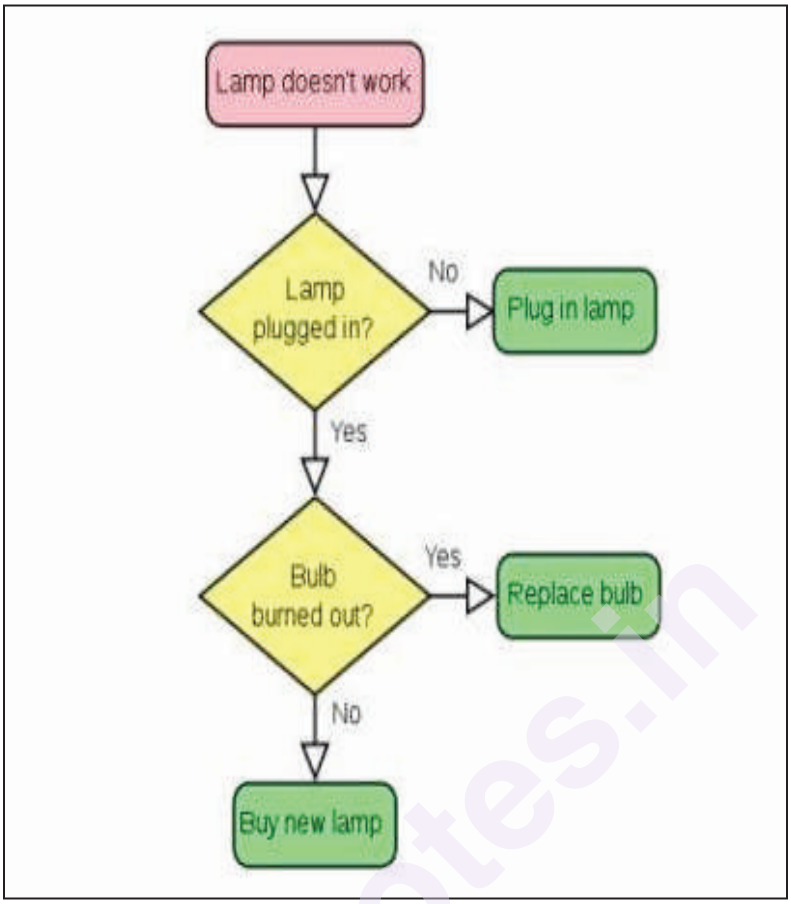
Flow Chart

Flow charts describe sequence of activities graphically in accomplishing a task. It must reflect the actual process rather than what the process owner wants it to be by generating flow charts, process owner can understand the process and working relationship between people and organization will be clarified.

Furthermore, flow charts will show the duplicated effort and other non-value added steps. So process owner can identify the target specific steps in order to make continuous improvement.

This is one of the basic quality tools that can be used for analyzing a sequence of events. The tool maps out a sequence of events that take place sequentially or in parallel. The flow chart can be used to understand a complex process in order to find the relationships and dependencies between events.

You can also get a brief idea about the critical path of the process and the events involved in the critical path. Flow charts can be used for any field to illustrate complex processes in a simple way. There are specific software tools developed for drawing flow charts, such as MS Visio.



7.6 LET US SUM UP

In this unit you have learnt the concepts of Statistical Quality Control, Causes of Variation in Quality and Techniques of SQC

7.7. EXERCISES

Q.1. What do you understand by the concept of quality and statistical quality control?

Q.2. Discuss different causes of variation in quality with examples from your field of experience or knowledge.

Q.3. Discuss in detail various techniques of statistical process control. Give some examples.

7.8 SUGGESTED READINGS

SQC section in any of the reference / text books

❖❖❖❖

Module - V
STATISTICAL QUALITY CONTROL

8

STATISTICAL QUALITY CONTROL CONTROL CHARTS

Control Charts, Control Charts for Variables, Control Charts for Attributes

Unit Structure :

8.1 INTRODUCTION

In this Unit-V - Chapter 5.2, we shall discuss the concept of control charts, control charts for variables and control charts for attributes.

8.2 OBJECTIVES

At the end of this unit the learners will be able to understand the

- concept of control charts
- control charts for variables
- control charts for attributes.

8.3 What are control charts?

The control chart is a graph used to study how a process changes over time. Data are plotted in time order. A control chart always has a central line for the average, an upper line for the upper control limit and a lower line for the lower control limit. These lines are determined from historical data. By comparing current data to these lines, you can draw conclusions about whether the process variation is consistent (in control) or is unpredictable (out of control, affected by special causes of variation).

Control charts for variable data are used in pairs. The top chart monitors the average, or the centering of the distribution of data from the process.

The bottom chart monitors the range, or the width of the distribution. If your data were shots in target practice, the average is where the shots are clustering, and the range is how tightly they are clustered. Control charts for attribute data are used singly.

When to Use a Control Chart

- When controlling ongoing processes by finding and correcting problems as they occur.
- When predicting the expected range of outcomes from a process.
- When determining whether a process is stable (in statistical control).
- When analyzing patterns of process variation from special causes (non-routine events) or common causes (built into the process).
- When determining whether your quality improvement project should aim to prevent specific problems or to make fundamental changes to the process.

Control Chart Basic Procedure

- Choose the appropriate control chart for your data.
- Determine the appropriate time period for collecting and plotting data.
- Collect data, construct your chart and analyze the data.
- Look for “out-of-control signals” on the control chart. When one is identified, mark it on the chart and investigate the cause. Document how you investigated, what you learned, the cause and how it was corrected.

Out-of-control signals

- A single point outside the control limits. In Figure 1, point sixteen is above the UCL (upper control limit).
- Two out of three successive points are on the same side of the centerline and farther than 2 σ from it. In Figure 1, point 4 sends that signal.

- Four out of five successive points are on the same side of the centerline and farther than 1 σ from it. In Figure 1, point 11 sends that signal.
- A run of eight in a row are on the same side of the centerline. Or 10 out of 11, 12 out of 14 or 16 out of 20. In Figure 1, point 21 is eighth in a row above the centerline.
- Obvious consistent or persistent patterns that suggest something unusual about your data and your process.

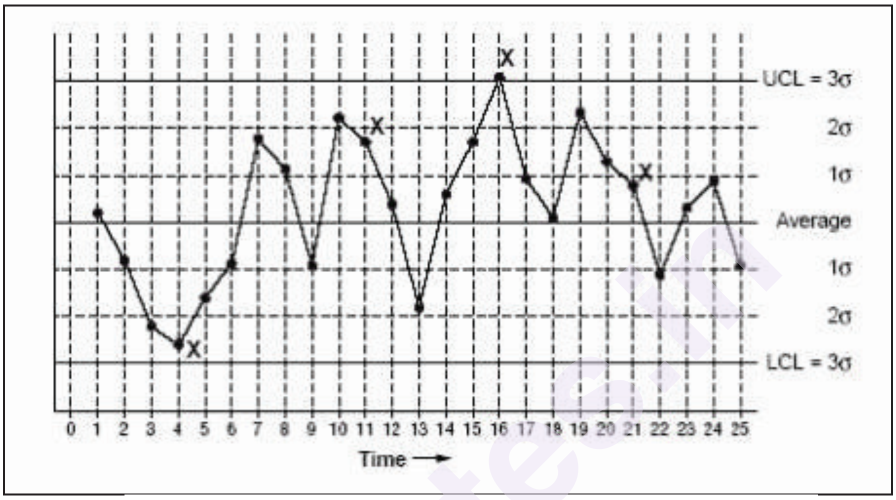
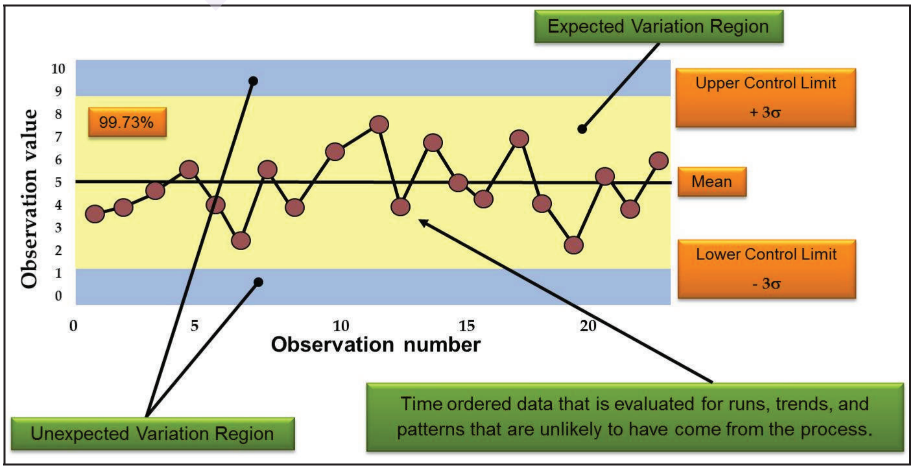
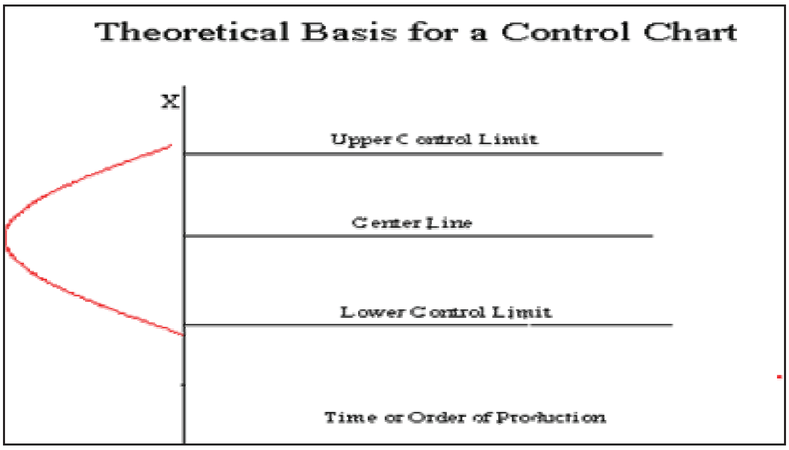
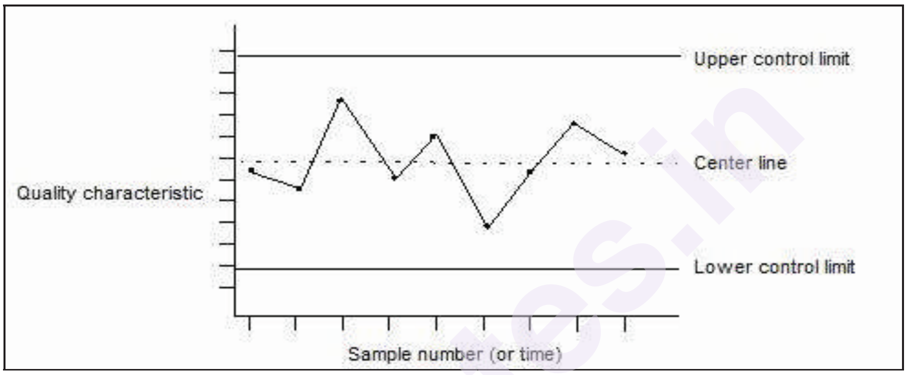


Figure 1: Control Chart: Out-of-Control Signals

- Continue to plot data as they are generated. As each new data point is plotted, check for new out-of-control signals.
- When you start a new control chart, the process may be out of control. If so, the control limits calculated from the first 20 points are conditional limits. When you have at least 20 sequential points from a period when the process is operating in control, recalculate control limits.



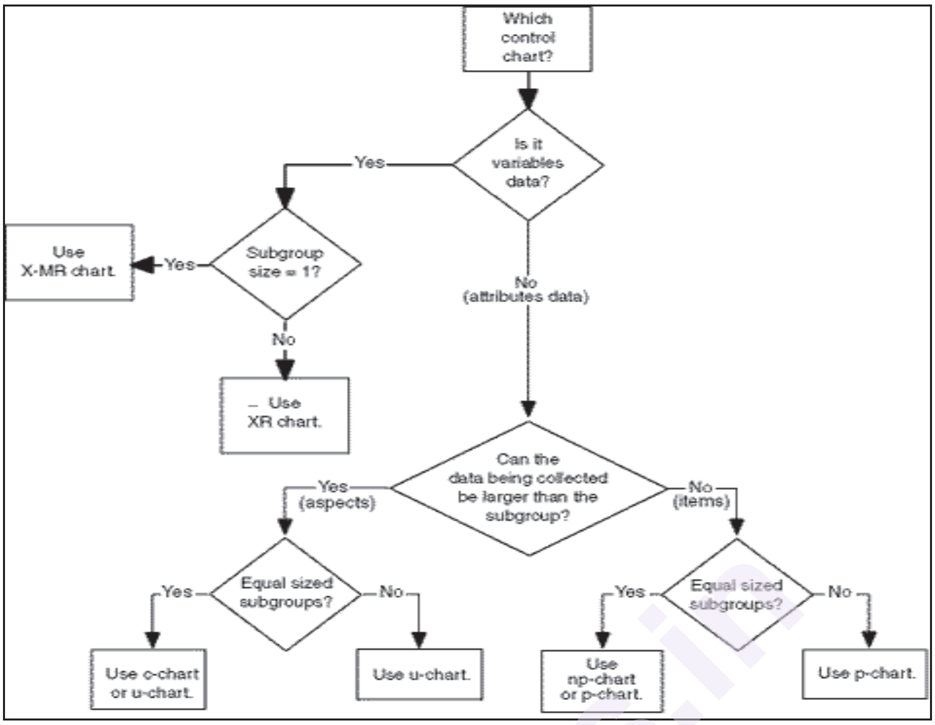

Theoretical Basis for a Control Chart
X
Upper Control Limit
Center Line
Lower Control Limit
Time or Order of Production



Types of Control Chart :

Control chart for variables are used to monitor characteristics that can be measured, e.g. length, weight, diameter, time, etc.

Control charts for attributes are used to monitor characteristics that have discrete values and can be counted, e.g. % defective, number of flaws in a shirt, number of broken eggs in a box, etc.



8.4 CONTROL CHARTS FOR VARIABLES

The different types of control charts for variables are as under:

- X-bar and R chart (also called averages and range chart)
- X-bar and s chart
- moving average–moving range chart (also called MA–MR chart)
- target charts (also called difference charts, deviation charts and nominal charts)
- CUSUM (cumulative sum chart)
- EWMA (exponentially weighted moving average chart)
- multivariate chart

In this unit we study only two charts for variables, i.e. X-bar and R-Chart.

X-bar Chart

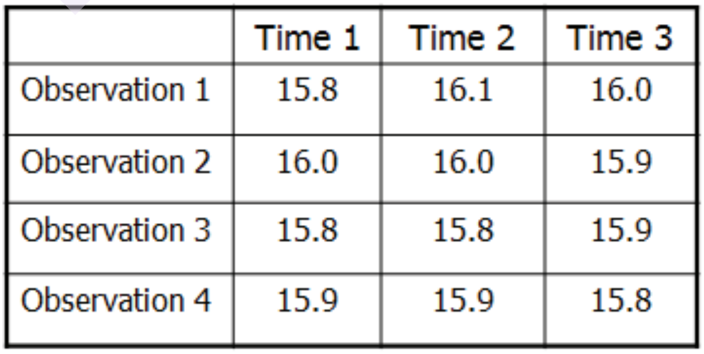
- The X-bar chart monitors the process location over time, based on the average of a series of observations, called a subgroup.
- X-bar / Range charts are used when you can rationally collect measurements in groups (subgroups) of between two and ten observations. Each subgroup represents a "snapshot" of the process at a given point in time. The charts' x-axes are time based, so that the charts show a history of the process. For this

reason, data should be time-ordered; that is, entered in the sequence from which it was generated. If this is not the case, then trends or shifts in the process may not be detected, but instead attributed to random (common cause) variation.

- For subgroup sizes greater than ten, use X-bar / Sigma charts, since the range statistic is a poor estimator of process sigma for large subgroups. In fact, the subgroup sigma is always a better estimate of subgroup variation than subgroup range. The popularity of the Range chart is only due to its ease of calculation, dating to its use before the advent of computers.
- For subgroup sizes equal to one, an Individual-X / Moving Range chart can be used, as well as EWMA or CuSum charts.
- X-bar Charts are efficient at detecting relatively large shifts in the process average, typically shifts of +-1.5 sigma or larger. The larger the subgroup, the more sensitive the chart will be to shifts, providing a Rational Subgroup can be formed. For more sensitivity to smaller process shifts, use an EWMA or CuSum chart.
- X-bar Charts track the tracks the central tendency (the average value observed) over time.

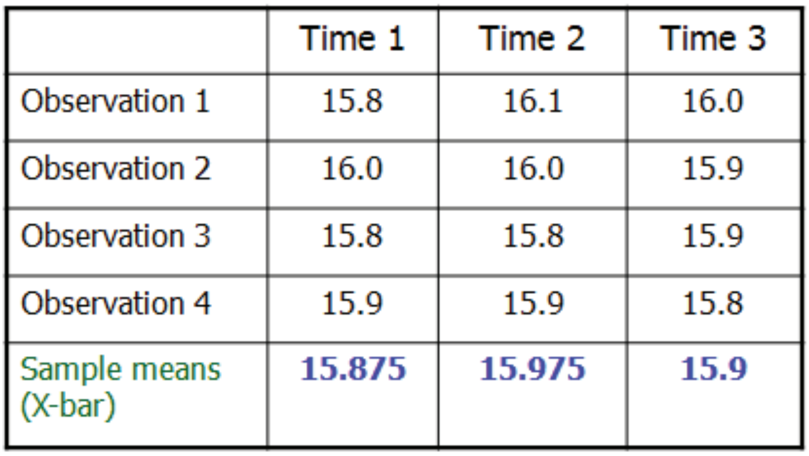
Illustrative Example on X-bar Chart

A quality control inspector at the Cocoa Fizz soft drink company has taken three samples with four observations each of the volume of bottles filled. If the standard deviation of the bottling operation is 0.2 ounces, use the data below to develop control charts with limits of 3 standard deviations for the 16 oz. bottling operation.



	Time 1	Time 2	Time 3
Observation 1	15.8	16.1	16.0
Observation 2	16.0	16.0	15.9
Observation 3	15.8	15.8	15.9
Observation 4	15.9	15.9	15.8

Step 1: Calculate the Mean of Each Sample



	Time 1	Time 2	Time 3
Observation 1	15.8	16.1	16.0
Observation 2	16.0	16.0	15.9
Observation 3	15.8	15.8	15.9
Observation 4	15.9	15.9	15.8
Sample means (X-bar)	**15.875**	**15.975**	**15.9**

Step 2: Calculate the Standard Deviation of the Sample Mean.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \left(\frac{.2}{\sqrt{4}}\right) = .1$$

Step 3: Calculate CL, UCL, LCL.

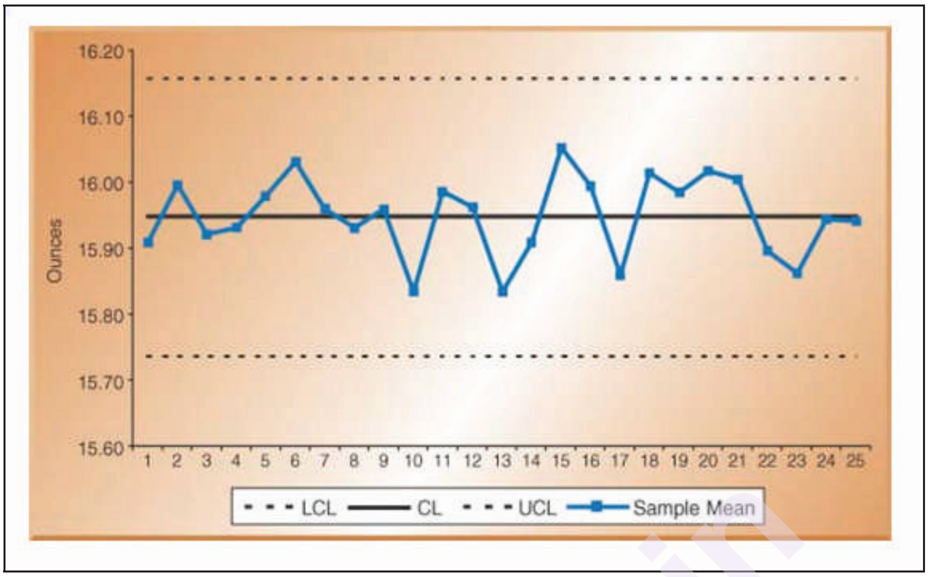
- Center line (x-double bar):

$$\bar{\bar{x}} = \frac{15.875 + 15.975 + 15.9}{3} = 15.92$$

- Control limits for ±3σ limits (z = 3):

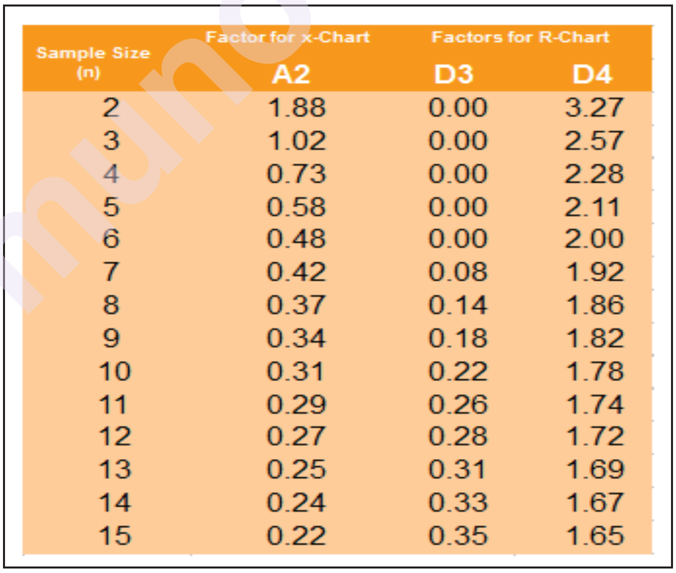
$$\text{UCL}_{\bar{x}} = \bar{\bar{x}} + z\sigma_{\bar{x}} = 15.92 + 3(.1) = 16.22$$
$$\text{LCL}_{\bar{x}} = \bar{\bar{x}} - z\sigma_{\bar{x}} = 15.92 - 3(.1) = 15.62$$

Step 4: Draw the Chart.



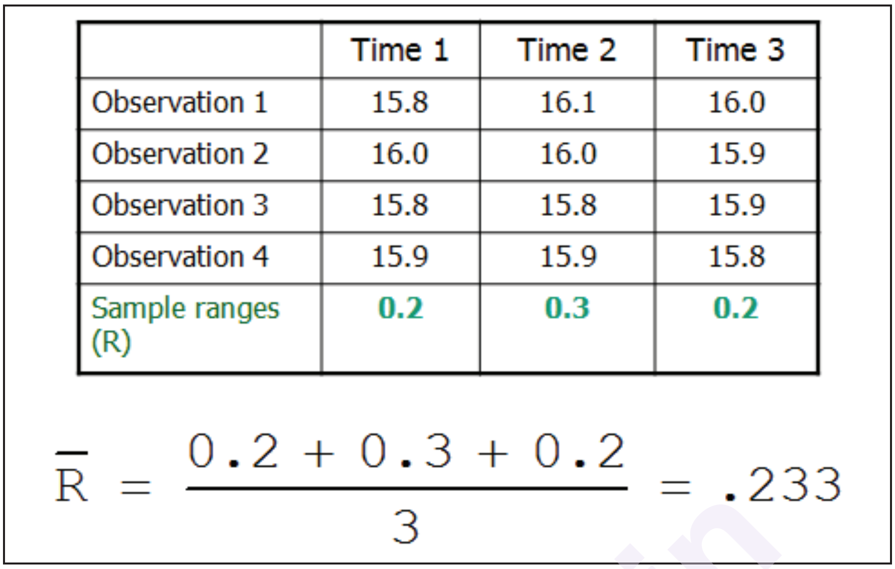
Illustrative Example on an Alternative Method for the X-bar Chart Using R-bar and the A2 Factor :

Use this method when sigma for the process distribution is not known. Use factor A2 from the below table:



Sample Size (n)	Factor for x-Chart	Factors for R-Chart	
	A2	D3	D4
2	1.88	0.00	3.27
3	1.02	0.00	2.57
4	0.73	0.00	2.28
5	0.58	0.00	2.11
6	0.48	0.00	2.00
7	0.42	0.08	1.92
8	0.37	0.14	1.86
9	0.34	0.18	1.82
10	0.31	0.22	1.78
11	0.29	0.26	1.74
12	0.27	0.28	1.72
13	0.25	0.31	1.69
14	0.24	0.33	1.67
15	0.22	0.35	1.65

Step 1: Calculate the Range of Each Sample and Average Range



	Time 1	Time 2	Time 3
Observation 1	15.8	16.1	16.0
Observation 2	16.0	16.0	15.9
Observation 3	15.8	15.8	15.9
Observation 4	15.9	15.9	15.8
Sample ranges (R)	**0.2**	**0.3**	**0.2**

$$\overline{R} = \frac{0.2 + 0.3 + 0.2}{3} = .233$$

Step 2: Calculate CL, UCL, LCL

- **Center line:**

$$CL = \overline{\overline{x}} = \frac{15.875 + 15.975 + 15.9}{3} = 15.92$$

- **Control limits for ±3σ limits:**

$$UCL_{\overline{x}} = \overline{\overline{x}} + A_2\overline{R} = 15.92 + (0.73).233 = 16.09$$
$$LCL_{\overline{x}} = \overline{\overline{x}} - A_2\overline{R} = 15.92 - (0.73).233 = 15.75$$

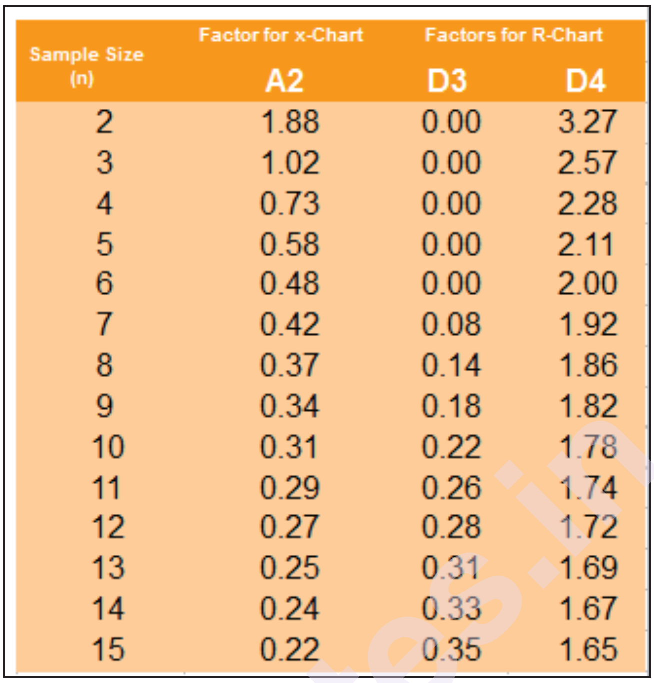
Illustrative Example on R Chart

Center Line and Control Limit calculations:

$$CL = \overline{R} = \frac{0.2 + 0.3 + 0.2}{3} = .233$$

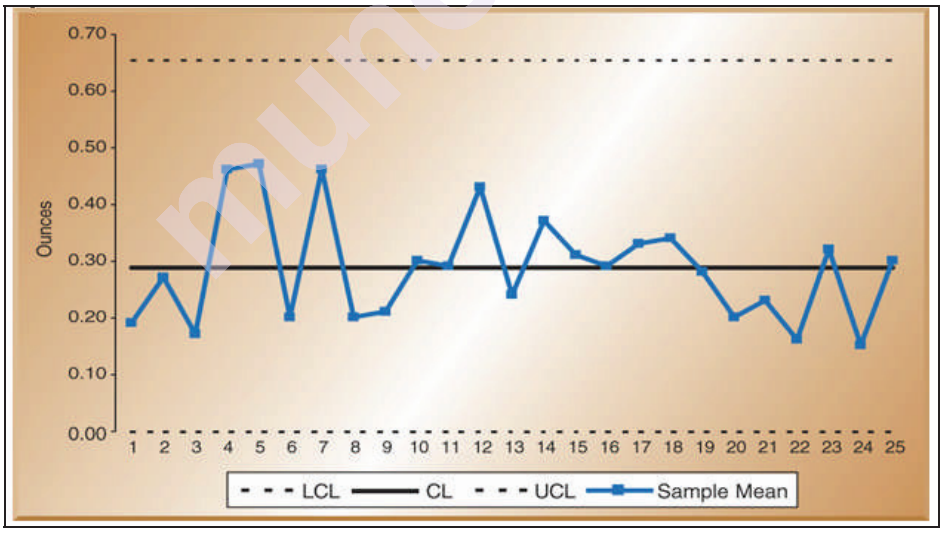
$$UCL = D_4\overline{R} = 2.28(.233) = .53$$
$$LCL = D_3\overline{R} = 0.0(.233) = 0.0$$

Following table can be used to get the values of D3 and D4 for a sample of size n. These constants are given in the examination.



Sample Size (n)	Factor for x-Chart	Factors for R-Chart	
	A2	D3	D4
2	1.88	0.00	3.27
3	1.02	0.00	2.57
4	0.73	0.00	2.28
5	0.58	0.00	2.11
6	0.48	0.00	2.00
7	0.42	0.08	1.92
8	0.37	0.14	1.86
9	0.34	0.18	1.82
10	0.31	0.22	1.78
11	0.29	0.26	1.74
12	0.27	0.28	1.72
13	0.25	0.31	1.69
14	0.24	0.33	1.67
15	0.22	0.35	1.65

R-Bar Control Chart



8.5 CONTROL CHARTS FOR ATTRIBUTES

P-Charts & C-Charts

Use P-Charts for quality characteristics that are discrete and involve yes/no or good/bad decisions.

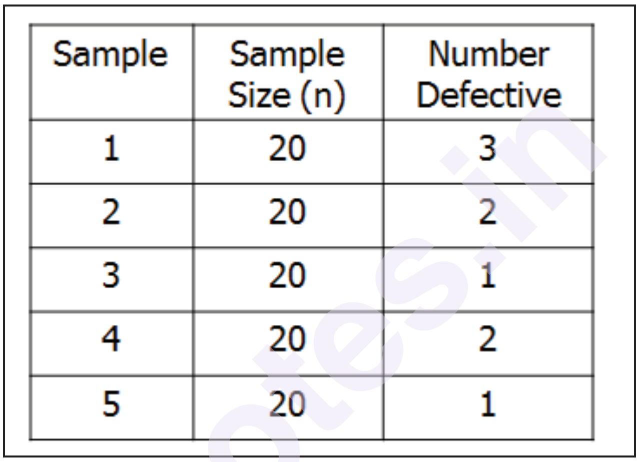
- Percent of leaking caulking tubes in a box of 48
- Percent of broken eggs in a carton

Use C-Charts for discrete defects when there can be more than one defect per unit.

- Number of flaws or stains in a carpet sample cut from a production run
- Number of complaints per customer at a hotel

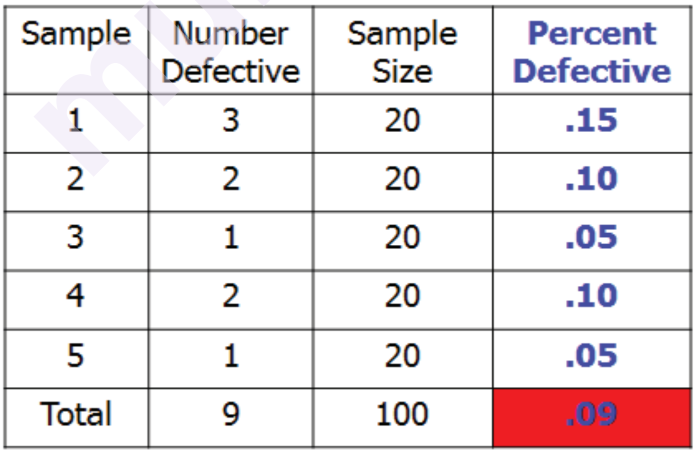
Illustrative Example on P Chart

A Production manager for a tire company has inspected the number of defective tires in five random samples with 20 tires in each sample. The table below shows the number of defective tires in each sample of 20 tires.



Sample	Sample Size (n)	Number Defective
1	20	3
2	20	2
3	20	1
4	20	2
5	20	1

Step 1: Calculate the Percent defective of Each Sample and the Overall Percent Defective (P-Bar).



Sample	Number Defective	Sample Size	**Percent Defective**
1	3	20	**.15**
2	2	20	**.10**
3	1	20	**.05**
4	2	20	**.10**
5	1	20	**.05**
Total	9	100	**.09**

Step 2: Calculate the Standard Deviation of P.

$$\sigma_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{(.09)\,(.91)}{20}} = 0.064$$

Step 3: Calculate CL, UCL, LCL

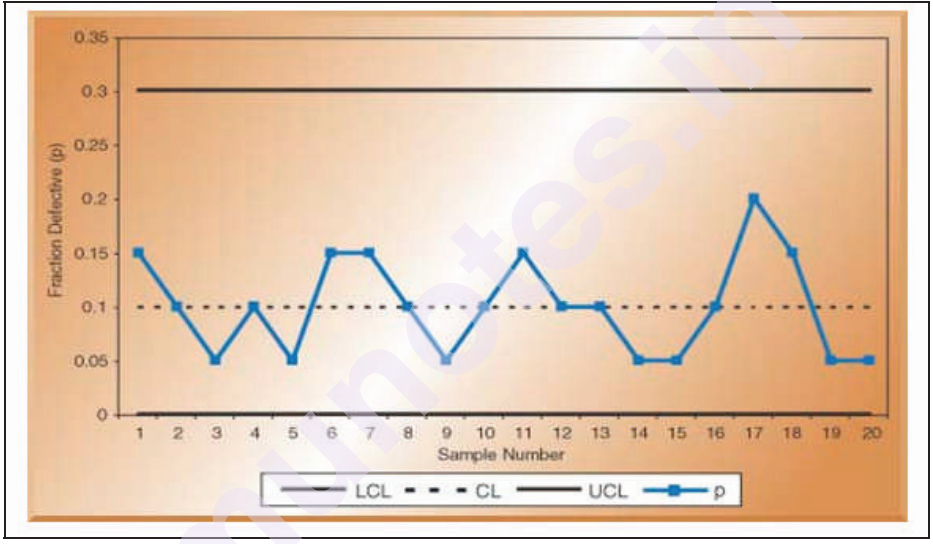
- **Center line (p bar):**

$$\boxed{CL = \bar{p} = .09}$$

- **Control limits for ±3σ limits:**

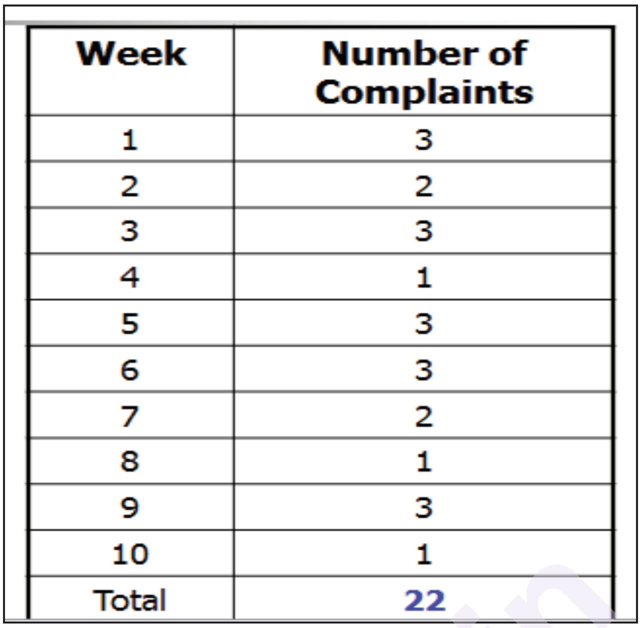
$$UCL = \bar{p} + z(\sigma_p) = .09 + 3(.064) = .282$$
$$LCL = \bar{p} - z(\sigma_p) = .09 - 3(.064) = -.102 = 0$$

Step 4: Draw the Chart



Illustrative Example on C Chart

The number of weekly customer complaints is monitored in a large hotel. Develop a three sigma control limits For a C-Chart using the data table below:



Week	Number of Complaints
1	3
2	2
3	3
4	1
5	3
6	3
7	2
8	1
9	3
10	1
Total	22

- **Center line (c bar):**

$$\mathrm{CL} = \frac{\#\text{complaints}}{\#\text{ of samples}} = \frac{22}{10} = 2.2$$

- **Control limits for ±3σ limits:**

$$\mathrm{UCL} = \bar{c} + z\sqrt{\bar{c}} = 2.2 + 3\sqrt{2.2} = 6.65$$

$$\mathrm{LCL} = \bar{c} - z\sqrt{\bar{c}} = 2.2 - 3\sqrt{2.2} = -2.25 = 0$$

8.6 LET US SUM UP

In this unit you have learnt the concepts of control charts, Control charts for variables and Control charts for attributes.

Service Organizations have lagged behind manufacturers in the use of statistical quality control.

Statistical measurements are required and it is more difficult to measure the quality of a service:

- Services produce more intangible products
- Perceptions of quality are highly subjective.

A way to deal with service quality is to devise quantifiable measurements of the service element:

- Check-in time at a hotel
- Number of complaints received per month at a restaurant
- Number of telephone rings before a call is answered
- Acceptable control limits can be developed and charted .

8.7. EXERCISES

Q.1. The following collection of data represents samples of the amount of force applied in a gluing process. Determine if the process is in control by calculating the appropriate upper and lower control limits of the X-bar and R charts:

Sample	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5
1	10.68	10.689	10.776	10.798	10.714
2	10.79	10.86	10.601	10.746	10.779
3	10.78	10.667	10.838	10.785	10.723
4	10.59	10.727	10.812	10.775	10.73
5	10.69	10.708	10.79	10.758	10.671
6	10.75	10.714	10.738	10.719	10.606
7	10.79	10.713	10.689	10.877	10.603
8	10.74	10.779	10.11	10.737	10.75
9	10.77	10.773	10.641	10.644	10.725
10	10.72	10.671	10.708	10.85	10.712
11	10.79	10.821	10.764	10.658	10.708
12	10.62	10.802	10.818	10.872	10.727
13	10.66	10.822	10.893	10.544	10.75
14	10.81	10.749	10.859	10.801	10.701
15	10.66	10.681	10.644	10.747	10.728

Ans:

$\bar{x}$ Chart Control Limits

$$UCL = \bar{\bar{x}} + A_2\bar{R}$$

$$LCL = \bar{\bar{x}} - A_2\bar{R}$$

R Chart Control Limits

$$UCL = D_4\bar{R}$$

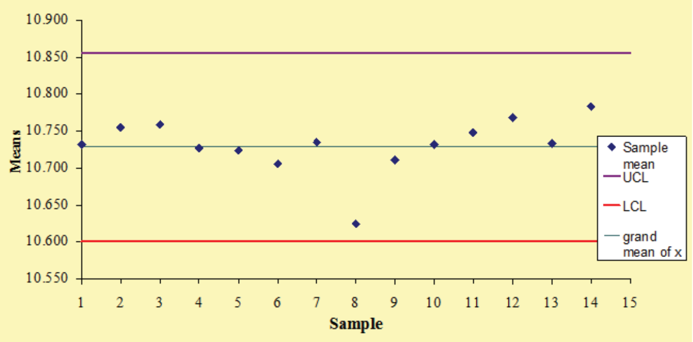
$$LCL = D_3\bar{R}$$

n	A2	D3	D4
2	1.88	0	3.27
3	1.02	0	2.57
4	0.73	0	2.28
5	0.58	0	2.11
6	0.48	0	2.00
7	0.42	0.08	1.92
8	0.37	0.14	1.86
9	0.34	0.18	1.82
10	0.31	0.22	1.78
11	0.29	0.26	1.74

Q.1. Ans:

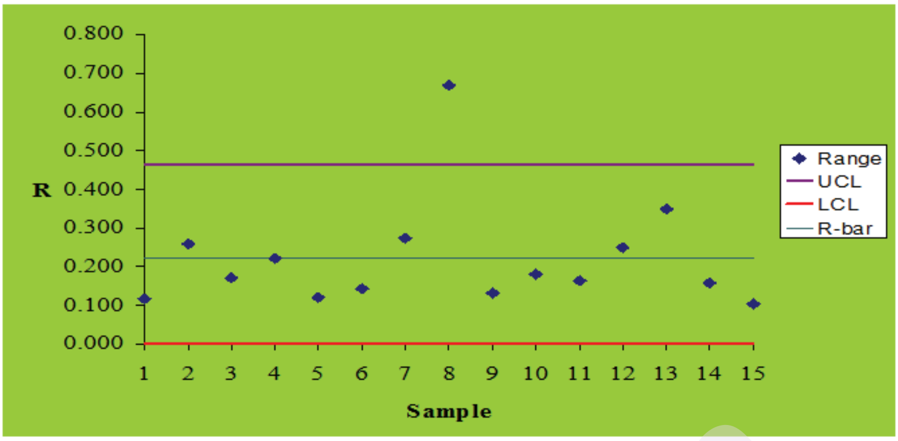
$$UCL = \bar{\bar{x}} + A_2\bar{R} = 10.728 + .58(0.2204) = 10.856$$

$$LCL = \bar{\bar{x}} - A_2\bar{R} = 10.728 - .58(0.2204) = 10.601$$



$$UCL = D_4\overline{R} = (2.11)(0.2204) = \mathbf{0.46504}$$

$$LCL = D_3\overline{R} = (0)(0.2204) = \mathbf{0}$$



Q.2. Table gives the number of units produced by a company each week during part of 1994 which require rework at certain stages of production because of faults. Rework is additional work required to bring a substandard unit up to standard and is an additional cost one wishes to avoid. The weeks are consecutive except for the last one (in 1995) when the factory closed down over the Christmas holiday period. Chart the percentages requiring rework over time to see whether the process is in control.

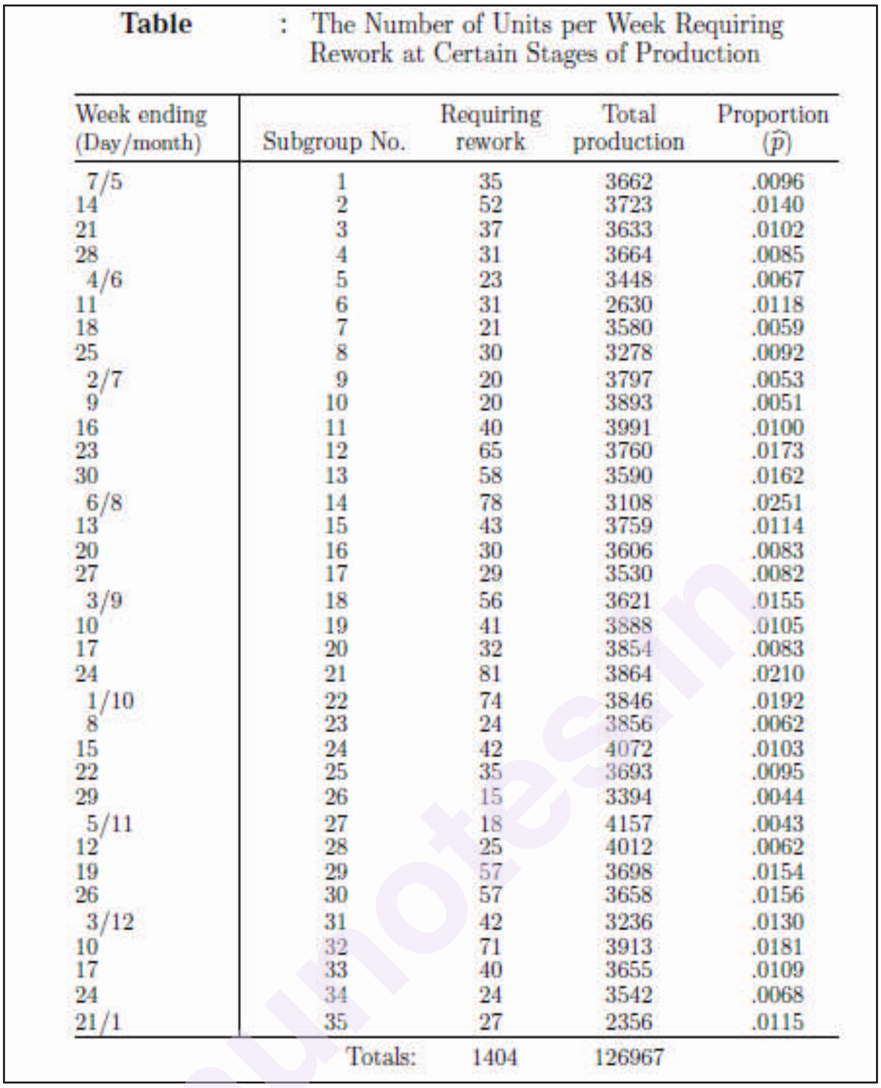
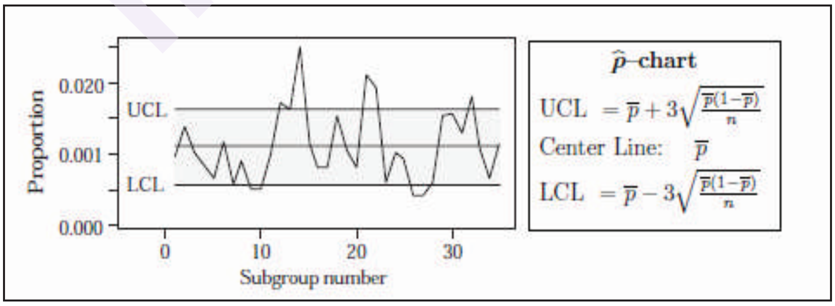


Table : The Number of Units per Week Requiring Rework at Certain Stages of Production

Week ending (Day/month)	Subgroup No.	Requiring rework	Total production	Proportion ($\widehat{p}$)
7/5	1	35	3662	.0096
14	2	52	3723	.0140
21	3	37	3633	.0102
28	4	31	3664	.0085
4/6	5	23	3448	.0067
11	6	31	2630	.0118
18	7	21	3580	.0059
25	8	30	3278	.0092
2/7	9	20	3797	.0053
9	10	20	3893	.0051
16	11	40	3991	.0100
23	12	65	3760	.0173
30	13	58	3590	.0162
6/8	14	78	3108	.0251
13	15	43	3759	.0114
20	16	30	3606	.0083
27	17	29	3530	.0082
3/9	18	56	3621	.0155
10	19	41	3888	.0105
17	20	32	3854	.0083
24	21	81	3864	.0210
1/10	22	74	3846	.0192
8	23	24	3856	.0062
15	24	42	4072	.0103
22	25	35	3693	.0095
29	26	15	3394	.0044
5/11	27	18	4157	.0043
12	28	25	4012	.0062
19	29	57	3698	.0154
26	30	57	3658	.0156
3/12	31	42	3236	.0130
10	32	71	3913	.0181
17	33	40	3655	.0109
24	34	24	3542	.0068
21/1	35	27	2356	.0115
	Totals:	1404	126967	

Q.2. Ans:



UCL $= \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$

Center Line: $\overline{p}$

LCL $= \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$

Q.3. A production manager for a tire company has inspected the number of defective tires in five random samples with 20 tires in each sample. The table below shows the number of defective tires in each sample of 20 tires. Calculate the control limits.

Sample	Number of Defective Tires	Number of Tires in each Sample	Proportion Defective
1	3	20	.15
2	2	20	.10
3	1	20	.05
4	2	20	.10
5	2	20	.05
Total	9	100	.09

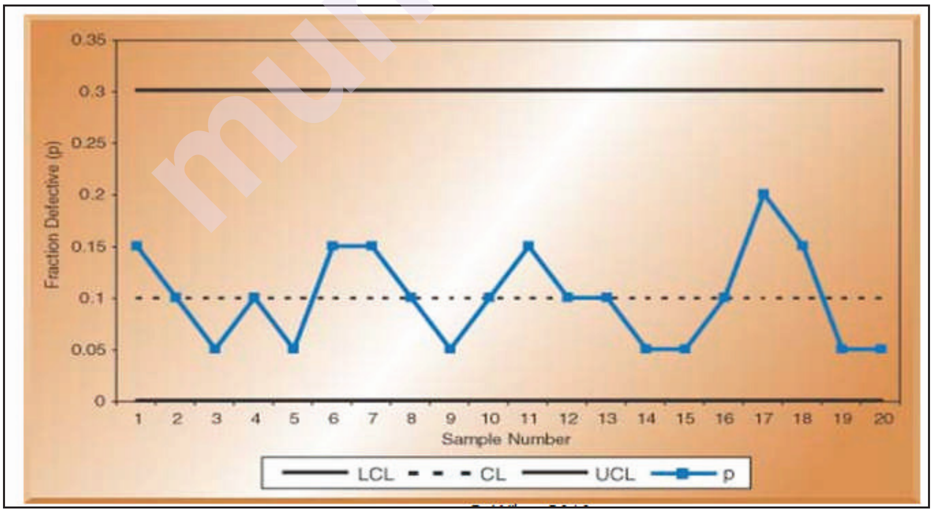
Q.3. Ans:

$$CL = \bar{p} = \frac{\#\text{Defectives}}{\text{Total Inspected}} = \frac{9}{100} = .09$$

$$\sigma_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{(.09)(.91)}{20}} = 0.64$$

$$UCL_p = \bar{p} + z(\sigma) = .09 + 3(.064) = .282$$

$$LCL_p = \bar{p} - z(\sigma) = .09 - 3(.064) = -.102 = 0$$



8.8 SUGGESTED READINGS

SQC section in any of the reference / text books

❖❖❖❖

Module - V
STATISTICAL QUALITY CONTROL

9

ACCEPTANCE SAMPLING AND INTRODUCTION TO SIX-SIGMA

Acceptance Sampling and, Introduction to Six-Sigma

Unit Structure :

9.1 INTRODUCTION

In this Unit-V - Chapter 5.2, we shall discuss the concept of acceptance sampling and six sigma.

9.2 OBJECTIVES

At the end of this unit the learners will be able to understand the

- Acceptance sampling
- Six-Sigma

9.3 WHAT IS ACCEPTANCE SAMPLING?

Acceptance sampling is an important field of statistical quality control that was popularized by Dodge and Romig and originally applied by the U.S. military to the testing of bullets during World War II. If every bullet was tested in advance, no bullets would be left to ship. If, on the other hand, none were tested, malfunctions might occur in the field of battle, with potentially disastrous results.

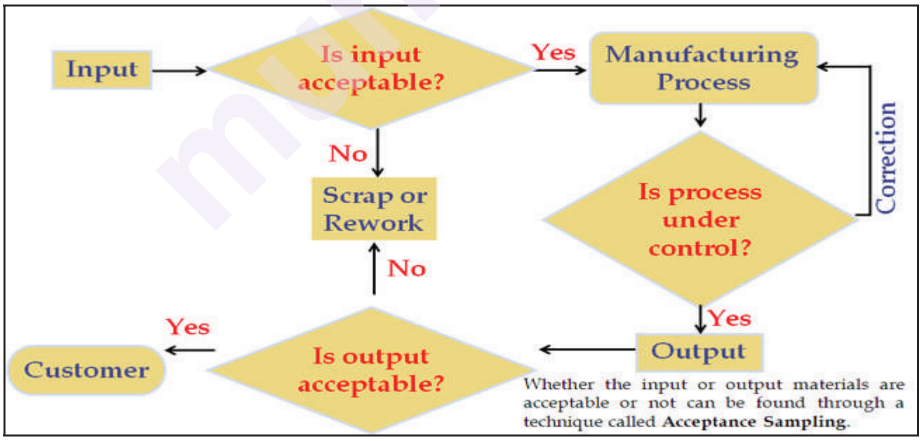
Dodge reasoned that a sample should be picked at random from the lot, and on the basis of information that was yielded by the sample, a decision should be made regarding the disposition of the lot. In general, the decision is either to accept or to reject the lot. This process is called Lot Acceptance Sampling or just Acceptance Sampling.

Acceptance sampling is "the middle of the road" approach between no inspection and 100% inspection. There are two major classifications of acceptance plans: by attributes ("go, no-go") and by variables. The attribute case is the most common for acceptance sampling. A point to remember is that the main purpose of acceptance sampling is to decide whether or not the lot is likely to be acceptable, not to estimate the quality of the lot.

Acceptance sampling is employed when one or several of the following hold:

- Testing is destructive
- The cost of 100% inspection is very high
- 100% inspection takes too long.

In situations where Producers and Consumers are willing to put up with shipments of goods which contain defects, the issue then becomes how to keep Lots (shipments) with an "excessive" number of defects from being accepted. In order to keep Appraisal costs down, sampling schemes are devised to sample items from the Lot and then to accept or reject the Lot based upon the number of defective items found in the sample.



Whether the input or output materials are acceptable or not can be found through a technique called **Acceptance Sampling**.

Important Definitions

Acceptance Sampling: An inspection procedure used to determine whether to accept or reject a specific quantity of materials.

Acceptable Quality Level (AQL): The quality level desired by the consumer.

Producer's risk (α): The risk that the sampling plan will fail to verify an acceptable lot's quality and, thus, reject it (a type I error).

Lot tolerance proportion defective (LTPD): The worst level of quality that the consumer can tolerate.

Single-sampling plan: A decision to accept or reject a lot based on the results of one random sample from the lot.

Double-sampling plan: A plan in which management specifies two sample sizes and two acceptance numbers; if the quality of the lot is very good or very bad, the consumer can make a decision to accept or reject the lot on the basis of the first sample, which is smaller than in the single-sampling plan.

Sequential-sampling plan: A plan in which the consumer randomly selects items from the lot and inspects them one by one.

Operating characteristic (OC) curve: A graph that describes how well a sampling plan discriminates between good and bad lots.

Notation

- N= number of items in the Lot.
- n= number of items in the random sample of items taken from the Lot.
- D= number (unknown) of items in the Lot which are defective.
- r= number of defective items in the random sample of size n.
- c= acceptance number; if r is less than or equal to c then the Lot is deemed to be of Acceptable Quality, otherwise the Lot is rejected.

Acceptance Sampling Plan

Assuming,

$$p = D/N$$

$$P_a(p) = \mathrm{P}(accept\_lot\_with\_p = proportion\_defective)$$

We can evaluate a sampling plan with its corresponding Operating Characteristic Curve or OC curve.

An Acceptance Sampling Plan is defined by:

- N= Lot size
- n=sample size
- c=acceptance number

Calculating probabilities for the OC curve

The probabilities used in calculating values for the OC curve are calculated using the Hyper geometric Distribution. Given that

$$p = D/N$$

then the probability that the number of defectives in the sample of size n is equal to r is given by

$$\frac{\begin{pmatrix} N-D \\ n-r \end{pmatrix}\begin{pmatrix} D \\ r \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Acceptable Quality Level-AQL

If the Lot has a proportion of defectives lower than the AQL, i.e.

$$p \le AQL$$

then the Lot is deemed to be of Acceptable Quality Level by Producer and Consumer.

OC curve and Producer and Consumer Risk

The Probability that we accept a Lot is

$$P_a(p) = P(r \le c)$$

for a given proportion of defectives

$$p = D/N$$

which is calculated using the Hyper geometric Distribution as given.

If the proportion defective is less than or equal to the AQL, then the Producer runs the risk of having the Lot rejected which is:

$$1 - P_a(p), p \le AQL$$

If the proportion of defectives is greater than the AQL, then the Consumer runs the risk of accepting a Lot of unacceptable quality which is:

$$P_a(p), p > AQL$$

Acceptance sampling is an inspection procedure used to determine whether to acceptor reject a specific quantity of material. As more firms initiate total quality management (TQM) programs and work closely with suppliers to ensure high levels of quality, the need for acceptance sampling will decrease. The TQM concept is that no defects should be passed from a producer to a customer, whether the customer is an external or internal customer. However, in reality, many firms must still rely on checking their materials inputs. The basic procedure is straightforward.

1. A random sample is taken from a large quantity of items and tested or measured relative to the quality characteristic of interest.
2. If the sample passes the test, the entire quantity of items is accepted.
3. If the sample fails the test, either (a) the entire quantity of items is subjected to 100 percent inspection and all defective items repaired or replaced or (b) the entire quantity is returned to the supplier.

Acceptance Sampling Plan Decisions :

Acceptance sampling involves both the producer (and supplier) of materials and the consumer (or buyer). Consumers need acceptance sampling to limit the risk of rejecting good-quality materials or accepting bad-quality materials. Consequently, the consumer, sometimes in conjunction with the producer through contractual agreements, specifies the parameters of the plan. Any company can be both a producer of goods purchased by another company and a consumer of goods or raw materials supplied by another company.

Quality and Risk Decisions :

Two levels of quality are considered in the design of an acceptance sampling plan. The first is the **acceptable quality level (AQL)**, or the quality level desired by the *consumer*. The producer of the item strives to achieve the AQL, which typically is written into a contract or purchase order. For example, a contract might call for a quality level not to exceed one defective unit in 10,000, or an AQL of 0.0001. The **producer's risk (α)** is the risk that the sampling plan will fail to verify an acceptable lot's quality and, thus, reject it—a type I error. Most often the producer's risk is set at 0.05, or 5 percent.

Although producers are interested in low risk, they often have no control over the consumer's acceptance sampling plan. Fortunately, the consumer also is interested in a low producer's risk because sending good materials back to the producer (1) disrupts the consumer’s production process and increases the likelihood of shortages in materials, (2) adds unnecessarily to the lead time for finished products or services, and (3) creates poor relations with the producer.

The second level of quality is the **lot tolerance proportion defective (LTPD)**, or the worst level of quality that the consumer can tolerate. The LTPD is a definition of bad quality that the consumer would like to reject. Recognizing the high cost of defects, operations managers have become more cautious about accepting materials of poor quality from suppliers.

Thus, sampling plans have lower LTPD values than in the past. The probability of accepting a lot with LTPD quality is the **consumer’s risk (β)**, or the type II error of the plan. A common value for the consumer’s risk is 0.10, or 10 percent.

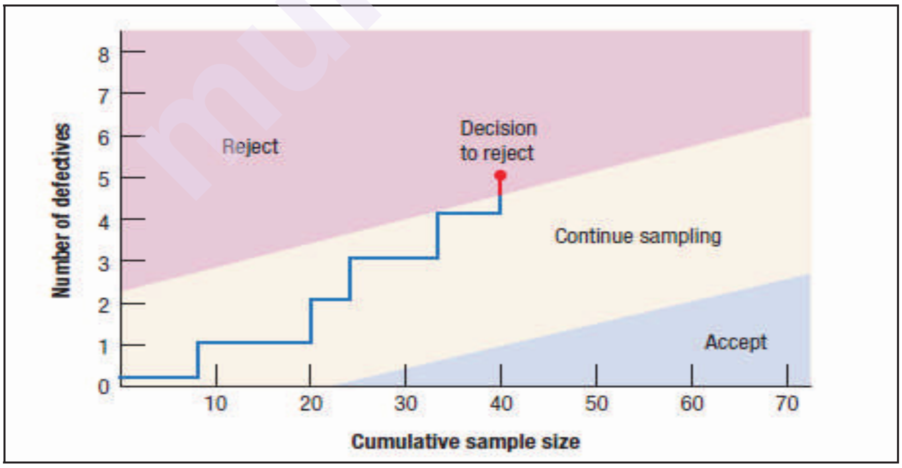
Sampling Plans

All sampling plans are devised to provide a specified producer’s and consumer’s risk. However, it is in the consumer’s best interest to keep the average number of items inspected (ANI) to a minimum because that keeps the cost of inspection low. Sampling plans differ with respect to ANI. Three often-used attribute sampling plans are the single-sampling plan, the double-sampling plan, and the sequential-sampling plan. Analogous plans also have been devised for variable measures of quality.

Single-Sampling Plan: The single-sampling plan is a decision rule to accept or reject a lot based on the results of one random sample from the lot. The procedure is to take a random sample of size (*n*) and inspect each item. If the number of defects does not exceed a specified acceptance number (*c*), the consumer accepts the entire lot. Any defects found in the sample are either repaired or returned to the producer. If the number of defects in the sample is greater than *c*, the consumer subjects the entire lot to 100 percent inspection or rejects the entire lot and returns it to the producer. The single-sampling plan is easy to use but usually results in a larger ANI than the other plans. After briefly describing the other sampling plans, we focus our discussion on this plan.

Double-Sampling Plan: In a double-sampling plan, management specifies two sample sizes (n1 and n2) and two acceptance numbers (c1 and c2). If the quality of the lot is very good or very bad, the consumer can make a decision to accept or reject the lot on the basis of the first sample, which is smaller than in the single-sampling plan. To use the plan, the consumer takes a random sample of size n1. If the number of defects is less than or equal to c1, the consumer accepts the lot. If the number of defects is greater than c2, the consumer rejects the lot. If the number of defects is between c1 and c2, the consumer takes a second sample of size n2. If the combined number of defects in the two samples is less than or equal to c2, the consumer accepts the lot. Otherwise, it is rejected. A double-sampling plan can significantly reduce the costs of inspection relative to a single-sampling plan for lots with a very low or very high proportion defective because a decision can be made after taking the first sample. However, if the decision requires two samples, the sampling costs can be greater than those for the single-sampling plan.

Sequential-Sampling Plan: A further refinement of the double-sampling plan is the sequential-sampling plan, in which the consumer randomly selects items from the lot and inspects them one by one. Each time an item is inspected, a decision is made to (1) reject the lot,(2) accept the lot, or (3) continue sampling, based on the cumulative results so far. The analyst plots the total number of defectives against the cumulative sample size, and if the number of defectives is less than a certain acceptance number (c1), the consumer accepts the lot. If the number is greater than another acceptance number (c2), the consumer rejects the lot. If the number is somewhere between the two, another item is inspected. Figure illustrates a decision to reject a lot after examining the 40th unit. Such charts can be easily designed with the help of statistical tables that specify the accept or reject cut-off values as a function of the cumulative sample size.



The ANI is generally lower for the sequential-sampling plan than for any other form of acceptance sampling, resulting in lower inspection costs. For very low or very high values of the proportion defective, sequential sampling provides a lower ANI than any comparable sampling plan. However, if the proportion of defective units falls between the AQL and the LTPD, a sequential-sampling plan could have a larger ANI than a comparable single- or double-sampling plan (although that is unlikely). In general, the sequential-sampling plan may reduce the ANI to 50 percent of that required by a comparable single-sampling plan and, consequently, save substantial inspection costs.

Operating Characteristic Curves :

Analysts create a graphic display of the performance of a sampling plan by plotting the probability of accepting the lot for a range of proportions of defective units. This graph, called an **operating characteristic (OC) curve**, describes how well a sampling plan discriminates between good and bad lots. Undoubtedly, every manager wants a plan that accepts lots with a quality level better than the AQL 100 percent of the time and accepts lots with a quality level worse than the AQL 0 percent of the time. This ideal OC curve for a single-sampling plan is shown in the figure below; however, such performance can be achieved only with 100 percent inspection.

A typical OC curve for a single-sampling plan shows the probability α of rejecting a good lot (producer's risk) and the probability β of accepting a bad lot (consumer's risk). Consequently, managers are left with choosing a sample size *n* and an acceptance number to achieve the level of performance specified by the AQL, α, LTPD, and β.

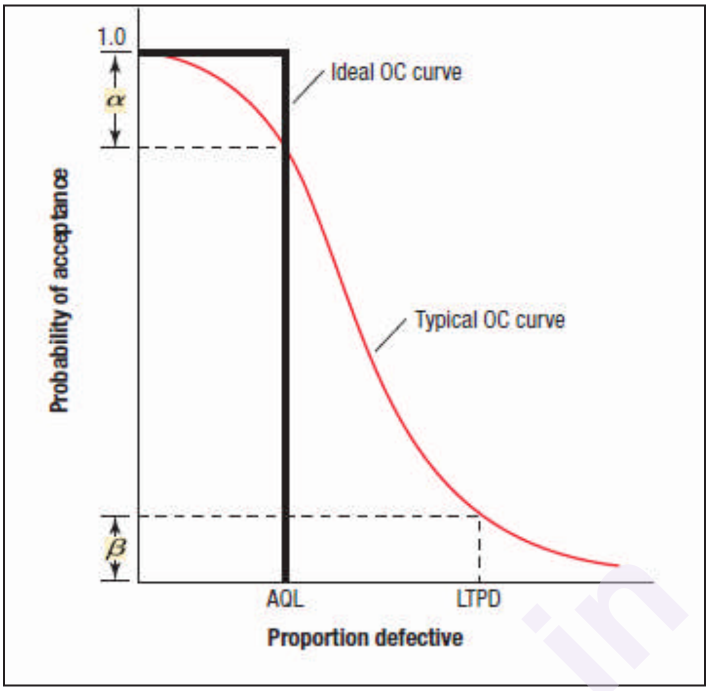
Drawing the OC Curve :

The sampling distribution for the single-sampling plan is the binomial distribution because each item inspected is either defective (a failure) or not (a success).

The probability of accepting the lot equals the probability of taking a sample of size *n* from a lot with a proportion defective of *p* and finding *c* or fewer defective items. However, if *n* is greater than 20 and *p* is less than 0.05, the Poisson distribution can be used as an approximation to the binomial to take advantage of tables prepared for the purpose of drawing OC curves. To draw the OC curve, look up the probability of accepting the lot for a range of values of *p*. For each value of *p*,

1. multiply *p* by the sample size *n*.
2. find the value of *n p*in the left column of the table.
3. move to the right until you find the column for *c.*
4. record the value for the probability of acceptance, *Pa*.

When *p* = AQL, the producer's risk, α, is 1 minus the probability of acceptance. When (*p* = LTPD), the consumer's risk, β, equals the probability of acceptance.



Illustrative Example 1

The Noise King Muffler Shop, a high-volume installer of replacement exhaust muffler systems, just received a shipment of 1,000 mufflers. The sampling plan for inspecting these mufflers calls for a sample size n = 60 and an acceptance number c = 1. The contract with the muffler manufacturer calls for an AQL of 1 defective muffler per 100 and an LTPD of 6 defective mufflers per 100. Calculate the OC curve for this plan, and determine the producer's risk and the consumer's risk for the plan.

Solution

Let p = 0.01. Then multiply n by p to get 60(0.01) = 0.60. Locate 0.60 in Table (Cumulative Poisson Probabilities).Move to the right until you reach the column for c=1. Read the probability of acceptance: 0.878. Repeat this process for arrange of p values. The following table (**Values for the Operating Characteristic Curve with n = 60 and c = 1)** contains the remaining values for the OC curve.

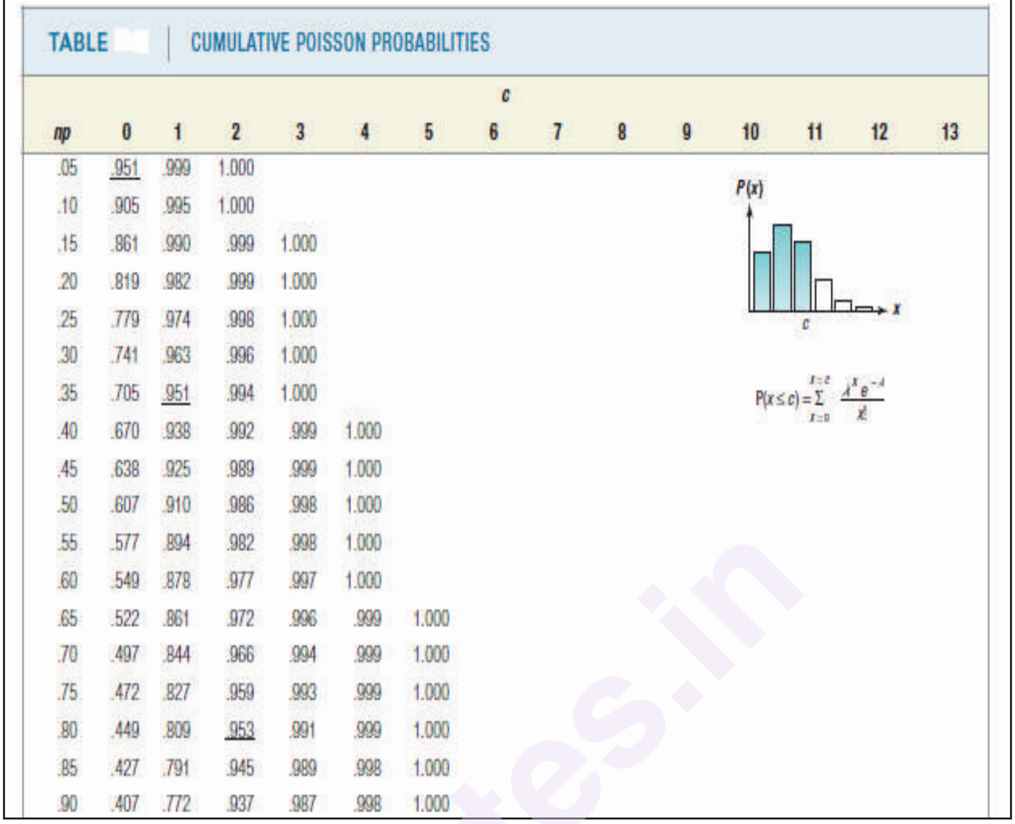
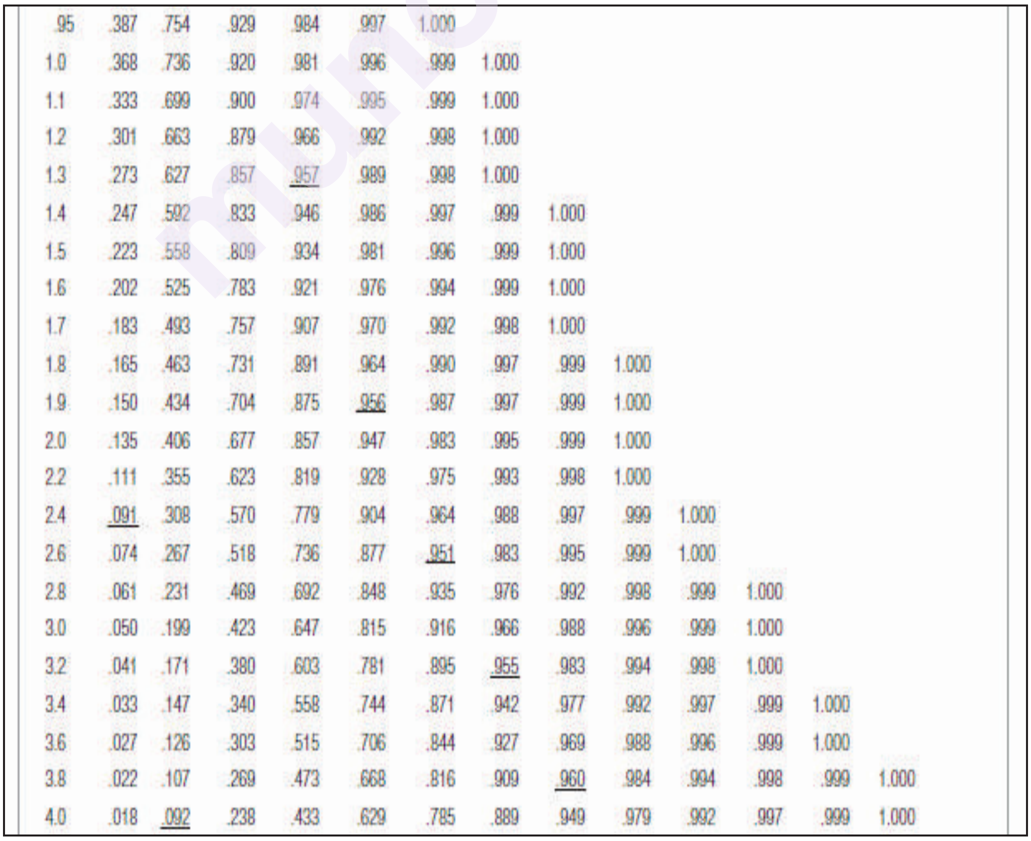


TABLE | CUMULATIVE POISSON PROBABILITIES

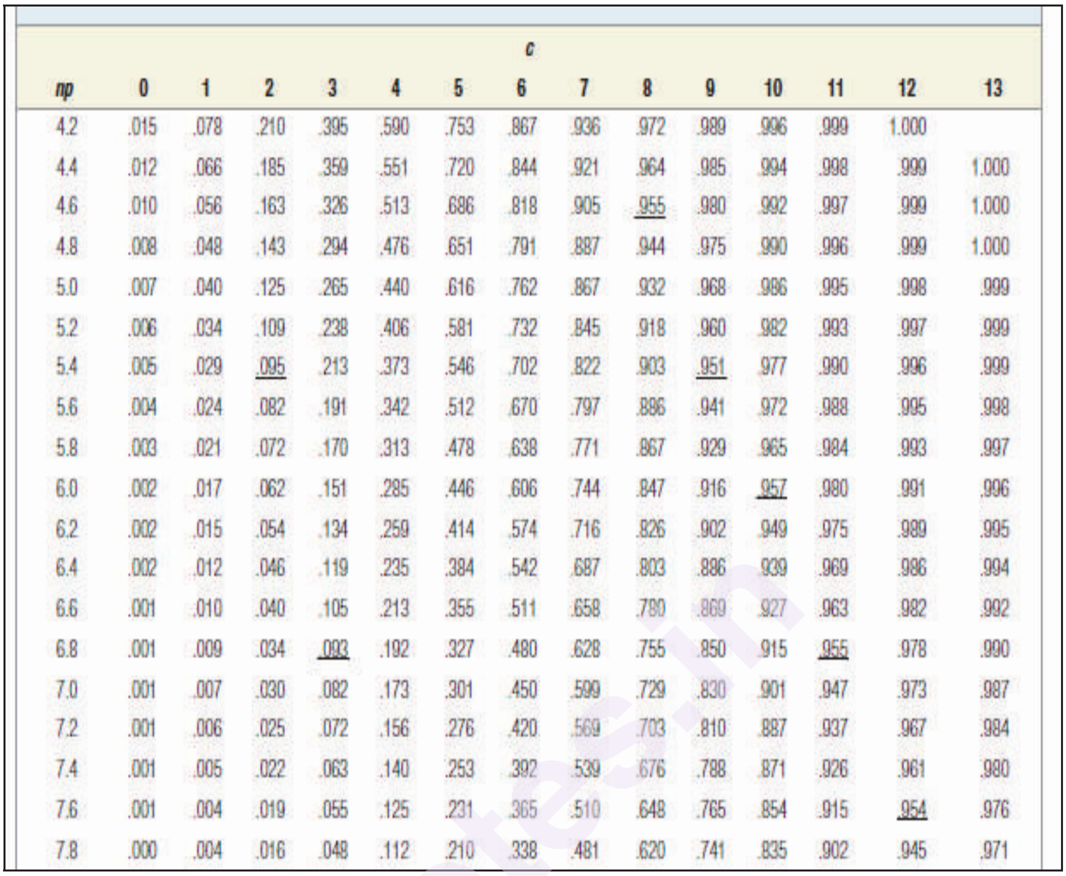
								c						
np	0	1	2	3	4	5	6	7	8	9	10	11	12	13
.05	.951	.999	1.000											
.10	.905	.995	1.000											
.15	.861	.990	.999	1.000										
.20	.819	.982	.999	1.000										
.25	.779	.974	.998	1.000										
.30	.741	.963	.996	1.000										
.35	.705	.951	.994	1.000										
.40	.670	.938	.992	.999	1.000									
.45	.638	.925	.989	.999	1.000									
.50	.607	.910	.986	.998	1.000									
.55	.577	.894	.982	.998	1.000									
.60	.549	.878	.977	.997	1.000									
.65	.522	.861	.972	.996	.999	1.000								
.70	.497	.844	.966	.994	.999	1.000								
.75	.472	.827	.959	.993	.999	1.000								
.80	.449	.809	.953	.991	.999	1.000								
.85	.427	.791	.945	.989	.998	1.000								
.90	.407	.772	.937	.987	.998	1.000								

$$P(x \le c) = \sum_{x=0}^{x=c} \frac{\lambda^x e^{-\lambda}}{x!}$$

Cumulative Poisson Probabilities …

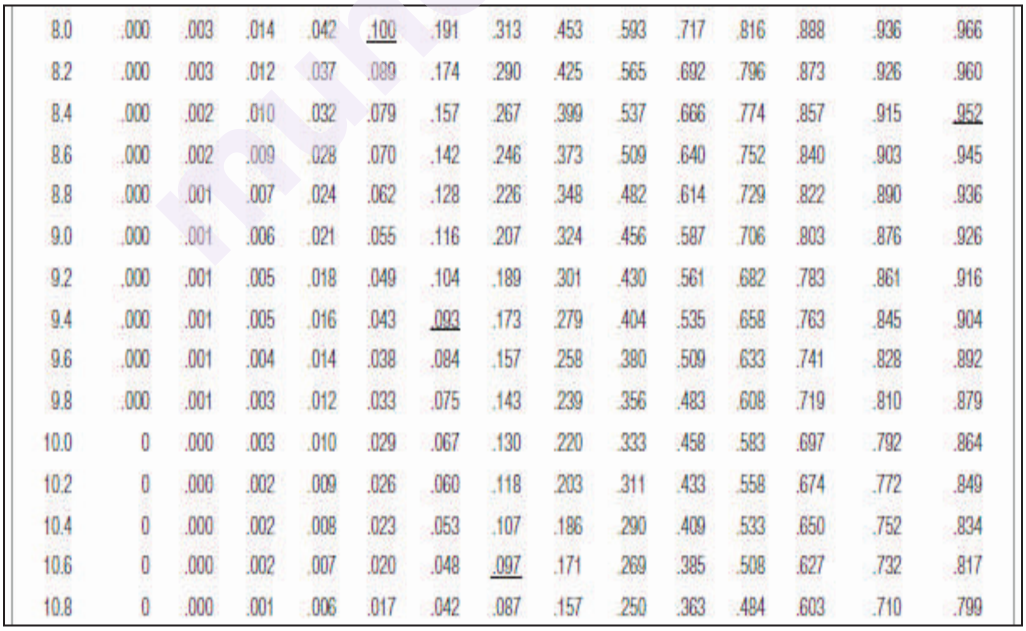


.95	.387	.754	.929	.984	.997	1.000								
1.0	.368	.736	.920	.981	.996	.999	1.000							
1.1	.333	.699	.900	.974	.995	.999	1.000							
1.2	.301	.663	.879	.966	.992	.998	1.000							
1.3	.273	.627	.857	.957	.989	.998	1.000							
1.4	.247	.592	.833	.946	.986	.997	.999	1.000						
1.5	.223	.558	.809	.934	.981	.996	.999	1.000						
1.6	.202	.525	.783	.921	.976	.994	.999	1.000						
1.7	.183	.493	.757	.907	.970	.992	.998	1.000						
1.8	.165	.463	.731	.891	.964	.990	.997	.999	1.000					
1.9	.150	.434	.704	.875	.956	.987	.997	.999	1.000					
2.0	.135	.406	.677	.857	.947	.983	.995	.999	1.000					
2.2	.111	.355	.623	.819	.928	.975	.993	.998	1.000					
2.4	.091	.308	.570	.779	.904	.964	.988	.997	.999	1.000				
2.6	.074	.267	.518	.736	.877	.951	.983	.995	.999	1.000				
2.8	.061	.231	.469	.692	.848	.935	.976	.992	.998	.999	1.000			
3.0	.050	.199	.423	.647	.815	.916	.966	.988	.996	.999	1.000			
3.2	.041	.171	.380	.603	.781	.895	.955	.983	.994	.998	1.000			
3.4	.033	.147	.340	.558	.744	.871	.942	.977	.992	.997	.999	1.000		
3.6	.027	.126	.303	.515	.706	.844	.927	.969	.988	.996	.999	1.000		
3.8	.022	.107	.269	.473	.668	.816	.909	.960	.984	.994	.998	.999	1.000	
4.0	.018	.092	.238	.433	.629	.785	.889	.949	.979	.992	.997	.999	1.000	

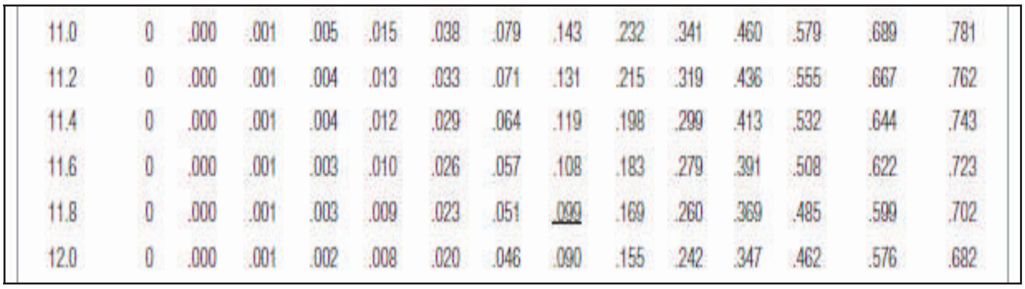


							c							
np	0	1	2	3	4	5	6	7	8	9	10	11	12	13
4.2	.015	.078	.210	.395	.590	.753	.867	.936	.972	.989	.996	.999	1.000	
4.4	.012	.066	.185	.359	.551	.720	.844	.921	.964	.985	.994	.998	.999	1.000
4.6	.010	.056	.163	.326	.513	.686	.818	.905	.955	.980	.992	.997	.999	1.000
4.8	.008	.048	.143	.294	.476	.651	.791	.887	.944	.975	.990	.996	.999	1.000
5.0	.007	.040	.125	.265	.440	.616	.762	.867	.932	.968	.986	.995	.998	.999
5.2	.006	.034	.109	.238	.406	.581	.732	.845	.918	.960	.982	.993	.997	.999
5.4	.005	.029	.095	.213	.373	.546	.702	.822	.903	.951	.977	.990	.996	.999
5.6	.004	.024	.082	.191	.342	.512	.670	.797	.886	.941	.972	.988	.995	.998
5.8	.003	.021	.072	.170	.313	.478	.638	.771	.867	.929	.965	.984	.993	.997
6.0	.002	.017	.062	.151	.285	.446	.606	.744	.847	.916	.957	.980	.991	.996
6.2	.002	.015	.054	.134	.259	.414	.574	.716	.826	.902	.949	.975	.989	.995
6.4	.002	.012	.046	.119	.235	.384	.542	.687	.803	.886	.939	.969	.986	.994
6.6	.001	.010	.040	.105	.213	.355	.511	.658	.780	.869	.927	.963	.982	.992
6.8	.001	.009	.034	.093	.192	.327	.480	.628	.755	.850	.915	.955	.978	.990
7.0	.001	.007	.030	.082	.173	.301	.450	.599	.729	.830	.901	.947	.973	.987
7.2	.001	.006	.025	.072	.156	.276	.420	.569	.703	.810	.887	.937	.967	.984
7.4	.001	.005	.022	.063	.140	.253	.392	.539	.676	.788	.871	.926	.961	.980
7.6	.001	.004	.019	.055	.125	.231	.365	.510	.648	.765	.854	.915	.954	.976
7.8	.000	.004	.016	.048	.112	.210	.338	.481	.620	.741	.835	.902	.945	.971

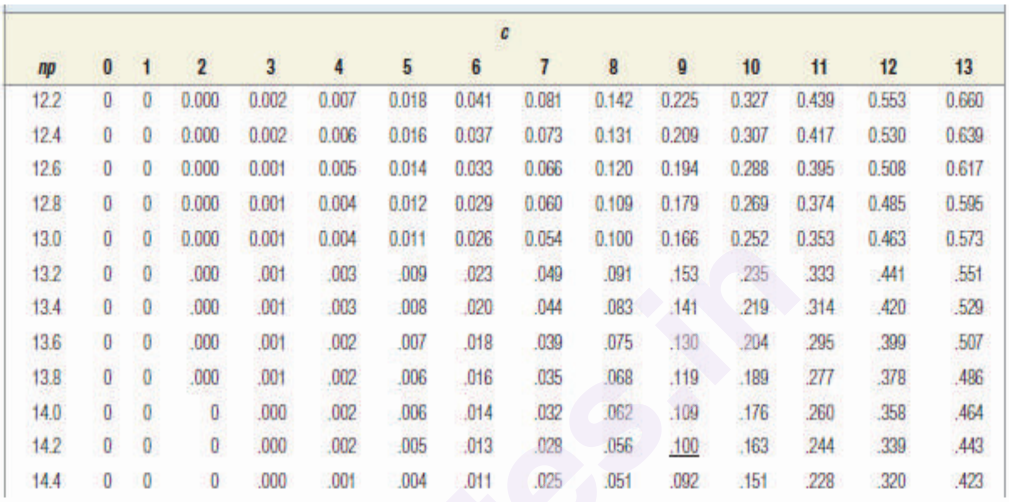
Cumulative Poisson Probabilities …



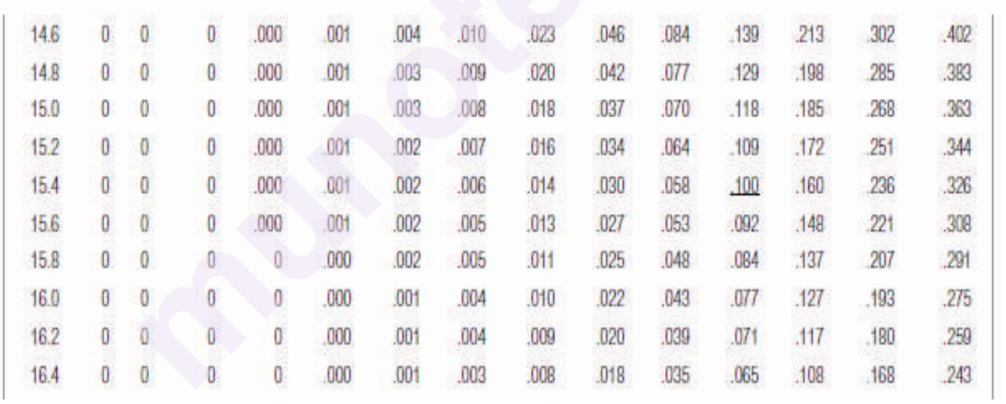
8.0	.000	.003	.014	.042	.100	.191	.313	.453	.593	.717	.816	.888	.936	.966
8.2	.000	.003	.012	.037	.089	.174	.290	.425	.565	.692	.796	.873	.926	.960
8.4	.000	.002	.010	.032	.079	.157	.267	.399	.537	.666	.774	.857	.915	.952
8.6	.000	.002	.009	.028	.070	.142	.246	.373	.509	.640	.752	.840	.903	.945
8.8	.000	.001	.007	.024	.062	.128	.226	.348	.482	.614	.729	.822	.890	.936
9.0	.000	.001	.006	.021	.055	.116	.207	.324	.456	.587	.706	.803	.876	.926
9.2	.000	.001	.005	.018	.049	.104	.189	.301	.430	.561	.682	.783	.861	.916
9.4	.000	.001	.005	.016	.043	.093	.173	.279	.404	.535	.658	.763	.845	.904
9.6	.000	.001	.004	.014	.038	.084	.157	.258	.380	.509	.633	.741	.828	.892
9.8	.000	.001	.003	.012	.033	.075	.143	.239	.356	.483	.608	.719	.810	.879
10.0	0	.000	.003	.010	.029	.067	.130	.220	.333	.458	.583	.697	.792	.864
10.2	0	.000	.002	.009	.026	.060	.118	.203	.311	.433	.558	.674	.772	.849
10.4	0	.000	.002	.008	.023	.053	.107	.186	.290	.409	.533	.650	.752	.834
10.6	0	.000	.002	.007	.020	.048	.097	.171	.269	.385	.508	.627	.732	.817
10.8	0	.000	.001	.006	.017	.042	.087	.157	.250	.363	.484	.603	.710	.799



11.0	0	.000	.001	.005	.015	.038	.079	.143	.232	.341	.460	.579	.689	.781
11.2	0	.000	.001	.004	.013	.033	.071	.131	.215	.319	.436	.555	.667	.762
11.4	0	.000	.001	.004	.012	.029	.064	.119	.198	.299	.413	.532	.644	.743
11.6	0	.000	.001	.003	.010	.026	.057	.108	.183	.279	.391	.508	.622	.723
11.8	0	.000	.001	.003	.009	.023	.051	.099	.169	.260	.369	.485	.599	.702
12.0	0	.000	.001	.002	.008	.020	.046	.090	.155	.242	.347	.462	.576	.682

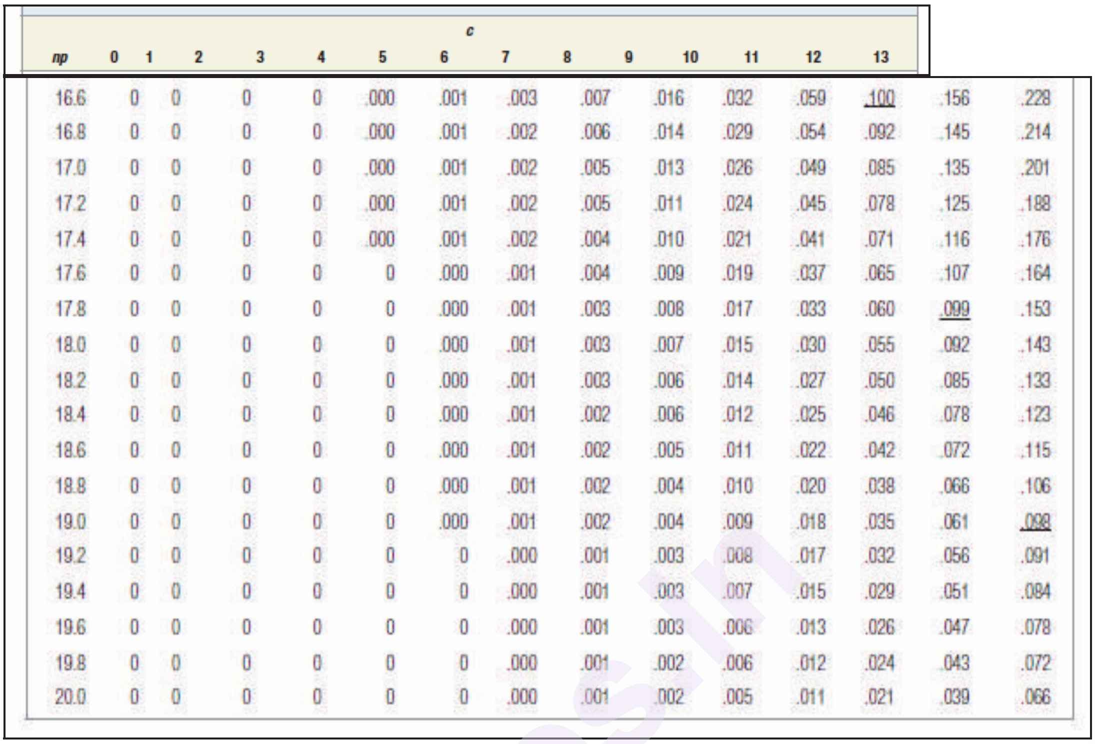


	c													
np	0	1	2	3	4	5	6	7	8	9	10	11	12	13
12.2	0	0	0.000	0.002	0.007	0.018	0.041	0.081	0.142	0.225	0.327	0.439	0.553	0.660
12.4	0	0	0.000	0.002	0.006	0.016	0.037	0.073	0.131	0.209	0.307	0.417	0.530	0.639
12.6	0	0	0.000	0.001	0.005	0.014	0.033	0.066	0.120	0.194	0.288	0.395	0.508	0.617
12.8	0	0	0.000	0.001	0.004	0.012	0.029	0.060	0.109	0.179	0.269	0.374	0.485	0.595
13.0	0	0	0.000	0.001	0.004	0.011	0.026	0.054	0.100	0.166	0.252	0.353	0.463	0.573
13.2	0	0	.000	.001	.003	.009	.023	.049	.091	.153	.235	.333	.441	.551
13.4	0	0	.000	.001	.003	.008	.020	.044	.083	.141	.219	.314	.420	.529
13.6	0	0	.000	.001	.002	.007	.018	.039	.075	.130	.204	.295	.399	.507
13.8	0	0	.000	.001	.002	.006	.016	.035	.068	.119	.189	.277	.378	.486
14.0	0	0	0	.000	.002	.006	.014	.032	.062	.109	.176	.260	.358	.464
14.2	0	0	0	.000	.002	.005	.013	.028	.056	.100	.163	.244	.339	.443
14.4	0	0	0	.000	.001	.004	.011	.025	.051	.092	.151	.228	.320	.423

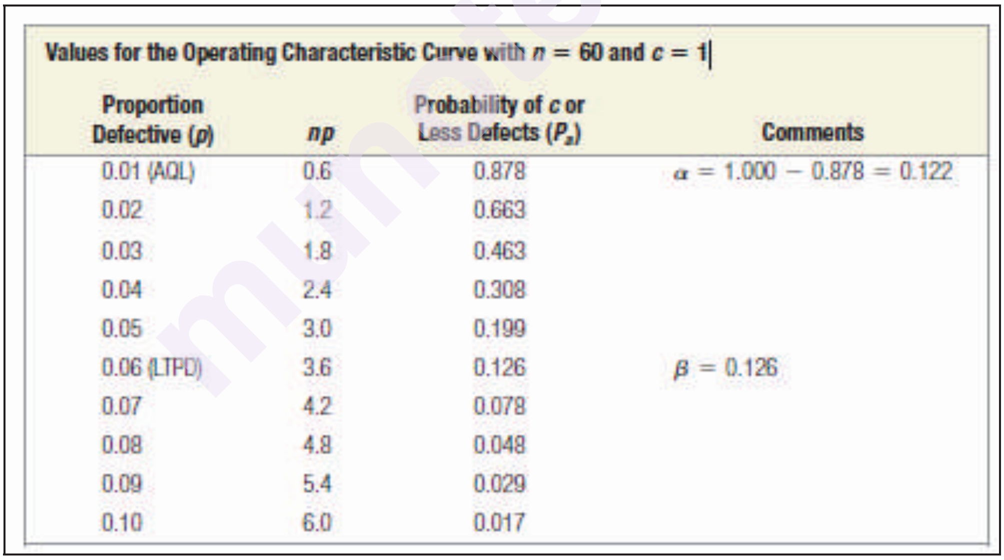


14.6	0	0	0	.000	.001	.004	.010	.023	.046	.084	.139	.213	.302	.402
14.8	0	0	0	.000	.001	.003	.009	.020	.042	.077	.129	.198	.285	.383
15.0	0	0	0	.000	.001	.003	.008	.018	.037	.070	.118	.185	.268	.363
15.2	0	0	0	.000	.001	.002	.007	.016	.034	.064	.109	.172	.251	.344
15.4	0	0	0	.000	.001	.002	.006	.014	.030	.058	.100	.160	.236	.326
15.6	0	0	0	.000	.001	.002	.005	.013	.027	.053	.092	.148	.221	.308
15.8	0	0	0	0	.000	.002	.005	.011	.025	.048	.084	.137	.207	.291
16.0	0	0	0	0	.000	.001	.004	.010	.022	.043	.077	.127	.193	.275
16.2	0	0	0	0	.000	.001	.004	.009	.020	.039	.071	.117	.180	.259
16.4	0	0	0	0	.000	.001	.003	.008	.018	.035	.065	.108	.168	.243

Cumulative Poisson Probabilities …



	c													
np	0	1	2	3	4	5	6	7	8	9	10	11	12	13
16.6	0	0	0	0	.000	.001	.003	.007	.016	.032	.059	.100	.156	.228
16.8	0	0	0	0	.000	.001	.002	.006	.014	.029	.054	.092	.145	.214
17.0	0	0	0	0	.000	.001	.002	.005	.013	.026	.049	.085	.135	.201
17.2	0	0	0	0	.000	.001	.002	.005	.011	.024	.045	.078	.125	.188
17.4	0	0	0	0	.000	.001	.002	.004	.010	.021	.041	.071	.116	.176
17.6	0	0	0	0	0	.000	.001	.004	.009	.019	.037	.065	.107	.164
17.8	0	0	0	0	0	.000	.001	.003	.008	.017	.033	.060	.099	.153
18.0	0	0	0	0	0	.000	.001	.003	.007	.015	.030	.055	.092	.143
18.2	0	0	0	0	0	.000	.001	.003	.006	.014	.027	.050	.085	.133
18.4	0	0	0	0	0	.000	.001	.002	.006	.012	.025	.046	.078	.123
18.6	0	0	0	0	0	.000	.001	.002	.005	.011	.022	.042	.072	.115
18.8	0	0	0	0	0	.000	.001	.002	.004	.010	.020	.038	.066	.106
19.0	0	0	0	0	0	.000	.001	.002	.004	.009	.018	.035	.061	.098
19.2	0	0	0	0	0	0	.000	.001	.003	.008	.017	.032	.056	.091
19.4	0	0	0	0	0	0	.000	.001	.003	.007	.015	.029	.051	.084
19.6	0	0	0	0	0	0	.000	.001	.003	.006	.013	.026	.047	.078
19.8	0	0	0	0	0	0	.000	.001	.002	.006	.012	.024	.043	.072
20.0	0	0	0	0	0	0	.000	.001	.002	.005	.011	.021	.039	.066



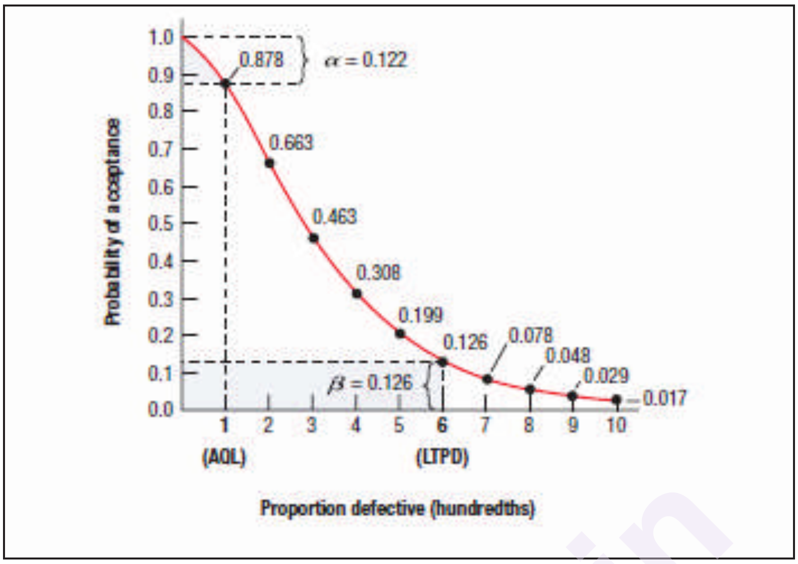
Values for the Operating Characteristic Curve with $n = 60$ and $c = 1$

Proportion Defective (p)	np	Probability of c or Less Defects (P_a)	Comments
0.01 (AQL)	0.6	0.878	$\alpha = 1.000 - 0.878 = 0.122$
0.02	1.2	0.663	
0.03	1.8	0.463	
0.04	2.4	0.308	
0.05	3.0	0.199	
0.06 (LTPD)	3.6	0.126	$\beta = 0.126$
0.07	4.2	0.078	
0.08	4.8	0.048	
0.09	5.4	0.029	
0.10	6.0	0.017	

Decision Point

Note that the plan provides a producer’s risk of 12.2 percent and a consumer’s risk of 12.6 percent. Both values are higher than the values usually acceptable for plans of this type (5 and 10 percent, respectively). Figure below shows the OC curve and the producer’s and consumer’s risks. Management can adjust the risks by changing the sample size.

The OC Curve for Single-Sampling Plan with *n* = 60 and *c* = 1

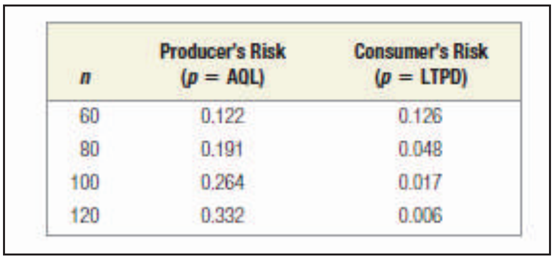


Explaining Changes in the OC Curve

This example raises the question: How can management change the sampling plan to reduce the probability of rejecting good lots and accepting bad lots? To answer this question, let uses how *n* and *c* affect the shape of the OC curve. In the Noise King example, a better single sampling plan would have a lower producer's risk and a lower consumer's risk.

Sample Size Effect What would happen if we increased the sample size to 80 and left the acceptance level, c, unchanged at 1?We can use Table **Cumulative Poisson Probabilities.**

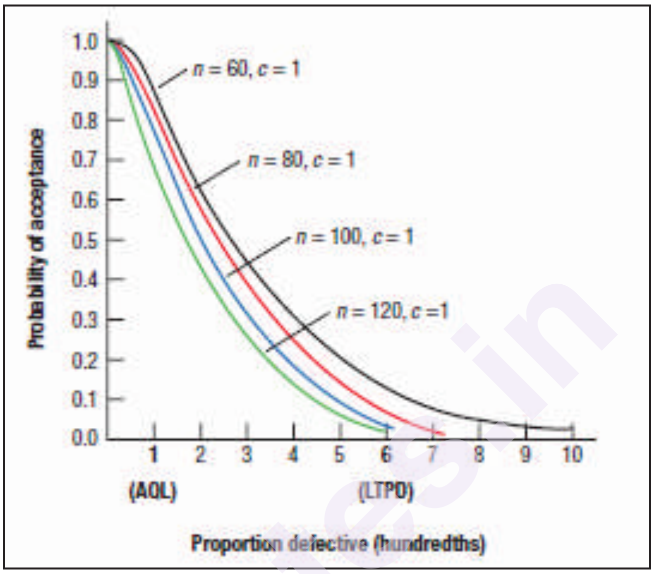
If the proportion defective of the lot is p = AQL = 0.01, the nnp=0.8 and the probability of acceptance of the lot is only 0.809. Thus, the producer's risk is 0.191. Similarly, if *p* = LTPD = 0.06, the probability of acceptance is 0.048. Other values of the producer's and consumer's risks are shown in the following table:



n	Producer's Risk (p = AQL)	Consumer's Risk (p = LTPD)
60	0.122	0.126
80	0.191	0.048
100	0.264	0.017
120	0.332	0.006

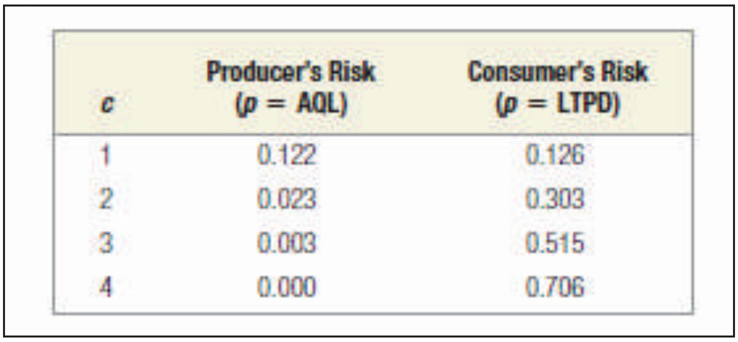
These results, shown in Figure below, yield the following principle: Increasing n while holding c constant increases the producer's risk and reduces the consumer's risk. For the producer of the mufflers, keeping c = 1 and increasing the sample size makes getting a lot accepted by the customer tougher—only two bad mufflers will get the lot rejected. And the likelihood of finding those 2 defects is greater in a sample of 120 than in a sample of 60.

Effects of Increasing Sample Size While Holding Acceptance Number Constant



Consequently, the producer's risk increases. For the management of Noise King, the consumer's risk goes down because a random sample of 120 mufflers from a lot with 6 percent defectives is less likely to have only 1 or fewer defective mufflers.

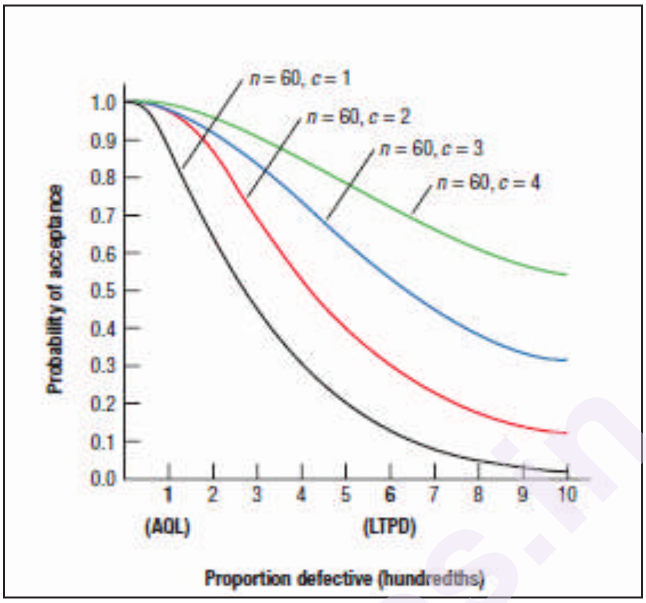
Acceptance Level Effect: Suppose that we keep the sample size constant at 60 but change the acceptance level. Again, we use Table Cumulative Poisson Probabilities.



c	Producer's Risk (*p* = AQL)	Consumer's Risk (*p* = LTPD)
1	0.122	0.126
2	0.023	0.303
3	0.003	0.515
4	0.000	0.706

The results are plotted in Figure below:

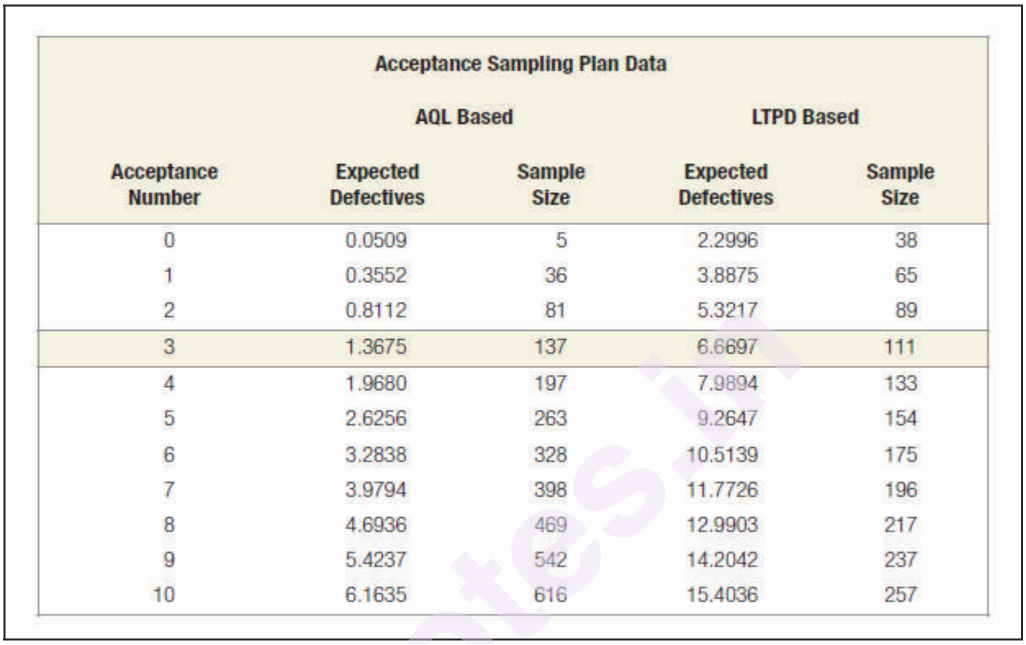
Effects of Increasing Acceptance Number While Holding Sample Size Constant



They demonstrate the following principle: *Increasing* c *while holding* n *constant decreases the producer's risk and increases the consumer's risk*. The producer of the mufflers would welcome an increase in the acceptance number because it makes getting the lot accepted by the consumer easier.

If the lot has only 1 percent defectives (the AQL) with a sample size of 60, we would expect only 0.01 (60) = 0.6 defect in the sample. An increase in the acceptance numbers from one to two lowers the probability of finding more than two defects and, consequently, lowers the producer's risk. However, raising the acceptance number for a given sample size increases the risk of accepting a bad lot. Suppose that the lot has 6 percent defectives (the LTPD). We would expect to have defectives in the sample. An increase in the acceptance number from one to two increases the probability of getting a sample with two or fewer defects and, therefore, increases the consumer's risk.

Thus, to improve Noise King's single-sampling acceptance plan, management should increase the sample size, which reduces the consumer's risk, and increase the acceptance number, which reduces the producer's risk. An improved combination can be found by trial and error using Table of Cumulative Poisson Probabilities. Alternatively, a computer can be used to find the best combination. For any acceptance number, the computer determines the sample size needed to achieve the desired producer's risk and compares it to the sample size needed to meet the consumer's risk. It selects the smallest sample size that will meet both the producer’s risk and the consumer’s risk. The following table shows that a sample size of 111 and an acceptance number of 3 are best. This combination actually yields a producer’s risk of 0.026 and a consumer’s risk of 0.10 (not shown). The risks are not exact because c and n must be integers.



Acceptance Sampling Plan Data

	AQL Based		LTPD Based	
Acceptance Number	Expected Defectives	Sample Size	Expected Defectives	Sample Size
0	0.0509	5	2.2996	38
1	0.3552	36	3.8875	65
2	0.8112	81	5.3217	89
3	1.3675	137	6.6697	111
4	1.9680	197	7.9894	133
5	2.6256	263	9.2647	154
6	3.2838	328	10.5139	175
7	3.9794	398	11.7726	196
8	4.6936	469	12.9903	217
9	5.4237	542	14.2042	237
10	6.1635	616	15.4036	257

Average Outgoing Quality

- **Average outgoing quality (AOQ):** The expressed proportion of defects that the plan will allow to pass.

- **Rectified inspection:** The assumption that all defective item sin the lot will be replaced with good items if the lot is rejected and that any defective items in the sample will be replaced if the lot is accepted.

- **Average outgoing quality limit (AOQL):** The maximum value of the average outgoing quality over all possible values of the proportion defective.

We have shown how to choose the sample size and acceptance number for a single-sampling plan, given AQL, α, LTPD, and β parameters. To check whether the performance of the plan is what we want, we can calculate the plan’s average outgoing quality (AOQ), which is the expected proportion of defects that the plan will allow to pass. We assume that all defective items in the lot will be replaced with good items if the lot is rejected and that any defective items in the sample will be replaced if the lot is accepted.

This approach is called rectified inspection. The equation for AOQ is

$$\text{AOQ} = \frac{p(P_a)(N - n)}{N}$$

where

$$\begin{aligned} p &= \text{true proportion defective of the lot} \\ P_a &= \text{probability of accepting the lot} \\ N &= \text{lot size} \\ n &= \text{sample size} \end{aligned}$$

The analyst can calculate AOQ to estimate the performance of the plan over a range of possible proportion defectives in order to judge whether the plan will provide an acceptable degree of protection.

The maximum value of the average outgoing quality over all possible values of the proportion defective is called the average outgoing quality limit (AOQL). If the AOQL seems too high, the parameters of the plan must be modified until an acceptable AOQL is achieved.

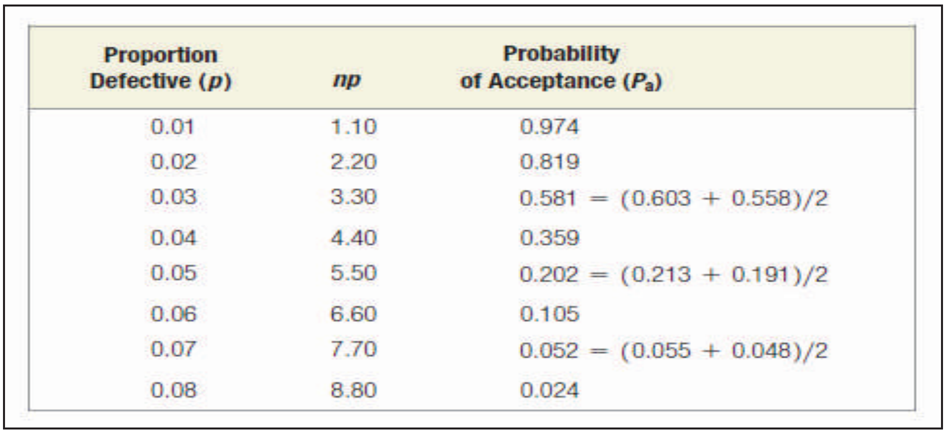
Illustrative Example 2

Suppose that Noise King is using rectified inspection for its single-sampling plan. Calculate the average outgoing quality limit for a plan with n=110, c=3, and N=100. Use the table of Cumulative Poisson Probabilities to estimate the probabilities of acceptance for values of the proportion defective from 0.01 to 0.08 in steps of 0.01.

Solution

Use the following steps to estimate the AOQL for this sampling plan:

Step 1: Determine the probabilities of acceptance for the desired values of p. These are shown in the following table. However, the values for p=0.03, 0.05 and 0.07 had to be interpolated because the table does not have them. For example, Pa for p=0.03, was estimated by averaging the Pa values for *np*= 3.2 and *np*= 3.4, or (0.603 + 0.558)/2 = 0.580.

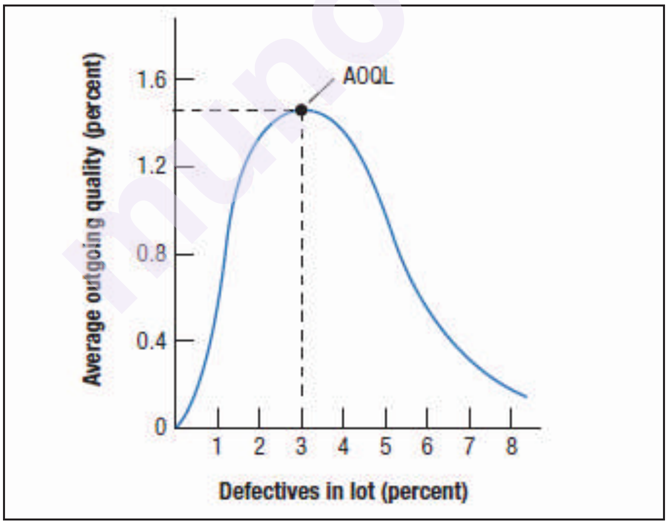


Proportion Defective (p)	np	Probability of Acceptance (P_a)
0.01	1.10	0.974
0.02	2.20	0.819
0.03	3.30	0.581 = (0.603 + 0.558)/2
0.04	4.40	0.359
0.05	5.50	0.202 = (0.213 + 0.191)/2
0.06	6.60	0.105
0.07	7.70	0.052 = (0.055 + 0.048)/2
0.08	8.80	0.024

Step 2: Calculate the AOQ for each value of p.

For p = 0.01: 0.01(0.974)(1000 - 110)/1000 = 0.0087
For p = 0.02: 0.02(0.819)(1000 - 110)/1000 = 0.0146
For p = 0.03: 0.03(0.581)(1000 - 110)/1000 = 0.0155
For p = 0.04: 0.04(0.359)(1000 - 110)/1000 = 0.0128
For p = 0.05: 0.05(0.202)(1000 - 110)/1000 = 0.0090
For p = 0.06: 0.06(0.105)(1000 - 110)/1000 = 0.0056
For p = 0.07: 0.07(0.052)(1000 - 110)/1000 = 0.0032
For p = 0.08: 0.08(0.024)(1000 - 110)/1000 = 0.0017

The plot of the AOQ values is shown below:



Step 3: Identify the largest AOQ value, which is the estimate of the AOQL. In this example, the AOQL is0.0155 at p = 0.03.

Illustrative Example 3

An inspection station has been installed between two production processes. The feeder process, when operating correctly, has an acceptable quality level of 3 percent. The consuming process, which is expensive, has a specified lot tolerance proportion defective of 8 percent.

The feeding process produces in batch sizes; if a batch is rejected by the inspector, the entire batch must be checked and the defective items reworked. Consequently, management wants no more than a 5 percent producer's risk and, because of the expensive process that follows, no more than a 10 percent chance of accepting a lot with 8 percent defectives or worse.

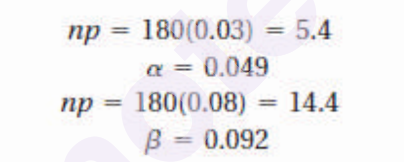
a. Determine the appropriate sample size, *n*, and the acceptable number of defective items in the sample, *c*.

b. Calculate values and draw the OC curve for this inspection station.

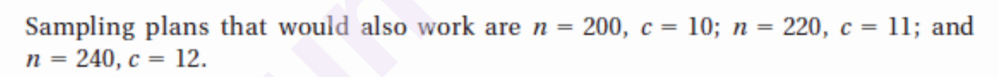
c. What is the probability that a lot with 5 percent defectives will be rejected?

Solution

a. For AQL=3 percent, LTPD=8 percent, α=5 percent, and β=10 percent, use Table of Cumulative Poisson Probabilities given above and trial and error to arrive at a sampling plan. If n=180 and c=9

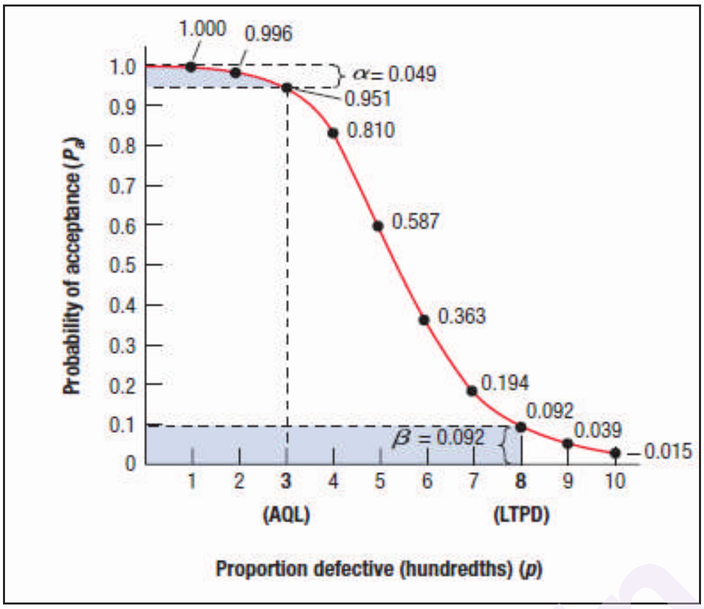


$$np = 180(0.03) = 5.4$$
$$\alpha = 0.049$$
$$np = 180(0.08) = 14.4$$
$$\beta = 0.092$$

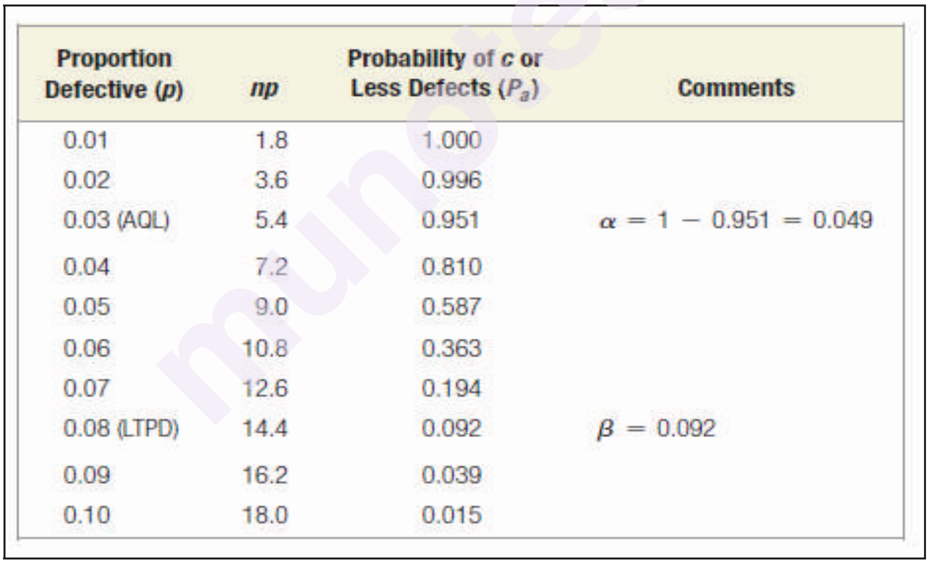


Sampling plans that would also work are $n = 200$, $c = 10$; $n = 220$, $c = 11$; and $n = 240$, $c = 12$.

b. The following table contains the data for the OC curve. Table of Cumulative Poisson Probabilities was used to estimate the probability of acceptance. Figure below shows the OC curve.



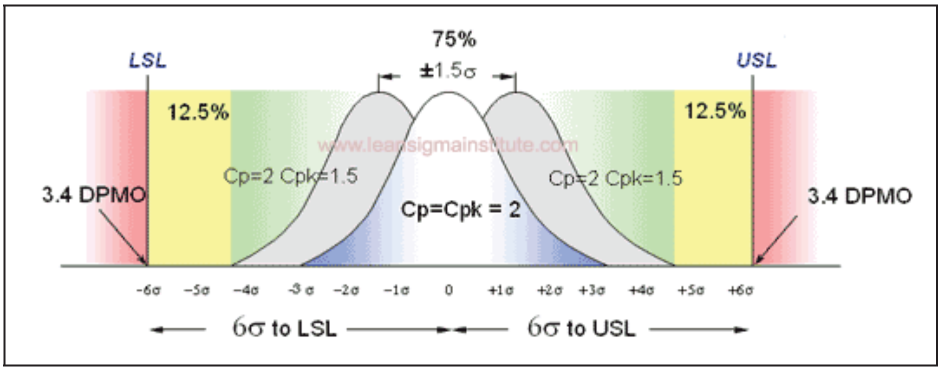
c. According to the table, the probability of accepting a lot with 5 percent defectives is 0.587. Therefore, the probability that a lot with 5 percent defects will be rejected is0.413, or 1.00 - 0.587.



Proportion Defective (p)	np	Probability of c or Less Defects (P_a)	Comments
0.01	1.8	1.000	
0.02	3.6	0.996	
0.03 (AQL)	5.4	0.951	$\alpha = 1 - 0.951 = 0.049$
0.04	7.2	0.810	
0.05	9.0	0.587	
0.06	10.8	0.363	
0.07	12.6	0.194	
0.08 (LTPD)	14.4	0.092	$\beta = 0.092$
0.09	16.2	0.039	
0.10	18.0	0.015	

9.4 SIX-SIGMA

Business world often describes Six Sigma as a highly technical method used by engineers and statisticians to fine-tune products and services. Another school defines Six Sigma as pursuing a goal near-perfection in meeting customer requirements by achieving 3.4 parts per million potential defects. Cultural change is also a valid way to describe Six Sigma. Motorola puts a lot of emphasis on cultural change such as breaking down the white space between departments, employee empowerment, etc.



Six Sigma Definitions:

- A Management driven, scientific methodology for product and process improvement which creates breakthroughs in financial performance and Customer satisfaction.

- A methodology that provides businesses with the tools to improve the capability of their business processes. This increase in performance and decrease in process variation lead to defect reduction and improvement in profits, employee morale and quality of product.

Six Sigma actually has its roots in a 19th Century mathematical theory, but found its way into today's mainstream business world through the efforts of an engineer at Motorola in the 1980s. Now heralded as one of the foremost methodological practices for improving customer satisfaction and improving business processes, Six Sigma has been refined and perfected over the years into what we see today.

No matter the setting, the goal remains the same: Six Sigma seeks to improve business processes by removing the causes of errors that lead to defects in a product or service. It accomplishes this by setting up a management system that systematically identifies errors and provides methods for eliminating them.

Those who learn Six Sigma practices achieve designations at each level of accomplishment, including Green Belt, Black Belt, Master Black Belt and Champion.

Beginnings of Six Sigma

The process that led to Six Sigma was originated in the 19th Century with the bell curve developed by Carl Fredrick Grauss. In the 1920s, statistician Carl Shewhart, a founding member of the Institute of Mathematical Statistics, showed that a process required correction after it had deviated by three sigma from the mean.

Move forward to the 1970s, when Motorola senior executive Art Sundry complained about the lack of consistent quality in the company's products, as per the 2006 book "Six Sigma" by Richard Schroeder and Harry Mikel.

According to the accepted story from numerous sources, Motorola engineer Bill Smith eventually answered the call to consistently manufacture quality products by working out the methodologies of Six Sigma in 1986. The system is influenced by, but different than, other management improvement strategies of the time, including Total Quality Management and Zero Defects.

Some off the Major Aspects of Six Sigma

In an effort to bring operations to a "six sigma" level – essentially 3.4 defects for every one million opportunities – the methodology calls for continuous efforts to get processes to the point where they produce stable and predictable results.

Deconstructing the manufacturing process down to its essential parts, Six Sigma defines and evaluates each step of a process, searching for ways to improve efficiencies in a business structure, improve the quality of the process and increase the bottom-line profit.

Toward that end, the methodology calls for the training of personnel in Six Sigma, including beginner Green Belts, Black Belts who often head up individual projects, and Master Black Belts who look for ways to apply Six Sigma across a business structure to make improvements.

The ultimate goal is to improve every process to a "six sigma" level or better. Does it work? Motorola reported in 2006 that the company had saved $17 billion using Six Sigma.

Methodologies of Six Sigma :

There are two major methodologies used within Six Sigma, both of which are composed of five sections, according to the 2005 book "JURAN Institute Six Sigma Breakthrough and Beyond" by Joseph A. De Feo and William Barnard.

DMAIC. This method is used primarily for improving existing business processes. The letters stand for :

- **D**efine the problem and the project goals
- **M**easure in detail the various aspects of the current process

- **A**nalyze data to, among other things, find the root defects in a process
- **I**mprove the process

- **C**ontrol how the process is done in the future.

DMADV. This method is typically used to create new processes and new products or services. The letters stand for:

- **D**efine the project goals
- **M**easure critical components of the process and the product capabilities
- **A**nalyze the data and develop various designs for the process, eventually picking the best one
- **D**esign and test details of the process
- **V**erify the design by running simulations and a pilot program, and then handing over the process to the client

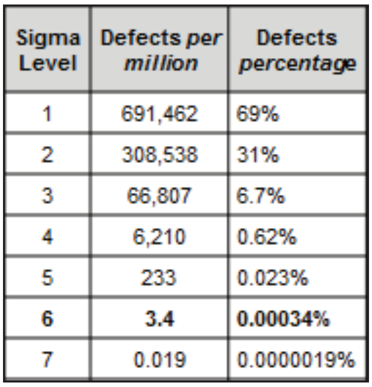
There are also many different management tools used within Six Sigma. While there are too many to list, here are details on a few of them.

Five Whys. This is a method that uses questions to get to the root cause of a problem. The method is simple: simply state the final problem (the car wouldn't start, I was late to work again today) and then ask the question "why," breaking down the issue to its root cause. In these two cases, it might be: because I didn't maintain the car properly and because I need to leave my house earlier to get to work on time. The process first came to prominence at Toyota.

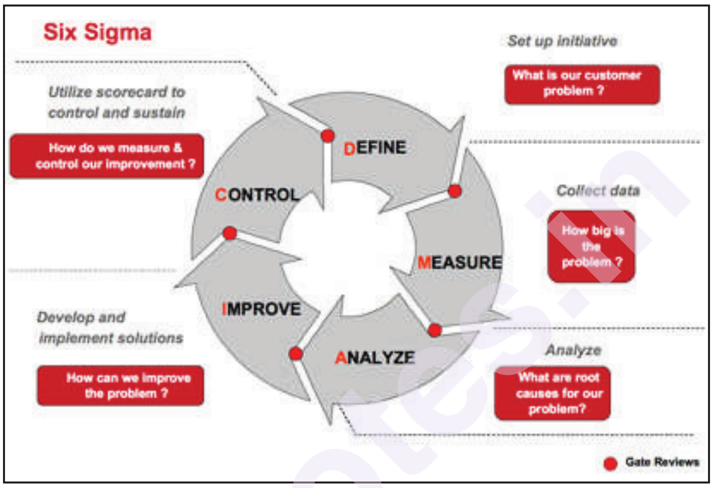
CTQ Tree. The Critical to Quality Tree diagram breaks down the components of a process that produces the features needed in your product and service if you wish to have satisfied customers.

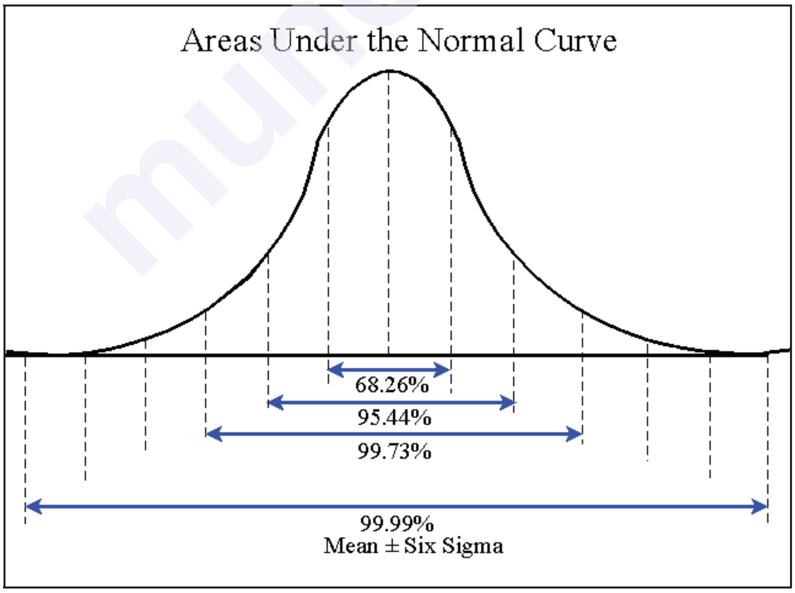
Root Cause Analysis. Much like the Five Whys, this is a process by which a business attempts to identify the root cause of a defect and then correct it, rather than simply correcting the surface "symptom" of the problem.

Ultimately, all of the tools and methodologies in Six Sigma serve one purpose: to streamline business processes in order to produce the best products and services possible with the smallest amount of defects. Its adoption by corporations around the globe is both an indicator of and testament to its remarkable success in today's business environment.



Sigma Level	Defects *per million*	Defects *percentage*
1	691,462	69%
2	308,538	31%
3	66,807	6.7%
4	6,210	0.62%
5	233	0.023%
6	**3.4**	**0.00034%**
7	0.019	0.0000019%


Six Sigma
Set up initiative
What is our customer problem ?
Utilize scorecard to control and sustain
How do we measure & control our improvement ?
DEFINE
CONTROL
MEASURE
IMPROVE
ANALYZE
Collect data
How big is the problem ?
Develop and implement solutions
How can we improve the problem ?
Analyze
What are root causes for our problem?
● Gate Reviews



9.5 LET US SUM UP

In this Unit V – Chapter 5.3 you have learnt acceptance sampling and six sigma philosophy.

9.6 EXERCISES

Q.1 Explain the concept of acceptance sampling with one or two examples.

Q.2. What is Six Sigma? Explain its importance.

Q.3. For c=10 and LTPD=5 percent, what value of n resultsin a 5 percent consumer's risk?

Q.4. The Pee Wee Cuisine Company manufactures gourmet food for babies. They receive turnips in batches of 600 and use MIL-STD-105D at general inspection level III. Recently, the company used double sampling. In its first sample, there were three unacceptable turnips. In the second sample from that same batch, there were three more unacceptable ones. If Pee Wee Cuisine uses an AQL of 1 percent, what decision would it make regarding that batch of turnips?

Q.5. A buyer of electronic components has a lot tolerance proportion defective of 20 parts in 5,000, with a consumer's risk of 15 percent. If the buyer will sample1,500 of the components received in each shipment, what acceptance number, c, would the buyer want? What is the producer's risk if the AQL is 10 parts per 5,000?

Q.6 For AQL=1 percent and c=2, what is the largest value of n that will result in a producer's risk of 5 percent? Using that sample size, determine the consumer's risk when LTPD=2 percent.

9.7 SUGGESTED READINGS

SQC section in any of the reference / text books

❖❖❖❖

Module - VI
ANOVA & CHI-SQUARE TEST

10

INTRODUCTION TO ANOVA

Introduction to ANOVA, One Way ANOVA, Two Way ANOVA without replication, Two Way ANOVA with replication

Unit Structure :

10.1 INTRODUCTION

In this Unit-VI - Chapter 6.1, we shall discuss the concepts of One Way ANOVA, Two Way ANOVA without replication and Two Way ANOVA with replication

10.2 OBJECTIVES

At the end of this unit the learners will be able to understand

- What is ANOVA?
- One Way ANOVA
- Two Way ANOVA without replication
- Two Way ANOVA with replication.

10.3 What is ANOVA?

ANOVA is a statistical technique that assesses potential differences in a scale-level dependent variable by a nominal-level variable having 2 or more categories. For example, an ANOVA can examine potential differences in IQ scores by Country (US vs. Canada vs. Italy vs. Spain). The ANOVA, developed by Ronald

Fisher in 1918, extends the *t* and the *z* test which have the problem of only allowing the nominal level variable to have just two categories. This test is also called the Fisher analysis of variance.

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups:

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.

The t-test is designed to test the hypothesis that 2 means could be from the same population of data. But what if we want to compare more than 2 means at the same time?

ANOVA is a general technique that can be used to test the hypothesis that the means among two or more groups are equal, *under the assumption that the sampled populations are normally distributed.*

Analysis of variance can be used to test differences among several means for significance without increasing the *Type I error rate.*

To begin, let us consider the effect of temperature on a passive component such as a resistor.

We select three different temperatures and observe their effect on the resistors. This experiment can be conducted by measuring all the participating resistors before placing *n* resistors each in three different ovens.

Each oven is heated to a selected temperature. Then we measure the resistors again after, say, 24 hours and analyse the responses, which are the differences between before and after being subjected to the temperatures.

The temperature is called a *factor*. The different temperature settings are called *levels*. In this example there are three levels or settings of the factor Temperature.

What Does “One-Way” or “Two-Way Mean?

One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test. One-way has one independent variable (with 2 levels) and two-way has two independent variables (can have multiple levels). For example, a one-way Analysis of Variance could have one IV (brand of cereal) and a two-way Analysis of Variance has two IVs (brand of cereal, calories).

What are “Groups” or “Levels”?

Groups or levels are different groups in the same independent variable. In the above example, your levels for “brand of cereal” might be Lucky Charms, Raisin Bran, Cornflakes — a total of three levels. Your levels for “Calories” might be: sweetened, unsweetened — a total of two levels.

Let’s say you are studying if Alcoholics Anonymous and individual counselling combined is the most effective treatment for lowering alcohol consumption. You might split the study participants into three groups or levels: medication only, medication and counseling, and counseling only. Your dependent variable would be the number of alcoholic beverages consumed per day.

What Does “Replication” Mean?

It’s whether you are replicating your test(s) with multiple groups. With a two way ANOVA *with replication*, you have two groups and individuals within that group are doing more than one thing (i.e. two groups of students from two colleges taking two tests). If you only have one group taking two tests, you would use **without replication.**

Types of Tests :

There are two main types: one-way and two-way. Two-way tests can be with or without replication.

- One-way ANOVA between groups: used when you want to test **two groups** to see if there’s a difference between them.
- Two way ANOVA without replication: used when you have **one group** and you’re **double-testing** that same group. For example, you’re testing one set of individuals before and after they take a medication to see if it works or not.
- Two way ANOVA with replication: **Two groups**, and the members of those groups are **doing more than one thing**. For example, two groups of patients from different hospitals trying two different therapies.

A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F-distribution. The null hypothesis for the test is that the two means are equal. Therefore, a significant result means that the two means are unequal.

When to use a one way ANOVA

Situation 1: You have a group of individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.

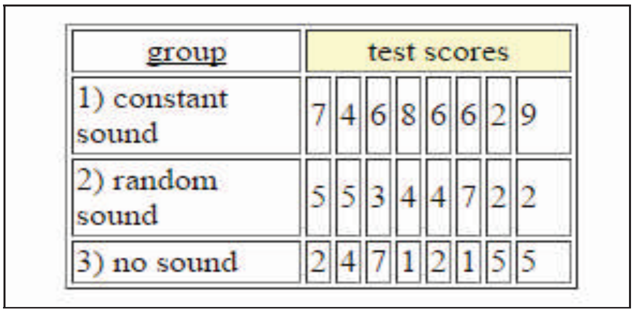
Situation 2: Similar to situation 1, but in this case the individuals are split into groups based on an attribute they possess. For example, you might be studying leg strength of people according to weight. You could split participants into weight categories (obese, overweight and normal) and measure their leg strength on a weight machine.

Limitations of the One Way ANOVA

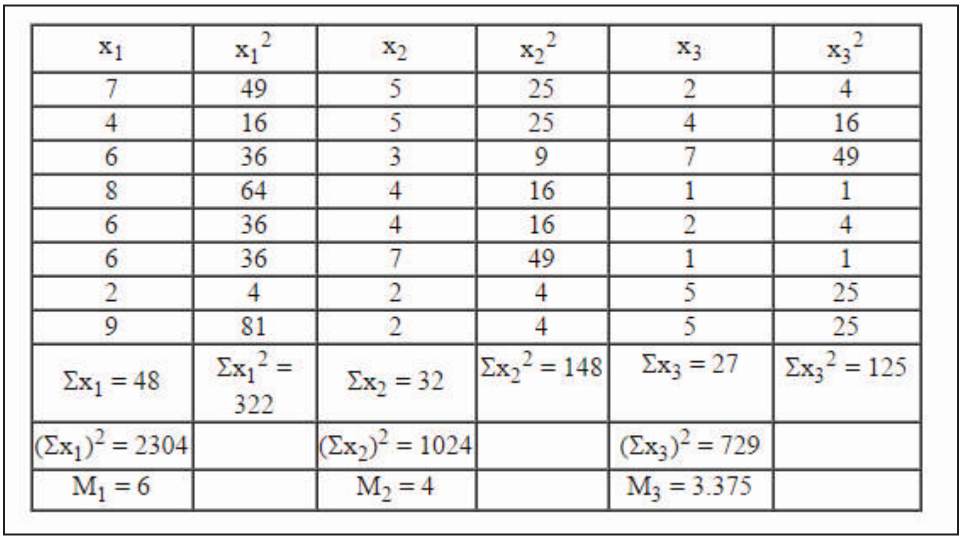
A one way ANOVA will tell you that at least two groups were different from each other. But it won't tell you what groups were different. If your test returns a significant F-statistic, you may need to run an ad hoc test (like the Least Significant Difference test) to tell you exactly which groups had a difference in means.

Illustrative Example 1

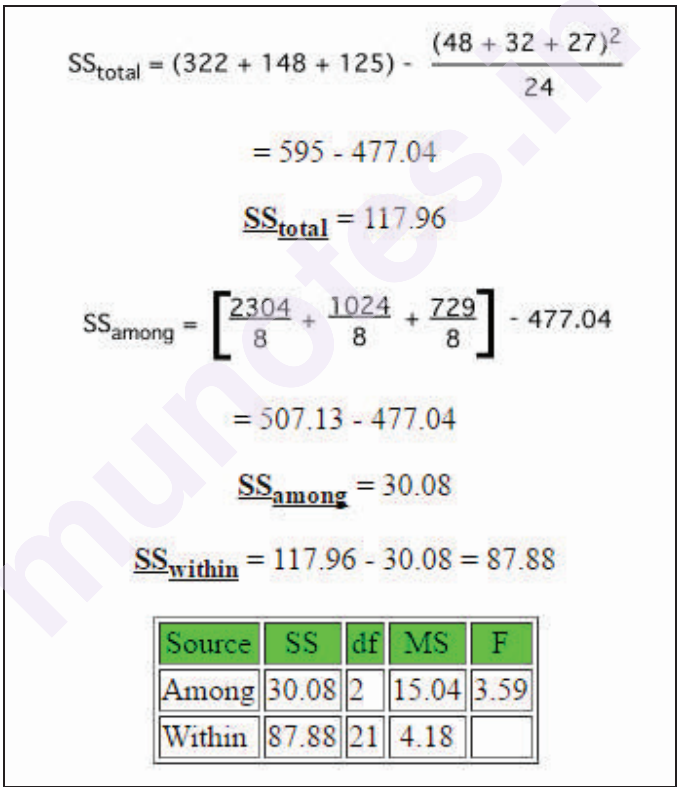
Susan Sound predicts that students will learn most effectively with a constant background sound, as opposed to an unpredictable sound or no sound at all. She randomly divides twenty-four students into three groups of eight. All students study a passage of text for 30 minutes. Those in group 1 study with background sound at a constant volume in the background. Those in group 2 study with noise that changes volume periodically. Those in group 3 study with no sound at all. After studying, all students take a 10 point multiple choice test over the material. Their scores follow:



group	test scores							
1) constant sound	7	4	6	8	6	6	2	9
2) random sound	5	5	3	4	4	7	2	2
3) no sound	2	4	7	1	2	1	5	5



x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
7	49	5	25	2	4
4	16	5	25	4	16
6	36	3	9	7	49
8	64	4	16	1	1
6	36	4	16	2	4
6	36	7	49	1	1
2	4	2	4	5	25
9	81	2	4	5	25
$\Sigma x_1 = 48$	$\Sigma x_1^2 = 322$	$\Sigma x_2 = 32$	$\Sigma x_2^2 = 148$	$\Sigma x_3 = 27$	$\Sigma x_3^2 = 125$
$(\Sigma x_1)^2 = 2304$		$(\Sigma x_2)^2 = 1024$		$(\Sigma x_3)^2 = 729$	
$M_1 = 6$		$M_2 = 4$		$M_3 = 3.375$	



$$SS_{total} = (322 + 148 + 125) - \frac{(48 + 32 + 27)^2}{24}$$

$$= 595 - 477.04$$

$$\underline{SS_{total}} = 117.96$$

$$SS_{among} = \left[\frac{2304}{8} + \frac{1024}{8} + \frac{729}{8}\right] - 477.04$$

$$= 507.13 - 477.04$$

$$\underline{SS_{among}} = 30.08$$

$$\underline{SS_{within}} = 117.96 - 30.08 = 87.88$$

Source	SS	df	MS	F
Among	30.08	2	15.04	3.59
Within	87.88	21	4.18	

Susan can conclude that her hypothesis may be supported. The means are as she predicted, in that the constant music group has the highest score. However, the significant F only indicates that at least two means are significantly different from one another, but she can't know which specific mean pairs significantly differ until she conducts a post-hoc analysis.

10.4 TWO WAY ANOVA WITHOUT REPLICATION

A Two Way ANOVA is an extension of the One Way ANOVA. With a One Way, you have one independent variable affecting a dependent variable. With a Two Way ANOVA, there are two independents. Use a two way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables. In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables, a two way ANOVA is appropriate.

For example, you might want to find out if there is an interaction between income and gender for anxiety level at job interviews. The anxiety level is the outcome, or the variable that can be measured. Gender and Income are the two categorical variables. These categorical variables are also the independent variables, which are called **factors** in a Two Way ANOVA.

The factors can be split into **levels**. In the above example, income level could be split into three levels: low, middle and high income. Gender could be split into three levels: male, female, and transgender with Treatment groups and all possible combinations of the factors. In this example there would be 3 x 3 = 9 treatment groups.

Main Effect and Interaction Effect

The results from a Two Way ANOVA will calculate a main effect and an interaction effect. The main effect is similar to a One Way ANOVA: each factor's effect is considered separately. With the interaction effect, all factors are considered at the same time. Interaction effects between factors are easier to test if there is more than one observation in each cell. For the above example, multiple stress scores could be entered into cells. If you do enter multiple observations into cells, the number in each cell must be equal.

Two null hypotheses are tested if you are placing one observation in each cell.

For this example, those hypotheses would be:

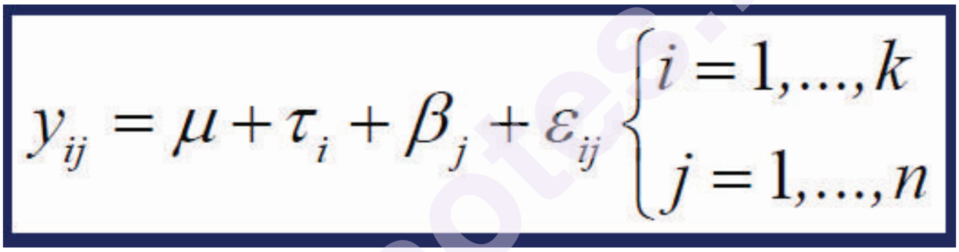
H01: All the income groups have equal mean stress.
H02: All the gender groups have equal mean stress.
For multiple observations in cells, you would also be testing a third hypothesis:
H03: The factors are independent *or* the interaction effect does not exist. An F-statistic is computed for each hypothesis you are testing.

Assumptions for Two Way ANOVA

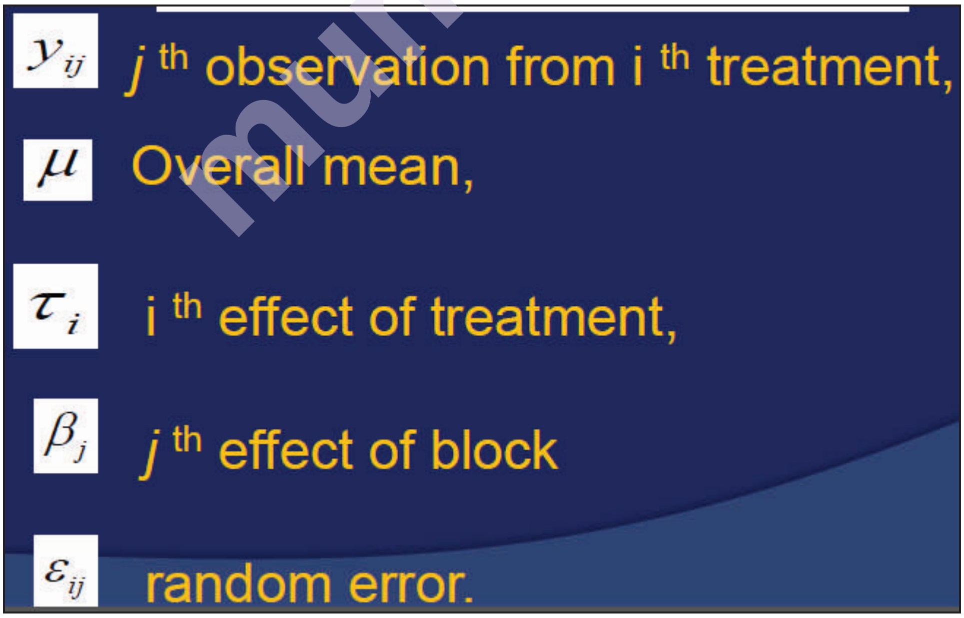
- The population must be close to a normal distribution.
- Samples must be independent.
- Population variances must be equal.
- Groups must have equal sample sizes.

Two-Way ANOVA without replication is also known as Randomized Block Design (RBD). In RBD there is one factor or variable that is of primary interest. However, there are also several other nuisance factors. Nuisance factors are those that may affect the measured result, but are not of primary interest. The way to control nuisance factor is by blocking them to reduce or eliminate the contribution to experimental error contributed by nuisance factors. A blocking variable is a second treatment variable that when included in ANOVA analysis will have the effect of reducing the SSE term.

The model for a randomized block design with one nuisance variable can be written as:



$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \begin{cases} i = 1,\ldots,k \\ j = 1,\ldots,n \end{cases}$$



y_{ij} *j* th observation from i th treatment,

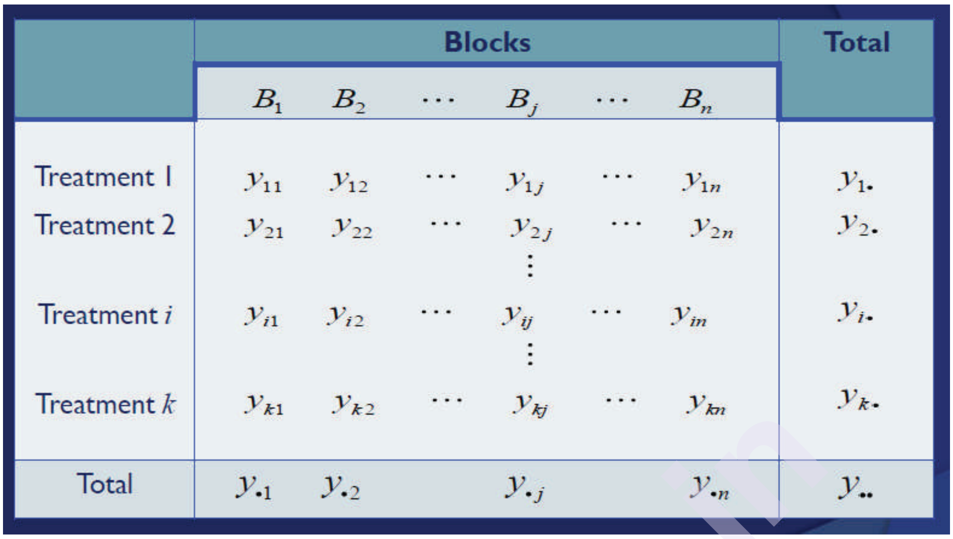
μ Overall mean,

τ_i i th effect of treatment,

β_j *j* th effect of block

ε_{ij} random error.

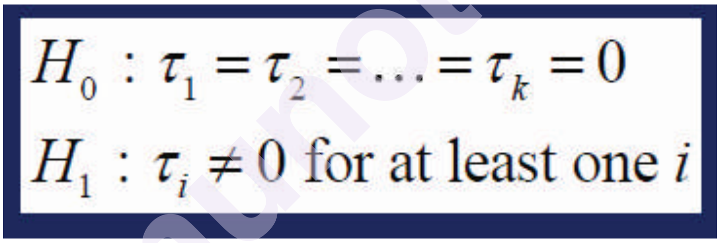
We can use the following layout for this kind of two-way classification:



	Blocks						Total
	B_1	B_2	$\cdots$	B_j	$\cdots$	B_n	
Treatment 1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n}	$y_{1\bullet}$
Treatment 2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n}	$y_{2\bullet}$
				$\vdots$			
Treatment *i*	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in}	$y_{i\bullet}$
				$\vdots$			
Treatment *k*	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn}	$y_{k\bullet}$
Total	$y_{\bullet 1}$	$y_{\bullet 2}$		$y_{\bullet j}$		$y_{\bullet n}$	$y_{\bullet\bullet}$

In the two way analysis of variance where each treatment is represented once in each block, the major objective is to test:

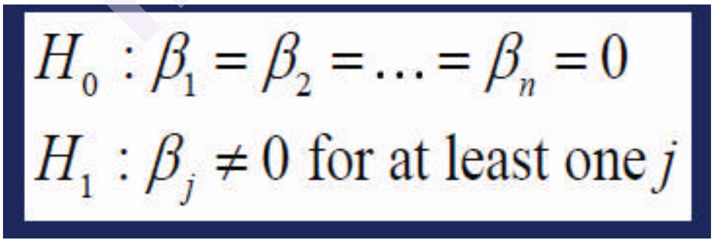
- The effect of treatment:



$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0$$

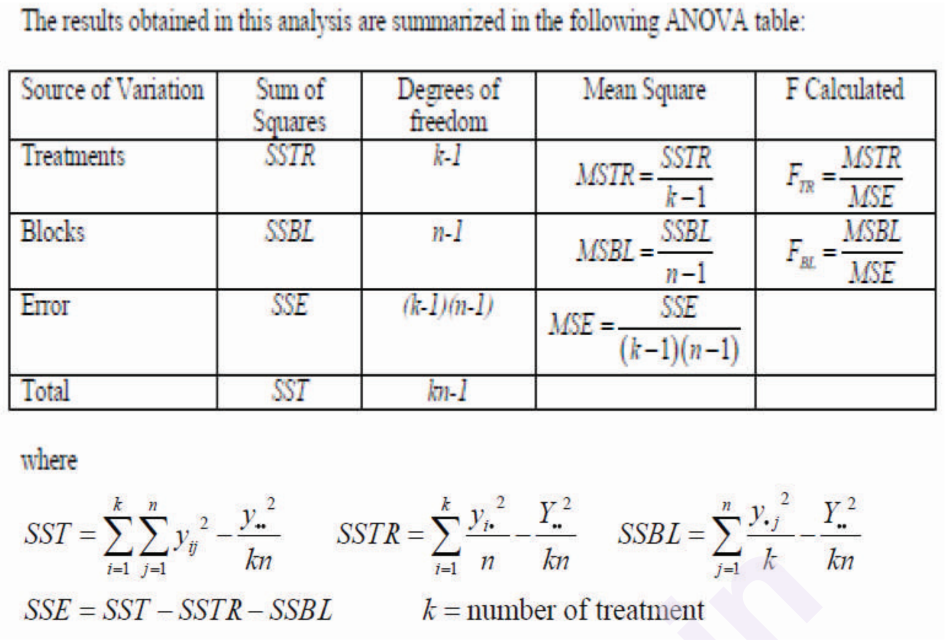
$$H_1 : \tau_i \neq 0 \text{ for at least one } i$$

- The effect of block:



$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_n = 0$$

$$H_1 : \beta_j \neq 0 \text{ for at least one } j$$



The results obtained in this analysis are summarized in the following ANOVA table:

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F Calculated
Treatments	SSTR	k-1	$MSTR = \frac{SSTR}{k-1}$	$F_{TR} = \frac{MSTR}{MSE}$
Blocks	SSBL	n-1	$MSBL = \frac{SSBL}{n-1}$	$F_{BL} = \frac{MSBL}{MSE}$
Error	SSE	(k-1)(n-1)	$MSE = \frac{SSE}{(k-1)(n-1)}$	
Total	SST	kn-1		

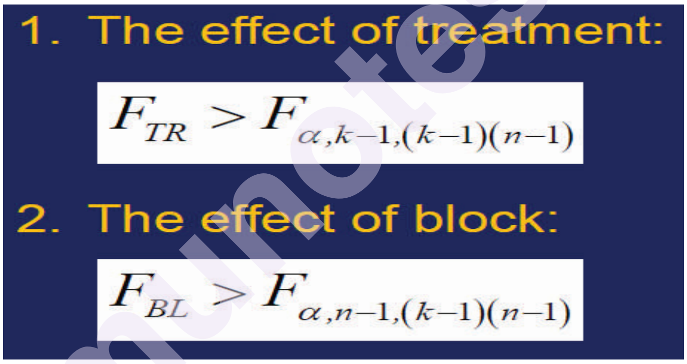
where

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^{2} - \frac{y_{..}^{2}}{kn} \qquad SSTR = \sum_{i=1}^{k} \frac{y_{i.}^{2}}{n} - \frac{Y_{..}^{2}}{kn} \qquad SSBL = \sum_{j=1}^{n} \frac{y_{.j}^{2}}{k} - \frac{Y_{..}^{2}}{kn}$$

$SSE = SST - SSTR - SSBL$ $\qquad k =$ number of treatment

n = number of block

We reject H0 if:



1. The effect of treatment:

$$F_{TR} > F_{\alpha,k-1,(k-1)(n-1)}$$

2. The effect of block:

$$F_{BL} > F_{\alpha,n-1,(k-1)(n-1)}$$

Illustrative Example 2 – ANOVA without replication

An experiment was designed to study the performance of 4 different detergents for cleaning fuel injectors. The following "cleanliness readings" were obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines:

	Engine 1	Engine 2	Engine 3
Detergent A	45	43	51
Detergent B	47	46	52
Detergent C	48	50	55
Detergent D	42	37	49

Looking on the detergents as treatments and the engines as the blocks, obtain the appropriate ANOVA table and test at 0.01 level of significance whether there are differences in the detergents or in the engines.

Solution

Step 1: Construct the table of calculation, we have k = 4 and n = 3:

	Engine 1	Engine 2	Engine 3	Totals
Detergent A	45	43	51	139
Detergent B	47	46	52	175
Detergent C	48	50	55	153
Detergent D	42	37	49	128
Totals	182	176	207	565

Step 2: Set up the hypothesis

$$H_0\colon\ \tau_1 = \tau_2 = \ldots = \tau_k = 0$$

$$H_1\colon\ \tau_i \neq 0 \text{ for at least one } i$$

or

$$H_0\colon\ \beta_1 = \beta_2 = \ldots = \beta_n = 0$$

$$H_1\colon\ \beta_j \neq 0 \text{ for at least one } j$$

Step 3: Construct the ANOVA table

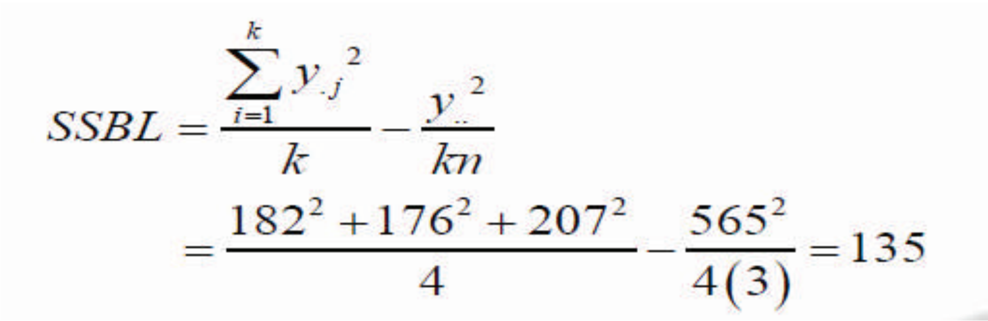
$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^{\ 2} - \frac{y_{..}^{\ 2}}{kn}$$

$$= \left(45^2 + 43^2 + 51^2 + \ldots + 42^2 + 37^2 + 49^2\right) - \frac{565^2}{4(3)}$$

$$= 26867 - 26602 = 265$$

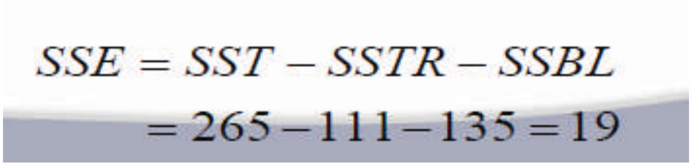
$$SSTR = \frac{\sum_{i=1}^{k} y_{i.}^{\ 2}}{n} - \frac{y_{..}^{\ 2}}{kn}$$

$$= \frac{139^2 + 145^2 + 153^2 + 128^2}{3} - \frac{565^2}{4(3)} = 111$$



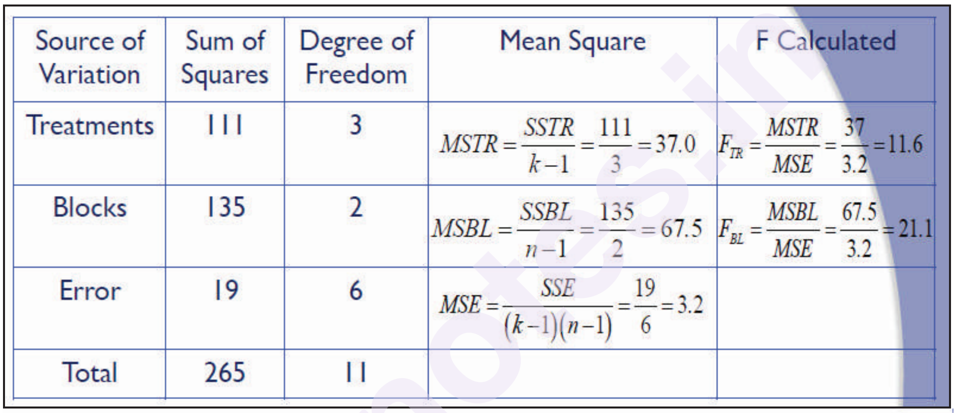
$$SSBL = \frac{\sum_{i=1}^{k} y_{.j}^{2}}{k} - \frac{y_{..}^{2}}{kn}$$

$$= \frac{182^2 + 176^2 + 207^2}{4} - \frac{565^2}{4(3)} = 135$$

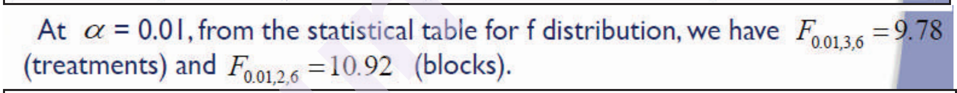


$$SSE = SST - SSTR - SSBL$$

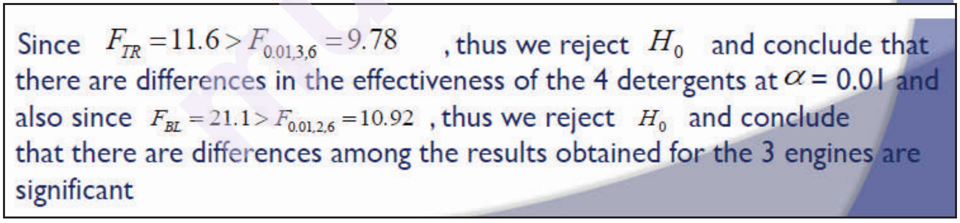
$$= 265 - 111 - 135 = 19$$



Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F Calculated
Treatments	111	3	$MSTR = \frac{SSTR}{k-1} = \frac{111}{3} = 37.0$	$F_{TR} = \frac{MSTR}{MSE} = \frac{37}{3.2} = 11.6$
Blocks	135	2	$MSBL = \frac{SSBL}{n-1} = \frac{135}{2} = 67.5$	$F_{BL} = \frac{MSBL}{MSE} = \frac{67.5}{3.2} = 21.1$
Error	19	6	$MSE = \frac{SSE}{(k-1)(n-1)} = \frac{19}{6} = 3.2$	
Total	265	11		



At $\alpha = 0.01$, from the statistical table for f distribution, we have $F_{0.01,3,6} = 9.78$ (treatments) and $F_{0.01,2,6} = 10.92$ (blocks).



Since $F_{TR} = 11.6 > F_{0.01,3,6} = 9.78$, thus we reject H_0 and conclude that there are differences in the effectiveness of the 4 detergents at $\alpha = 0.01$ and also since $F_{BL} = 21.1 > F_{0.01,2,6} = 10.92$, thus we reject H_0 and conclude that there are differences among the results obtained for the 3 engines are significant

10.5 TWO WAY ANOVA WITH REPLICATION

Two Way ANOVA with replication is also called a Factorial Experiment. Factorial Experiment is used to evaluate 2 or more factors simultaneously. Replication means an independent repeat of each factor combination.

The purpose of factorial experiment is to examine:

- The effect of factor A on the dependent variable, y.
- The effect of factor B on the dependent variable, y along with
- The effects of the interactions between different levels of the factors on the dependent variable, y.

Interaction exists when the effect of a level for one factor depends on which level of the other factor is present.

Factorial Experiment over one factor at a time (one-way ANOVA) is more efficient & allows interactions to be detected.

The effect model for a factorial experiment can be written as:

$$y_{ijk} = \mu + \tau_i + \beta_j + \left(\tau\beta\right)_{ij} + \varepsilon_{ijk} \begin{cases} i = 1,\dots,a \\ j = 1,\dots,b \\ k = 1,\dots,r \end{cases}$$

y_{ijk} : The response from the kth experimental unit receiving the *i*th level of factor *A* and the *j*th level of factor *B*

μ : Overall mean

τ_i : An effect due to the *i*th level of factor *A*

β_j : An effect due to the *j*th level of factor *B*

$\tau\beta_{ij}$: An interaction effect of the *i*th level of factor *A* with jth level of factor *B*

ε_{ijk} : A random error associated with the response from the *k*th experimental unit receiving the *i*th level of factor *A* combined with jth level of factor *B*

There are three sets of hypothesis:

- Factor A Effect:

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$$
$$H_1 : \text{ at least one } \tau_i \neq 0$$

- Factor B Effect:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_b = 0$$
$$H_1 : \text{ at least one } \beta_j \neq 0$$

- Interaction Effect:

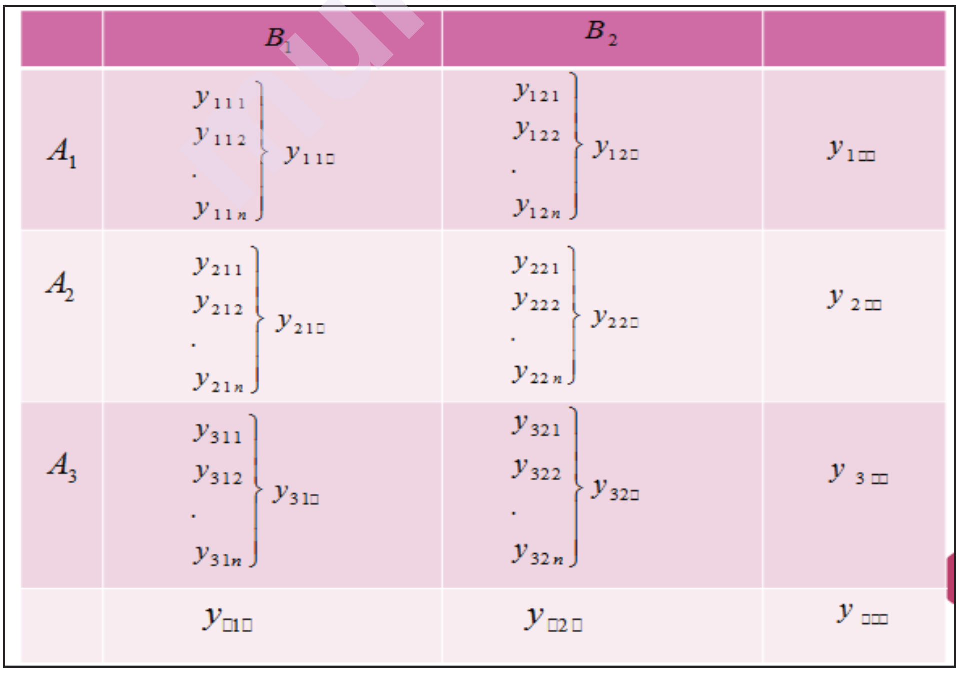
$$H_0 : (\tau\beta)_{ij} = 0 \text{ for all i,j}$$

$$H_1 : \text{ at least one } (\tau\beta)_{ij} \neq 0$$

The results obtained in this analysis are summarized in the following ANOVA table:

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F Calculated
Factor A	*SSA*	*a-1*	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$
Factor B	*SSB*	*b-1*	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Interaction	*SSAB*	*(a-1)(b-1)*	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error	*SSE*	*ab(r-1)*	$MSE = \frac{SSE}{ab(r-1)}$	
Total	*SST*	*abr-1*		

Two way Factorial Treatment Structure



	B_1	B_2	
A_1	$\left.\begin{matrix} y_{111} \\ y_{112} \\ . \\ y_{11n} \end{matrix}\right\} y_{11\square}$	$\left.\begin{matrix} y_{121} \\ y_{122} \\ . \\ y_{12n} \end{matrix}\right\} y_{12\square}$	$y_{1\square\square}$
A_2	$\left.\begin{matrix} y_{211} \\ y_{212} \\ . \\ y_{21n} \end{matrix}\right\} y_{21\square}$	$\left.\begin{matrix} y_{221} \\ y_{222} \\ . \\ y_{22n} \end{matrix}\right\} y_{22\square}$	$y_{2\square\square}$
A_3	$\left.\begin{matrix} y_{311} \\ y_{312} \\ . \\ y_{31n} \end{matrix}\right\} y_{31\square}$	$\left.\begin{matrix} y_{321} \\ y_{322} \\ . \\ y_{32n} \end{matrix}\right\} y_{32\square}$	$y_{3\square\square}$
	$y_{\square 1\square}$	$y_{\square 2\square}$	$y_{\square\square\square}$

$$SSA = \frac{\sum y_{i\square\square}^{2}}{br} - \frac{y^{2}_{\square\square\square}}{abr}$$

$$SSB = \frac{\sum y_{\square j\square}^{2}}{ar} - \frac{y^{2}_{\square\square\square}}{abr}$$

$$SSAB = \frac{\sum y_{ij\square}^{2}}{r} - \frac{y^{2}_{\square\square\square}}{abr} - SSA - SSB$$

$$SST = \sum\sum\sum y_{ijk}^{2} - \frac{y^{2}_{\square\square\square}}{abr}$$

$$SSE = SST - SSA - SSB - SSAB$$

Reject H_o if :

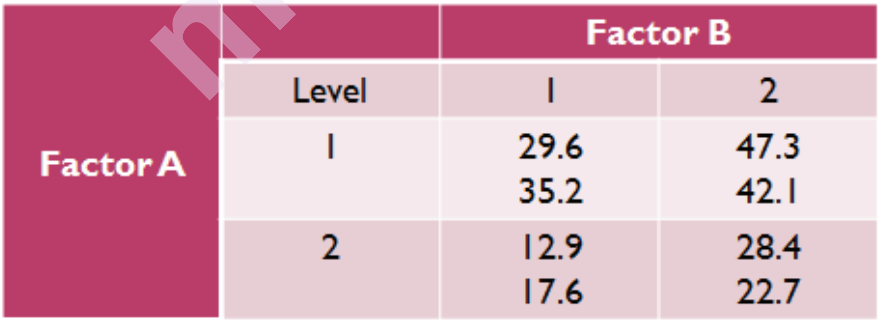
(1) $F_A > F_{\alpha, a-1, ab(r-1)}$ **for effect of factor A.**

(2) $F_B > F_{\alpha, b-1, ab(r-1)}$ **for effect of factor B.**

(3) $F_{AB} > F_{\alpha, (a-1)(b-1), ab(r-1)}$ **for effect of interaction.**

Illustrative Example 3 – ANOVA with replication

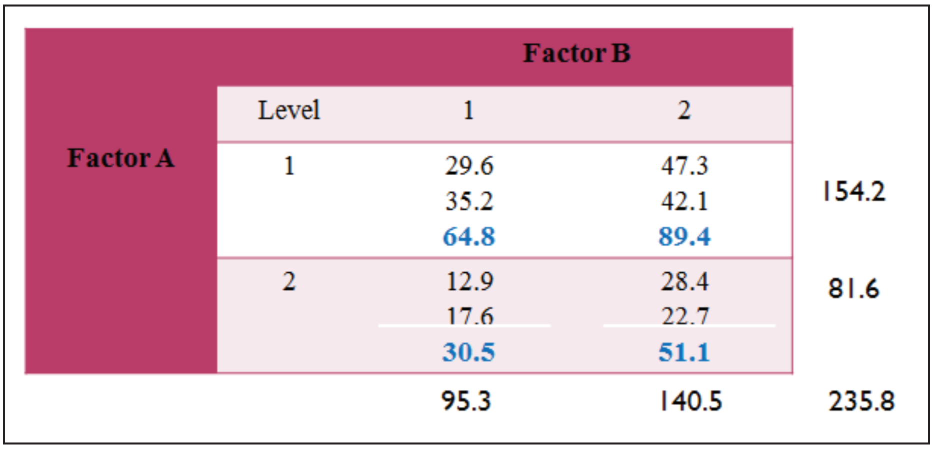
The two-way table gives data for a 2x2 factorial experiment with two observations per factor – level combination.



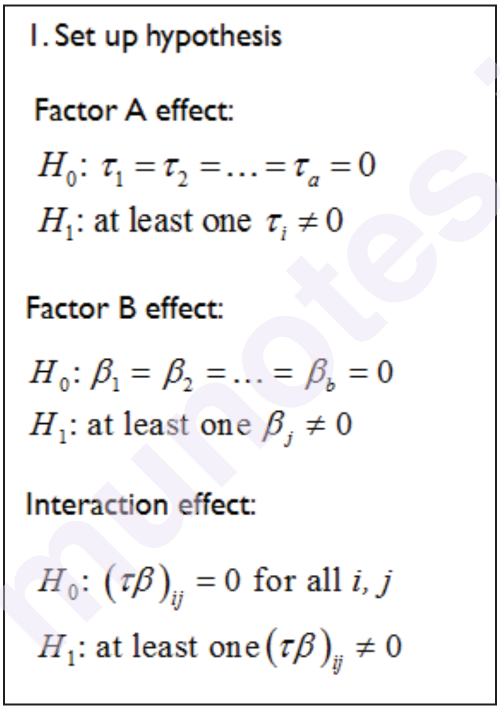
		Factor B	
	Level	1	2
Factor A	1	29.6 35.2	47.3 42.1
	2	12.9 17.6	28.4 22.7

Construct the ANOVA table for this experiment and do a complete analysis at a level of significance 0.05.

Solution:



		Factor B		
Factor A	Level	1	2	
	1	29.6 35.2 64.8	47.3 42.1 89.4	154.2
	2	12.9 17.6 30.5	28.4 22.7 51.1	81.6
		95.3	140.5	235.8



1. Set up hypothesis

Factor A effect:

H_0: $\tau_1 = \tau_2 = \ldots = \tau_a = 0$

H_1: at least one $\tau_i \neq 0$

Factor B effect:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_b = 0$

H_1: at least one $\beta_j \neq 0$

Interaction effect:

H_0: $(\tau\beta)_{ij} = 0$ for all i, j

H_1: at least one $(\tau\beta)_{ij} \neq 0$

$$SST = \sum\sum\sum y_{ijk}^{\ 2} - \frac{y^2_{\square\square\square}}{abr}$$

$$= \left(29.6^2 + 35.2^2 + \cdots + 22.7^2\right) - \frac{235.8^2}{(2)(2)(2)}$$

$$= 972.715$$

$$SSA = \frac{\sum y_{i\square\square}^{\ 2}}{br} - \frac{y^2_{\square\square\square}}{abr}$$

$$= \left(\frac{154.2^2 + 81.6^2}{4}\right) - \frac{235.8^2}{8}$$

$$= 658.845$$

$$SSB = \frac{\sum y_{\square j\square}^{\ 2}}{ar} - \frac{y^2_{\square\square\square}}{abr}$$

$$= \left(\frac{95.3^2 + 140.5^2}{4}\right) - \frac{235.8^2}{8}$$

$$= 255.38$$

$$SSAB = \frac{\sum\sum y_{ij\square}^{\ 2}}{r} - \frac{y^2_{\square\square\square}}{abr} - SSA - SSB$$

$$= \left(\frac{64.8^2 + 89.4^2 + 30.5^2 + 51.1^2}{2}\right) - 658.845 - 255.38 - \frac{235.8^2}{8}$$

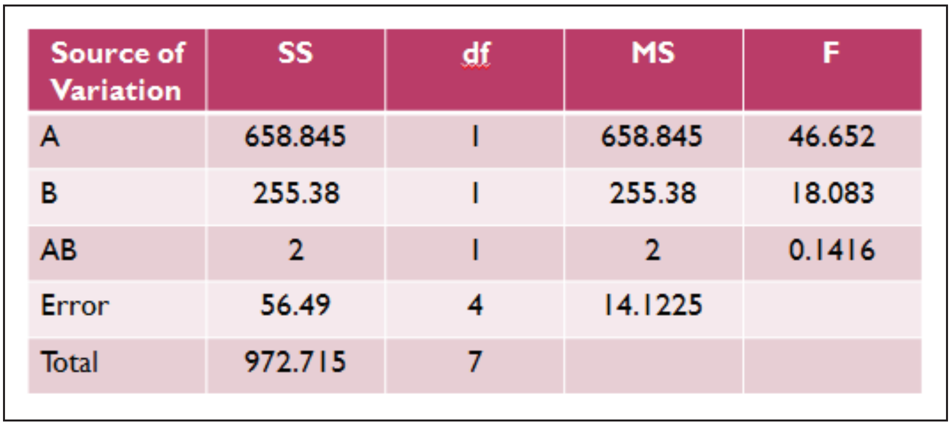
$$= 2$$

$$SSE = SST - SSA - SSB - SSAB$$

$$= 972.715 - 658.845 - 255.38 - 2$$

$$= 56.49$$

Calculation (given the ANOVA table is as follows):



Source of Variation	SS	df	MS	F
A	658.845	1	658.845	46.652
B	255.38	1	255.38	18.083
AB	2	1	2	0.1416
Error	56.49	4	14.1225	
Total	972.715	7		

With α = 0.05 we reject H_0 if :

$$F_A > F_{\alpha,a-1,ab(r-1)} \text{ for effect of factor A}$$

$$F_B > F_{\alpha,b-1,ab(r-1)} \text{ for effect of factor B}$$

$$F_{AB} > F_{\alpha,(a-1)(b-1),ab(r-1)} \text{ for effect of interaction}$$

From ANOVA/table F, the critical and F effects are given as follow:

$$F_A = 46.652 \text{ and } F_{\alpha,a-1,ab(r-1)} = F_{0.05,1,4} = 7.71$$

$$F_B = 18.083 \text{ and } F_{\alpha,b-1,ab(r-1)} = F_{0.05,1,4} = 7.71$$

$$F_{AB} = 0.1416 \text{ and } F_{\alpha,(a-1)(b-1),ab(r-1)} = F_{0.05,1,7} = 7.71$$

Factor A : since $F_A = 46.652 > F_{0.05,1,4} = 7.71$, thus we reject H_0

We conclude that the difference level of A effect the response

Factor B : since $F_B = 18.083 > F_{0.05,1,4} = 7.71$, thus we reject H_0

We conclude that the difference level of B effect the response

Interaction: since $F_{AB} = 0.1416 < F_{0.05,1,4} = 7.71$, thus we failed to reject H_0

We conclude that no interaction between factor A and factor B.

10.6 LET US SUM UP

In this unit you have learnt the concepts the concepts of One Way ANOVA, Two Way ANOVA without replication and Two Way ANOVA with replication.

10.7. EXERCISES

Q.1: A chemical engineer is studying the effects of various reagents and catalysts on the yield of a certain process. There is a combination among three reagents and four catalysts. The results are presented below:

	Reagent 1	Reagent 2	Reagent 3
Catalyst A	84.85	89.13	85.28
Catalyst B	75.35	79.40	84.65
Catalyst C	70.30	76.65	78.20
Catalyst D	73.18	81.10	77.23

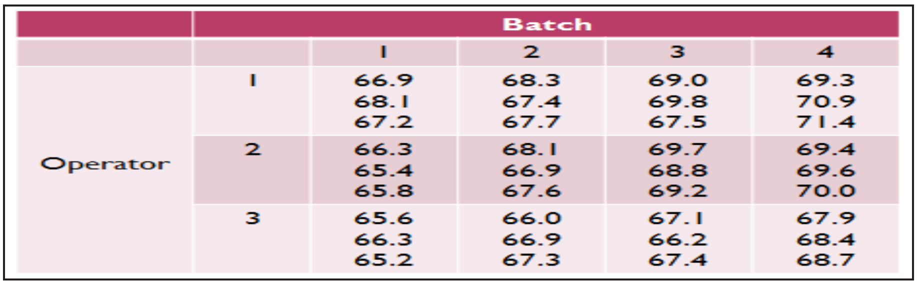
Develop a complete analysis of variance table for two factor and it is possible the main effects of catalyst (Treatments) and reagent (Blocks) are different at α=0.01?

Q.2: WARTA, the Warren Area Regional Transit Authority, is expanding bus service from the suburb of Star brick into the central business district of Warren. There are four routes being considered from Star brick to downtown Warren. WARTA conducted several tests to determine whether there was a difference in the mean travel times along the four routes. Because there will be many different drivers, the test was set up so each driver drove along each of the four routes. Below is the travel time, in minutes for each driver-route combination.

	Route 1	Route 1	Route 1	Route 1
Driver 1	18	17	21	22
Driver 2	16	23	23	22
Driver 3	21	21	26	22
Driver 4	23	22	29	25
Driver 5	25	24	28	28

At α=0.05, is there any possible differences in treatments and also in blocks?

Q.3. In a study to determine which the important source of variation is in an industrial process, 3 measurements are taken on yield for 3 operators chosen randomly and 4 batches a raw materials chosen randomly. It was decided that a significance test should be made at the0.05 level of significance to determine if the variance components due to batches, operators, and interaction are significant. In addition, estimates of variance components are to be computed. The data are as follows with the response being percent by weight.



		Batch			
		1	2	3	4
Operator	1	66.9 68.1 67.2	68.3 67.4 67.7	69.0 69.8 67.5	69.3 70.9 71.4
	2	66.3 65.4 65.8	68.1 66.9 67.6	69.7 68.8 69.2	69.4 69.6 70.0
	3	65.6 66.3 65.2	66.0 66.9 67.3	67.1 66.2 67.4	67.9 68.4 68.7

Perform the analysis of variance of this experiment at level of significance 0.05. State your conclusion

10.8 SUGGESTED READINGS

ANOVA section in any of the reference / text books

❖❖❖❖

Module - VI
ANOVA & CHI-SQUARE TEST

11

INTRODUCTION TO ANOVA

Introduction to Chi Square, Defining Chi-Square Statistics
Chi-Square Goodness of Fit Test

Unit Structure :

11.1 INTRODUCTION

In this Unit-VI - Chapter 6.2, we shall discuss the concepts of Chi Square; define Chi-Square Statistics and Chi-Square Goodness of Fit Test.

6.2.2 OBJECTIVES

At the end of this unit the learners will be able to understand:

- What is Chi Square?
- Chi-Square Statistics
- Chi-Square Goodness of Fit Test.

11.3 INTRODUCTION TO CHI SQUARE

The chi square distribution is a theoretical or mathematical distribution which has wide applicability in statistical work. The term `chi square' (pronounced with a hard `ch') is used because the Greek letter χ is used to define this distribution. It will be seen that the elements on which this distribution is based are squared, so that the symbol χ^2 is used to denote the distribution.

The Chi-Square Statistic :

Suppose we conduct the following statistical experiment. We select a random sample of size n from a normal population, having a standard deviation equal to σ. We find that the standard deviation in our sample is equal to s. Given these data, we can define a statistic, called chi-square, using the following equation:

$$\boxed{X^2 = [(n - 1) * s^2] / \sigma^2}$$

The distribution of the chi-square statistic is called the chi-square distribution. The **chi-square distribution** is defined by the following probability density function:

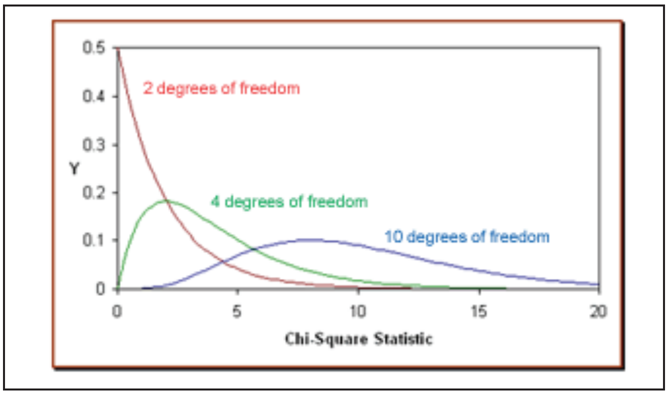
$$\boxed{Y = Y_0 * (X^2)^{(v/2 - 1)} * e^{-X^2 / 2}}$$

where Y_0 is a constant that depends on the number of degrees of freedom, X^2 is the chi-square statistic, $v = n - 1$ is the number of degrees of freedom, and *e* is a constant equal to the base of the natural logarithm system (approximately 2.71828). Y_0 is defined, so that the area under the chi-square curve is equal to one.

An example of the chi squared distribution is given in the figure below. Along the horizontal axis is the χ^2 value. The minimum possible value for a χ^2 variable is 0, but there is no maximum value. The vertical axis is the probability, or probability density, associated with each value of χ^2. The curve reaches a peak not far above 0, and then declines slowly as the χ^2 value increases, so that the curve is asymmetric. As with the distributions introduced earlier, as larger χ^2 values are obtained, the curve is asymptotic to the horizontal axis, always approaching it, but never quite touching the axis.

The red curve shows the distribution of chi-square values computed from all possible samples of size 3, where degrees of freedom is *n* - 1 = 3 - 1 = 2.

Similarly, the green curve shows the distribution for samples of size 5 (degrees of freedom equal to 4); and the blue curve, for samples of size 11 (degrees of freedom equal to 10).



The chi-square distribution has the following properties:

- The mean of the distribution is equal to the number of degrees of freedom:
  $\mu = v$.

- The variance is equal to two times the number of degrees of freedom:
  $\sigma^2 = 2 * v$

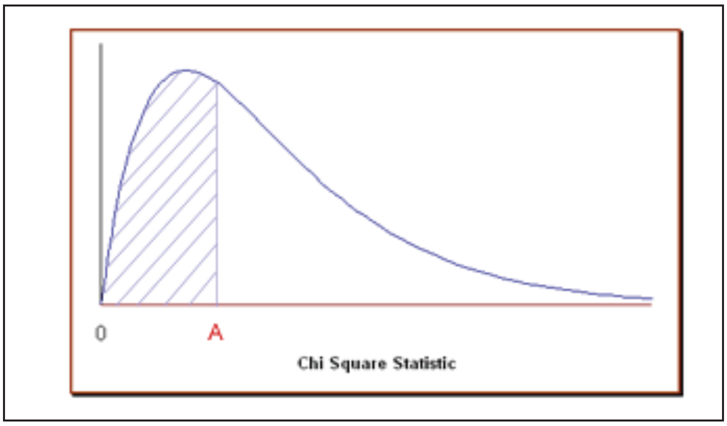
- When the degrees of freedom are greater than or equal to 2, the maximum value for Y occurs when $X^2 = v - 2$.

- As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

11.4 DEFINING CHI-SQUARE STATISTICS

Cumulative Probability and the Chi-Square Distribution

The chi-square distribution is constructed so that the total area under the curve is equal to 1. The area under the curve between 0 and a particular chi-square value is a cumulative probability associated with that chi-square value. For example, in the figure below, the shaded area represents a cumulative probability associated with a chi-square statistic equal to A; that is, it is the probability that the value of a chi-square statistic will fall between 0 and A.



The aim of the Chi^2 test is to test a theory by comparing observed numbers with those expected, we asses if any observed discrepancies from our theory can be reasonably put down to chance.

- From a random sample, we observe the numbers Oi, falling into the ith of k categories.
- We calculate the expected numbers Ei, in each category i, assuming the theory or null hypothesis.
- We then test if the differences Oi - Ei large enough to be inconsistent with the theory.
- Chi-Square Statistic,

$$\mathbf{G} = \Sigma_{i=l}^{k} \frac{(O_i - E_i)^2}{E_i}$$

For a large sample size and assuming the theory (null hypothesis) is true G has an approximate Chi-square distribution with (k – 1) degrees of freedom.

$$\mathbf{G} \approx \chi_{k-1}^{2}$$

Illustrative Example 1

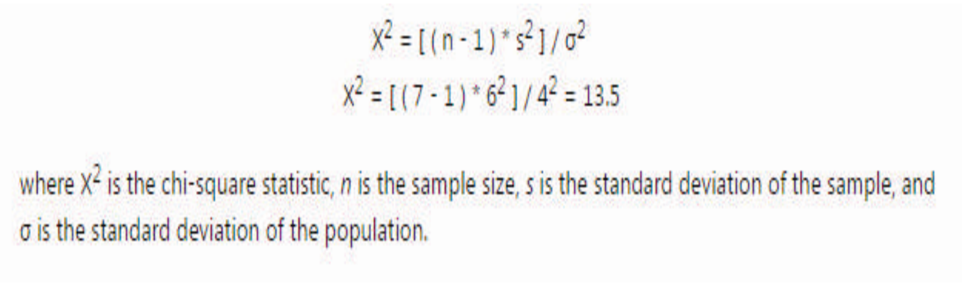
The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

We know the following:

- The standard deviation of the population is 4 minutes.
- The standard deviation of the sample is 6 minutes.
- The number of sample observations is 7.

To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.



$$X^2 = [\,(n - 1) * s^2\,] / \sigma^2$$

$$X^2 = [\,(7 - 1) * 6^2\,] / 4^2 = 13.5$$

where X^2 is the chi-square statistic, *n* is the sample size, *s* is the standard deviation of the sample, and σ is the standard deviation of the population.

Illustrative Example 2

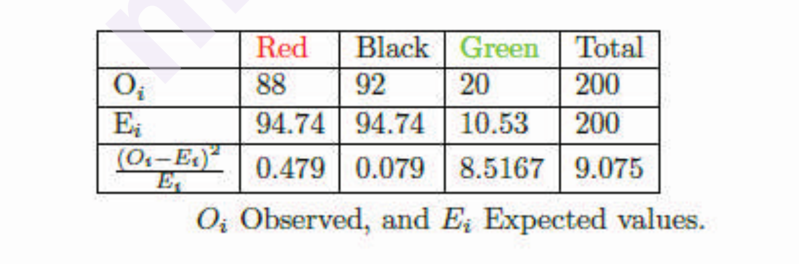
Let's revisit the problem presented above. The manufacturing department ran a quality control test, using 7 randomly selected batteries. In their test, the standard deviation was 6 minutes, which equated to a chi-square statistic of 13.5.

Suppose they repeated the test with a new random sample of 7 batteries. What is the probability that the standard deviation in the new test would be greater than 6 minutes?
We know the following:

- The sample size n is equal to 7.
- The degrees of freedom are equal to n - 1 = 7 - 1 = 6.
- The chi-square statistic is equal to 13.5 (see Example 1 above).

Illustrative Example 3

- 200 plays
- Wheel has 18 reds, 18 balcks, and 2 greens.
- All 38 numbers are equally likely.



	Red	Black	Green	Total
O_i	88	92	20	200
E_i	94.74	94.74	10.53	200
$\frac{(O_i - E_i)^2}{E_i}$	0.479	0.079	8.5167	9.075

O_i Observed, and E_i Expected values.

Hence G = 9:075, and from the Chi-Square distribution on 2 degrees of freedom we obtain the p - value = P (G >9:075) = 0:0107 (i.e. the probability of observing this test statistic or something more extreme given the data). This is strong evidence against the numbers being equally likely, we would therefore reject the theory that all numbers are equally likely.

11.5 CHI-SQUARE GOODNESS OF FIT TEST

The chi-squared test can be used to test for the independence or goodness of fit of a distribution to the given data set, that is you will be given observed values and you will test to see if it fits a certain probability distribution such as normal, binomial, Poisson, or any other.

The procedure to test for the goodness of fit

1. State a hypotheses based on the fit of the data, e.g.
   H0: the data fits a binomial distribution.
   H1: the data does not fit a binomial distribution.

2. Make a table of the observed and expected values. You will most likely be given the observed values.

3. Calculate the chi-squared test statistic, this is
   $$\mathbf{G} = \Sigma_{i=l}^{k} \frac{(O_i - E_i)^2}{E_i}$$

4. Look up the chi-squared critical value from your chi-squared tables in the information booklet.

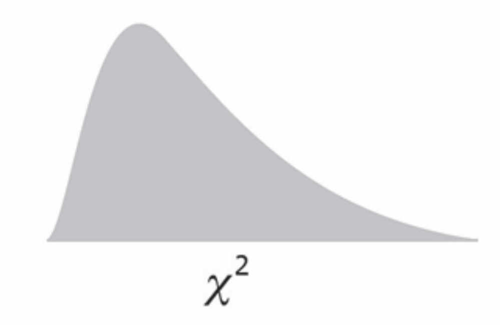
5. Compare your test statistic with your critical value and make a conclusion. If the test statistic lies in the critical region then reject H0 in favor of H1. Otherwise do not reject H0 in favor of H1

Degrees of freedom, *v* :

When undertaking a chi-squared test you will have a table of observed and expected values. The degrees of freedom will be defined as:

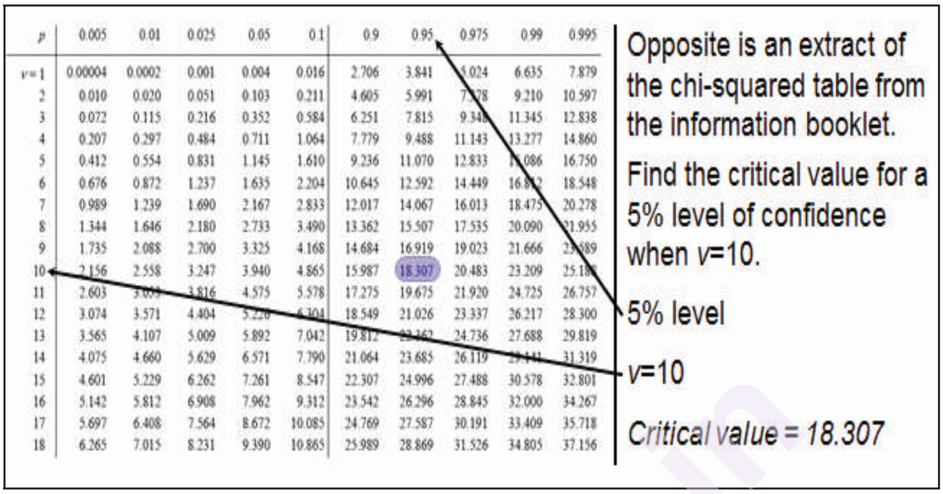
v=(number of rows-1)(number of columns-1)

The chi-squared distribution



The distribution will alter depending on the value of *v*. The general curve is shown opposite.

Finding the critical values from a table



p	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
v=1	0.00004	0.0002	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156

Opposite is an extract of the chi-squared table from the information booklet.

Find the critical value for a 5% level of confidence when *v*=10.

5% level

v=10

Critical value = 18.307

Find the critical value for:

1. *v*=4 at the 1% level of confidence
2. *v*=15 at the 10% level of confidence
3. *v*=8 at the 5% level of confidence

Answers:

1. 13.277
2. 22.307
3. 15.507

Illustrative Example 3 - Chi-Square Goodness of Fit – Binomial Distribution

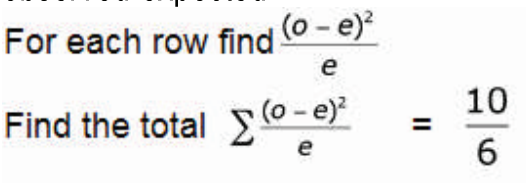
A group of students are testing 5 dice. The students split into 5 groups and each roll a die 36 times, counting the number of sixes. Their results are shown below. Test at the 5% level of significance that the dice follow a binomial distribution.

Group	Observed 6s	Expected 6s	o-e	$\frac{(o-e)^2}{e}$
1	5	6	-1	$\frac{1}{6}$
2	7	6	1	$\frac{1}{6}$
3	8	6	2	$\frac{4}{6}$
4	6	6	0	0
5	4	6	-2	$\frac{4}{6}$
			$\sum\frac{(o-e)^2}{e}$	$\frac{10}{6}$

Make your hypotheses:

H0: the die fit a binomial distribution
H1: the die do not fit a binomial distribution.

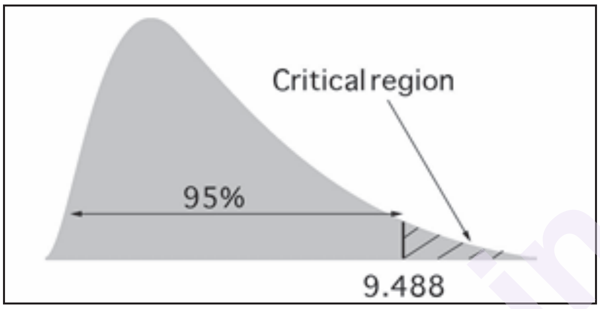
Make an expected column in your table.For each row find observed-expected



For each row find $\frac{(o-e)^2}{e}$

Find the total $\sum \frac{(o-e)^2}{e} \quad = \quad \frac{10}{6}$

This is your chi-squared test statistic. Look up the critical value in your chi-squared table. v = 4, 95% level:

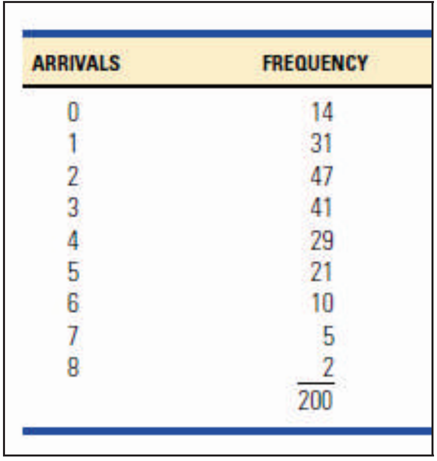


As the test statistic does not fall in the critical region we do not reject H0, e.g. at the 95% level of significance the dice follow a binomial model

Illustrative Example 4 - Chi-Square Goodness of Fit – Poisson Distribution :

Poisson distribution is used to model the number of arrivals per minute at a bank located in the central business district of a city. Suppose that the actual arrivals per minute were observed in 200 one-minute periods over the course of a week. The results are summarized in the table below:

Frequency distribution of arrivals per minute during a lunch period



ARRIVALS	FREQUENCY
0	14
1	31
2	47
3	41
4	29
5	21
6	10
7	5
8	2
	200

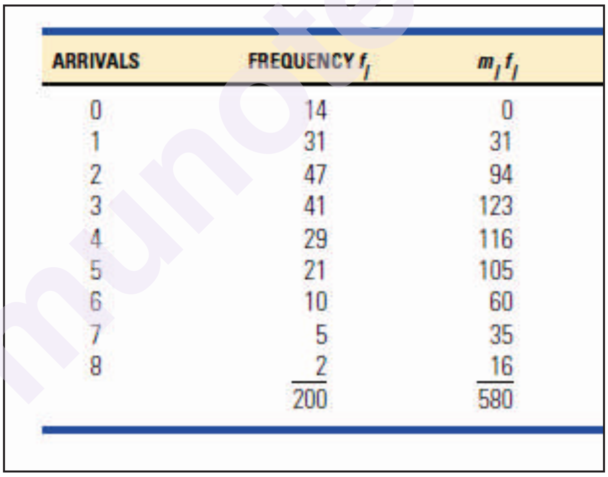
To determine whether the number of arrivals per minute follows a Poisson distribution, the null and alternative hypotheses are as follows:
H0: The number of arrivals per minute follows a Poisson distribution
H1: The number of arrivals per minute does not follow a Poisson distribution

Since the Poisson distribution has one parameter, its mean λ, either a specified value can beincluded as part of the null and alternative hypotheses, or the parameter can be estimated from thesample data.

$$\overline{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

$$\overline{X} = \frac{580}{200} = 2.90$$

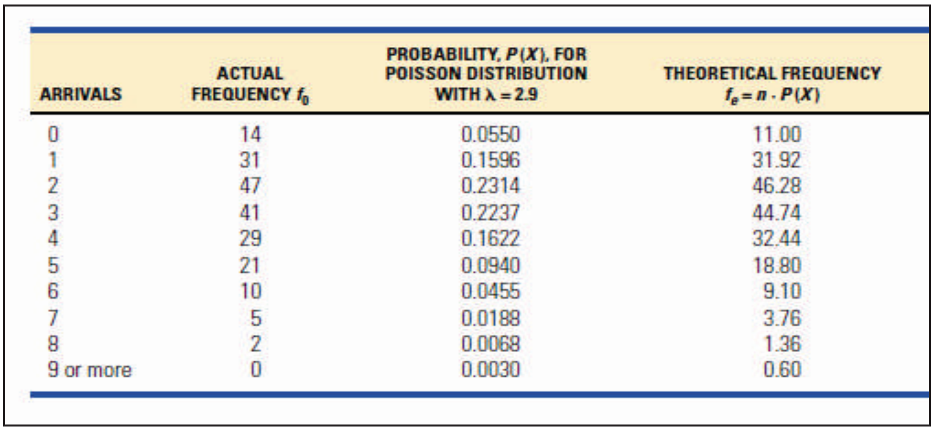
Computation of the sample average number of arrivals from the frequency distribution of arrivals per minute



ARRIVALS	FREQUENCY f_j	$m_j f_j$
0	14	0
1	31	31
2	47	94
3	41	123
4	29	116
5	21	105
6	10	60
7	5	35
8	2	16
	200	580

This value of the sample mean is used as the estimate of λ for the purposes of finding the probabilities from the tables of the Poisson distribution.

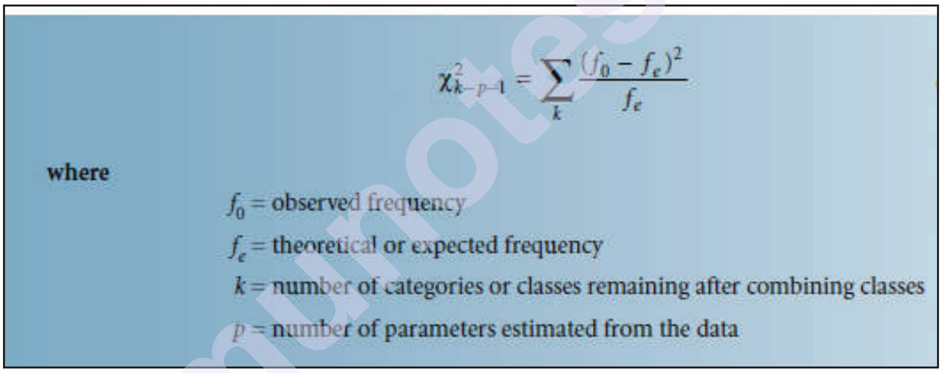
From Poisson distribution table, for λ = 2.9, the frequency of X successes (X = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or more) can be determined. The theoretical frequency for each is obtained by multiplying the appropriate Poisson probability by the sample size n. These results are summarized in the table below:



ARRIVALS	ACTUAL FREQUENCY f_0	PROBABILITY, $P(X)$, FOR POISSON DISTRIBUTION WITH $\lambda = 2.9$	THEORETICAL FREQUENCY $f_e = n \cdot P(X)$
0	14	0.0550	11.00
1	31	0.1596	31.92
2	47	0.2314	46.28
3	41	0.2237	44.74
4	29	0.1622	32.44
5	21	0.0940	18.80
6	10	0.0455	9.10
7	5	0.0188	3.76
8	2	0.0068	1.36
9 or more	0	0.0030	0.60

Observe from the table that the theoretical frequency of 9 or more arrivals is less than 1.0. In order to have all categories contain a frequency of 1.0 or greater, the category 9 or more is combined with the category of 8 arrivals.

The chi-square test for determining whether the data follow a specific probability distribution is computed using Equation



$$\chi^2_{k-p-1} = \sum_k \frac{(f_0 - f_e)^2}{f_e}$$

where

f_0 = observed frequency

f_e = theoretical or expected frequency

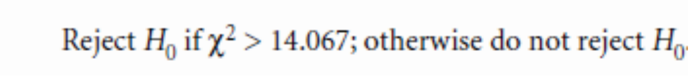
k = number of categories or classes remaining after combining classes

p = number of parameters estimated from the data

Returning to the example concerning the arrivals at the bank, nine categories remain (0, 1, 2, 3, 4, 5, 6, 7, 8 or more). Since the mean of the Poisson distribution has been estimated from the data, the number of degrees of freedom are:

k - p - 1 = 9 – 1 - 1 = 7 degrees of freedom

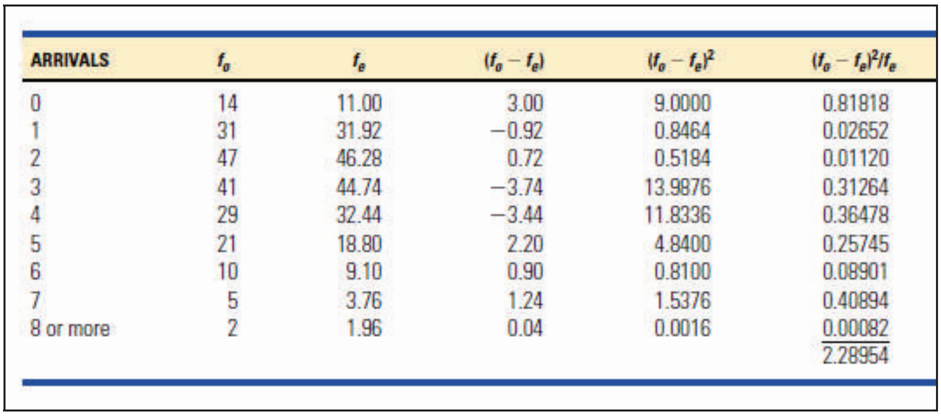
Using the 0.05 level of significance, from Table E.4, the critical value of χ2 with 7 degrees of freedom is 14.067. The decision rule is



Reject H_0 if $\chi^2 > 14.067$; otherwise do not reject H_0.

Since χ2 = 2.28954 < 14.067, the decision is not to reject H0. There is insufficient evidence to conclude that the arrivals per minute do not fit a Poisson distribution.

Computation of the chi-square test statistic for the arrivals per minute



ARRIVALS	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
0	14	11.00	3.00	9.0000	0.81818
1	31	31.92	−0.92	0.8464	0.02652
2	47	46.28	0.72	0.5184	0.01120
3	41	44.74	−3.74	13.9876	0.31264
4	29	32.44	−3.44	11.8336	0.36478
5	21	18.80	2.20	4.8400	0.25745
6	10	9.10	0.90	0.8100	0.08901
7	5	3.76	1.24	1.5376	0.40894
8 or more	2	1.96	0.04	0.0016	0.00082
					2.28954

The Chi-Square Goodness of Fit Test for a Normal Distribution

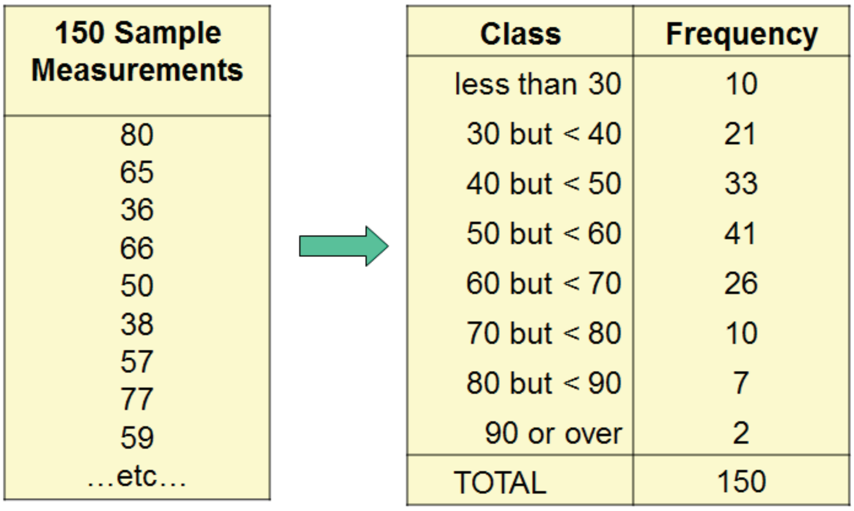
When testing hypotheses about numerical variables, the assumption is made that the underlying population was normally distributed. While graphical tools such as the box-and-whisker plot and the normal probability plot can be used to evaluate the validity of this assumption, an alternative that can be used with large sample sizes is the chi-square goodness-of-fittest for a normal distribution.

Do measurements from a production process follow a normal distribution with μ = 50 and σ = 15?

Process:

- Get sample data
- Group sample results into classes (cells)
- (Expected cell frequency must be at least 5 for each cell)
- Compare actual cell frequencies with expected cell frequencies

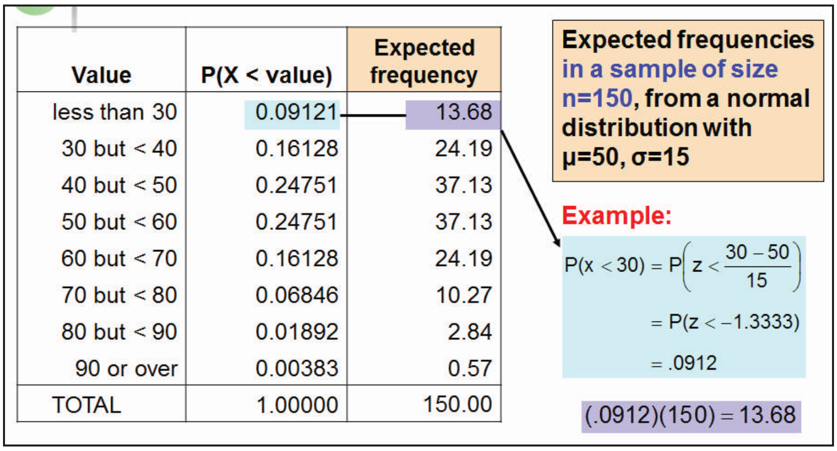
Sample data and values grouped into classes.



150 Sample Measurements
80
65
36
66
50
38
57
77
59
…etc…

Class	Frequency
less than 30	10
30 but < 40	21
40 but < 50	33
50 but < 60	41
60 but < 70	26
70 but < 80	10
80 but < 90	7
90 or over	2
TOTAL	150

What are the expected frequencies for these classes for a normal distribution with μ = 50 and σ = 15?



Value	P(X < value)	Expected frequency
less than 30	0.09121	13.68
30 but < 40	0.16128	24.19
40 but < 50	0.24751	37.13
50 but < 60	0.24751	37.13
60 but < 70	0.16128	24.19
70 but < 80	0.06846	10.27
80 but < 90	0.01892	2.84
90 or over	0.00383	0.57
TOTAL	1.00000	150.00

Expected frequencies in a sample of size n=150, from a normal distribution with μ=50, σ=15

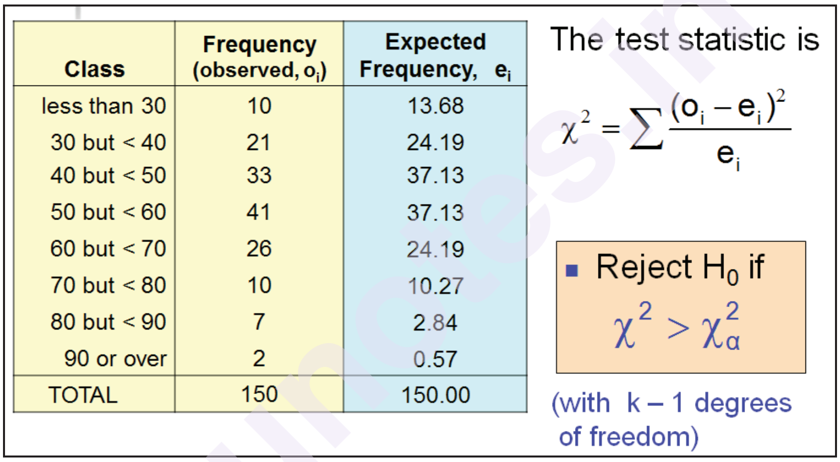
Example:

$$P(x < 30) = P\left(z < \frac{30-50}{15}\right)$$

$$= P(z < -1.3333)$$

$$= .0912$$

$$(.0912)(150) = 13.68$$



Class	Frequency (observed, o_i)	Expected Frequency, e_i
less than 30	10	13.68
30 but < 40	21	24.19
40 but < 50	33	37.13
50 but < 60	41	37.13
60 but < 70	26	24.19
70 but < 80	10	10.27
80 but < 90	7	2.84
90 or over	2	0.57
TOTAL	150	150.00

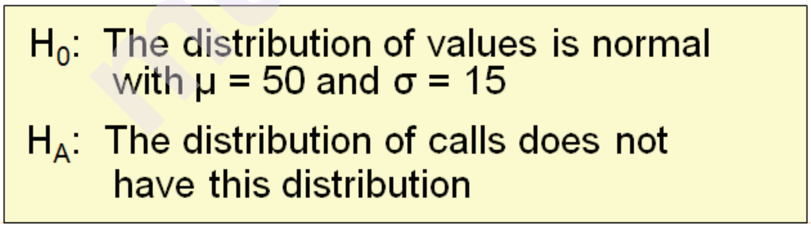
The test statistic is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

- Reject H_0 if

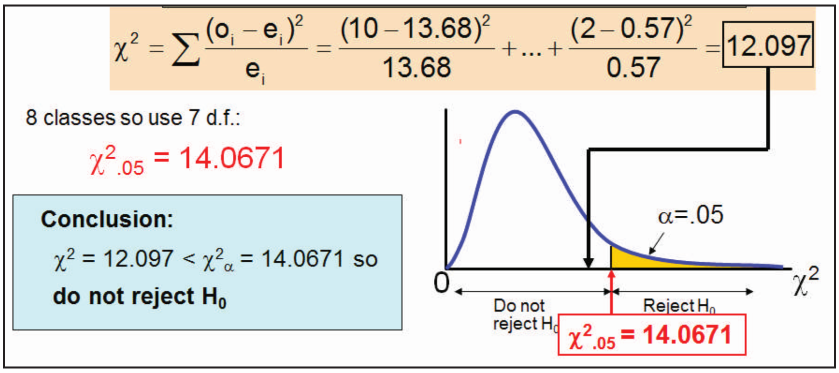
$$\chi^2 > \chi^2_\alpha$$

(with k – 1 degrees of freedom)



H_0: The distribution of values is normal with μ = 50 and σ = 15

H_A: The distribution of calls does not have this distribution



$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(10-13.68)^2}{13.68} + \ldots + \frac{(2-0.57)^2}{0.57} = \boxed{12.097}$$

8 classes so use 7 d.f.:

$\chi^2_{.05}$ = 14.0671

Conclusion:

$\chi^2 = 12.097 < \chi^2_\alpha = 14.0671$ so

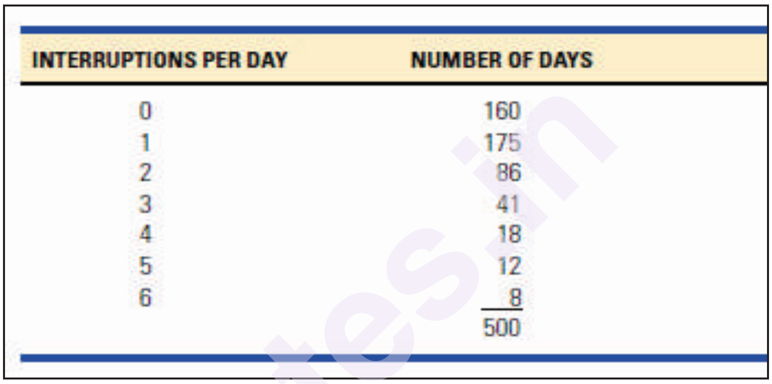
do not reject H_0

11.6 LET US SUM UP

In this Unit-VI - Chapter 6.2, you have learnt the concepts of Chi Square, Chi-Square Statistics and Chi-Square Goodness of Fit Test for Bionomial, Poisson and Normal distributions.

11.7 EXERCISES

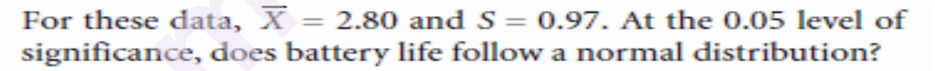
Q.1. The manager of a computer network has collected data on the number of times that service has been interrupted on each day over the past 500 days. The results are as follows:



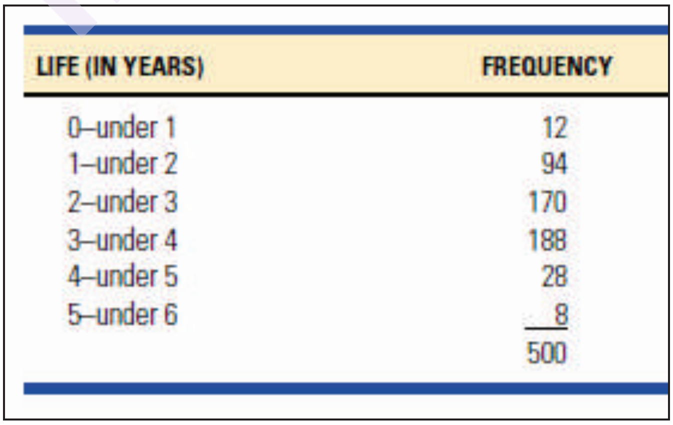
INTERRUPTIONS PER DAY	NUMBER OF DAYS
0	160
1	175
2	86
3	41
4	18
5	12
6	8
	500

Does the distribution of service interruptions follow aPoisson distribution? (Use the 0.01 level of significance.)

Q.2 A random sample of 500 car batteries revealed the following distribution of battery life (in years).

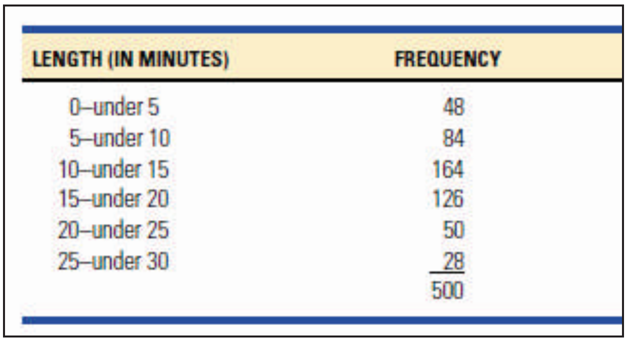


For these data, $\overline{X} = 2.80$ and $S = 0.97$. At the 0.05 level of significance, does battery life follow a normal distribution?



LIFE (IN YEARS)	FREQUENCY
0–under 1	12
1–under 2	94
2–under 3	170
3–under 4	188
4–under 5	28
5–under 6	8
	500

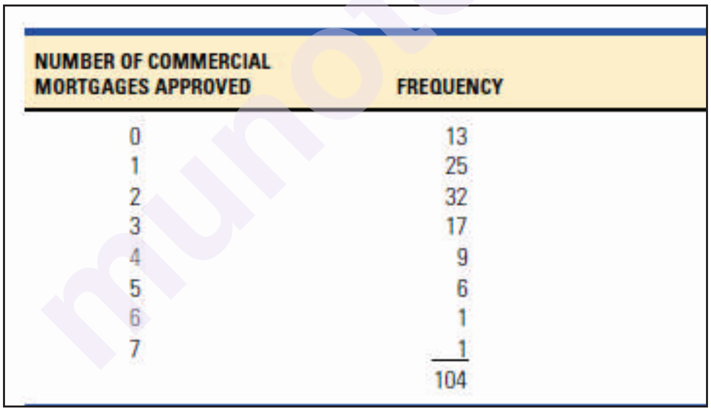
Q.3 A random sample of 500 long distance telephone calls revealed the following distribution of call length (in minutes).



LENGTH (IN MINUTES)	FREQUENCY
0–under 5	48
5–under 10	84
10–under 15	164
15–under 20	126
20–under 25	50
25–under 30	28
	500

a. Compute the mean and standard deviation of this frequency distribution.
b. At the 0.05 level of significance, does call length follow a normal distribution?

Q.4 The manager of a commercial mortgage department of a large bank has collected data during the past two years concerning the number of commercial mortgages approved per week. The results from these two years (104 weeks) indicated the following:



NUMBER OF COMMERCIAL MORTGAGES APPROVED	FREQUENCY
0	13
1	25
2	32
3	17
4	9
5	6
6	1
7	1
	104

Does the distribution of commercial mortgages approved per week follow a Poisson distribution? (Use the 0.01 level of significance.)

11.8 SUGGESTED READINGS

Chi Square section in any of the reference / text books

❖❖❖❖

Module - VI
ANOVA & CHI-SQUARE TEST

12

INTRODUCTION TO ANOVA

Chi-Square Test of Independence: Two Way Contingency Analysis

Unit Structure :

12.1 INTRODUCTION

In this Unit-VI - Chapter 6.3, we shall discuss the concepts of Chi-Square Test of Independence and Two Way Contingency Analysis.

12.2 OBJECTIVES

At the end of this unit the learners will be able to understand:

- Chi-Square Test of Independence
- Two Way Contingency Analysis.

12.3 CHI-SQUARE TEST OF INDEPENDENCE

Chi-square test for independence test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference. The sample problem at the end of the lesson considers this example.

When to Use Chi-Square Test for Independence :

The test procedure described in this section is appropriate when the following conditions are met:

- The sampling method is simple random sampling
- The variables under study are each categorical
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

What is simple random sampling?

Simple random sampling refers to any sampling method that has the following properties.

- The population consists of *N* objects.
- The sample consists of *n* objects.
- If all possible samples of *n* objects are equally likely to occur, the sampling method is called simple random sampling.

An important benefit of simple random sampling is that it allows researchers to use statistical methods to analyze sample results. For example, given a simple random sample, researchers can use statistical methods to define a confidence interval around a sample mean. Statistical analysis is not appropriate when non-random sampling methods are used.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the *N* population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects *n* numbers. Population members having the selected numbers are included in the sample.

What is a categorical Variable?

Variables can be classified as **categorical** (aka, qualitative) or **quantitative** (aka, numerical).

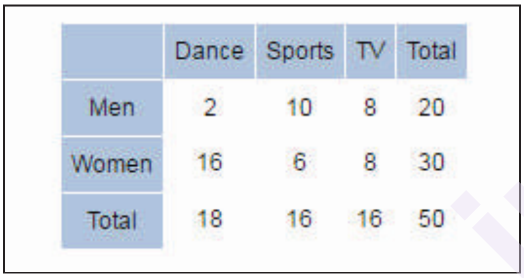
- ***Categorical***. Categorical variables take on values that are names or labels. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, and terrier) would be examples of categorical variables.
- ***Quantitative.*** Quantitative variables are numerical. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.

12.4 TWO WAY CONTINGENCY ANALYSIS

What is a contingency table?

A two-way table (also called a contingency table) is a useful tool for examining relationships between categorical variables. The entries in the cells of a two-way table can be frequency counts or relative frequencies (just like a one-way table).

Illustrative Example 1 :



	Dance	Sports	TV	Total
Men	2	10	8	20
Women	16	6	8	30
Total	18	16	16	50

The two-way table above shows the favorite leisure activities for 50 adults - 20 men and 30 women. Because entries in the table are frequency counts, the table is a frequency table.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses :

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

Suppose that Variable A has *r* levels, and Variable B has *c* levels. The null hypothesis states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

H_0: Variable A and Variable B are independent
H_a: Variable A and Variable B are not independent.

The alternative hypothesis is that knowing the level of Variable A *can* help you predict the level of Variable B.

Support for the alternative hypothesis suggests that the variables are related; but the relationship is not necessarily causal, in the sense that one variable "causes" the other.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the chi-square test for independence to determine whether there is a significant relationship between two categorical variables.

Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to:

$$DF = (r - 1) * (c - 1)$$

where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

- **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute r * c expected frequencies, according to the following formula.

  $$E_{r,c} = (n_r * n_c) / n$$

  where $E_{r,c}$ is the expected frequency count for level *r* of Variable A and level *c* of Variable B, $n_r$is the total number of sample observations at level r of Variable A, n_c is the total number of sample observations at level *c* of Variable B, and n is the total sample size.

- **Test statistic.** The test statistic is a chi-square random variable (X^2) defined by the following equation.

  $$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

  where $O_{r,c}$ is the observed frequency count at level *r* of Variable A and level *c* of Variable B, and $E_{r,c}$ is the expected frequency count at level *r* of Variable A and level *c* of Variable B.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Use the degrees of freedom computed above.

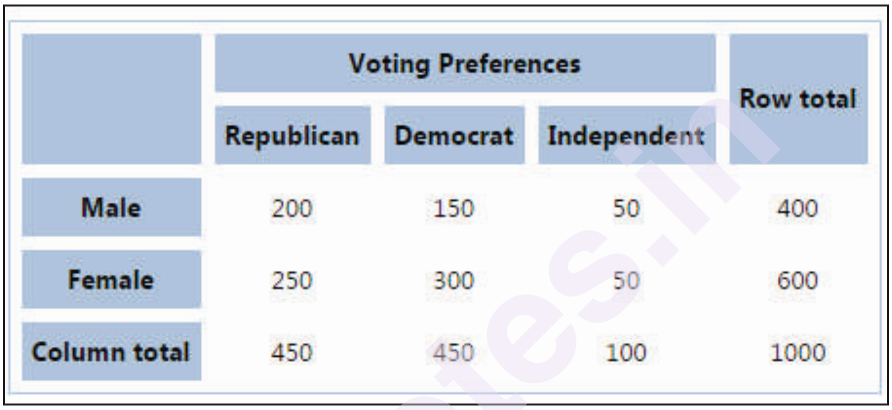
Interpret Results :

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Illustrative Example 2 :

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below:

Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.



	Voting Preferences			Row total
	Republican	Democrat	Independent	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Solution :

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

  H_0: Gender and voting preferences are independent.
  H_a: Gender and voting preferences are not independent.

- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square test for independence.

- **Analyze sample data**. Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

  $DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$
  $E_{r,c} = (n_r * n_c) / n$

$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$
$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$
$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$
$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$
$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$
$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$
$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$

$X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40$
$+ (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/60$
$X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270$
$+ 100/60$
$X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$

Where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, n_r is the number of observations from level *r* of gender, n_c is the number of observations from level *c* of voting preference, n is the number of observations in the sample, $E_{r,c}$ is the expected frequency count when gender is level *r* and voting preference is level *c*, and $O_{r,c}$ is the observed frequency count when gender is level *r* voting preference is level *c*.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

We use the Chi-Square Distribution table to find P(X2 > 16.2) = 0.0003.

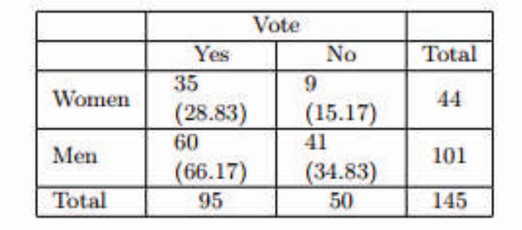
Interpret results. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

12.5 LET US SUM UP

In this Unit-VI - Chapter 6.3, you have learnt the concepts of Chi-Square Test of Independence andTwo Way Contingency Analysis.

12.6 EXERCISES

Q.1. In 2000 the Vermont State legislature approved a bill authorizing civil unions. The vote can be broken down by gender to produce the following table, with the expected frequencies given in parentheses. The expected frequencies are computed as Ri × Cj/N, where Ri and Cj represent row and column marginal totals and N is the grand total.



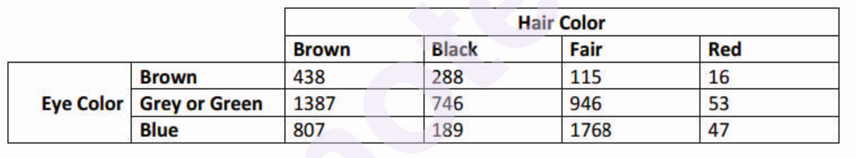
	Vote		
	Yes	No	Total
Women	35 (28.83)	9 (15.17)	44
Men	60 (66.17)	41 (34.83)	101
Total	95	50	145

Can we accept the null hypothesis that voting behavior is independent of gender at 5% level of significance?

Q.2. The relationship between disease and exposure may be displayed in a contingency table.

	Disease		
Exposure	**Yes**	**No**	**Total**
Yes	**37**	**13**	**50**
No	**17**	**53**	**70**
Total	**54**	**66**	**120**

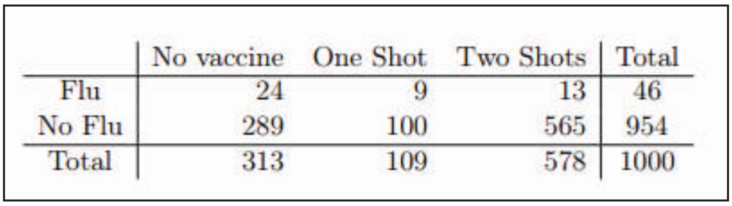
Q.3. To study the association of hair color and eye color in a German population, an anthropologist observed a sample of 6,800 men.



		Hair Color			
		Brown	Black	Fair	Red
Eye Color	Brown	438	288	115	16
	Grey or Green	1387	746	946	53
	Blue	807	189	1768	47

Test, at the α = 0.05 significance level, whether hair color is associated with eye color in this population of German men.

Q.4. A survey was conducted to evaluate the effectiveness of a new flu vaccine that had been administered in a small community. It consists of a two–shot sequence in two weeks. A survey of 1000 residents the following spring provided the following information:



	No vaccine	One Shot	Two Shots	Total
Flu	24	9	13	46
No Flu	289	100	565	954
Total	313	109	578	1000

Test at α = 0.05 significance level, whether getting the flu and vaccine usage are independent?

12.7 SUGGESTED READINGS

Chi Square section in any of the reference / text books

❖❖❖❖