APPLIED STATISTICS

Complementary Course of

BSc Mathematics

IV Semester

CUCBCSS

2014 Admission Onwards

CALICUT UNIVERSITY

SCHOOL OF DISTANCE EDUCATION

985

UNIVERSITY OF CALICUT

SCHOOL OF DISTANCE EDUCATION

Study Material

APPLIED STATISTICS

Complementary Course of

BSc Mathematics IV Semester

Prepared & Scrutinised By:

Dr.K.K.Joseph,

Director,

Academic Staff College,

University of Calicut.

settings & Lay Out

SDE

CONTENTS

Chapter 1

STATISTICAL SURVEY

Population and Sample

An aggregate of individual items relating to a phenomenon under investigation is technically termed as 'population'. In other words a collection of objects pertaining to a phenomenon of statistical enquiry is referred to as population or universe. Suppose we want to collect data regarding the income of college teachers under University of Calicut, then, the totality of these teachers is our population. The population can be finite or infinite.

When few items are selected for statistical enquiry from a given population it is called a 'sample'. A sample is the small part or subset of the population.

Census and Sample Method

In any statistical investigation, one is interested in studying the population characteristics. This can be done either by studying the entire items in the population or on a part drawn from it. If we are studying each and every element of the population, the process is called *census method* and if we are studying only a sample, the process is called *sample survey*, *sample method* or *sampling*. For example, the Indian population census or a socio economic survey of a whole village by a college planning forum are examples of census studies. The national sample survey enquiries are examples of sample studies.

Organization of Statistical Survey

A proper planning is essential before an enquiry is conducted. Planning must precede execution. The several stages of a survey, starting from planning and ending with writing of the final report are considered under two major heads.

1. Planning the survey
2. Execution of the survey.

Planning the Survey

Preliminary steps before data collection are the steps thought about during planning. They are enumerated as follows.

1. Purpose of the survey
2. Scope of the survey
3. Nature of information required
4. Units to be used
5. Sources of data
6. Techniques to be adopted
7. Choice of frame
8. Accuracy aimed
9. Other considerations

Execution of the Survey

The plan of any survey is to be followed by proper execution of the survey. The various phases are as follows:

1. Setting up an administrative organization .
2. Designing of forms.
3. Selecting, training and supervising the field investigators.
4. Controlling the accuracy of the fieldwork.
5. Reducing non-response.
6. Presenting the information.
7. Analyzing the information.
8. Preparing the reports .

Chapter 2

METHODS OF SAMPLING

Principles of Sampling

There are two important principles upon which the theory of sampling is based. They are

i. Principle of Statistical Regularity

ii. Principle of Inertia of Large Numbers.

Sampling method can be conveniently grouped into two; viz., random sampling and nonrandom sampling. The random sampling is also known as 'probability sampling' or 'chance sampling'.

Non probability sampling procedure is that sampling method which does not afford any basis for estimating the probability that each item in the population has of being included in the sample. In this method samples are selected according to a fixed sampling rule and not by assigning probabilities.

Here we consider only probability sampling. This is the scientific method of selecting samples according to some laws of chance in which each unit in the population has some definite pre-assigned probability of being selected in the sample. The different types of probability sampling are

i. where each unit has an'equal chance' of being selected.

ii. sampling units have different probabilities of being selected.

iii. probability of selection of a unit is proportional to the sample size.

Methods of Random Sampling

There are various methods of random sampling and non-random sampling. In our present study we are confined ourselves to the discussion of the following types of random sampling.

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

1. Simple random sampling

Here each unit in the population gets an equal chance of being represented in a sample, and such a sample is called 'random sample' and the technique of selecting a random sample is called simple random sampling (srs).

This is the common method of sampling. The sample so obtained is called a *random sample*. By a random sample we mean, we should give equal chance or opportunity to every unit in the population to be included in the sample. A random sample can be selected by two methods (i) lottery method (ii) random number method.

i. Lottery Method

The lottery method is practicable when the population size is comparatively small. To select a random sample by this method, first assign serial numbers as 1, 2, 3,... to each item. Write these numbers on pieces of paper of equal size and of the same quality. After this, roll the papers called 'lots' and shuffle them thoroughly. Take one lot at random. Note the number on it. Now select the item corresponding to this number into the sample. Repeat the process till we get the required number of items into the sample.

ii. Random Number Method

This method is adopted with the aid of random number tables. These are table of numbers in which digits selected by a mechanical process of randomization are tabulated. They are tabled as 2 digital, 3, 4 or 5 digital numbers. After assigning serial numbers to each sampling unit, open any page of the random number table. Then a blind-fold selection of a number is made. Starting from that number we can proceed along a row, column, or diagonally the successive numbers there occur will give a random selection of items. This method can be used to select a random sample even when the population is large.

Merits

1. It saves money and time.
2. It eliminates the possibility of biased errors.
3. As the random sample becomes extensive, the errors neutralise among themselves and it becomes more representative of the aggregate.
4. As in the case of deliberate selection, there is no need for a detailed plan for the selection of the samples.
5. There is the possibility of testing the accuracy of a sample by examining another sample from the same universe.

Demerits

1. It cannot be applied where some units of the universe are so important that their inclusion in the sample is essential.
2. If the sample is not large, it may not be truly representative of the whole universe.
3. It will not always be possible or practical to identify every population member, if a sample from a churn of milk is required, it is obviously impossible to distinguish every molecule.
4. This method however, is expensive and time consuming, especially when the population is large.

Now we are going to explain methods of sampling in which the samples are selected partly according to some laws of chance and partly according to a fixed sampling rule with no arrangement of probabilities, they are termed as mixed samples and the technique of selecting such a sample is called mixed sampling. They are the following.

2. Startified Random Sampling

When the population is heterogeneous with respect to some characteristic, but can be divided into a number of homogeneous subgroups known as 'strata' this method can be applied. Here, we divide the population into different strata and from each stratum select the items proportionally using lottery method. This will constitute a stratified random sample.

For example, suppose we want to study the academic level of 2000 students in a college. Let us assume that this consists of 600 IDC, 500 II DC, 400 III DC, 300 I PG and 200 II PG students. In this situation we can select a stratified random sample of 200 students. So divide the whole students into 5 strata as I DC, II DC, III DC, I PG and II PG. Now from each stratum select the students proportionally using any of the random sampling methods. That means we have to select 601 DC, 50 II DC, 40 III DC, 30 IPG and 20 II PG students respectively from each stratum. This will constitute a stratified random sample of 200 students.

Merits

1. The sample becomes more representative since each stratum is adequately represented in the sample.
2. Since the strata are homogenous the precision will be more.
3. Stratification many times leads to administrative convenience.

Demerits

1. If there is any wrong in the formation of different strata, the results will be quite unreliable.
2. Difficulties are also faced while deciding the basis and the number of stratifications.
3. The procedure is tedious and time consuming.

3. Systematic Sampling

If we can arrange the sampling units is a definite order, say, alphabetical, chronological, geographical etc., this method can be employed. Once the sampling units are arranged in a definite order, give them serial numbers as 1, 2, 3,..... Divide them into a number of groups which is equal to the required sample size. From the first group choose an item at random using lottery method. If we select the 4^{th} item, choose the 4^{th} item of every group systematically at equally spaced intervals. This will constitute a systematic sample.

For example, if we want to select a systematic sample of 100 students from a group of 1000, first arrange the students

alphabetically or chronologically. The order of arrangement should not be related to the variable under study. Now give them serial numbers as 1, 2, 3,..., 1000. Divide them into 100 groups of size 10. From the first group choose a student at random. If we choose the 7^{th} student from the first group choose the 7^{th} student of every group systematically at equally spaced intervals. That means, we will choose 7^{th}, 17^{th}, 27^{th}, ..., 997^{th} students respectively into the sample. This will constitute the systematic sample of 100 students. This method is also called quasi-random sampling method.

4. Cluster Sampling

Cluster sampling is the selection of sample units in two stages. In the first stage, certain groups or clusters called **primary sampling units** are selected from the population. These units, say, might be companies selected from an industry. In the second stage, individual items called **elementary sampling units** are drawn from each of these clusters. These units might be employees-either a sample or all of the employees in a company. This second step in the sampling process is called **sub-sampling**. When each cluster is contained in a compact geographic area, cluster sampling is also called **area sampling**.

Advantages of Sampling

1. The sample method is comparatively more economical.
2. The sample method ensures completeness and a high degree of accuracy due to the small area of operation.
3. It is possible to obtain more detailed information in a sample survey than complete enumeration.
4. Sampling is also advocated where census is neither necessary nor desirable.
5. A sample survey is much more scientific than census because in it the extent of the reliability of the results can be known where as this is not always possible in census.

Limitations of Sampling

1. In order to obtain accurate results it is indispensable that a sample survey has been properly planned and executed otherwise incorrect and misleading results may be obtained.
2. Most sampling requires the services of experts and if there is a paucity of such people, sampling may give unsatisfactory results owing to the use of faulty methods of selection, inappropriate sampling design, or inefficient methods of estimation.
3. Where one is interested in minute details in the characteristics of individual constituent of a universe, sampling is ruled out
4. There are various sources of errors in a sample survey. Every attempt must be made to minimize the chances of such errors; otherwise right inferences cannot be entailed.
5. If the sample is not truly representative and wrong type of sampling method is selected, then the sample will fail to give the true characteristics of the population.

These limitations of a sample survey as stated above clearly show that sampling in no case can be regarded as a completely fool proof method of statistical investigations.

Thus the choice between sample and census method of enquiry must be carefully made. If population is small and precise information is needed concerning it, a census will be appropriate. But when population is very large or field of enquiry is very wide, and quick results are needed, sampling is to be resorted to.

Sampling and Non Sampling Errors

The errors that may occur due to the collection, classification, processing and analysis of data may be broadly classified into two as given below.

1. Sampling Errors and 2. Non sampling errors.

1. Sampling Errors

Here, we should be familiar with the terms, namely, ‘statistic’ and ‘parameter’. Any value computed using sample observations is called a ‘statistic’ where as any value computed from a population is called ‘parameter’. So sampling error can

also be defined as the error between statistic and parameter,

2. Non Sampling Errors

We see that sampling errors occur due to the inductive process of inferring about the population on the basis of a sample. But non sampling errors primarily arise at the stages of data collection, classification, analysis and interpretation of the data. So non sampling errors present in both the complete enumeration survey and sample survey.

Preparation of Questionnaire

A distinction is often made between questionnaire and schedule. The questionnaire is ordinarily filled up by the informant, while the schedule is ordinarily filled up by a trained enumerator who gathers the information by questioning informants. However, the difference between the two is one of degree and not of kind.

When data are collected by means of questionnaires or schedules it is necessary to exercise great care in drafting it, since the success of any statistical investigation is determined by the quality of the questionnaire and the response it evokes from the informants. The following are the points to be borne in mind while drafting a good questionnaire or schedule.

1. There should be as few questions as possible.
2. The questions should be simple and easy to understand.
3. The questions should not be ambiguous.
4. As far as possible 'Yes' or 'No' questions should be given.
5. The questions should be such that they will be answered without bias.
6. The questions should not be unnecessarily inquisitorial.
7. Specific *information questions* should be included.
8. Open questions ought to be avoided.
9. Instructions to the informants should be given.
10. Questions should be capable of objective answer.
11. Questionnaire should look attractive.
12. Pre testing the questionnaire.
13. Questions requiring calculations should not be asked.

EXERCISES

Very Short Questions

1. Distinguish between population and sample.
2. Define a random sample.
3. Define a systematic sample.
4. Define a stratified random sample.
5. Define a cluster sample.
6. State the different types of probability sampling
7. Define sampling error
8. Define non sampling error
9. Define statistics.
10. Define parameter.

Short Answer Questions

11. Explain the concept of population and sample.
12. Distinguish between census and sampling.
13. What are the advantages of sampling over census?
14. What is a random sample? How will you select it?
15. Explain the method of selecting a systematic sample.
16. How will you select a stratified random sample?
17. Distinguish between schedule and questionnaire.

Essay Questions

18. Describe the different stages of statistical enquiry.
19. Distinguish between questionnaire and schedule. What are the points to be remembered while drafting a good questionnaire.
20. Describe simple random sampling and stratified random sampling.
21. Distinguish between systematic sampling and stratified random sampling.

Module II

ANALYSIS OF VARIANCE

Chapters

Chapter 1

INTRODUCTION

Analysis of Variance

The technique of analysis of variance was first devised by Sir. Ronald Fisher, an English Statistician who is also considered to be the father of modern statistics as applied to social and behavioral sciences. It was first reported in 1923 and its early applications were in the field of agriculture. Since then it has found wide applications in many areas of experimentation. The analysis of variance, as the name indicates, deals with variance rather than with standard deviations and standard errors.

The ANOVA is a procedure, which separates the variation assignable to one set of causes from the variation assignable to other set of causes. For example, four varieties of wheat are sown in plots and their yield per acre was recorded. We wish to test the null hypothesis that the four varieties of wheat produce an equal yield. This can be done by taking all possible pairs of means ($4C_2$ in number) and testing the significance of their difference. But, the variation in yield may be due to differences in varieties of wheat seeds used, fertility of the soil of the various plots and the different kinds of fertilizers used. We are interested in testing whether the variation in yield is due to differences in wheat varieties, the differences in types of the fertilizers or differences in both. In ANOVA we can estimate the contribution made by each *factor* to the total variation. Now we can split the total variation into the two components (i) variation between the samples and (ii) variation within the samples. If the variance between the sample means is significantly greater than the variance within the samples, it can be concluded that the samples are drawn from different populations. However, the total variation may be due to the experimental error and the variation due to the difference in varieties of wheat or may be due to the experimental error and the variation due to the fertilizers used.

When we classify data on the basis of variety of wheat alone

it is known as one-way classification. On the other hand when we classify observations on the basis of the variety of wheat and type of fertilizer used it is called two-way classiffication. The procedure followed in the analysis of variance would be explained separately for

1. One-way classification, and
2. Two-way classification

The technique of analysing the variance in case of one factor and two factors is similar. In the case of one way ANOVA the total variance is divided into two parts namely; (i) variance between the samples and (ii) variance within the samples. The variance within the samples is also known as the residual variance. In the case of two factor analysis, the total variance is divided into three parts namely (i) variance due to factor one, (ii) variance due to factor two and (iii) the residual variance.

However, irrespective of the type of classification the analysis of variance is a technique of partitioning the total sum of squared deviations of all sample values from the grand mean and is divided into two parts-

1. *Sum of squares between the samples and*
2. *Sum of squares within the samples.*

Individual observations in the same treatment samples, however, can differ from each other only because of chance variation, since each individual within the group receives exactly the same treatment.

Assumptions in Analysis of Variance

The analysis of variance technique is based on the following assumptions:

1. Population from which samples have been drawn are normally distributed.
2. Populations from which the samples are drawn have same variance.
3. The observations, in the sample, are randomly selected from the population.
4. The observations are non correlated random variables.
5. Any observation is the sum of the effects of the factors influencing it.
6. The random errors are normally distributed with mean 0, and a common variance $\pm^2$.

Chapter 2

ANOVA MODELS

One-way Classification Model

The term one-factor analysis of variance refers to the fact that a single variable or factor of interest is controlled and its effect on the elementary units is observed. In other words, in one-way classification the data are classified according to only one criterion. Suppose we have independent samples of $n_1, n_2, ..., n_k$ observations from k populations. The population means are denoted by $«_1, «_2,..., «_k$. The one-way analysis of variance is designed to test the null hypothesis:

$$H_0 : «_1 = «_2 = «_3 = ... = «_k$$

i.e., the arithmetic means of the population from which the k samples are randomly drawn are equal to one another. The steps involved in carrying out the analysis are:

1. Calculate the Variance Between the Samples

Sum of squares is a measure of variation. The sum of squares between samples is dented by SSC. For calculating variance between samples, we take the total of the square of the variations of the means of various samples from the grand average and divide this total by the degrees of freedom. Thus the steps in calculating variance between samples will be:

a. Calculate the mean of each sample, ie., $\bar{X}_1, \bar{X}_2, ..., \bar{X}_k$;

b. Calculate the grand average $\bar{\bar{X}}$. Its value is obtained as follows: $\bar{\bar{X}} = \dfrac{\bar{X}_1 + \bar{X}_2 + ... + \bar{X}_k}{N_1 + N_2 + ... + N_k}$

c. Take the difference between the means of the various samples and the grand average;

d. Square these deviations and obtain the total which will The Analysis of Variance Table

Since there are several steps involved in the computation of both the between and within sample variances, the entire set of results may be organised into an analysis of variance (ANOVA) table. This table is summarized and shown below.

Source of variation	Sum of Squares SS	Degrees of Freedom	Mean Square MS	Variance Ratio F
Between Samples	SSC	$c-1$	$MSC = \dfrac{SSC}{c-1}$	$F = \dfrac{MSC}{MSE}$
Within Samples	SSE	$n-c$	$MSE = \dfrac{SSE}{n-c}$	
Total	SST	$n-1$		

Shortcut method

To use ANOVA table it is convenient to use the shortcut computation formula. The steps are given below.

1. Assume the means of the populations from which the k samples are randomly drawn are equal.
2. Compute Mean squares between the samples, say MSC and Mean square within the samples say MSE.

For computing MSC and MSE, following calculations are made,

a. T = sum of all the observations in rows and columns.

b. SST = sum of squares of all observations $-T^2/n$. Here T^2/n is called correction factor.

c. SSC = $\dfrac{(\sum x_1)^2}{n_1} + \dfrac{(\sum x_2)^2}{n_2} + ... - T^2/n$ where $\sum x_1, \sum x_2, ...$ are the column totals.

d. $SSE = SST - SSC$

e. Then, $MSC = \dfrac{SSC}{n-c}$

f. Calculate: $MSE = \dfrac{SSE}{n-c}$

3. Calculate F-ratio = $\frac{MSC}{MSE}$

4. Obtain the table value of F for $(c-1, n-c)$ degrees of freedom. If the calculated value of F<table value accept the hypothesis that the sample means are equal. That is, the factors influence inthe same manner.

Example 1

Below are given the yield (in kg) per acre for 5 trial plots of 4 varieties of treatment.

	Treatment			
Plot No.	1	2	3	4
1	42	48	68	80
2	50	66	52	94
3	62	68	76	78
4	34	78	64	82
5	52	70	70	66

Carry out an analysis of variance and state your conclusions.

Solution

$I(x_1)$	$II(x_2)$	$III(x_3)$	$IV(x_4)$
42	48	68	80
50	66	52	94
62	68	76	78
34	78	64	82
52	70	70	66
240	330	330	400

T = Sum of all observations = 42 + 50 + ...+ 66 = 1300

$$\text{Correction factor} = \frac{T^2}{n} = \frac{(1300)^2}{20} = 84500$$

SST = Sum of the squares of all the observations $-T^2/n$.

$$= \left(42^2 + 50^2 + 62^2 + ... + 66^2\right) - 84500 = 4236$$

$$\text{SSC} = \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + ... - T^2/n$$

$$= \frac{(240)^2}{5} + \frac{(240)^2}{5} + \frac{(330)^2}{5} + \frac{(330)^2}{5} + \frac{(400)^2}{5} - 84500 = 2580$$

$$\text{SSE} = SST - SSC = 4236 - 2580 = 1656$$

$$\text{MSC} = \frac{SSC}{c-1} = \frac{2580}{3} = 860,\ MSE = \frac{SSE}{n-c} = \frac{1656}{20-4} = 103.5$$

The degree of freedom = $(c-1, n-c) = (3, 16)$

The Analysis of Variance Table

Source of variation	Sum of Squares SS	Degrees of Freedom	Mean Square MS
Between Samples	SSC = 2580	$c-1$ = 3	$MSC = 860$
Within Samples	SSE = 1656	$n-c$ = 16	$MSE = 103.5$
Total	SST = 4236	$n-1$ = 19	

$$F = \frac{MSC}{MSE} = \frac{860}{103.5} = 8.3$$

The table value of F at 5% level of significance for (3, 16) degrees of freedom is 3.24. The calculated value of F is more than the table value of F. Therefore the null hypothesis is rejected. ∴ The treatments do not have same effect.

Coding Method

Coding method is, in fact, a furtherance of the short cut method. This method is based on a very important property of the variance ratio or the F-coefficient that its value does not change if all the given item values are either multiplied or divided by a common figure or if a common figure is either added or subtracted from each of the given item values. The main advantage of this method is that big figures are reduced in magnitude by division or subtraction and work is simplified without any disturbance on the variance ratio. This method should be used specially when given figures are big or otherwise inconvenient. Once the given figures are converted with the help of some common figure or the common factor, then all the steps of the shortcut method outlined above may be adopted for obtaining and interpreting the variance ratio.

Example 2

For the data given in example 1 carry out the analysis of variance technique using coding method.

Solution

Applying the coding method let us subtract 20 from each observation. Then the coded data is obtained as shown below:

X_1	X_2	X_3	X_4
0	+5	+4	+3
–1	+3	0	0
+1	+1	+2	0

To compute diffrent quantities, let us make the following table:

x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
0	0	+5	25	+4	16	+3	9
– 1	1	+3	9	0	0	0	0
+1	1	+1	1	+2	4	0	0
$\Sigma x_1 = 0$	$\Sigma x_1^2 = 2$	$\Sigma x_2 = 9$	$\Sigma x_2^2 = 35$	$\Sigma x_3 = 6$	$\Sigma x_3^2 = 20$	$\Sigma x_4 = 3$	$\Sigma x_4^2 = 9$

T = Sum of all observations = 0 + 5 + 4 + 3 + ... + 0 = 18

$$T^2/n = \frac{18 \times 18}{12} = 27$$

SST = Sum of squares of all observations $-T^2/n$

$$= \left(0^2 + 5^2 + 4^2 + ... + 0^2\right) - 27 = 66 - 27 = 39$$

$$SSC = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} + \frac{(\Sigma x_4)^2}{n_4} - T^2/n$$

$$= \frac{0^2}{3} + \frac{9^2}{3} + \frac{6^2}{3} + \frac{3^2}{3} - 27 = 42 - 27 = 15$$

$$SSE = SST - SSC = 39 - 15 = 24$$

$$MSC = \frac{SSC}{c-1} = \frac{15}{3} = 5,\ MSE = \frac{SSE}{n-c} = \frac{24}{8} = 3$$

Analysis of Variance Table

Source of variation	Sum of Squares SS	Degrees of Freedom	Mean Square MS	Variance Ratio F
Between Samples	15	$4 - 1 = 3$	$MSC = \frac{15}{3} = 5$	$F = \frac{5}{3} = 1.67$
Within Samples	24	$12 - 4 = 8$	$MSE = \frac{24}{8} = 3$	
Total	39	$12 - 1 = 11$		

Table value of F for (3, 8) at 5% level of significance is 4.07. Since the calculated value of F is less than the table value, the null hypothesis is accepted. Therefore we can infer that the average life time of different brands of bulbs are equal.

Two-way Classification Model

In one-factor analysis of variance explained above the treatments constitute different levels of a single factor which is controlled in one experiment. There are, however many situations in which the response variable of interest may be affected by more than one factor. For example, sales of Maxfactor cosmetics, in addition to being affected by the point of sale display, might also be affected by the price charged, the size and/or location of the store or the number of competitive products sold by the store. Similarly petrol mileage may be affected by the type of car driven, the way it is driven, road conditions and other factors in addition to the brand of petrol used.

Thus with the two-factor analysis of variance, we can test two sets of hypothesis with the same data at the same time.

The procedure for analysis of variance is somewhat different from the one followed while dealing with problems of one-way classification.

The steps are as follows:

1. (a) Assume that the mean of all columns are equal. Choose the hypothesis as $H_0 = \sim_1 = \sim_2 = \ldots = \sim_c$

   (b) Assume that the means of all rows are equal. Choose the hypothesis as $H_0 = {}_1 = {}_2 = \ldots = {}_r$

2. Find T, the sum of all observations

3. Calculate SST = Sum of squares of all observations $-T^2/n$

4. Find SSR = $\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \ldots - T^2/n$ where $\sum x_1, \sum x_2, \ldots$ are row totals.

5. Find SST = $\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \ldots - T^2/n$ where $\sum x_1, \sum x_2, \ldots$ are column totals.

6. SSE = SST – SSC – SSR.

7. MSC = $\frac{SSC}{c-1}$, $MSR = \frac{SSR}{r-1}$, $MSE = \frac{SSE}{(c-1)(r-1)}$

   where c = number of columns, r = number of rows

8. $F_c = \frac{MSC}{MSE}$ and $F_r = \frac{MSR}{MSE}$

9. Determine the table value of F

   If F_c < table value, accept H_0 as $\sim_1 = \sim_2 = \ldots = \sim_c$

   If F_r < table value, accept H_0 as ${}_1 = {}_2 = \ldots = {}_r$

The Analysis of Variance Table

Source of variation	Sum of Squares SS	Degrees of Freedom	Mean Square MS	Variance Ratio F
Between Columns	SSC	$c-1$	MSC	F_c = MSC/MSE
Between rows	SSR	$r-1$	MSR	F_r = MSR/MSE
Residual	SSE	$(c-1)(r-1)$	MSE	
Total	SST	$n-1$		

Example 3

The following represent the number of units of production per day turned out by 4 different workers using 5 different types of machines.

Machine Types

Worker	A	B	C	D	E	Total
1	4	5	3	7	6	25
2	6	8	6	5	4	29
3	7	6	7	8	8	36
4	3	5	4	8	2	22
Total	20	24	20	28	20	112

On the basis of this information, can it be concluded that (a) the mean productivity is the same for different machines, (b) the mean productivity is different with respect to different workers.

Solution

Let us take the hypothesis that (a) the mean productivity of diffrent machines is same, ie., H_0 ; $\sim_1 = \sim_2 = \sim_3 = \sim_4 = \sim_5$ and that (b) the 4 workers don't differ in respect of productivity. to carryout the analysis of variance. ie H_0 ; $_1 = _2 = _3 = _4$

Correction factor = $T^2 / n = (112)^2 / 20 = 627.2$

Sum of squares between machines

$$SSC = \frac{(20)^2}{4} + \frac{(24)^2}{4} + \frac{(20)^2}{4} + \frac{(28)^2}{4} + \frac{(20)^2}{4} - C.F.$$

$$= 100+144+100+196 \quad +100 - 627.2 = 640 - 627.2 = 12.8$$

d.f.= (c – 1) = (5 – 1) = 4

Sum of squares between workers

$$SSR = \frac{(25)^2}{5} + \frac{(29)^2}{5} + \frac{(36)^2}{5} + \frac{(22)^2}{5} - C.F.$$

$$= 125+168.2+259.2+98.8 - 627.2$$

$$= 649.2 - 627.2 = 22$$

d.f.= (r – 1) = (4 – 1) = 3

Total sum of squares = SST

$$= 4^2 + 6^2 + 7^2 + 3^2 + 5^2 + 8^2 + 6^2 + 5^2 + 3^2 + 6^2 + 7^2 + 4^2 + 7^2 + 5^2 + 8^2 + 8^2 + 6^2 + 4^2 + 8^2 + 2^2 - 627.2$$

$$= 692 - 627.2 = 64.8$$

SSE = SST – SSC – SSR = 64.8 – 12.8 – 22.0 = 30

df = $(c-1)(r-1) = (5-1)(4-1) = 12$

Source of variation	Sum of Squares SS	d.f.	Mean Square MS	Variance Ratio F
Between machines	12.8	4	3.20	$F_c = \frac{3.2}{2.5} = 1.28$
Between workers	22.0	3	7.33	$F_r = \frac{7.33}{2.5} = 2.93$
Residual	30.0	12	2.50	
Total	64.8	19		

For d.f. (4, 12) = $F_{0.05}$ = 3.26. The calculated value of F_c is less than the table value. Our hypothesis is true. There is no significant difference in the mean productivity of five machines. The table values of F_r for (2, 12) d.f; at 5% level is 3.49.

The calculated value of F is less than table value. Our hypothesis is true. Hence there is no significant difference in the mean productivity of different workers.

EXERCISES

Very Short Answer Questions

1. Give a practical situation where ANOVA can be applied.
2. What is the use of ANOVA technique?
2. State three assumptions of ANOVA technique.
4. What is one-way classification in ANOVA.
5. What do you mean by two-way classification model?

Short Essay Questions

6. What is 'analysis of variance' and where is it used? Give two suitable examples.
7. State some applications of the analysis of variance.
8. In order to determine whether there are significant differences in the durability of three makes of computer, samples of size n = 5 are selected from each make and the frequency of repair during the first year of purchase is observed. The results are as follows. In view of the above data, what conclusion can you draw?

Make A	:	5	6	8	9	7
MakeB	:	8	10	11	12	4
Make C	:	7	3	5	4	1

Long Essay Questions

9. The following figures relate to production in kg. of three varieties A, B and C of wheat sown in 12 plots

   A 14, 16, 18

   B 14, 13, 15, 22

   C 18, 16, 16, 19, 20

   Is there any significant difference in the production of 3 varieties?

10. Set a table of analysis of variance for the following data

Plots /Variety	A	B	C	D
1	200	230	250	300
2	190	270	300	270
3	240	150	145	180

Test whether varieties are different.

11. Following table gives the number of refrigerators sold by 4 salesmen in three months:

	Salesman			
Month	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

a. Determine whether there is any difference in the average sales made by four salesmen.

b. Determine whether the sales differ with respect to different month.

Module III

ANALYSIS OF TIME SERIES & INDUX NUMBERS

Chapters

Chapter 1

COMPONENTS OF TIME SERIES

A *time series* may be defined as a set of values of a variable collected and recorded in chronological order of the time intervals. In short, time series refers to the data depending on time. According to Morris Hamburg. "A time series is a set of statistical observations arranged in chronological order". Therefore time series is also called *historical data or historical series*. For example, if we observe production, sales, population, profit, national income, import, export etc. at different points of time, say, over the last 5 or 10 years, the set of observations formed shall constitute time series. The study of movement of quantitative data through time is referred to as 'time series analysis'.

Utility of Time Series Analysis

Analysis of time series is of relevance whenever a variable is found to vary over time. Variables such as sales, production, profit, population and employment opportunity assume different values at different points of time. The importance can be given under the following four heads:

(i) *The analysis of time series helps to know the past conditions.*

(ii) It helps in assessing the present achievements.

(iii) It helps to predict the future.

(iv) It enables comparison. History repeats. It may be worth to watch one series to know the future of a similar series.

(v) It forewarns.

Components of Time Series

If we observe a time series, we can categories the various forces affecting a time series into four components. They are

1. Secular Trend
2. Seasonal Variations
3. Cyclical Variations
4. Irregular Variations

1. Secular Trend

These are changes that have occurred as a result of general tendency of the data to increase or decrease over a long period of time. This is also called *long term trend or simply trend*.

2. Seasonal Variations

The change that have taken place within a year as a result of change in climate, weather conditions, festivals etc., are called seasonal variations.

3. Cyclical Variations

These are changes that have taken place as a result of booms and depressions. Normally the period of cyclical variation is more than a year.

Cyclical variations are similar to seasonal variations. The difference is in the period. If changes take place periodically and if the period is more than one year, the variations are said to be cyclical.

4. Irregular Variations

These are changes that have taken place as a result of such forces that could not be predicted like floods, earthquakes, famines etc.

Time Series Models

There are two types of mathematical models for a time series.

1. *Additive model* : When the changes in the data are the result of the combined impact of the four components, we can write the original data (Y) as the sum of the four components,

   i.e., $Y = T + S + C + 1$, where

   T - Secular trend, S - Seasonal variation

   C - Cyclical variation, I - Irregular variation.

2. *Multiplicative model*: In this model, original data,

$$Y = T \times S \times C \times I.$$

Chapter 2

MEASUREMENT OF TREND

We know that the term trend implies secular trend. It measures long term changes occurring in a time series without bothering about short term fluctuations, occurring in between. There are four methods to estimate the secular trend. They are.

1. Graphic Method
2. Semi-Average Method
3. Method of Moving Averages
3. Method of Least Squares

1. Graphic Method

This is the simplest method of measuring trend of a time series. In this method, firstly, we have to draw the time series graph. To draw the time series graph draw the X axis and Y axis on a graph paper. The independent variable time, is taken along the X axis. On the Y axis, the variable which is depending on time is measured. Plot the time values against the corresponding values of the dependent variable. Join these points by means of a smooth line. This is called ‘time series graph’ or ‘historiegram’

Then to depict the trend behaviour a line is drawn through the points that is judged by the analyst to represent adequately the long term movement in the series.

Merits:

1. It is a simple method to estimate trend.
2. It is flexible and either a curve or a straight line could be found as trend on the basis of the positions of the points. But under the method of semi-averages, straight line trend alone could be fit for any given data.
3. It may give better estimates if it is used by an experienced statistician who knows the changes which have taken place subsequent to the collection of each item of data.

Limitations

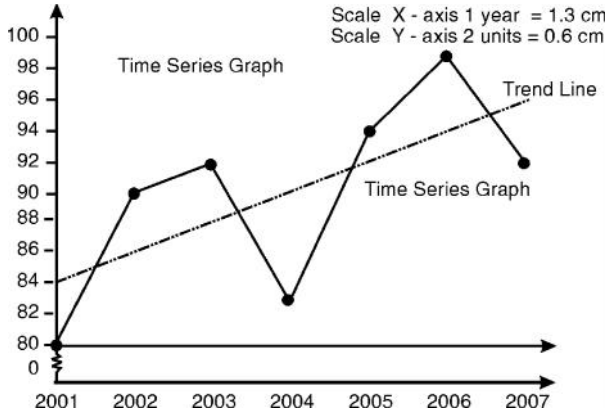
1. It is subjective. Each person may get one line (or one curve) and the trend lines (curves) drawn by different persons might be different although the data are the same.
2. It is not used for prediction because of its subjective character.
3. It is not as simple as it looks or as the method of semi-averages. It needs experience and careful approach.

Example 1

Draw a time series graph for the following data. Also draw the trend line

Yer :	2001	2002	2003	2004	2005	2006	2007
Production :	80	90	92	83	94	99	92

Solution



2. Semi Averages Method

As per its title, the data is divided into two parts equally and the average of the values of each half together with the mid period is plotted as a point on a graph sheet. The two points so plotted are joined by a straight line and the line is extended on either side till the end of the graph sheet. The line is called the trend line and the trend of any period could be read out from it.

Merits:

1. It is not a subjective but an objective method. Under graphic method, different persons may get different trend lines. The trend lines obtained by the different persons using the method of semi-averages are one and the same.
2. It involves very simple calculations. It is easy to adopt.

Limitations

1. It is not flexible and so is not suitable sometimes. Trend curve, if any, could not be drawn.
2. It is based on arithmetic mean. The limitations of arithmetic mean may prevent its unsuspected usage sometimes.

Example 2

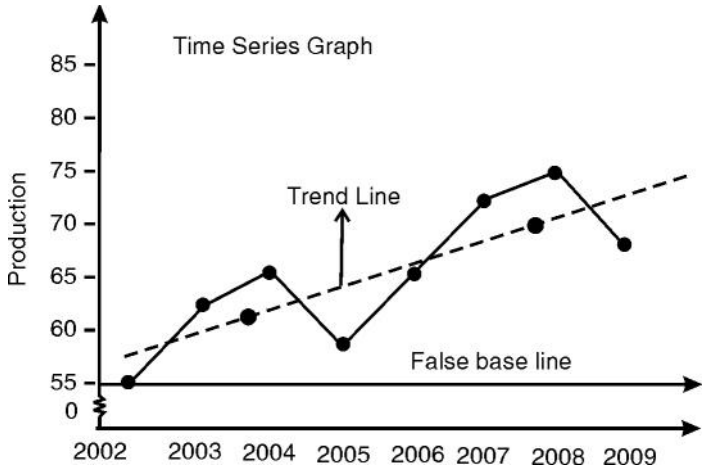
Draw a trend line by the semi-average method for the following data.

Year :	2002	2003	2004	2005	2006	2007	2008	2009
Production (in tons) :	55	62	65	58	65	72	75	68

Solution

Divide the data into two equal parts and the average of each part is obtained as follows.

Year	Production (tons)	
2002	55	$\frac{240}{4} = 60$
2003	62	
2004	65	
2005	58	
2006	65	$\frac{280}{4} = 70$
2007	72	
2008	75	
2009	68	



Note:

If we are given data for odd number of years, divide the data into two equal parts by leaving the middle year and then the average of each part is obtained.

3. Method of Moving Averages

In this method, the short term variations are eliminated by finding the moving averages. If the time series shows variation with period m(may be seasonal or cyclical variation), the moving averages with period m are obtained. These moving averages indicate the trend.

If $y_1, y_2, y_3,....y_n$ are n observations and if m = 3 is the period, the successive moving averages are-

$$\frac{y_1+y_2+y_3}{3}, \frac{y_2+y_3+y_4}{3}, \frac{y_3+y_4+y_5}{3}, ...$$

Here the first moving average is the average of the first, the second and the third observations. It is written against the second time point which is the middle most of the first, the second and the third time points. Such an association of moving averages with time points is possible only if the period m is an odd number. If m is an even number, firstly, moving averages with period m are found. And then, moving averages with period 2 of these moving averages are found. The second set of moving averages can be associated with given time points. This procedure is called *centering*. However, for finding trend values when m is even, firstly, moving totals with period m are obtained. Then, moving totals with period 2, of these moving totals are obtained, and the resulting totals are divided by 2m. Thus we get the trend values.

Merits

1. It is a simple method. The calculations are easy.
2. It is an objective method unlike the graphic method.
3. It is a flexible method in the sense that if the data of a few more years are available, the earlier calculations are not to be redone and the trends of a few more years are additionally available,
4. If the period of moving average coincides with the period of cyclical fluctuation, the cyclical fluctuation is completely eliminated.

Limitations

1. The very purpose of analysis of a time series is defeated by the method of moving averages. It can not be used for finding the trend of any period in future. Further, the trend of a few periods in the beginning of the series and that of equal number of periods in the end cannot be found out.
2. It may turn out to be tedious when the period is large as well as an even number.
3. Arithmetic mean has certain limitations consequently, moving average which is based on arithmetic mean is likely to be affected.
4. The period of moving average is to be decided carefully, otherwise, a distorted picture of the time series will emerge.

Example 3

Compute the trend values by finding three-yearly moving averages for the following times series.

Year	: 2000	2001	2002	2003	2004	2005	2006
Population (in millions)	: 412	438	446	454	470	483	490

Solution

Here, m = 3 is an odd number. To find the moving averages, firstly, the moving totals are obtained. Then, each of these are divided by 3. These moving averages are the trend values.

Year	Population (millions)	3-yearly moving totals	Trend values (millions)
2000	412	–	–
2001	438	1296	432.00
2002	446	1338	446.00
2003	454	1370	456.67
2004	470	1407	469.00
2005	483	1443	481.00
2006	490	–	–

Example 4

Compute the trend values by finding four - yearly moving averages for the following time series. Also, graph the observed values and the trend values.

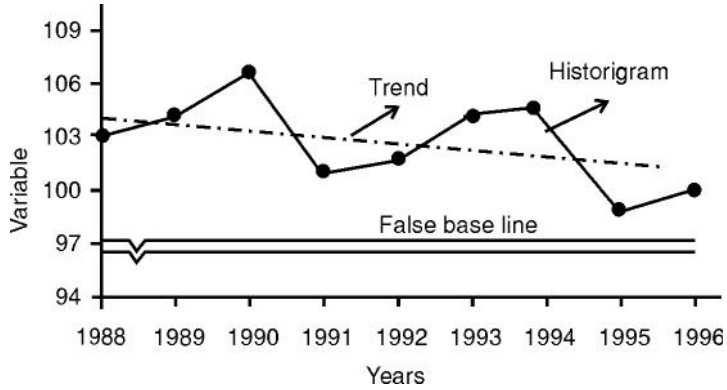
Year :	1988	1989	1990	1991	1992	1993	1994	1995	1996
Value :	103	104	107	101	102	104	105	99	100

Solution

Here, m = 4 is an even number. Therefore to find the trend values, firstly the moving totals should be obtained. Then two-item moving totals of these moving totals should be obtained. Finally, these totals are divided by 2m = 8.

Year	Value	4-yearly moving total	2-item moving total	Trend (previous co/8)
1988	103	–		–
1989	104	–		–
		415		
1990	107		829	103.6
		414		
1991	101		828	103.5
		414		
1992	102		826	103.3
		412		
1993	104		822	102.8
		410		
1994	105		818	102.3
		408		
1995	99	–		–
1996	100	–		–

The graph of the observed and trend values is as follows



4. Method of Least Squares

In this method, a mathematical relation is established between time and the variable which is depending on time. The relation may be

1. Linear (Straight line)
2. Quadratic (Parabolic)
3. Exponential

Here, the relation is so derived that the sum of the squared deviations (errors) of the observed values from the theoretical values is the least. The process of minimisation of the sum of the squared errors results in some equations called normal equations. The normal equations are the equations, which are used for finding the coefficients of the relation, which is fitted by the method of least squares.

We have already discussed the procedure of fitting a straight line, parabola and exponential trend in previous lessons.

1. Straight line trend by the method of least squares

In this method, a relation of the type y = a + bx is fitted to the time series. Here y denotes the values and x denotes the points of time. The constants a and b are obtained by solving the normal equations.

$$na + b\Sigma x = \Sigma y, \qquad a\Sigma x + b\Sigma x^2 = \Sigma xy$$

Here, y denotes the observed values and n denotes the number of observations in the series. Here, if x is chosen such that $\Sigma x = 0$ then, the values a and b are:

$$y = \frac{\Sigma y}{n} \text{ and } b = \frac{\Sigma xy}{\Sigma x^2}$$

By substituting the values of x in the trend equation, the corresponding trend values can be obtained. The graph of the trend equation can be obtained by plotting any two-trend values and joining them by a straight line.

Example 5

The following are the figures of production (in thousand quintals) of a sugar factory.

Year:	1992	1994	1996	1998	2000	2002	2004
Production:	77	81	88	94	94	96	98

i. Fit a straight line trend to the data

ii. Graph the observed values and the trend values

iii. Estimate the production in the year 2006.

Solution

Let y = a + bx be the trend equation. Then, the constants a and b can be obtained from the given data. To simplify the calculations, x is chosen such that $\Sigma x = 0$.

Here n = 7 is and **odd** number. Therefore, the middle most point of time is taken as zero. And the other values of x are written down as shown below.

Year X	y	x(time) = X – 1998	x^2	xy	Trend Values (thousand quintals
1992	77	– 3	9	– 231	78.87
1994	81	– 2	4	– 162	82.48
1996	88	– 1	1	– 88	86.09
1998	94	0	0	0	89.70
2000	94	1	1	94	93.31
2002	96	2	4	192	96.92
2004	98	3	9	296	100.53
Total	628	0	28	101	

Thus, $a = \frac{\Sigma y}{n} = \frac{628}{7} = 89.7$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{101}{28} = 3.61$$

Thus, the trend equations is y = **89.7 + 3.61x**

The trend values are obtained by substituting the different values of x.

x = – 3 gives, y = 89.7 + 3.61 × (– 3) = 89.7 – 10.83 = **78.87**

x = – 2 gives, y = 89.7 + 3.61 × (– 2) = **82.48**

x = – 1 gives, y = 89.7 + 3.61 × (– 1) = **86.09**

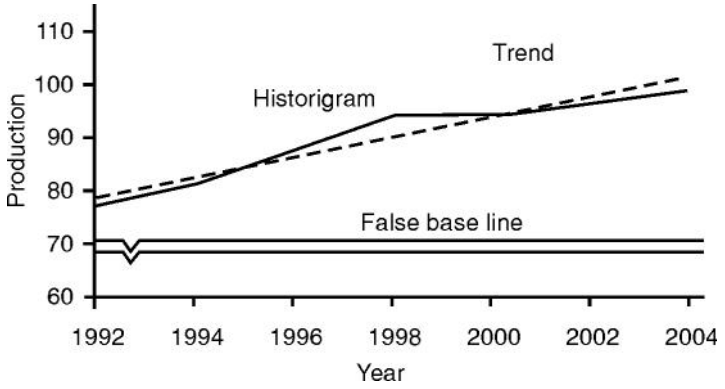
x = 0 gives, y = 89.7 + 3.61 × 0 = **89.7**

x = 1 gives, y = 89.7 + 3.61 × 1 = **93.31**

x = 2 gives, y = 89.7 + 3.61 × 2 = **96.92**

x = 3 gives, y = 89.7 + 3.61 × 3 = **100.53**

The graph of the observed and trend values is as follows.



The value of x corresponding to the year 2006 is 4. Therefore, the estimate of production for the year 2006 is -

Y = 89.7 + 3.61 x4 = **104.14** thousand quintals.

Note

Here, *b* = 3.61 thousand quintals is the rate of increase in the production per unit time (two years). Also, it should be noted that the successive trend values can be obtained by adding b = 3.61 to the preceeding trend value.

Example 6

Fit a straight-line trend for the following data by the method of least squares. Estimate the probable value for 2006.

Year :	1999	2000	2001	2002	2003	2004
Value :	10	8	12	9	11	12

Solution

Here, n = 6 is an even number. Take the transformation as $x = \frac{X - 2001.5}{0.5}$. Therefore, the two middle most time points are taken as x = – 1 and x = 1. The other values of x are taken as shown below.

Year (X)	y	x	x^2	x y
1999	10	– 5	25	– 50
2000	8	– 3	9	– 24
2001	12	– 1	1	– 12
2002	9	1	1	9
2003	11	3	9	33
2004	12	5	25	60
Total	62	0	70	16

Here, since $\Sigma x = 0$

$$a = \frac{\Sigma y}{n} = \frac{62}{6} = 10.33; \quad b = \frac{\Sigma xy}{\Sigma x^2} = \frac{16}{70} = 0.23$$

Thus, the trend equation is y = a + bx

That is, y = **10.33 + 0.23x**

The trend values can be obtained by putting x = – 5, – 3, etc.

But here, we are not required to calculate the trend values.

The probable value for 2006 can be obtained by putting x = 9 in the trend equation. Thus, the probable value is

Y = 10.33 + 0.23 x9

= 10.33+ 2.07 = **12.**4

Thus, the probable value for 2006 is y = **12.4**

2. Fitting of a parabola $y = a+bx+cx^2$

Here, a relation of the type $y = a + bx + cx^2$ is fitted to the time series. The constant a, b and c are obtained by solving the normal equations-

$$n a + b\Sigma x + c\Sigma x^2 = \Sigma y$$

$$a\Sigma x + b\Sigma x^2 + c\Sigma x^3 = \Sigma x y$$

$$a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 = \Sigma x^2 y$$

As in the case of linear trend, here also, x is so chosen that $\Sigma x = 0$.

Example 7

Fit a quadratic trend for the following time series. Estimate the population for the year 2001.

Year :	1951	1961	1971	1981	1991
Population (Crore) :	36	44	55	68	84

Solution

Here, n = 5 is an odd number. Therefore, the middle most point of time is taken as x = 0. The other values are taken as shown below.

Year	y	x	x^2	x^3	x^4	xy	x^2y
1951	36	– 2	4	– 8	16	– 72	144
1961	44	– 1	1	– 1	1	– 44	44
1971	55	0	0	0	0	0	0
1981	68	1	1	1	1	68	68
1991	84	2	4	8	16	168	336
Total	287	0	10	0	34	120	592

Let the quadratic trend equation be $y = a + bx + cx^2$. Then the normal equation are-

$$n a + b\Sigma x + c\Sigma x^2 = \Sigma y$$

$$a\Sigma x + b\Sigma x^2 + c\Sigma x^3 = \Sigma x y$$

$$a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 = \Sigma x^2 y$$

On Substitution, the equations are-

$$5a + 0 + 10c = 287 \quad \rightarrow \quad (1)$$

$$0 + 10b + 0 = 120 \quad \rightarrow \quad (2)$$

$$10a + 0 + 34c = 592 \quad \rightarrow \quad (3)$$

By equation (2), b = 120/10 = 12

Equation (1) when multiplied by 2 is

$$10a + 20c = 574$$

Subtracting this from equation (3) the result is

$$14c = 18. \text{ That is, } c = 18/14 = 1.29$$

Therefore by equation (1)

$$5a = 287 \text{ J } 10c = 287 - 12.9 = 274.1$$

$$\therefore a = \frac{274.1}{5} = 54.82$$

Thus, a = 54.82, b = 12 and c = 1.29

The trend values can be obtained by putting x = – 2, – 1, 0 etc. in this equation.

The value of x corresponding to the year 2001 is x = 3. Therefore, the estimate of population for 2001 is-

$$y = 54.82 + 12x3 + 1.29\ x3^2 = \mathbf{102.43} \text{ crore}$$

3. Exponential trend by the method of least squares

Here, a relation of the type $y = ab^x$ is fitted to the time series. The relation can be reduced to the linear form by taking logarithm. Thus, by taking logarithm, $y = ab^x$ reduces to -

$\log y = \log a + x.\log b$

That is, $Y = A + Bx$

where $Y = \log y$, $A = \log a$ and $B = \log b$

The normal equation for fitting $Y = A + Bx$ are -

$$nA + B\Sigma x = \Sigma Y$$

$$A\Sigma x + B\Sigma x^2 = \Sigma xY$$

On solving we get A and B. Then, the values of a and b are

a = antilog A and b = antilog B. Substituting a and b in the given equation we get the required trend.

Note: While fitting exponential curve, since logarithm are to be made use of, it is better to take x = 1, 2, 3 ..... as the time points.

Example 8

Fit an exponential trend to the following time series.

Year :	1990	1991	1992	1993	1994	1995
Value :	2	3	4	6	9	13

Solution

Let the trend equation be $y = ab^x$. On taking logarithm, this equation reduces to

$\log Y = \log a + (\log b)\, x$

That is, $Y = A + Bx$

where $y = \log y$, $A = \log a$ and $B = \log b$

The normal equations are

$$nA + B\Sigma x = \Sigma Y$$

$$A\Sigma x + B\Sigma x^2 = \Sigma xY$$

Year	Value (y)	time (x)	x^2	Y = log y	xY
1990	2	1	1	0.3010	0.3010
1991	3	2	4	0.4771	0.9542
1992	4	3	9	0.6021	1.8063
1993	6	4	16	0.7782	3.1128
1994	9	5	25	0.9542	4.7710
1995	13	6	36	1.1139	6.6934
Total	–	21	91	4.2265	17.6287

Thus, the normal equations are-

$6A + 21B = 4.2265 \quad \rightarrow \quad (1)$

$21A + 91B = 17.6287 \quad \rightarrow \quad (1)$

Multiplying equation (2) by 2 equation (1) by 7-

(2) $\times$ 2 $\rightarrow$ 42A + 182 B = 35.2574

(1) $\times$ 7 $\rightarrow$ 42A + 147 B = 29.5855

Subtracting: 35B = 5.6719

$$B = \frac{5.6719}{35} = 0.1621$$

Substituting this value in equation (1)

$6A + 21 \times 0.1621 = 4.2265$

$6A + 3.4041 = 4.2265$

$6A = 4.2265 - 3.4041 ; 6A = 0.8224$

$$A = \frac{0.8224}{6} = 0.1371$$

Thus A = 0.1371 and B = 0.1621

$\therefore$ a = antilog A = antilog 0.1371 = 1.371

b = antilog B = antilog 0.1621 = 1.452

Thus, the equation to the curve is-

$y = (1.371) \times (1.452)^x$

The value of x corresponding to the year 1998 is 9. Therefore, the estimate for 1998 is obtained by putting x = 9 in the trend equation. Thus, the estimate is $y = (1.371) \times (1.452)^9 = 39.33$

Merits

1. It is an objective method. It is mathematical and impersonal.
2. It is flexible in the sense that the trend line or the trend curve can be drawn depending on which one suits a given series.
3. The trend of each of the given periods can be determined. Similarly, the trend of any other period, past or future, can be calculated.
4. It is the best method. Trend line or trend curve is uniquely and mathematically obtained unlike in graphic method. Trend curve, if found suitable to a series, is obtained unlike in semi average method. Trend of any future or past or given period can be found unlike in moving averages method.

Limitations

1. It is neither simple nor easy.
2. If the data for a few more periods are available, calculations are to be done from the beginning unlike in the moving averages method.
3. Extreme values affect the results unduly unlike in the moving averages method.

Chapter 3

MEASUREMENT OF SEASONAL VARIATIONS

We have seen that, short term fluctuations observed in a time series data, particularly in a specified period, usually with in a year, are called seasonal variations.

The four methods mainly used to determine seasonal variations are

1. Method of Simple Averages
2. Ratio-to-Moving Average method
3. Ratio-to-Trend Method
4. Link Relative Method

1. Method of Simple Averages

For determining the seasonal variations, seasonal data are necessary. The steps in the calculation of seasonal indices are the following.

i. The data are arranged season-wise in chrnonological order.

ii. The total of the values of each season is found.

iii. From each total, average of the season is found.

iv. Seasonal averages are totalled and the total is divided by the number of seasons to get the grand average.

v. Seasonal indices of the different seasons are calculated using the formula.

$$\text{Seasonal Index} = \frac{\text{Seasonal Average}}{\text{Grand Average}} \times 100$$

Note

1. Total of the seasonal indices = 100 $\times$ No. of seasons.

Example 9

Calculate seasonal variation indices from the data given below.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2005	3.7	4.1	3.1	3.5
2006	3.7	3.9	3.6	3.6
2007	4.0	4.1	3.3	3.1
2008	3.3	4.4	4.0	4.0

Solution

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2005	3.7	4.1	3.1	3.5
2006	3.7	3.9	3.6	3.6
2007	4.0	4.1	3.3	3.1
2008	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.0	14.2
Average	3.675	4.125	3.5	3.55
Seasonal Variation Index adjusted	98.98	111.11	94.27	95.62

$$\text{Grand Average} = \frac{3.675 + 4.125 + 3.5 + 3.55}{4}$$

$$= \frac{14.85}{4} = 3.7125$$

$$\text{Seasonal Variation Index} = \frac{\text{Quarterly Average}}{\text{Grand Average}} \times 100$$

$$\text{Seasonal Variation Index for I Quarter} = \frac{3.675}{3.7125} \times 100 = 98.98$$

$$\text{Seasonal Variation Index for II Quarter} = \frac{4.125}{3.7125} \times 100 = 111.11$$

$$\text{Seasonal Variation Index for III Quarter} = \frac{3.5}{3.7125} \times 100 = 94.27$$

$$\text{Seasonal Variation Index for IV Quarter} = \frac{3.55}{3.7125} \times 100 = 95.62$$

Merits

1. It is the easiest method
2. It is the simplest and the least time consuming method.

Demerits

1. It assumes absence of trend which is not true in many series. It assumes that cyclical variations are compensated among themseleves which is also not true.
2. Trend values are calculated by the method of moving averages. Period of moving average is equal to number of seasons. Assuming additive model or multiplicative model, the seasonal effects are assessed.

2. Ratio - to - Moving Average Method

The method assumes multiplicative model. The following are the steps.

1. Trend values are calculated by the method of moving averages. Period of moving average = Number of seasons per year
2. Original values are expressed as the percentages of the trend values.

   ie., $\frac{\text{Original Value}}{\text{Trend}} \times 100$ are calculated
3. The percentages are tabulated, are totalled and are averaged.
4. Correction factor is calculated by remembering the following formula.

   $$\text{Correction Factor} = \frac{\text{No of seasons}}{\text{Total of the averages}} \times 100$$

v. Seasonal averages are multiplied by the correction factor to get the seasonal indices.

Note

The total of seasonal indices = 100 × no. of seasons

Merits

1. It considers trend and is better than the method of simple averages.
2. It eliminates the effect of cyclical variations and is found to agree with the nature of many series. It is claimed to be better than any other method.
3. The fluctuations of the ratios are less in this method when compared to ratio-to-trend method.
4. It is flexible in the sense that different methods are available under additive and multiplicative models.
5. It is the most popular method. Its popularity is due to the ease with which it is calculated as well as to its satisfactory estimates.

Demerits

1. Although it is more simple than other methods, it is more difficult than the method of simple averages.
2. While calculating trend, the moving averages of a few periods in the beginning and of equal number of periods in the end are not available. That makes one to feel that some information is lost during calculation.

Example 10

Calculate seasonal indicies for the following quarterly data.

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2001	30	81	62	119
2002	33	104	86	171
2003	42	133	99	221
2004	56	172	129	335
2005	67	201	136	302

Solution

Year and Quarter	Trend (Y)	4 quarterly moving total	Centered 4 quarterly moving total	Trend (Y)
2001 I	30		–	–
II	81		–	–
		292		
III	62		587	73.375
		295		
IV	119		613	76.625
		318		
2002 I	33		660	82.500
		342		
II	104		736	92.000
		394		
III	86		797	99.625
		403		
IV	171		835	104.375
		432		
2003 I	42		877	109.625
		445		
II	133		940	117.500
		495		
III	99		1004	125.500
		509		
IV	221		1057	132.125
		548		
2004 I	56		1126	140.750
		578		
II	172		1270	158.750
		692		
III	129		1395	174.375
		703		
IV	335		1435	179.375
		732		
2005 I	67		1471	183.875
		739		
II	201		1445	180.625
		706		
III	136		–	–
IV	302		–	–

After calculating the trend values, $\frac{\text{Original Value}}{\text{Trend}} \times 100$ are calculated and are tabulated season-wise as in the table.

Other steps are indicated above.

Percentages of the values to trend

Year	I	II	III	IV
2001	–	–	84.50	155.30
2002	40.00	113.04	86.32	163.83
2003	38.31	113.19	78.88	167.27
2004	39.79	108.35	73.98	186.76
2005	36.44	111.28	–	–
Total	154.54	445.86	323.68	673.16
Average	38.64	111.47	80.92	168.29
Seasonal Indices	38.71	111.66	81.06	168.58

EXERCISES

Multiple choice questions

1. A time series is a set of data recorded:
   a) periodically
   b) at time or space intervals
   c) at successive points of time
   d) all the above

4. The component of a time series which is attached to short-term fluctuations is:
   a) seasonal variation b) cyclic variation
   c) irregular variation d) all the above

6. The general decline in sales of cotton clothes is attached to the component of the time series:
   a) secular trend b) cyclical variation

   c) seasonal variation d) all the above

8. Method of least squares to fit in the trend is applicable only if the trend is:
   a) linear b) parabolic
   c) both (a) and (b) d) neither (a) nor (b)

10. Cyclic variations in a time series are caused by:
   a) lockouts in a factory b) war in a country -
   c) floods in the states d) none of the above

11. Irregular variations in a time series are caused by:
   a) lockouts and strikes b) epidemics
   c) floods d) all the above

12. Trend in a time series means:
   a) long-term regular movement
   b) short-term regular movement
   c) both (a) and (b)
   d) neither (a) nor (b)

13. An additive model of time series with the components *T, S, C* and I is:
   a) $Y = T + S + C \times I$
   b) $Y = T + S \times C \times I$
   c) $Y = T + S + C + I$
   d) $Y = T + S \times C + I$

14. A multiplicative model of a time series with compoennts T, S, C and I is
   a) $Y = T \times S \times C \times I$
   b) $T = Y / (S \times C \times I)$
   c) $C \times I = Y / (S \times I)$
   d) all the above

21. Least square estimates of parameter of a trend line:
   a) have minimum variance
   b) are unbiased

c) can exactly be obtained

d) all the above

28. Simple average method is used to calculate:

a) trend values b) cyclic variations

c) seasonal indices d) none of these

29. In case of multiplicative model, the sum of seasonal indices is:

a) 100 times the number of seasons

b) zero c) 100 d) any of the above

36. For the given five values 15, 24, 18, 33, 42, the three years moving averages are:

a) 19, 22, 33 b) 19,25,31

c) 19,30,31 d) none of the above

37. If the slope of the trend line is positive, it shows:

a) rising trend b) declining trend

c) stagnation d) any of the above

Very Short Answer Questions

41. Define time series.
42. Define secular trend of a time series.
43. Give the meaning of the term 'seasonal variation'. What is meant by trend?
48. What are cyclical variations?
49. What is the principle of least squares?
50. Define moving average.
51. What are the normal equations to fit a straight line $y = a + bx$?

Short Essay Questions

55. What is a time series? Explain with examples.
56. Explain the components of a time series with examples.
60. Define trend of time series. Give a graphical method of obtaining it.
61. Describe briefly the various methods of determining trend in a time series.
62. Explain the method of moving averages of obtaining trend. What are the merits and demerits of this method?
63. What is 'Method of least squares'? What are 'Normal equations'?

Long Essay Questions

70. Explain the meaning of time series analysis. Mention the four components into which a time series may be analysed.
71. Write a short essay on 'Analysis of Time Series'.
75. Calculate the trend values by finding 3-yearly moving averages. Show the trend on a graph.

Year :	'80	'81	82	'83	'84	'85	'86	'87	'88	'89
Sales :	1230	1060	1240	1300	1450	1160	1430	1320	1260	1120

76. The following data reveal the number of televisions of a particular model sold at different shops of a particular locality in Chennai. Compute 4-yearly moving averages. Plot the original and trend values on a graph.

Year :	'82	'83	84	'85	'86	'87	'88	'89	'90	'91	'92	93
T.V. sold :	580	540	570	680	690	700	612	692	697	620	589	501

79. Find the linear trend values by the method of least squares. Plot the original values and the trend values on the same graph.

Year :	1960	1965	1970	1975	1980	1985
Production (000kgs):	16	20	18	15	18	21

80. Fit an equation of the type $y = a + b\,x$ for the following data and estimate the production in 1993.

Year :	1985	1986	1987	1988	1989	1990	1991
Production :	142	180	150	127	140	171	140

87. For the following time series, obtain (i) Quadratic trend (ii) Exponential trend by the method of least squares.

Year :	1980	1982	1984	1986	1988	1990	1992
Value :	4	6	10	18	32	51	70

Chapter 4

INDEX NUMBERS

Meaning of Index Numbers

Index numbers are the devices to measure the relative movements in variables which are capable of being measured directly. They are used to compare the level of certain phenomena at different times or at different places on the same date. Index numbers are measures of relative changes and can show only a general tendency. In this sense they are techniques of estimating the general trends in prices, production and other economic variables. They are used to feel the pulse of the economy and they indicate the inflationary and deflationary tendencies. So index numbers are generally called 'economic barometers'.

'Index numbers can be defined as a statistical measure designed to show changes in a variable or a group of related variables with respect to time, place or other characteristics such as income, profession etc.'

Price Relatives

In order to study the relative change in the level of price of a commodity over a period of time we can find the relationship between the prices of that commodity at two times. The simplest way will be to find the price of that commodity for a particular year assuming the base year price as 100, ie. to express the price of a particular year as percentage of the corresponding price in the base year. This price percentage generally known as the rice relative is the simplest type of index numbers.

For example, the price of rice in 1990 is Rs. 750/- per quintal as compared to Rs.500 per quintal in 1985. To obtain the price index for 1990, we shall express 1990 price as percentage of the 1985 price; ie. Index number of price for 1990 with 1985 as base is.

$$\text{Price relative} = \frac{750}{500} \times 100 = 150$$

$$= \frac{\text{Current year price}}{\text{Basic year price}} \times 100$$

Let us denote the price in the current year (period) by p_1 and the price in the base year (period) by p_0. Here current year (period) is the year (period) for which the index number is to be constructed and base year (period) is the year (period) to which comparison is to be made. Similarly q_1 and q_0 denote the quantities for the current year (period) and base year respectively.

$$\text{So price relative} = \frac{\text{Current year price}}{\text{Basic year price}} \times 100$$

Problems in the Construction of Index Numbers

In the process of construction of price index numbers, one encounters several problems which are to be tackled and solved very carefully. We shall discuss some of these problems in the following lines. The major problems generally encountered are:

1. Purpose of index number
2. Selection of base period
3. Selection of items
4. Price quotations
5. Choice of an average
6. Selection of appropriate weights
7. Selection of formula

1. Purpose of Index Number

It is absolutely necessary that the purpose of the index number should be clearly and unambiguously defined, since most of the other problems will depend upon the purpose. Once the purpose is clear, the rest of the procedure is easy to apprehend.

2. Selection of Base Period

Index numbers are specially designed to compare the levels of certain phenomena at a given time to the levels of those phenomena on some previous time known as base period or base date. Again the choice is to be made by the statistician keeping the purpose of the index in his view. The selection of base period is essentially arbitrary. Yet he has to follow certain guidelines in its selection. The base period may be a single date, week, month, or year.

This base period should be fairly normal one, free from fluctuations and disturbances,

A base year should not be too remote or too distant in the past. Because as time passes (i) the depression of price relations may become so great that no average is reliable, (ii) Market conditions, ie. tastes and habits of people undergo some change resulting in the replacement of old goods by new ones, (iii) The quantity of many commodities may change progressively with time.

3. Selection of Items

In the answers to the questions, 'what items should be included in a sample and how many items should be included in a sample' lies the solution of this problem. The items to be selected should be relevant, representative, reliable and comparable.

4. Price Quotations

Collection of price quotations for the commodities selected is a difficult problem. The price of a commodity varies from market to market and even at one market from shop to shop. As it is not possible to obtain price quotations from all the markets, we have to select representative markets. When the market from where the price quotations are to be obtained are decided the next thing is to select a suitable price reporting agency, like business firms, chamber of commerce, trade associations, government agent etc.

Further the question arises as to the price, ie., whether wholesale price or retail price should be taken into account. By and large it depends upon the purpose of the index number. Whole sale prices should be preferred to the retail ones as the

former, fluctuate less as compared to the latter and are more sensitive to the conditions of demand and supply.

5. Choice of an Average

Since an index number is a technique of averaging all the changes in a group of values over a period of time, the problem is to select an average which summarises the changes in the component items adequately. Among the averages, generally arithmetic mean or geometric mean employed for constructing index numbers.

6. Selection of Appropriate Weights

The items included in the index number are not of equal importance. So the different items included in the index number must be weighted according to their importance. The allotment of degree of relative importance to each item is known as *weighting*. Thus we have to choose weights to each item.

7. Selection of Formula

A large number of formula has been devised for constructing index numbers. The problem very often is that of selecting the most appropriate formula. The choice of the formula would depend not only on the purpose of the index but also on the data available.

Methods of Constructing Index Numbers

A wide variety of methods are used in the construction of index numbers. Broadly speaking they can be classified under two heads.

1. Simple index numbers
2. Weighted index numbers

1. Simple Index Numbers

They are of two types of simple index numbers.

(i) Simple aggregative index number

$$\text{Simple aggregative Index No} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

The steps in the calculations are:

a. Express the item in some common unit of measurement.

b. Total the prices of items for both base period and current period, which will give Σp_0 and Σp_1.

c. Divide the total of the current period by the base period total and multiply it by 100. The result is the simple aggregative index number.

(ii) Simple Average of Price Relatives

Under this method,

The simple average Index No = $\frac{1}{n}\ \Sigma\frac{p_1}{p_0}\times 100$

where 'n' denotes the no: of commodities The steps in its construction are:

a. Determine the base period.

b. To obtain price relatives, divide the current year prices of each commodity by their base year prices.

c. Obtain the total of the price relatives.

d. Divide the total by the number of commodities.

e. Multiply this average by 100, which will give the simple average of price relatives.

The following example illustrate the procedure.

Example 1

For the following data compute simple index numbers.

Commodity	:	Rice	Wheat	Ghee	Sugar
Price in 1985	:	20	30	20	10
Price in 1990	:	40	45	50	30

Solution

Commodity	Price in 1985 (p_0)	Price in 1990 (p_1)	Price relative p_1/p_0
Rice	20	40	2.0
Wheat	30	45	1.5
Ghee	20	50	2.5
Sugar	10	30	3.0
	80	165	9.0

1. Simple Aggregative Index Number = $\Sigma\frac{p_1}{p_0}\times 100$

$$=\frac{165}{80}\times 100 = \mathbf{206.25}$$

2. Simple Average of Relatives Index Number = $\frac{\Sigma\frac{p_1}{p_0}}{n}\times 100$

$$=\frac{9.0}{4}\times 100 = \mathbf{225.0}$$

2. Weighted Index Numbers

In the calculation of simple index numbers we assign equal importance to all the items included in the index. But construction of useful index numbers requires a conscious effort to assign to each commodity a weight in accordance with its importance. There are two types of weights namely quantity weights and value weights. In the case of aggregative method we use quantities (q) as weights but in the case of price relatives value weights are used. Value implies rupee value and symbolised by pq.

Weighted index numbers are of two types.

1. Weighted Aggregative Index Number

As weights can be assigned in many ways different methods of constructing index numbers have been devised of which some of the more important ones are:

1. Laspeyre' s method
2. Paasche's method
3. Fisher's ideal method
4. Marshall - Edgeworth method.

1. Laspeyre's Method

Laspeyre's Index Number = $\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0}\times 100$

Here q_0 = base year quantity, q_1 - current year quantity

This is called Laspeyre's formula

2. Paasche's Method

$$\text{Paasche's Index number} = \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0} \times 100$$

This is called Paasche's formula.

3. Fisher's ideal method

Prof. Irving Fisher has given a number of formulae for the construction of index numbers. According to him,

Fisher's ideal Index number = $\sqrt{L \times P}$ where L denotes Laspeyre's index and P denotes Paasche's index. Substituting L and P, we get the Fishers index number as,

$$\textit{Fishers ideal Index number} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}} \times 100$$

4. Marshal-Edgeworth method

In the method also both the current year as well as base year prices and quantities are considered. The formula for constructing the index is

$$\text{Marshall-Edgeworth Index Number} = \frac{\Sigma (q_0 + q_1) p_1}{\Sigma (q_0 + q_1) p_0} \times 100$$

$$= \frac{\Sigma q_0 p_1 + \Sigma q_1 p_1}{\Sigma q_0 p_0 + \Sigma q_1 p_0} \times 100$$

Example 2

Calculate Laspeyre's, Paasche's, Fisher's index number of prices for the following data.

Commodities	Base year		Current year	
	Price (p_0)	Quantity (q_0)	Price (p_1)	Quantity (q_1)
A	10	12	12	15
B	7	15	5	20
C	5	24	9	20
D	16	5	14	5

Solution

The price index numbers of current year w.r.t. base year is denoted as P_{01}.

q_0p_0	q_0p_1	q_1p_0	q_1p_1
120	144	150	180
105	75	140	100
120	216	100	180
80	70	80	70
425	505	470	530

Laspeyre's index number, $P_{01} = \frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times 100$

$$= \frac{505}{425} \times 100 = \mathbf{118.82}$$

Paache's index number, $P_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0} \times 100$

$$= \frac{530}{470} \times 100 = \mathbf{112.76}$$

Fisher's index number, $P_{01} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}} \times 100$

$$= \sqrt{\frac{505}{425} \times \frac{530}{470}} \times 100 = \mathbf{115.8}$$

Example 3

It is stated that the Marshall-Edgeworth index number is a good approximation to the ideal index number. Verify, using the following data:

Commodities	1990		1994	
	Price	Quantity	Price	Quantity
A	2	74	3	82
B	5	125	4	140
C	7	40	6	33

Solution

Nothing is indicated to decide which is base year and which is current year. In such cases the recent year is to be taken as the current year and the other as the base year and the procedure is as follows:

Com.	1990		1994		q_0p_0	q_0p_1	q_1p_0	q_1p_1
	$Pri(p_0)$	$Qty(q_0)$	$Pri(p_1)$	$Qty(q_1)$				
A	2	74	3	82	148	222	164	246
B	5	125	4	140	625	500	700	560
C	7	40	6	33	280	240	231	198
Total					1053	962	1095	1004

Fisher's Ideal I.N., $P_{01} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}} \times 100$

$$= \sqrt{\frac{962}{1053} \times \frac{1004}{1095}} \times 100 = \mathbf{91.524}$$

Marshall-Edgeworth I.N.,

$$P_{01} = \frac{\Sigma q_0 p_1 + \Sigma q_1 p_1}{\Sigma q_0 p_0 + \Sigma q_1 p_0} \times 100$$

$$= \frac{962 + 1004}{1053 + 1095} \times 100 = \mathbf{91.527}$$

The values show that the Marshall-Edgeworth index number is a good approximation to Fisher's ideal index number.

Example 4

Using Fisher's Ideal Formula compute price and quantity index numbers for 1984 with 1982 as base year, given the following:

Year	Commodity A		Commodity B		Commodity C	
	Price (Rs.)	Quantity (Kg.)	Price (Rs.)	Quantity (Kg.)	Price (Rs.)	Quantity (Kg.)
1982	5	10	8	6	6	3
1984	4	12	7	7	5	4

Solution

The prices and the quantities of the three commodities are to be taken one below the other before necessary products are found.

Com.	$Pri(p_0)$	$Qty(q_0)$	$Pri(p_1)$	$Qty(q_1)$	q_0p_0	q_0p_1	q_1p_0	q_1p_1
A	5	10	4	12	50	40	60	48
B	8	6	7	7	48	42	56	49
C	6	3	5	4	18	15	24	20
Total	-	-	-	-	116	97	140	117

By Fisher's ideal formula,

$$P_{01} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}} \times 100$$

$$= \sqrt{\frac{97}{116} \times \frac{117}{140}} \times 100$$

$$= \mathbf{83.60}$$

$$Q_{01} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}} \times 100$$

$$= \sqrt{\frac{140}{116} \times \frac{117}{97}} \times 100$$

$$= \mathbf{120.65}$$

2. Weighted Averages of Relatives Method

Price Indices (P_{01})	Quantity Indices (Q_{01})
i. Using A.M.; $P_{01} = \frac{\Sigma WP}{\Sigma W}$	$Q_{01} = \frac{\Sigma WQ}{\Sigma W}$
ii. Using G.M.; $P_{01} = \text{Antilog}\left[\frac{\Sigma W \log P}{\Sigma W}\right]$	$Q_{01} = \text{Antilog}\left[\frac{\Sigma W \log Q}{\Sigma W}\right]$

where $P = \frac{p_1}{p_0} \times 100$, $Q = \frac{q_1}{q_0} \times 100$ and $W = q_0 p_0$

This method is better than the corresponding unweighted method in showing the relative change. From the data available under this, unweighted averages of relatives also could be calculated. This method facilitates replacing one or more items at a later stage.

Tests for an Ideal Index Number

Index numbers are constructed to study the relative changes in prices, quantities, etc. of one time in comparison with another. Many formulae are available. The present consideration is whether each of the formulae is as it is expected under the following three test conditions.

1. Time Reversal Test
2. Factor Reversal Test

1. Time Reversal Test

This requires the formula to be such that $P_{01} \times P_{10} = 1$, after ignoring the factor 100 in each index. In the words of Prof. Irving Fisher who proposed the test condition, *"...the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base or putting it in another way, the index number reckoned forward should be reciprocal of the one reckoned backward".*

P_{10} is the price index number of the base year in comparison with the current year. That is, base year figure will be in numerator and current year figure will be in denominator. Hence, it is expected to be the reciprocal of P_{01} In other words, the product of P_{01} and P_{10} is expected to be unity.

This test is not satisfied by Laspeyre's method and Paasche's method. We can see that, under Laspeyre's method,

$$P_{01} = \frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \quad \text{and} \quad P_{10} = \frac{\Sigma q_1 p_0}{\Sigma q_1 p_1}$$

(obtained by interchanging 0 by 1 and 1 by 0 in P_{01})

$$\therefore P_{01} \times P_{10} = \frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_0}{\Sigma q_1 p_1} = 1$$ and the test is not satisfied.

Similar is the case of Paasche's index.

But, besides Fisher's method, Marshal-Edgeworth and weighted and unweighted geometric mean of relatives methods satisfy this test. We can verify this in the case of Fisher's method.

Here $$P_{01} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}}$$

$$P_{10} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_1 p_1} \times \frac{\Sigma q_0 p_0}{\Sigma q_0 p_1}}$$

$$\therefore P_{01} \times P_{10} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}} \times \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_1 p_1} \times \frac{\Sigma q_0 p_0}{\Sigma q_0 p_1}}$$

$$= \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0} \times \frac{\Sigma q_1 p_0}{\Sigma q_1 p_1} \times \frac{\Sigma q_0 p_0}{\Sigma q_0 p_1}}$$

$= \sqrt{1} = 1$ and the test is satisfied.

2. Factor Reversal Test

This requires the formula to be such that, $P_{01} \times Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_0}$ after ignoring the factor 100 in each index. In the words of Prof. Irving Fisher who proposed this condition also, *"just as each formula should permit the inter change of two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results-that is; the two results multiplied together should give the true value ratio, except for a constant of proportionality."*

P_{10} gives the relative change in price while Q_{01} gives the relative change in quantity, which is obtained by interchanging p by q and q by p in P_{01}. Hence $P_{01} \times P_{10}$ should give the relative change in price multiplied by quantity [ie., value] and so should be equal to $\frac{\Sigma q_1 p_1}{\Sigma q_0 p_0}$.

Fisher's is the only formula which satisfies this test, as can be seen below.

$$P_{10} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}}$$

$$Q_{01} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$$

$$\therefore P_{01} \times Q_{01} = \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0}} \times \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$$

$$= \sqrt{\frac{\Sigma q_0 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_1 p_0} \times \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$$

$= \sqrt{\frac{(\Sigma q_1 p_1)^2}{(\Sigma q_0 p_0)^2}} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_0}$ and the test is satisfied.

Fisher's formula is found to satisfy the time reversal and factor reversal tests while no other formula does. Hence, Fisher's formula is called as ideal index number formula.

Uses of Index Numbers

We have seen that index numbers are devices for measuring differences in the magnitude of a group of related variables. The following are the uses of index numbers.

Limitations of Index Numbers

We have seen that index numbers study only those problems which are capable of quantitative expression and which change with time variation. Without the knowledge of time variation and quantitative version of the problem it is impossible to apply the technique of index number to study such problems. The following are the limitations of index numbers.

1. They are only approximate indicators of the relative changes, due to the fact that one cannot conceive of absolute accuracy in their construction.
2. An index number is generally based on a sample.
3. It is very difficult to select a base year which is normal in all respects.
4. The index numbers of prices may not take into account variations in quantity which may be significant.
5. Index numbers are specialised type of averages and as such they also suffer from a weakness common to most averages.
6. Many times unit of comparison are different, so that same

data may yield different index numbers at the hands of different individuals.

7. An index number can be calculated in so many different ways and different methods give different answers.

8. An index number is useful only for the purpose for which it is designed.

EXERCISES

Very Short Answer Questions

11. Define an index number
12. What is price relative?
13. Give the formula for Fisher's ideal index.
14. What is time reversal test?
15. What is factor reversal test?
16. Explain the precautions that we have to take to fix 'base year' to calculate index number.
17. What is quantity relative?
18. Explain briefly the concept of whole sale price index number.
19. Give the formula for Laspeyre's and Paasche's index.
20. Give the formula for Marshall Edgeworth index Number?
21. Give any three limitations index number.
22. What are the uses of index numbers?

Short Essay Questions

23. What are index numbers? What are their uses? What are their uses?
24. Explain the different steps in the construction of cost of living index numbers.
25. What do you understand by time reversal test. Test whether Laspeyre's price index number satisfy this test.

26. Examine analytically the relationship between Laspeyre's and Paasche's index numbers.
27. What are time reversal and factor reversal tests in index numbers?
28. Why is Fisher's Index Number known as Ideal Index Number.
29. Why are index numbers called economic barometers?
30. What are the limitations of Index Numbers?

Long Essay Questions

31. What is time reversal test? With the help of the following data show that Laspeyre's index does not satisfy this test.

Commodity	Price		Quantity	
	1970	1980	1970	1980
X	4	10	50	40
Y	2	8	10	5
Z	3	7	5	5

32. Compute Fisher's and Marshal Edgeworth's price index numbers for the following data

Items	1981		1991	
	Price	Quantity	Price	Quantity
I	5	62	6	71
II	7	43	8	100
III	9	93	12	65

33. Calculate simple aggregative index and Fisher's index based on the following information.

Items	Production in tons		Price per ton	
	1986	1990	1986	1990
A	25	30	150	300
B	10	12	120	200
C	2	3	600	1000
D	1	2	200	300

34. Compute Fisher's ideal index from the following data.

Commodity	Price		Quantity	
	Base year	Current year	Base year	Current year
A	10	12	12	15
B	7	5	15	20
C	5	9	24	20
D	16	14	5	5

Module IV

STATISTICAL QUALITY CONTROL

Chapters

Chapter 1

QUALITY CONTROL

Introduction

With the development in theory of statistics and sampling methods, there are a number of statistical devices which are helpful in maintaining and improving the quality of the products produced and this process is known as Statistical Quality Control [SQC]. SQC has its sound foundation in the theory of sampling and in probability distributions and tests of significance. In applying SQC the number of rejections are minimized, the cost of production reduced and good quality products are assured even before the products are produced by the industry.

Quality

The term *quality* in Statistical Control does not always mean the highest standards of manufacture, but it refers to the good quality item which conforms to the standards specified for measurement.

Statistical quality control is a statistical method for determining the extent to which quality goals are being met without necessarily checking every item produced and for indicating whether or not the variations which occur are exceeding the normal expectations of SQC.

Here we mention some terminologies related to quality variations.

Specification Limits: It is customary for any buyer to expect the product to be at a particular size, quality, etc. For example, a carpenter may find screws of length 50mm most appropriate for a particular work. If he wants screws of length 50mm alone, he may buy accordingly. Or he may find that screws of length between 49mm and 51mm are useful for the work and so is prepared to accept such of them. In the later case, the lower and the upper specification limits are 49mm and 51mm.

Specifications limits are stipulated by the buyers. The buyers are expected to accept those items which fall within the specification limits.

Chance Causes: Certain variations are random in nature and are beyond the control of the human beings. Those variations may be reduced but cannot be eliminated. This kind of variation is called *allowable variation.* The causes for this kind of variation are called chance causes or non-assignable causes.

Assignable Causes: The causes of certain variations can be identified. By taking appropriate steps the variations can be eliminated, if necessary. This kind of variation is said to be *preventive variation.* The causes for this kind of variation are called *assignable causes.*

Natural Tolerance Limits: It is known that variation due to chance causes is inherent in the outputs under any scheme of production. As to be seen later under $\bar{x}$ chart almost all the variation due to chance causes lie within $\mu \pm 3\sigma$ limits and those limits are called natural tolerance limits. The width 6σ is called natural tolerance.

Process Control and Product Control: A production process is said to be under statistical control when the manufactured items lie within the natural tolerance limits. That is, the assignable causes of variation are absent and the process is governed by the chance causes alone. The process control is achieved through the technique of control charts pioneered *by W.A* Shewart.

By product control we mean controlling the quality of product by critical examination at strategic points and this is achieved by 'Sampling Inspection Plans' developed by Dodge and Romig.

Acceptance Sampling: A buyer draws a sample of a few units from a lot that is to be accepted by him if the number of defective is less. If the number of defectives is more, the lot is to be rejected. This technique of sampling inspection and decision making is called acceptance sampling.

Variables and Attributes

Variables are those quality characteristics of a product or item which are measurable to any extent (theoretically) and can be expressed in their respective units of measurements.

Attributes are the characteristics of products which are not amenable to measurement. Such characteristics can be identified by their presence or absence of them. This applies to many things that may be judged by visual examination.

Control Charts

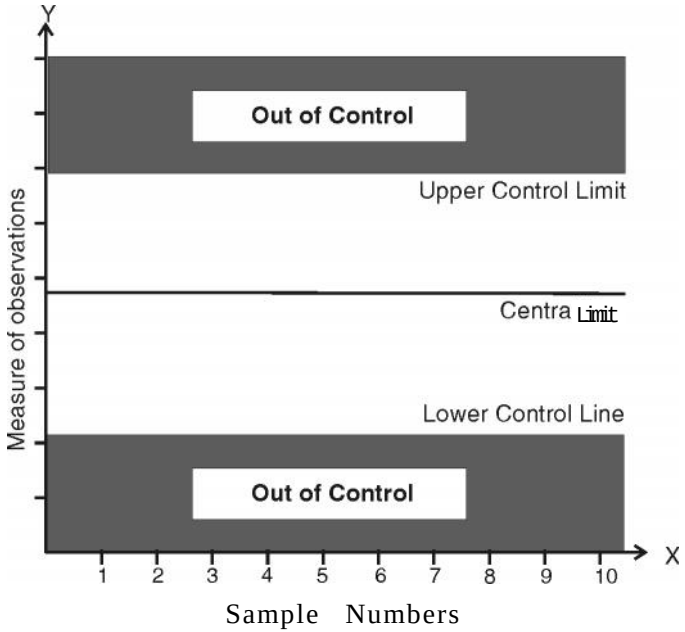
Dr. Walter A. Shewhart propounded a technique for finding out whether a manufacturing process is governed by chance causes of variation alone and so is under statistical control or is affected by assignable causes of variation also and so needs setting right. For that purpose he used charts called control charts.

Control charts consist of three horizontal lines besides the usual axes. The lines are

1. *Central line (C.L.).* The central line is generally drawn at the mean of the observations.
2. *Upper Control Limit (U.C.L).* The upper control limit is at three sigma distance above the mean.
3. *Lower Control Limit (L.C.L.)* The lower control limit is either at $3\pm$ distance below the mean or the X axis when it is to be below the X axis.

Model of a Control Chart

A chart is drawn for one statistical measure of observations If the measure follows normal probability distribution, 99.73% of the observations are expected to lie within the control limit. Otherwise 8/9 of the observations are expected to lie in those control limits. Hence in either case, the control charts protect against both the types of errors-not noticing the trouble when there is and looking for trouble when there is not.



Types of control charts

There are two types of control charts

(i) Control charts for variables and

(ii) Control charts for attributes

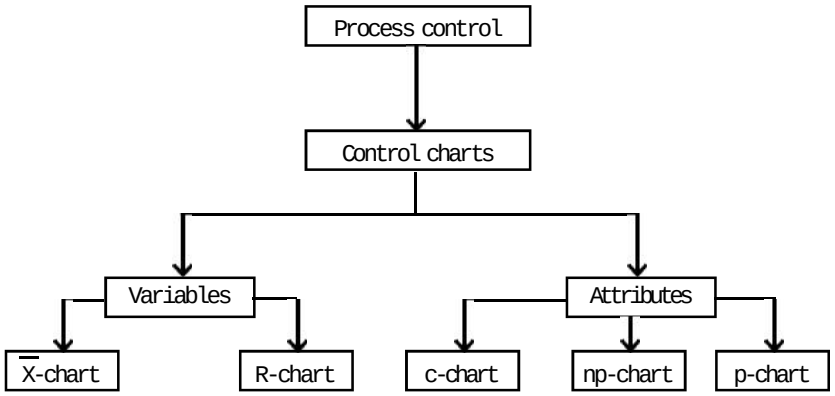
(i) Control chart for variables: As we have defined earlier, variables are the quality characteristics of manufactured products which are measurable. It can be expressed in their respective units of measurements. For example, diameters of wires, tensile strength of steel rods, breaking strength of yarn, life of electric bulbs, pitch diameters of screws, capacity of tins and cans etc. represent variables.

These types of variables are continuous in nature and they follow the Normal distribution. For this purpose we use two charts: (1) the Mean chart known as $\overline{X}$-charts and (2) the Range chart known as R-chart.

(ii) Control chart for attributes: The quality characteristics of products which are not measurable, but which can be identified by their presence or absence in the products, are

termed as attributes. By inspection one can judge whether a product is good or bad. For example, a glass-tumbler is cracked or not?, the number of imperfections in a piece of cloth, number of defective items in a very big lot. To control the quality of products which are governed by the attributes we use, the p-chart for proportion of defectives, the np-chart for number of defectives and the c-chart for number of defects per item or unit.

To summarise we have the following diagram.



3_S Control limits

$3\pm$ limits were proposed by Dr. Shewhart for his control charts from various considerations, the main being probabilistic considerations. Consider the statistic $t = t(x_1, x_2, ..., x_n)$, a function of the sample observations $x_1, x_2, ..., x_n$.

Let E(t) = « *and* $Var(t) = \pm^2$

If the statistic t is normally distributed, then from the fundamental area property of the normal distribution, we have

$$P[\ll -3\pm < t < \ll +3\pm] = 0.9973 \quad \text{or} \quad P[|t - \ll| > 3\pm] = 0.0027$$

Here $\ll -3\pm$ is called lower control limit and $\ll +3\pm$ is called upper control limit. If for the i-th sample, the observed t lies between LCL and UCL, there is nothing to worry, otherwise a danger signal is indicated.

Chapter 2

CONTROL CHARTS FOR VARIABLES

These charts may be applied to any quality characteristic that is measurable. Many quality characteristic of a product are measurable and can be expressed in specific units of measurements such as diameter of a screw, weight of soaps, tensile strength of steel pipe, life of an electric bulb etc. Here control charts are employed to control the mean and standard deviation respectively of the measurable characteristic. Usually $\overline{X}$ and R charts are employed for this purpose.

Mean Chart or $\overline{X}$ Chart

No production process is perfect enough to produce all the items exactly alike. Some amount of variation is always expected in any production scheme. This variation may be due to chance causes and/or assignable causes. The $\overline{X}$ chart is prepared to show the fluctuations of the means of samples. The steps involved in the construction of $\overline{X}$ chart are the following.

Procedure for the construction

1. Compute the mean of each sample say $\bar{x}_1, \bar{x}_2, ...\bar{x}_k$ where k denote the number of samples.
2. Compute $\bar{\bar{x}} = \dfrac{\bar{x}_1 + \bar{x}_2 + ... + \bar{x}_k}{k}$
3. Choose the central line as X.
4. Fix the UCL and LCL using any of the appropriate formula

   a) $UCL_{\bar{x}} = «' + A±'$

   $LCL_{\bar{x}} = «' - A±'$

This is used when the standards « and ± are known. «′ and ±′ are specified values of « and ± respectively and A = $3/\sqrt{n}$, which is tabulated for different values of n from 2 to 25.

b) $UCL_{\bar{x}} = \overline{\overline{X}} + A_2\overline{R}$

$LCL_{\bar{x}} = \overline{\overline{X}} - A_2\overline{R}$

This is used when standards are not given. A_2 can be obtained from table for different values of n from 2 to 25.

c) $UCL_{\bar{x}} = \overline{\overline{X}} + A_1\overline{S}$

$LCL_{\bar{x}} = \overline{\overline{X}} - A_1\overline{S}$

This is also used when the standards are not given. A_1 is tabulated for different values of n from 2 to 25.

5. After fixing LCL, CL and UCL the sample means are plotted on the chart.

Range Chart or R-Chart

R-chart is also drawn for measurable characteristics such as hardness, breadth, thickness etc. It shows variability within the process.

The R-chart is constructed using the following steps

Procedure for the construction

1. Calculate the range R for each sample.

2. Compute $\overline{R} = \frac{R_1 + R_2 + ... + R_k}{k}$, k being the sample number.

3. Set the control limits as follows.

a) When standards are given, say SD = $\pm'$

$CL = d_2\pm', \quad UCL_R = D_2\pm', \quad LCL_R = D_1\pm'$

b) When standards are not given

$CL = \overline{R}, \quad UCL_R = D_4\overline{R}, \quad LCL_R = D_3\overline{R}$

The values d_2, D_1, D_2, D_3 and D_4 are obtained from the table.

c) The $3\pm$ control limits for R-chart are

$CL = \overline{R}, \quad UCL_R = \overline{R} + 3\pm_R, \quad LCL_R = \overline{R} - 3\pm_R$

Control chart for Standard Deviation ($\pm$-chart)

When the standard deviation $\pm$ is specified, then R chart is constructed as follows:

Central line = $d_2\pm$

Upper control limit = $D_2\pm$

Lower control limit = $D_1\pm$

where $\pm$ is the value of standard deviation D_1, D_2 and d_2 values are obtained from tables depending on the sample size n. When $\pm$ is not known, it can be replaced by the appropriate estimate ($\pm = \overline{R}/d_2$).

Interpretation of $\overline{X}$ and R Charts

1. A process is said to be under statistical control if both $\overline{X}$ and R charts show control. That is if all the sample points fall with in the control limits, we say that the process is in control.
2. If one or more of the prints in $\overline{X}$ chart or in R chart or in both fall outside the control limits, we say that the process is out of control. This will be due to some assignable causes of random fluctuations and can be rectified.
3. $\overline{X}$ chart shows the undesirable variations between samples while R chart reveals any undesirable variations with in the samples. The two charts must be studied simultaneously to decide whether the process is under control or not. It is advisable to construct R chart first. If R charts shows a lack of control it must be rectified before constructing $\overline{X}$ chart.

Example

Construct a control chart for the mean and range for the data in which samples of 5 are taken

19	36	42	51	60	18	42	42	19	36	42	51
24	54	51	74	60	20	65	45	34	50	50	74
80	69	57	75	72	27	75	68	70	69	58	75
81	77	59	78	95	42	78	72	81	81	59	78
81	84	78	132	138	60	87	90	81	84	78	60

Comment whether the production seems to be under control.

Solution

Mean = $\overline{\overline{X}} = \frac{757.6}{12} = 63.13$ and $\overline{R} = \frac{613}{12} = 51.08$

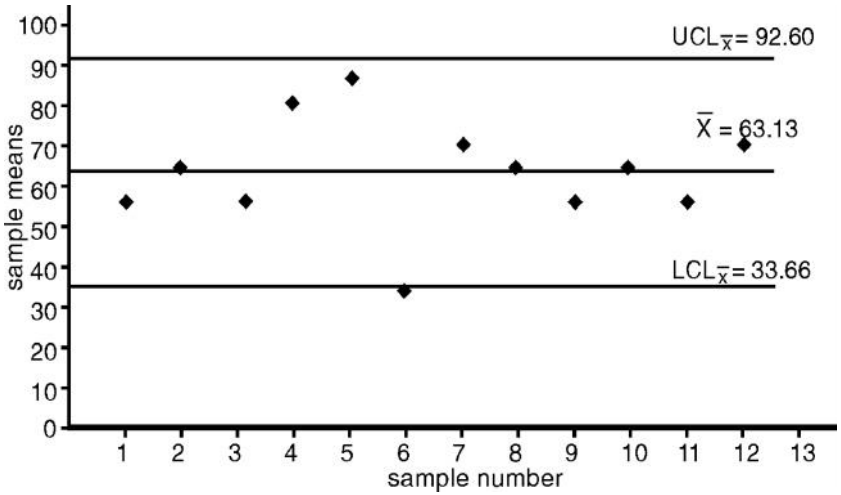
From table, for n = 5, A_2 = 0.577, D_3 = 0 and D_4 =2.115

Sample No.	Sample Values					Mean	Range
	X_1	X_2	X_3	X_4	X_5	X	R
1	19	24	80	81	81	57.0	62
2	36	54	69	77	84	64.0	48
3	42	51	57	59	78	57.4	36
4	51	74	75	78	132	82.0	81
5	60	60	72	95	138	85.0	78
6	18	20	27	42	60	33.4	42
7	42	65	75	78	87	69.4	45
8	42	45	68	72	90	63.4	48
9	19	34	70	81	81	57.0	62
10	36	50	69	81	84	64.0	48
11	42	50	58	59	78	57.4	36
12	51	74	75	78	60	67.6	27
	Total					757.6	613

Hence for $\overline{X}$ chart, CL: $\overline{X}$ = 63.13

$$UCL_{\bar{x}} = \overline{\overline{X}} + A_2\overline{R} = 63.13 + 0.577 \times 51.08 = 92.60$$

$$LCL_{\bar{x}} = \overline{\overline{X}} - A_2\overline{R} = 63.13 - 0.577 \times 51.08 = 33.66$$

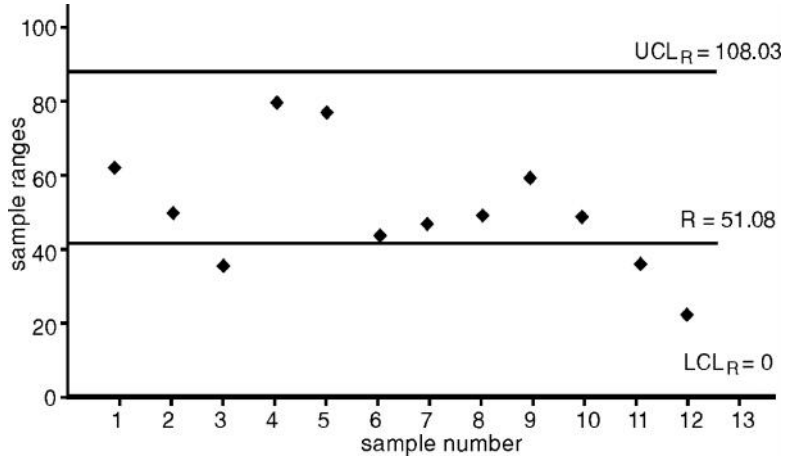


Sample number 6 lies outside the control limits. It indicates that the production process is out of control.

For R chart, C.L.: $\overline{R}$ = 51.08

U.C.L.: $D_4\overline{R}$ = 2.115 × 51.08 = 108.03

L.C.L.: $D_3\overline{R}$ = 0 × 51.08 = 0



Since all the points lie within that control limits, the production process is in control as far as the measure of dispersion R is concerned.

Chapter 3

CONTROL CHARTS FOR ATTRIBUTES

1. c-chart (Control chart for number of defects)

One of the important control chart for attributes is the c-chart. It is designed to control the number of defects per unit. There are many situations in industry where the data are obtained by counting and are of discrete type. For example, the number of idle machines, number of defects in castings, number of air bubbles in glasses, number of blemishes in a sheet of paper, number of blemishes in galvinised sufaces, number of weak spots in a given length of wire, and number of breakdowns at this weak spots, number of defects in body alignments of aircrafts, and buses, number of rust spots in steel sheet, number of mould marks on fibre glass canoes, are the examples where c-chart is usually of one unit. It may be of fixed time, length, area, a single unit or a group consisting of different units. In the case of textile yarn, a fixed length is taken as a unit. In the case of glass, a given area is taken as a unit. In a glass vessel, one vessel is taken as a unit. In the vase of wool or textile cloth, a given area is taken as one unit. In the case of surface blemishes, a given area of the surface is the unit of sample.

The theoretical basis for the c-chart is derived from the poisson distribution. In case of defects, in units, the opportunity for the occurrence is very large (large number of trials) while the actual occurrence of defects tend to be small (the probability of success p is very small). This situation can be very well represented by a Poisson Distribution.

Procedure for the construction of c-chart

Let c denote the number of defects counted in one unit of cloth of paper or any material. Find the mean $\bar{c} = (c_1 + c_2 + ... + c_n / n)$ where $c_1, c_2, ..., c_n$ are the defects

counted in several such units. The expected standard or central line is $\bar{c}$. In a poisson distribution, the variance is equal to mean ie., $\pm^2 = \bar{c}$ or $\pm = \sqrt{\bar{c}}$. Based on this, $3\pm$ limits for the upper and lower control limits are obtained.

$$\text{UCL} = \bar{C} + 3\pm = \bar{C} + 3\sqrt{\bar{C}} \quad \text{and}$$

$$\text{LCL} = \bar{C} - 3\pm = \bar{C} - 3\sqrt{\bar{C}}$$

The use of c-chart is valid and made appropirate when the opportunities for the occurrence of a defect in each production unit are infinite but the probability of a defect at any point is very small and is constant. While using c-chart uniform sample size is taken or unit sizes are taken.

Applications of c-Chart

In spite of the limited field of application of c-chart, there are instances in industry where c-chart is definitely needed. Some of the typical defects to which c-chart can be applied with advantage are the following.

1. The number of imperfections observed in a bale of cloth.
2. The number of surface defects observed in the roll of coated paper or a sheet of photographic film.
3. Number of defects of all types observed in air craft subass emblies of final assembly.
4. Number of breakdowns at weak spot in insulation in a given length of insulated wire subject to a specified test voltage.

2. p-chart (Chart for fraction defectives)

The most versatile and widely used control chart for attributes is the *p*-chart. This chart is for the fraction rejected as nonconforming to specifications. This is also called the *chart for fraction defective*. This chart is applied to quality characteristics that are considered as attributes.

Fraction defective or Fraction rejected p, is defined as the ratio of the number of nonconforming articles in any inspection or series of inspections to the total number of articles actually inspected.

The theoretical basis for the *p*-chart is the binomial distribution. This distribution gives better results when the sample size is atleast 50 or above. About 20 to 25 samples are sufficient to get an idea of the performance of the process.

Procedure for the construction of p-chart

The different steps for constrcting a p-chart is given below:

1. Compute the fraction defective for each sample.

$$p = \frac{\text{Number of defective units in the sample}}{\text{Sample size or Total number of units inspected in the sample}}$$

2. Obtain the average fraction defective p from all the given samples

$$\bar{p} = \frac{\text{Total number of defectives in all the samples combined}}{\text{Total number of items inspected in all the sample combined}}$$

3. The expected standard or central line CL = $\bar{p}$

4. The upper and lower control limits are given by

$$UCL_p = \bar{p} + 3\pm_p$$

$$= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

and $$LCL_p = \bar{p} - 3\pm_p$$

$$= \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

After setting the control limits and marking on the graph sheet we can plot the sample points. If all the points lie within the sample limits the process is under control, otherwise it is not.

3. *np* chart (Control chart for number of defectives): In order to simplify the work of the person in the shops and factories to plot the necessary points in the control chart, the p-chart is modified and instead of plotting the fraction defectives

we plot the number of defectives np in each sample which is always an integer. If the number of items inspected on each occasion is the same, then the plotting of the actual number of defectives is more convenient and meaningful than plotting the fraction defective p. In the np chart if n is taken as 100 then it represents the percentage defective chart or 100 p-chart. The central line or the expected standard is

$$CL = n\bar{p} = \frac{\text{Total number of defectives of all samples}}{\text{Number of samples inspected}}$$

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

and $$LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

We have to plot the values of the number of defectives $d = n$ « in the chart and find the nature of the process. Here d will never be negative. Hence whenever LCL is less than zero it is taken as zero.

Note

To obtain the control limits *for np* chart, multiply the control limits *of p* chart by *n*.

Application

The preparation of the *p*-chart helps the management in the following way:

1. It helps to arrive at the average proportion of nonconforming articles or products submitted for inspection over a period of time.
2. It attracts the attention of authorities over any changes in the average quality level.
3. It guides to discover the out of control danger spots that call for action and identify and correct the causes for bad quality. Further if points, fall below LCL, it reflects the slackness in the inspection authorities because of relaxed inspection standards.

4. The *p*-chart locates the source of the difficulty, though the preparation of it is less expensive. However the mean chart ($\overline{X}$) and range chart (R), though expensive finds the cause.
5. The p-chart gives a good basis for judgement whether successive lots may be considered as a representative of a process. Knowledge of the fraction rejects gives the administration with information on whether or not to release an order, before shipment to customer.

Example 3

12 samples of 200 bulbs each were examined in a fortnights production. The number of defective bulbs in each sample was recorded as below.

Sample No.:	1	2	3	4	5	6	7	8	9	10	11	12
No. of defectives:	23	32	40	30	43	27	28	24	10	12	9	10

i) Draw the control chart for faction defective.

ii) What do you find out from the chart?

Solution

(i)

Sample No.	1	2	3	4	5	6	7	8	9	10	11	12
No. of defectives	23	32	40	30	43	27	28	24	10	12	9	10
Fraction defective	.115	.160	.200	150	.215	.135	.140	.120	.050	.060	.045	.050

$$p_i = d_i \div 200$$

$$\text{C.L.:} = \bar{p} = \frac{\sum p_i}{12} = \frac{1.440}{12} = 0.12$$

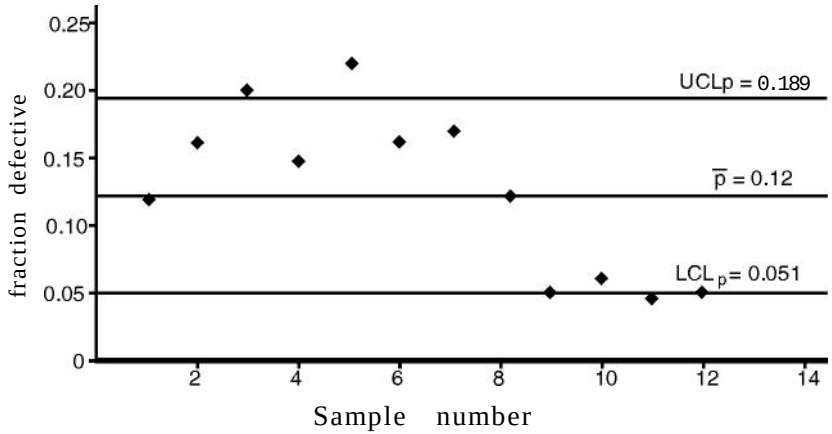
$$\text{U.C.L.} = \bar{p} + 3\sqrt{\bar{p}(1-\bar{p})/n}$$

$$= .12 + 3\sqrt{.12 \times .88/200} = 0.1889$$

$$\text{L.C.L.} = \bar{p} - 3\sqrt{\bar{p}(1-\bar{p})/n}$$

$$= 0.12 - 3\sqrt{0.12 \times 0.88/200} = 0.0511$$

p-Chart



(ii) Many points are found to lie outside the control limits. Hence, the production process of bulb is not under statistical control.

Example 4

Ten pieces of cloth out of different rolls of equal length contained the following number of defects 1, 3, 5, 0, 6, 0, 9, 4, 4, 3. Draw a control chart for the number of defects and state whether the process is in a state of statistical control.

Solution

1, 3, 5, 0, 6, 0, 9, 4, 4, 3 are denoted by c_i

Their mean $\bar{c}$ = 35/10 = 3.5 Hence,

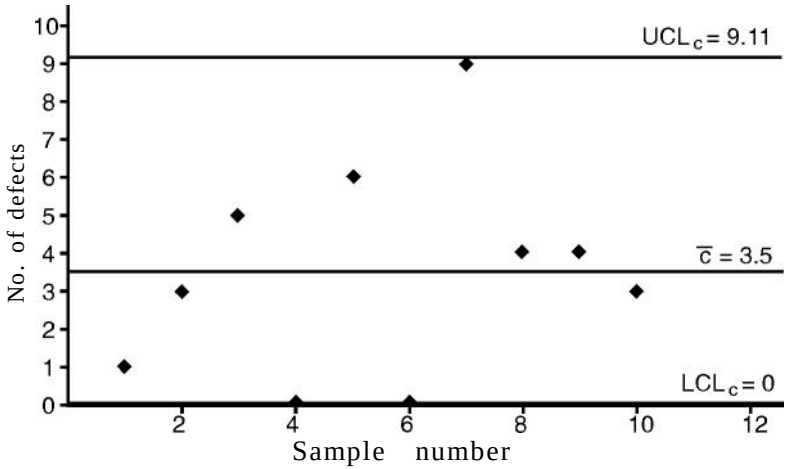
$$\text{C.L.} = \bar{c} = 3.50$$

$$\text{U.C.L.} = \bar{c} + 3\sqrt{c} = 3.50 + 3\sqrt{3.50} = 9.11$$

$$\text{L.C.L.} = \bar{c} - 3\sqrt{c} = 3.50 - 3\sqrt{3.50} = -2.11$$

It is taken as 0.

***c* -Chart**



All the points lie within the control limits in the above control chart for the number of defects. Hence, the process is in statistical control.

Example 5

A sample of 100 items was examined each hour from a production process. The numbers of defectives so found on a day are reproduced below:

16, 18, 12, 4, 10, 15, 13, 6, 7, 12, 10, 10,
2, 3, 13, 4, 1, 6, 5, 8, 4, 2, 5, 6

Draw the control chart for number of defectives and comment on the state of control of the process.

Solution

Here we have to draw an np-chart or d-chart. The given numbers of defectives are the values of d_i

$$\text{So} \quad \bar{d} = \frac{192}{24} = 8, \qquad \bar{p} = \bar{d}/n = 8/100 = 0.08$$

$$\text{C.L.} = n\bar{p} = 100 \times 0.08 = 8$$

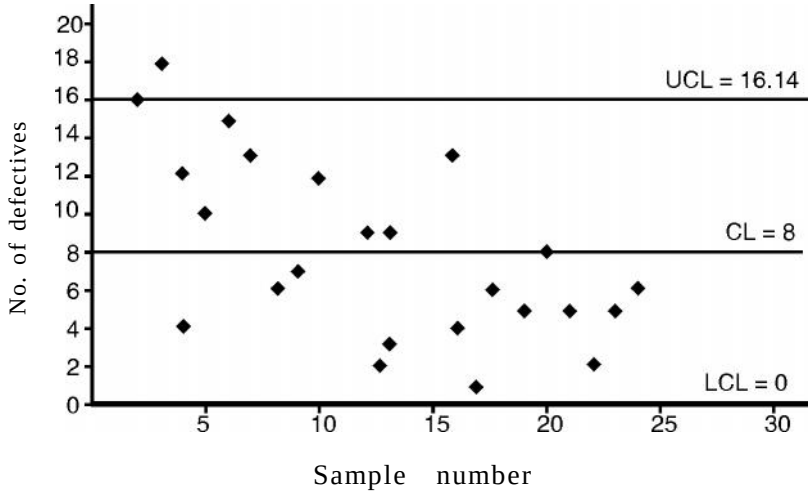
$$\text{U.C.L} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$= 18 + 3\sqrt{8(1-0.08)} = 16.1388$$

$$\text{LCL} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 8 - 3\sqrt{8(1-0.08)} = -0.1388$$

It is taken as 0.

***d* -Chart**



Second point lies outside the control limits. Hence, the production process is not under statistical control.

Defectives and Defects

An item on inspection may be found to be defective or non-defective. Further, when it is defective, the number of defects in it can be found out. It is obvious that the number of defects in each non-defective item is zero. For example, consider car tyres. First tyre inspected may be non-defective. Second tyre may have 2 defects and so, is defective. Third one may be defective with 1 defect, fourth with no defect. In such a case, the numbers of defects per unit are 0, 2, 1, 0 ,...,. p and d charts are drawn on the basis of data on defectives. c-Chart is drawn for defects per unit.

The following table may be used as a ready reckoner for drawing the control charts when the standards are not given.

Type of chart	C.L.	U.C.L.	L.C.L.
$\bar{X}$	$\bar{\bar{x}}$	$\bar{\bar{x}} + A_2\bar{R}$	$\bar{\bar{x}} - A_2\bar{R}$
R	$\bar{R}$	$D_4\bar{R}$	$D_3\bar{R}$
p	$\bar{p}$	$\bar{p} + A\sqrt{\bar{p}(1-\bar{p})}$	$\bar{p} - A\sqrt{\bar{p}(1-\bar{p})}$
d	$n\bar{p}$	$n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$	$n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$
c	$\bar{c}$	$\bar{c} + 3\sqrt{\bar{c}}$	$\bar{c} - 3\sqrt{\bar{c}}$

Uses of Statistical Quality Control

By considering the aspects discussed above we can summarise the advantages or uses of statistical quality control as given below:

1. An objective check is maintained on the quality of the product through statistical quality control.
2. Through SQC, we can know that whether the manufacturing process is under control or not. If it is not under control, remedial measures can be expiadated which saves materials and also ensures quality products.
3. SQC has a healthy influence on workers.
4. SQC enhances the goodwill of the producer if he is adopting an efficient and strict SQC system.
5. The quality of the product can be guaranteed before any governmental agency, on the basis of SQC records.
6. As a by-product of SQC, good deal of statistical data are made available.

EXERCISES

Very Short Answer Questions

1. What is statistical quality control?
2. Define specification limits.
3. What are chance causes?
4. What are assignable causes?
5. Define natural tolerance limits.
6. Define process control.
7. Define product control.
8. What do you mean by acceptance sampling?
9. Distinguish between variables and attributes.
10. Sketch the model of a control chart.

Short Essay Questions

11. Distinguish between random variation and assignable variation.
12. What are the objectives of statistical quality control?
13. What are the uses of statistical quality control?
14. What are process control and product control?
15. How do you construct control charts?
16. Distinguish between control chart for variables and control chart for attributes.
17. Explain how you will construct (i) $\bar{X}$ chart (ii) R chart (iii) P-chart (iv) C-chart.
18. Explain how the $\bar{X}$ and R charts are used to control the quality of products in the industry.
19. Differentiate between p-chart and c-chart in the context of statistical quality control.
20. Distinguish between the control chart for number of defectives and that for number of defects per unit.
21. Distinguish between (a) defect and defective (b) chance causes and assignable causes of variation.

Long Essay Questions

22. You are given the values of sample means ($\bar{X}$) and the ranges (R) for ten samples of size 5 each. Draw mean and range charts and comment on the state of control of the process

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean :	43	49	37	44	45	37	51	46	43	47
Range :	5	6	5	7	7	4	8	6	4	6

You may use the following control chart constants.

For n = 5: $A_2 = 0.58$, $D_3 = 0$, $D_4 = 2.115$

23. Construct a control chart for mean and range for the samples of 5 being taken every hour.

Sample No.	Sample Values				
1	42	65	75	78	87
2	42	45	68	72	90
3	19	24	80	81	81
4	16	54	69	77	84
5	42	51	57	59	78
6	51	74	75	78	60

24. During an examination of equal lengths of cloths, the following number of defects were observed. 2, 3, 4, 0, 5, 6, 7, 4, 3, 2. Draw a control chart for the number of defects and comment whether the process is under control.

25. The following is the number of defective items observed in 15 consecutive samples of size 50 each

12, 9, 15, 14, 10, 8, 6, 12, 9, 5, 12, 10, 11, 9, 10

Draw the control chart for fraction defective and comment upon the state of control of the manufacturing process.

MULTIPLE CHOICE QUESTIONS

1. Statistical population may consists of
   (a) an infinite number of items
   (b) a finite number of items
   (c) either of (a) or (b) (d) none of (a) and (b)
2. A study based on complete enumeration is known as
   (a) sample survey (b) pilot survey
   (c) census survey (d) none of the above
3. Sampling frame refers to
   (a) the number of items in the sample
   (b) the number of items in the population
   (c) listing of all items in the population
   (d) listing of all items in the sample.
4. Startified random sampling is used when
   (a) population is heterogeneous w.r.t some characteristics
   (b) When population is homogeneous
   (c) When population is finite
   (d) None of these
5. Simple random sample can be drawn with the help of
   (a) random number tables (b) chit method
   (c) roulette wheel (d) all the above
6. Random sampling is also termed as
   (a) probability sampling (b) chance sampling
   (c) both (a) and (b) (d) None of these
7. The difference between statistic and parameter is called
   (a) Non sampling error (b) standard error
   (c) Sampling error (d) None of these
8. Questionnaires and schedule are
   (a) Same in its kind and degree
   (b) Same in its degree and vary in its kind

(c) Vary in its kind and degree

(d) Same in its kind and vary in its degree.

9. Equality of several normal population means can be tested by

(a) Bartlett's test (b) F-test

(c) χ^2 test (d) t-test

10. Analysis of variance utilizes

(a) F-test (b) χ^2 test

(c) Z-test (d) t-test

11. The error degrees of freedom for two-way ANOVA with r rows and c columns is

(a) $r - 1$ (b) $c - 1$

(c) $(r - 1)(c - 1)$ (d) $rc - 1$

12. The analysis of variance technique is based on the assumption

(a) Populations from which the samples have been drawn are normal.

(b) The populations have the same variance

(c) The random errors are normally distributed

(d) All the above

13. The technique of Analysis of variance was first devised by

(a) Karl Pearson (b) R.A. Fisher

(c) Irwing Fisher (d) W.Z. Gosset

14. Index number is a:

a) measure of relative changes

b) a special type of an average

c) a percentage relative d) all the above

15. Index numbers are expressed:

a) in percentages b) in ratios

c) in terms of absolute value d) all the above

16. Index numbers help:

a) in framing of economic policies

b) in assessing the purchasing power of money

c) for adjusting national income

d) all the above

17. Index numbers are also known as:

a) economic barometers b) signs and guide posts

c) both (a) and (b) d) neither (a) nor (b)

18. The first and fore most step in the construction of index numbers is:

a) choice of base period b) choice of weights

c) to delineate the purpose of index numbers

d) all the above

19. Base period for an index number should be:

(a) a year only b) a normal period

c) a period at distant past d) none of the above

20. Laspeyre's index formula uses the weight of the:

a) base year b) current year

c) average of the weights of a number of years

d) none of the above

21. The weights used in Paasche's formula belong to:

a) the base period b) the given period

c) to any arbitrary chosen period

d) none of the above

22. The geometric mean of Laspeyre's and Paasche's price indices is also known as:

a) Fisher's price index b) Kelly's price index

c) Drobish-Bowley price index d) Walsh price index

23. If the index number is independent of the units of measurements, then it satisfies:
    a) times reversal test b) factor reversal test
    c) unit test d) all the above

24. Variation in the items produced in a factory may be due to:
    (a) chance factors (b) assignable causes
    (c) both (a) and (b) (d) none of the above

25. Chance or random variation in the manufactured product is
    (a) controlable (b) not controlable
    (c) both (a) and (b) (d) none of the above

26. Chance variation in respect of quality control of a product is
    (a) tolerable (b) not effecting the quality of a product
    (c) uncontrolable (d) all the above

27. The causes leading to vast variation in the specifications of a product are usually due to:
    (a) random process (b) assignable causes
    (c) non-traceable causes (d) all the above

28. Variation due to assignable causes in the product occurs due to:
    (a) faulty process (b) carelessness of operators
    (c) poor quality of raw material (d) all the above

29. The faults due to assignable causes:
    (a) can be removed (b) cannot be removed
    (c) can sometimes be removed (d) all the above

30. Control charts in statistical quality control are meant for:
    (a) describing the pattern of variation
    (b) checking whether the variability in the product is within the tolerance limits or not
    (c) uncovering whether the variability in the product is due to assignable causes or not
    (d) all the above

31. Control charts consist of:
    (a) three control lines
    (b) upper and lower control limits
    (c) the level of the process
    (d) all the above

32. Main tools of statistical quality control are:
    (a) shewhart charts (b) acceptance sampling plans
    (c) both (a) and (b) (d) none of the above

33. The relation between expected value of R and S.D. $\pm$ with usual constant factors is.
    (a) E(R) = $d_1\pm$ (b) E (R) = $d_2\pm$
    (c) E(R) = $D_1\pm$ (d) E (R) = $D_2\pm$

34. The control limits for R-chart with a known specified range R' and usual constant factors are:
    (a) $U.C.L._{R'} = (d_2 - 3d_3)\pm_{R'}$, $C.L._{R'} = d_1\pm_{R'}$ and
    $L.C.L._{R'} = (d_2 - 3d_3)\pm_{R'}$
    (b) $U.C.L._{R'} = D_2 - R'$, $C.L._{R'} = d_2\pm_{R'}$ and
    $L.C.L._{R'} = D_1 - \pm_{R'}$
    (c) either (a) or (b) (d) neither (a) nor (b)

35. When the value of the population range R is not known, then for $\overline{X}$-chart, the trial control limits with usual constant factors are:
    (a) $U.C.L. = \overline{X} + A_3\overline{R}$, $C.L. = \overline{X}$ and $L.C.L. = \overline{X} - A_2\overline{R}$
    (b) $U.C.L. = \overline{X} + A_3\overline{R}$, $C.L. = \overline{X}$ and $L.C.L. = \overline{X} - A_2\overline{R}$
    (c) $U.C.L. = A_3\overline{R}$, $C.L. = \overline{X}$ and $L.C.L. = A_2\overline{R}$
    (d) $U.C.L. = \overline{X} + A_3\overline{R}$, $C.L. = \overline{X}$ and $L.C.L. = \overline{X} - A_3\overline{R}$

36. The trial control limits for R-chart with usual constant factors are:

(a) U.C.L.= D_4R, C.L.= R and L.C.L. = D_3R

(b) U.C.L.= $D_4\overline{R}$, C.L.= $\overline{R}$ and L.C.L. = $D_3\overline{R}$

(c) U.C.L.= $D_4\overline{R}$, C.L.= $\overline{R}$ and L.C.L. = $D_4\overline{R}$

(d) all the above

37. 3-sigma trial control limits with p' as mean number of defectives based on a sample of size n are:

(a) $U.C.L. = n\overline{p} + \sqrt{n\overline{p}(1-\overline{p})}, C.L. = \overline{p}$

and $L.C.L. = n\overline{p} + \sqrt{n\overline{p}(1-\overline{p})}$

(b) $U.C.L. = n\overline{p} + 3\sqrt{n\overline{p}(1-\overline{p})}, C.L. = n\overline{p}$

and $L.C.L. = n\overline{p} - 3\sqrt{n\overline{p}(1-\overline{p})}$

(c) $U.C.L. = n\overline{p} + 3\sqrt{n\overline{p}(1-\overline{p})}, C.L. = \overline{p}$

and $L.C.L. = \overline{p} - 3\sqrt{n\overline{p}(1-\overline{p})}$

(d) none of the above.

SYLLABUS

SEMESTER IV

Complimentary Course iv

APPLIED STATISTICS

Module I: Census and Sampling, Principal steps in a sample survey, different types of sampling, Organisation and execution of large scale sample surveys, errors in sampling (Sampling and nonsampling errors) preparation of questionnaire, simple random sampling with and without replacement, Systematic stratified and cluster sampling (concept only)

20 hours

Module II: Analysis of variance; one way, two way classifications. Null hypothesis, total, between and within sum of squares. Assumptions-ANOVA table.

15 hours

Module III: Time series Components of time series-additive and multiplicative models, measurement of trend, moving averages, seasonal indices-simple average-ratio to moving average. Index numbers: meaning and definition-uses and types-problems in the construction of index numbers - different types of simple and weighted index numbers. Test for an ideal index number- time and factor reversal test.

30 hours

Module IV: Statistical Quality Control: Concept of statistical quality control, assignable causes and chance causes, process control. Construction of control charts, 3sigma limits. Control chart for variables-Mean chart and Range chart. Control chart for attributes- p chart, d or np chart and c chart **25 hours**