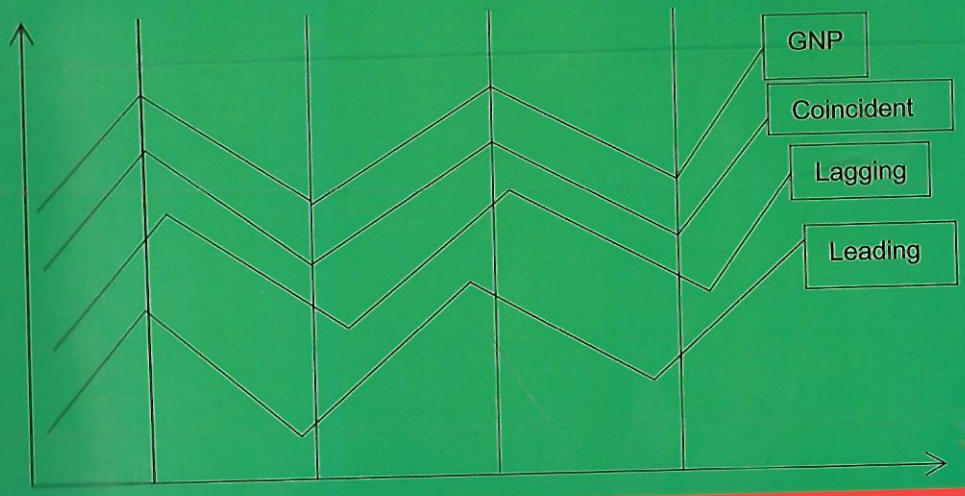
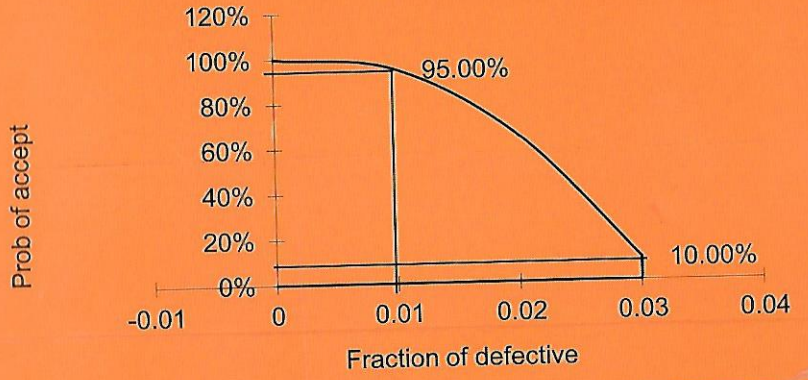


Applied Statistics for Business and Economic Analysis

Venkatesh SESHAMANI
Oliver KAONGA



9 789982 702089

Table of Contents

List of tables

List of figures

Abbreviations

GNP	Gross National Product
SQC	Statistical Quality Control
LCMS	Living Conditions Monitoring Survey
ZHDS	Zambia Demographic and Health Survey
LFS	Labour force Survey
OC	Operating Characteristic curve
LCL	Lower Control Limit
UCL	Upper Control Limit
AQL	Acceptable Quality Level
LTPD	Tolerance Percent Defective
CL	Centre Line
SEA	Standard Enumeration Area
FHH	Female Headed Household
MHH	Male Headed Household
CSO	Central Statistical Office
NSS	National Statistical System
ZCPH	Zambia Census of Population and Housing

Preface

This book is a sequel to the book "Statistical Methods for Economic Analysis" by Venkatesh Seshamani and Obrian Ndhlovu. It deals with some specialized topics such as Sampling, Time Series, Forecasting, Index Numbers, Statistical Quality Control and Zambian Statistics.

The two books together adequately cover the content of the undergraduate course in Statistics for those majoring in economics. They will also be useful to economic policy makers and practitioners who would like to familiarize themselves with the basic statistical tools available for economic analysis.

The statistical theories and techniques have been profusely illustrated with examples relating to Zambian and African contexts that will hopefully enhance the understanding and appreciation of the users.

We thank all those who encouraged and supported our efforts in preparing this book, notably the HSS Dean Dr. Felix Masiye and the Economics Department Head Dr. Chrispin Mphuka.

Authors

CHAPTER ONE

1 Sampling techniques

1.1 Samples and sampling

A sample is a finite part of a statistical population whose properties are studied to gain information about the whole population. Sampling is, therefore, a statistical method of obtaining representative data or information from a population. Sampling is used when a census i.e. collecting data from every unit or person in a population, is cost-prohibitive.

1.2 Characteristics of a good sample design

Due to time and cost involved in field study, often, only a section of the population is studied. A sample design is a definite plan for obtaining a sample from a population; it is a technique or procedure for obtaining a sample from the target population.

A good sample design must fulfil the following characteristics:

1. A Sample design must be truly representative:

In research, a relatively small number of study units is selected, it is therefore pertinent that the selected sample closely match all the characteristics of the entire population. The representativeness of the sample is key because then the findings from an experiment on a representative sample can be generalised to the large universe being studied.

2. Sample design should have a small sampling error:

Sampling error is the error caused by taking a small number of units instead of the whole population for study. In other words, sampling error is the discrepancy that may come about as a result of judging all population units on the basis of a sample. We can reduce error by ensuring that large enough samples are selected. Additionally, one has to employ efficient sample design and estimation strategies.

3. Sample design should be viable in the context of available funds:

Obviously, a researcher would only embark on a field study with a view to complete it. This may not be possible if the study cannot be accommodated within a limited research budget. The sampling should, therefore, be done in such a way that it is within the research budget. It shouldn't be too expensive to be replicated.

4. Sample design should have good control over systematic bias:

Systematic bias results from errors in the sampling procedures which cannot be reduced or eliminated by increasing the sample size. For example, a survey of university students to measure voter apathy among youths will be a biased sample because it does not include the youths who are not enrolled in the university. The best bet for researchers is to detect the causes and correct them at the design stage.

5. Results of study must be applicable to the population with reasonable level of confidence:

This implies that the sampling design should be created bearing in mind that the interest is not only to get results about the sample but about the whole universe of the study.

1.3 Restricted Versus Unrestricted Sampling

Sampling techniques can be classified into two broad categories, *unrestricted sampling* and *restricted sampling*. In unrestricted sampling each and every sampling unit has an equal chance of being selected. Restricted sampling is applied when the population has heterogeneous sampling units, the population is first categorised into homogenous groups and samples are drawn independently from each group.

Table 0.1 Types of sampling

Element selection technique	**Representation bias**	
	Probability sampling	Non probability sampling
Unrestricted sampling	Simple random sampling	Haphazard or convenience sampling
Restricted sampling	Complex random sampling; cluster sampling, systematic, stratified, multistage etc.	Purposive sampling, quota sampling

1.4 Probability sampling or random sampling

1.1.1 Simple random sampling

It is a sampling procedure which:

- Gives each element of a population an equal probability of getting into the sample, and all choices are independent of each other.
- Gives each possible sample combination an equal probability of being chosen. For example, suppose a population has 6 elements a, b, c, d, e, f which implies N=6, and we want to select a sample of 3 elements, n = 3.

  Then there are ${}^{6}C_{3} = \frac{6!}{3!3!} = 20$ sample combinations (abc, abd, abe, abf, bcd, bcf, acd, ace, acf, ade, adf, aef, bec, bde, bdf, bef, cde, cdf, cef, def).

If we choose any of these samples, each has a probability of 1/20 of being chosen.

A practical way of selecting a simple random sample is by using a random number table. There are several such tables such as the ones provided by Tippet, Yates and Fisher. Table 1.2 shows the first 30 sets of Tippett's numbers (are in 4 digits).

Table 0.2 Tippett's random numbers

2952	6698	3992	9792	7979	5911
3170	5624	4167	9525	1545	1396
7203	5356	1300	2693	2370	7483
3408	2769	3563	6107	6913	7691
0560	5246	1112	9025	6008	8129

1.1.2 Stratified sampling

Population allocation using simple random sampling is okay only if strata differ only in size. But if they also differ in variability, then one must take larger samples from more variable data and smaller samples from less variable strata; that is, one must use *disproportionate sampling design.*

Suppose there are K strata with N= $N_1+N_2+N_3+\ldots\ldots N_K$, and we want to select a sample n = $n_1+n_2+n_3+\ldots\ldots n_K$.

Then we require $\frac{n_1}{N_1\sigma_1}=\frac{n_2}{N_2\sigma_2}=\ldots\ldots\ldots\ldots\frac{n_K}{N_K\sigma_K}$, where

$\sigma_1,\sigma_2,\sigma_3\ldots\ldots\ldots\ldots\sigma_K$ is the size of the K respective strata.

Allocation will be done by the following formula:

$$n_i=\frac{n.N_i\sigma_i}{N_1\sigma_1+N_2\sigma_2+\ldots\ldots+N_K\sigma_K}$$

For example, for a population with 3 strata: N_1 =5000, N_2 =2000, N_3 =3000; and σ_1 =15, σ_2 =18, σ_3 =5. We want a sample of size n=84.

For stratum which has got N_1 =5000,

$$n_1=\frac{84.(5000)(15)}{(5000)(15)+(2000)(18)+(3000)(5)}$$

$$=\frac{6300000}{126000}=50$$

For stratum N_2 =2000,

$$n_2=\frac{84.(2000)(18)}{(5000)(15)+(2000)(18)+(3000)(5)}$$

$$=\frac{3024000}{126000}=24$$

For stratum N_3 =3000,

$$n_3=\frac{84.(3000)(5)}{(5000)(15)+(2000)(18)+(3000)(5)}$$

$$=\frac{1260000}{126000}=10$$

In addition to differences in stratum size, and stratum variability, we may have difference in stratum sampling cost. The formula then becomes

$$n_i=\frac{n\,.\,N_i\,Q_i/\sqrt{C_i}}{N_1\,Q_1/\sqrt{C_K}+N_2\,Q_2/\sqrt{C_K}\ldots\ldots\ldots\ldots\ldots+N_K\,Q_K/\sqrt{C_K}}$$

Where c_i =cost of sampling a unit in stratum i.

Table 0.3 Stratified Sampling

Marketing legumes	Population frequency	Population % distribution	Cost per unit(K)	Variability σ	$\sigma/\sqrt{c}$	Sample size optimizing cost	Sample size optimising variable	Sample size optimising cost & variable
District 1	18000	33 %	18	4.3	1.014	300	190	203
District 2	600	1 %	10	6.4	2.024	358	282	405
District 3	12000	22 %	39	9.4	1.505	138	415	302
District 4	24000	44 %	24	7.1	1.449	224	313	290
Total	**54600**	**100 %**				**1200**	**1200**	**1200**

Stratified sampling is most effective when 3 conditions are met:

1. Variability within strata are minimised
2. Variability between strata are maximised
3. Variables stratified are strongly correlated with the desired dependent variable

What then are the advantages and disadvantages of stratified sampling? We present them now.

Advantages:

1. It focusses on important subpopulations and ignores irrelevant ones. In this way, we can have more precise information inside the subpopulations about the variables we are studying.
2. It Improves accuracy and efficiency of estimation.

Disadvantages:

1. Selection of relevant stratification variables maybe difficult.
2. It is not useful when there are no homogenous groups.
3. Can be expensive.

Generally, stratified sampling provides better results than simple random sampling when the strata are more different among themselves and more homogeneous internally.

1.1.3 Cluster sampling

In cluster sampling, the population is divided into groups or clusters which are relatively homogenous. A random sample of clusters is then selected. In single-stage cluster sampling, all the elements in each of the selected clusters are used. In two-stage cluster sampling, random sampling is applied to the elements from each of the selected clusters.

Illustration of two-stage cluster sampling: The following table shows the selection of a sample of Zambian households from the national population of households in two stages:

- First stage: 1000 Standard Enumeration Areas (SEAs) are selected with probability proportional to size within the respective strata.
- Second stage: From each rural and urban SEA, 15 and 25 Households are selected respectively.

Table 0.4 First stage cluster sampling results

Province	Rural SEA		Urban SEA		Total	
	2006	2010	2006	2010	2006	2010
Central	56	55	30	41	86	96
Copper belt	44	48	100	114	144	162
Eastern	98	76	24	24	122	100
Luapula	64	54	22	22	86	76
Lusaka	28	32	78	84	106	116
Northern	106	97	38	47	144	144
North western	60	62	24	28	84	90
Southern	100	93	44	53	144	146
Western	62	48	22	22	84	70
Zambia	618	565	382	435	1000	1000

Source: Zambia Living Conditions Monitoring Surveys 2006 &2010

1.5 Non probability sampling: Disproportionate sampling

Suppose a researcher wants to study living conditions in Male Headed Households (MHH) and Female Headed Households (FHH) in a constituency in which there are 900 MHH and 100 FHH. He wants a sample of 100 HHs. Then according to proportional sampling, he must select 90 MHH and 10 FHH. But this small number of FHH may not provide adequate representation for drawing conclusions. Disproportionate sample will allow the researcher to select adequate size from the two strata. This can be done by;

1. Choosing equal number (50) of MHH and FHH or
2. A higher proportion of FHH than in proportional sample, for instance 65 MHH and 35 FHH.

This is not random sampling since MHHs have better chances of being selected in the sample.

1.1.4 Sequential sampling

Sequential sampling is a non-probability sampling technique wherein the researcher picks a single subject or a group of subjects in a given time interval, conducts the study, analyses the results, then picks another subject or group of subjects if needed, and so on.

This technique gives the researcher limitless chances of fine tuning the research methods and gain vital insights into the subject of the study.

Illustration 1:
Consider the following scenario: Accept or reject a lot with a pre-set rule;

$$X_a = -h_1 + sn \;\; ; \;\; X_r = h_2 + sn$$

You sample a lot,

$$X_a = -1 + 2n \;\; ; \;\; X_r = 2 + 2n$$

- Reject if you get 2+2 = 4 defectives or more
- Accept if you get -1+2 =1 defectives or less.

Figure 0.1 Rejection and Acceptance region

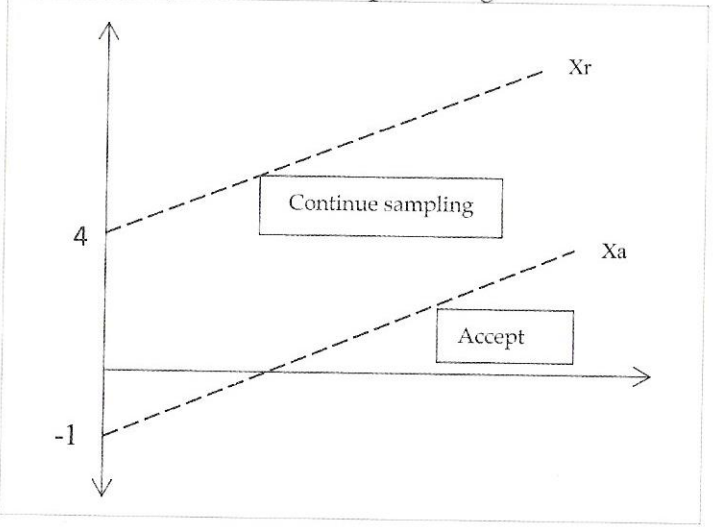


Table 0.5 Sequential sampling

0	0✓	0	0✓	0	0×	0	0	0	0	0	0	0×	0	0	0	0	0	0	0
0✓	0	0	0	0	0✓	0✓	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0✓	0	0	0	0	0×	0	0✓	0	0	0	0	0	0	0×	0	0	0
0	0	0	0✓	0	0	0	0	0×	0	0	0✓	0	0	0	0	0	0	0	0×
0	0	0	0	0✓	0	0	0×	0	0	0	0	0	0	0	0×	0	0	0	0

✓= 1st sample ×= 2nd sample

Illustration 2: Bootstrap samples

Suppose we want to test the hypothesis (H) that 90 % of FHH in Shang'ombo district are extremely poor and suppose a pre-set rule is as follows:

If probability is 1% or less that the statement is false, then accept H and if probability is greater than 1%, reject H.

But suppose on the basis of the selected sample, we do not get enough evidence to accept or reject H. Then the way out is to test H on the basis of another sample. But selecting another sample may be a very costly exercise and one may also not have the time to go in the field to collect the sample. For example, consider table 1.5. Suppose the initial sample whose units are indicated by ✓ does not yield any definite result. Then one may have to select another sample such as the one indicated by ×. But doing this by going back to the field may be costly and time consuming.

To overcome this constraint, an ingenious method called the *bootstrap method* has been devised. In this method, you draw repeated samples from the initial sample itself with replacement. The initial sample acts as a surrogate population. We then create a large number of "phantom samples" known as bootstrap samples. One produces a large number of copies of a sample statistic computed from the phantom bootstrap samples. You then get a bootstrap distribution of the statistic. Since a large number of samples are required, they are generated using a computer whose cost is quite low.

Bootstrap samples: A mini example

Take an initial sample of numbers 2, 4, 5, 6, 6. The following are possible bootstrap samples: 2, 5, 5, 6, 6; 4, 5, 6, 6, 6; 2, 2, 4, 5, 5; 2, 2, 2, 4, 6; 2, 2, 2, 2, 2; 4, 6, 6, 6, 6.

The moot question arising in this method is: Is it really possible to improve the sample estimate of the population parameter by reusing the same sample again and again? And the answer is yes. Bootstrapping can in fact accomplish this. We shall not go into the proof of this statement here.

1.1.5 Haphazard or Convenience Sampling

This is one of the most common methods under non-probability sampling. A sample of convenience refers to elements that have been selected from the target population on the basis of their accessibility or convenience to the researcher. For example, a public opinion poll conducted on radio to get a quick response on the most popular presidential candidate. Such a sample is non-representative. Convenience sampling assumes that research units in the population are homogeneous, implying that there would be no differences in the samples regardless of the method used to obtain it.

1.1.6 Purposive sampling

In purposive or judgemental sampling, the researcher samples with a 'purpose' in mind; to select research units which represent a 'typical sample'. This approach is used when a sample is taken based on certain judgements about the overall population. The underlying assumption is that the investigator will select units that are characteristic of the population. Often, a specific predefined group is targeted. For example, in market research, a researcher would be positioned near the entrance of Shoprite stores looking for females aged 20-25 years old. The critical issue here is objectivity in arriving at a typical sample; no two researchers will agree upon the exact composition of a typical sample. Purposive sampling can be very useful for situations where you need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern.

1.1.7 Quota Sampling

This type of sampling involves selecting research units non-randomly according to some fixed quota. Two types of quota

sampling can be identified: *proportional and non-proportional*. In proportional quota sampling, the objective is to have the major characteristics of the population represented. This is achieved by sampling a proportional amount from each quota. For instance, suppose you have a population which has 70% women and 30% men. To obtain a sample of 100, you will continue sampling until you get a sample with proportions of men and women matching those in the population. *Proportional Quota sampling* is often erroneously referred to as 'representative sampling' because numbers of elements are drawn from various target population strata in proportion to the size of these strata.

Non-proportional quota sampling is a bit more flexible. In this technique, you specify the minimum number of sampled units you want in each category. Having numbers that match the proportions in the population is not of primary concern. This method is the non-probabilistic analogue of stratified random sampling in that it is typically used to ensure that smaller groups are adequately represented in the sample.

CHAPTER TWO

2 Elementary analysis of time series

2.1 Meaning and rationale of time series

A time series is a set of magnitudes of a variable or an index of variables obtained over time, usually at regular intervals such as every month, every year and so on. The variables can be income, agricultural production, Gross National Product, price of a commodity, general price level, etc. The fluctuations or movements in the values of such variables over time are caused both by systematic and random factors operating in the business and economic environment. Mathematically, then, a time series may be defined as a function $X = f(t)$ where X is the variable whose magnitudes are observed at specific times and t stands for time. It is clear that in a time series function, while X may vary with the context of study, the independent variable is always the same, namely t or time.

The rationale for dealing with time series lies in its value as a base for statistical forecasting. If a time series of past observations of a variable displays a persistent_pattern, it may be reasonable to assume, ceteris paribus, that the pattern or trend may continue at least for some time in the future. In other words, the future course of the variable, to a significant degree, may be regarded as an extension of the past pattern and if it is possible to decipher this pattern, it will also be possible to predict, on its basis, future likely values of the variable.

2.2 Composition and decomposition of time series

The characteristic movements of a time series are affected by a host of factors which may be economic, political, social, institutional and natural. Some of these factors may influence the

movement only over a long period of time while others may affect the short-run movements.

Some factors like an earthquake or war may occur briefly and sporadically but their impact may linger for a long time.

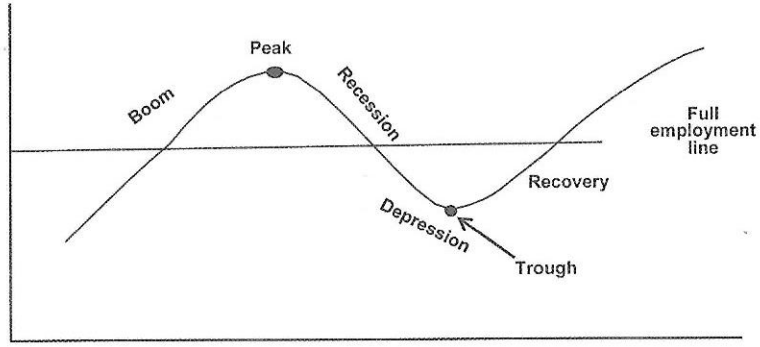
The patterns in a time series are yielded by four types of movements of which the time series is composed. They are:

1. Secular or long term trend
2. Cyclical movements
3. Seasonal variations
4. Random or irregular fluctuations

The secular trend refers to the general direction in which the variable in question appears to be moving over a long period of time. The actual time series graph may display a lot of upward and downward movements from period to period, but over a large sequence of periods, a steadily rising or declining trend may be perceptible.

Cyclical movements refer to the long term oscillations about the trend curve. The cycles need not be periodic, i.e. they need not reveal similar patterns after equal time intervals. The most popular example of cyclical fluctuations is the business cycle. A business cycle consists of various phases in business activity such as prosperity, recession, recovery, etc. A typical business cycle is portrayed in the figure 2.1.

Figure 2.1 Phases of a business cycle



A cycle, to be properly so-called, must be of at least a year's duration. In economics, one speaks of cycles of various durations like Kondratieff's cycles which are supposed to be of nearly 100 year's duration, Juggler's cycles of 11 year's duration, and so on. The nature of a few economic characteristics during the different phases of a business cycle is listed in table 2.1

Table 2.1 Economic characteristics

Characteristic	Prosperity	Recession	Depression	Recovery	Peak
Employment	High	Sudden	Low	Slow rise	High
Industrial Output	High	Decreasing	Low	Slow increase	High but bottlenecks
Prices	High	Rapid fall	Low	Slow rise	Very high
Cost of production	High	Falling	Low	Slow rise	Very high
Profits	High but falling	Disappear	Appear toward end	Rise	Begin to fall
Feeling	Optimism	Hesitation and fear	Pessimism and despair	Cautious hope	Over optimism

Seasonal variations refer to the patterns that manifest themselves at particular times during the corresponding periods. For example, every year, the sale of garments and sweets may be high during Christmas and New year; the sale of raincoats may go up during the monsoons; food-grain prices may fall immediately after the harvest season and may be very high just before harvest; and so on. Patterns may also be visible every month during certain weeks or days and every week during certain days. For example, every month, deposits in the banks may go up during the first week and withdrawals may be high during the last week. Or, every week, theatres may have heavy booking on Saturdays and Sundays. All these are examples of seasonal-type fluctuations.

While the above three components of a time series are caused by systematic forces, the last component, vis. irregular movements are caused by random factors like changes in weather, floods, earth-quakes, elections etc. These factors, being random, cannot, as a rule, be anticipated and hence movements caused by them cannot be systematically analysed. In a few instances, such as elections, however, variations can be anticipated and suitable policies may be adopted to compensate for them. For instance, suppose a firm gets its supply of a raw material from four suppliers who are supplying nearly equal quantities of the material and if during a certain period, one of the suppliers is not able to provide the material, the production chart of the firm will show a downward variation for that period. However, had the firm been able to anticipate this contingency, it could have ordered additional quantities in advance either from that supplier himself or others. Thereby production during that period would have been unaffected.

In the following figures 2.2 to 2.5, the four components of the time series are graphed.

Figure 2.2 Trend

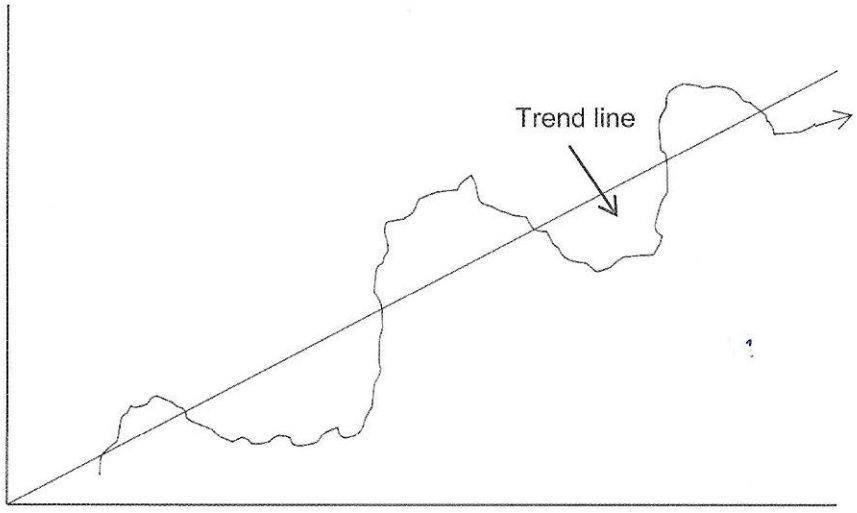


Figure 2.3 Cyclical fluctuations

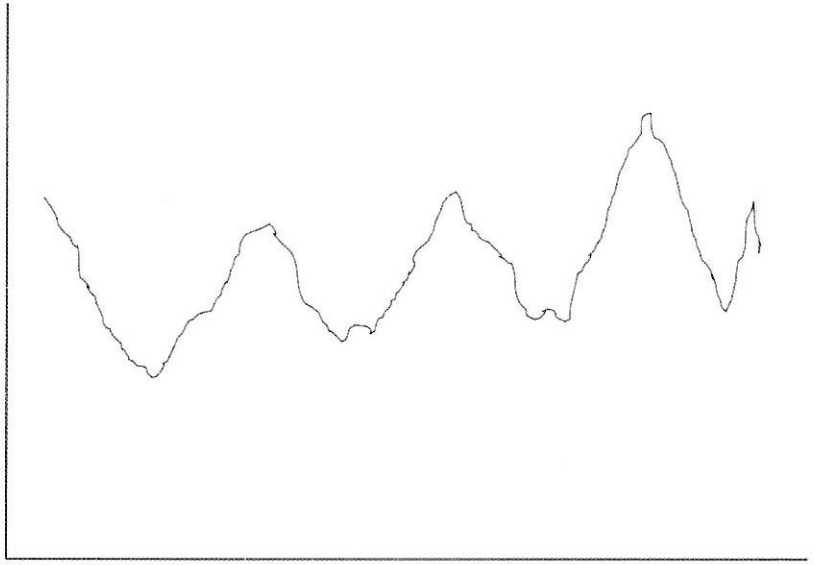


Figure 2.4 Cyclical fluctuations

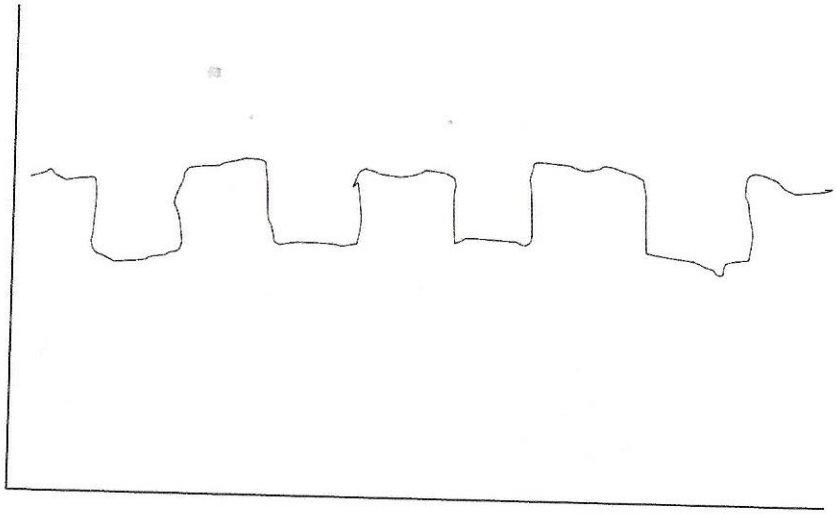
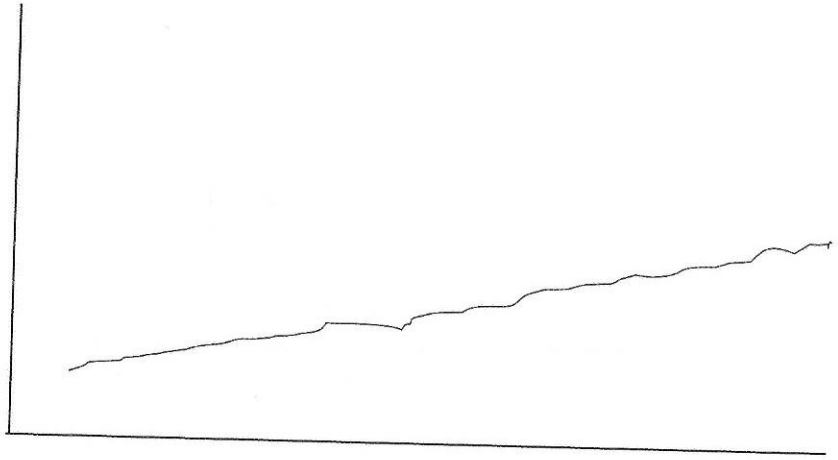


Figure 2.5 Irregular variations



The analysis of time series involves the decomposition of the series into its four major constituent parts to enable study of each

component separately. The study of each component has its own specific purposes. For instance, if a firm wishes to decide whether it should or not expand on its plant size, it should be guided by the secular trend in demand. If the long-term trend in demand is in the nature of a decline and the firm, misled by a temporary spurt in demand opts for a plant of a larger size, then over the long period, it will be faced with excess capacity. On the other hand, if there is an increasing trend in demand for the firm's product, but the relatively low current demand makes the firm pessimistic and makes it decide not to expand plant capacity, then again, over a long period, it will lose its share of the market to the competitors.

Similarly, in planning inventories or scheduling production which is usually done on a monthly or quarterly basis, knowledge of seasonal fluctuations is relevant. Demand for durable consumption goods and producer goods are influenced by the cyclical factor and hence the specific study of cyclical factors is necessary.

There are two types of time series models:

1. The *additive* model which assumes the time series to be the sum of the four components which are independent of each other. Whatever be the magnitude of one component, it will not affect the magnitudes of the other three.
2. The *multiplicative* model which assumes the time series to be the product of the four components. While this model permits relationship between the four components, it assumes that the four components are due to different causes so that any one component can be isolated by dividing the time series by the other three components.

Using notation, let

Y= actual value of the time series;

T = value of the trend;

C = value of cycle;

S = value of seasonal factor;

I = value of irregular fluctuation.

The additive model states that

$$Y = T+C+S+I$$

The multiplicative model states that

$$Y = TCSI$$

2.3 Analysis of secular trends

There are several techniques for measuring the secular trend of a time series. We shall discuss three broad methods: (a) The method of moving averages; (b) the method of semi averages; (c) curve fitting method.

2.3.1 The method of moving average

For a set of time series data: X_1 X_2, X_3....., a moving average of order N is defined by the sequence of arithmetic means

$X_1 + X_2 + X_3 + \ldots\ldots + X_N$ $\quad$ $X_1 + X_2 + X_3 + \ldots\ldots + X_{N+1}$

$X_1 + X_2 + X_3 + \ldots\ldots + X_{N+2}$

The sums in the numerators are the *moving totals* of order N.

Consider the following data:

Table 2.2 Moving average-data

Year	Time series value	Year	Time series value
1	1	15	3
2	2	16	4
3	3	17	5
4	4	18	6
5	3	19	5
6	2	20	4
7	1	21	3
8	2	22	4
9	3	23	5
10	4	24	6
11	5	25	5
12	4	26	4
13	3	27	3
14	2	28	4

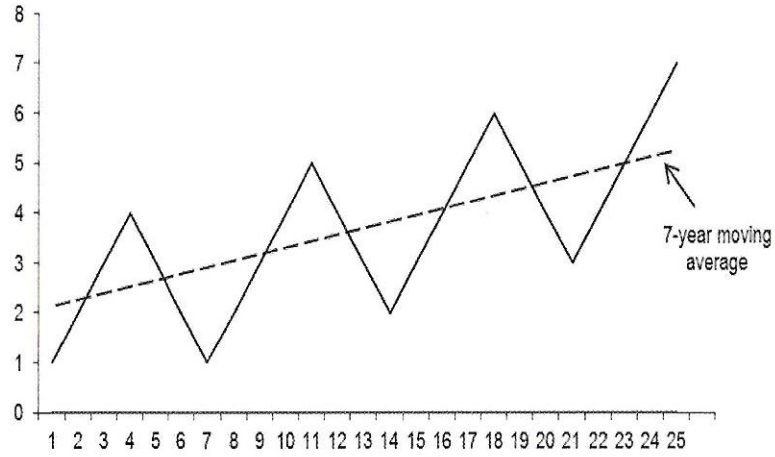
The above hypothetical data has 7-year cycles with uniform duration and amplitude. The results of calculating a moving average of order 5 and a moving average of order 7 are shown in the table 2.3.

Table 2.3 Computing the moving average

Year	Actual value	5 -Year Moving total	5 -Year Moving average	7-Year Moving total	7-Year Moving average
1	1	_	_	_	_
2	2	_	_	_	_
3	3	13	2.6	_	_
4	4	14	2.8	16	2.29
5	3	13	2.6	17	2.43
6	2	12	2.4	18	2.57
7	1	11	2.2	19	2.71
8	2	12	2.4	20	2.86
9	3	13	3	21	3.00
10	4	18	3.6	22	3.14
11	5	19	3.8	23	3.29
12	4	18	3.6	26	3.43
13	3	17	3.4	25	3.57
14	2	16	3.2	26	3.71
15	3	17	3.4	27	3.86
16	4	20	4	28	4.00
17	5	23	4.6	29	4.14
18	6	24	4.8	30	4.29
19	5	23	4.6	31	4.43
20	4	22	4.4	32	4.57
21	3	21	4.2	33	4.71
22	4	22	4.4	34	4.86
23	5	25	5.0	35	5.00
24	6	28	5.6	36	5.14
25	7	29	5.8	37	5.29
26	6	28	5.6	_	_
27	5	_	_	_	_
28	_	_	_	_	_

The actual data and the 7-year moving average are graphed in the figure 2.3.

Figure 2.6 7-year moving average



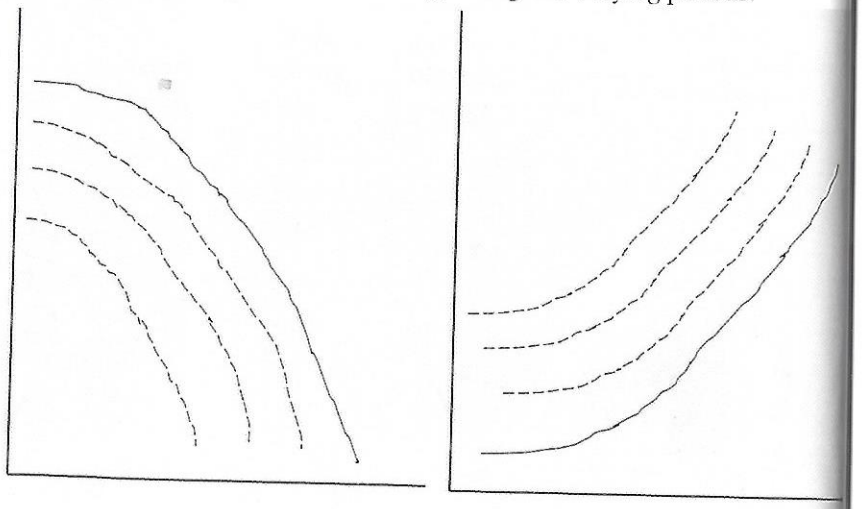
It is seen in the above figure that the moving average of order 7 has completely eliminated the periodic fluctuations and is linear.

The following observations can be made about the moving average:

I. It reproduces the linear trend. Longer is the period of the moving average, greater is the difference between the trend values and the actual data value. This can be seen by comparing the five-year moving average graph with the actual time series graph.

II. It reduces the curvature of the curvi-linear trend. Longer is the order of the moving average, greater is the difference between trend values and actual data. Look at the diagrams in figure 2.4 below.

Figure 2.7 Trends obtained from moving averages of varying periods.



III. If the moving average is taken of the order equal to the period of the cycle or its multiple, it will eliminate completely all cyclical fluctuations.

IV. It cannot completely eliminate irregular variations or cycles of varying durations.

V. It has a serious disadvantage of generating cycles or other movements which were not present in the original data. This is known as the *Slutsky-Yule effect.*

VI. Being nothing but a sequence of arithmetic means it has the drawback of being affected by extreme values. This problem can be overcome by using a weighted moving average with central items being given larger weights and extreme items smaller weights.

VII. Data at the beginning and at the end of a series are lost. Greater is the order of the moving average, greater is this disadvantage. In fact, if there are M values in the original data and moving average of order N is computed, there will be only M-N+1 trend values.

2.3.2 Method of the semi-averages

This method can be applied only where the trend is known to be linear or where the data can be broken up into parts in each of which the trend is linear. The procedure for obtaining the trend by this method is as follows:

1. Divide the data into two equal parts. If the series consists of an odd number of observations, the central observation may be omitted.
2. Find the average of the data in each part using arithmetic mean or median.
3. Plot the points corresponding to these averages and join them by a straight line. Extend the line both sides to get the trend values corresponding to the entire data.

Consider the following hypothetical data on agricultural output for 11 years (in millions of tons):

Table 2.4 Method of semi-averages

Year	1	2	3	4	5	6	7	8	9	10	11
Output	80	85	88	77	90	95	98	100	104	103	110

To divide the data into 2 equal parts, we omit year 6. The average for year 1 to 5 is given by:

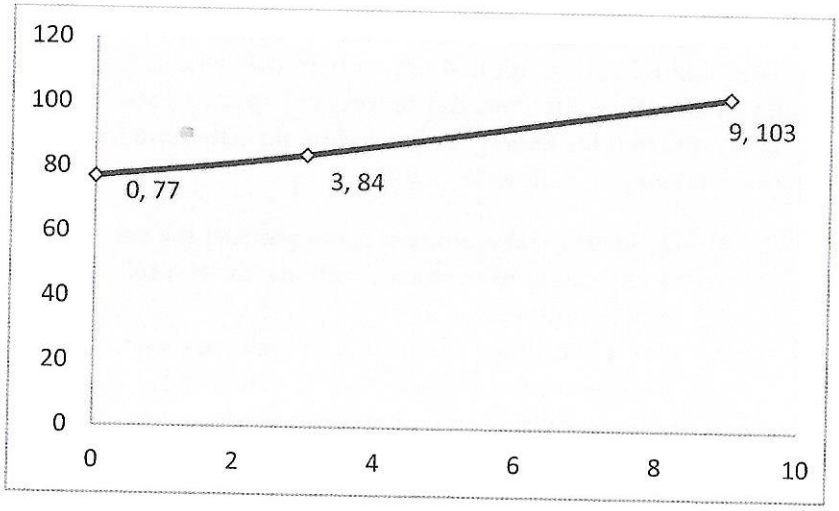
$$\frac{80+85+88+77+90}{5} = \frac{420}{5} = 84 \text{ mn tonnes}$$

This average is the trend value corresponding to year 3. The average for years 7 to 11 is given by:

$$\frac{98+100+104+103+110}{5} = \frac{515}{5} = 103 \text{ mn tonnes}$$

This average is the trend value corresponding to year 9. These two values are plotted and the points are joined to obtain the trend line as depicted in Figure 2.5

Figure 2.8 Method of Semi-average trend



Alternatively the other trend values can be estimated as follows: From year 3 to year 9, there has been an increase in the agricultural output from 84 mn tonnes to 103 mn tonnes giving an average increase of (103-84)/6 = 3.167 mn tonnes. 3.167 is the slope of the trend line. Hence the trend value for year 4 is equal to 84 + 3.167 mn tonnes, for year 5, is equal to 84+2(3.167) mm tonnes, for year 2, is equal to 84 - 3.167 mn tonnes, and so on.

The method of semi-averages is a simple technique to use but it has limited applicability since it can be used for the estimation of linear trends only.

2.3.3 Curve-fitting method

Form the scatter diagram or graph of the trend. The trend for different data may assume different forms, some of which are listed below:

(i) Straight line: $Y = a_0 + a_1X$
(ii) Parabola of 2nd degree curve: $Y = a_0 + a_1X + a_2X^2$
(iii) Cubic or 3rd degree curve: $Y = a_0 + a_1X + a_2X^2 + a_3X^3$
(iv) Quadratic or 4th degree curve: $Y = a_0 + a_1X + a_2X^2 + a_3X^3 + a_4X^4$
(v) Hyperbola: $Y = 1/(a_0 + a_1X)$
(vi) Geometric curve: $Y = ab^x$
(vii) Exponential curve: $Y = aX^b$

(viii) Gompertz curve: $Y = ab^{cx}$
(ix) Logistic curve: $Y = 1/(a + bc^X)$

We shall be concentrating only on the linear trend and also make references to a few other important types of trends.

2.3.4 Least squares estimation of linear trend

When the trend in a time series data is known to be linear, its mathematical form is $T_t = a_0 + a_1X$. The question then is which straight line would best estimate the trend. In other words, the question is of specifying the values of the parameters a_0 and a_1. There are several methods available for estimating the parameters of the trend line. According to one of the most familiar methods, the method of least squares, the values of a_0 and a_1 should be such as to minimize the sum of the squared deviations of the actual values from the values estimated on the basis of the fitted line. These values of a_0 and a_1 are obtained by the two normal equations.

$$\sum Y = na_0 + a_1\sum X$$

$$\sum XY = a_0\sum X + a_1\sum X^2$$

These two equations get simplified in the case of time series data. Since the units of time series are of uniform duration and denoted by consecutive numbers, one can take the middle period as the origin. This would make $\sum X=0$. The normal equations thus reduce to:

$$\sum Y = na_0$$

$$\sum XY = a_1\sum X^2$$

We then have simply

$$a_0 = \sum Y/n$$

$$a_1 = \sum XY/\sum X^2$$

Let us obtain trend values by the least squares method for the data on the agricultural output in table 2.4.

Table 2.5 Least squares estimation

Year	X	Output (Y)	XY	x^2
1	-5	80	-400	25
2	-4	85	-340	16
3	-3	88	-264	9
4	-2	77	-154	4
5	-1	90	-90	1
6	0	95	0	0
7	1	98	98	1
8	2	100	200	4
9	3	104	312	9
10	4	103	412	16
11	5	110	550	25
	0	1030	324	110

$$a_0 = 1030/11=93.6$$

$$a_1 = 324/110=2.9$$

The least squares trend line is therefore given by;

$$Y=93.6 + 2.9X$$

The trend values are presented in table 2.6.

Table 2.5 Trend values obtained by Least squares method

Year	1	2	3	4	5	6	7	8	9	10	11
Trend	79.1	82	84.9	87.8	90.7	93.6	96.5	99.4	102	105	108.1

2.3.5 Other trend forms

The least squares method can be applied for estimating the parameters of second-degrees, third-degree and, in general, n^{th}-

degree curve. The equation of the parabola or 2nd degree curve is given by:

$$Y = a_0+a_1x +a_2x^2$$

There are 3 parameters a_0, a_1 and a_2 to be estimated for which we have the following three normal equations of the least squares parabola.

$$\sum Y = na_0 + a_1\sum X + a_2\sum X^2$$
$$\sum XY = a_0\sum X + a_1\sum X^2 + a_2\sum X^3$$
$$\sum X^2Y = a_0\sum X^2 + a_1\sum X^3+a_2\sum X^4$$

When the middle period of the time series is taken as the origin, all sums of odd powers of X vanish and the three normal equations get simplified to

$$\sum Y = na_0 + a_2\sum X^2$$
$$\sum XY = a_1\sum X^2$$
$$\sum X^2Y = a_0\sum X^2+a_2\sum X^4$$

We then get

$$a_1 = \frac{\sum XY}{\sum X^2}$$

$$a_2 = \frac{n\sum X^2Y - \sum Y\sum X^2}{n\sum X^4 - (\sum X^2)^2}$$

$$a_0 = \frac{\sum Y - a_2\sum X^2}{n}$$

An exponential trend is defined by the equation $Yt = ab^x$. We can make the equation linear by taking logarithms on both sides

$$\log Y = \log a + x\log b$$

The method of least squares estimation can then be applied to yield to normal equations

$$\sum \log Y = n \log a + \log b\sum X$$
$$\sum X \log Y = \log a\sum X + \log b\sum X^2$$

Again, by taking the origin as the middle period, we get

$$\sum \log Y = n\log a$$
$$\sum X \log Y = \log b\sum X^2$$

Which gives the values of log a and log b as

$$\log a = \frac{\sum \log Y}{n}$$

$$\log b = \frac{\sum X \log Y}{\sum X^2}$$

We then get a and b by taking the anti-logarithm of these values.

Though curve-fitting techniques are indeed valuable on account of their objectivity, they have one drawback. If, after the trend has been computed on the basis of this technique, some additional data become available for later periods, the trend will have to be calculated all over again after incorporating the additional data. This is not necessary in the case of the moving average method.

2.4 Analysis of seasonal variations

Seasonal patterns are sought to be calculated by eliminating trends, cycles and irregular fluctuations from the time series data and preparing a seasonal index which compares each month's figures with the average of all the months of the year. The sum of all the seasonal index numbers is obviously 1200 with the average equal to 100.

There are several methods available for constructing the seasonal index. We shall merely outline a few of them here.

2.4.1 Average percentage method

For each month, the actual figure is expressed as a percentage of the average for the corresponding year. These percentages for corresponding months of different years are averaged by using arithmetic mean or median. The resulting 12 percentages constitute the seasonal index,

In finding the percentage for each month (dividing actual figure by the year average) the trend is sought to be eliminated. Averaging these percentages for corresponding months of different years eliminates irregular fluctuations. This leaves S and C. Thus, this method is not completely satisfactory.

2.4.2 Percentage trend method

The average for each year is calculated by dividing the sum of all monthly value by twelve. This average is taken as the trend value corresponding to June of that year. Other monthly trend values are obtained by using, say, the least squares technique. For each month the actual value is expressed as a percentage of its trend value. These percentages for corresponding months of different years are averaged to get the required seasonal index.

This method has the same defect as the previous one. Dividing the actual value by the trend value T leaves CSI. Subsequent averaging may adjust for I. This still leaves CS.

2.4.3 Percentage moving average method

We first compute a moving average of order twelve for the entire data. When this is done it will be found that each moving average value corresponds not to any month but in-between two months. For instance, the first value will correspond to between June and July of first year, the second value to between July and August, and so on. We, therefore, take a moving average of order two of the twelfth order moving averages. Now, the first value will correspond to July of first year, the second value to August and so on. Obviously, data for the first six months of the first year and last six months of the last year are lost in the process.

The original value for each month is expressed as a percentage of its centred moving average value. These percentages for corresponding months are then averaged to give the required seasonal index.

Logically, this is one of the best methods for analysing S. The centred moving average removes an S and I and what remains is TC. Further, dividing Y by TC gives back SI. The final averaging eliminates I leaving only S.

2.4.4 Link relative method

The actual value for each month is expressed as a percentage of the value for the previous month. These percentages are called

link relatives. The link relatives for the corresponding months are then averaged to get 12 average link relatives. The value for January is equated to 100 and the relative percentage values for all other months with respect to January are taken. The percentage for the next January will then be found to be higher or lower than 100 depending on whether there has been an increasing or decreasing trend. The various percentages are then adjusted for this trend and these adjusted percentages provide the seasonal index.

2.5 Analysis of cyclical and random fluctuations

There are several methods that can be employed for obtaining the cyclical component. All these methods give the same result since they all essentially seek to arrive at an estimate of *C* by eliminating *T, S* and I from the actual data *Y*. They differ only in the sequence in which *T, S* and I are eliminated. The choice of the method depends mainly on the nature of the available data and computational convenience. For instance, one may first *detrend* the data by dividing Y by *T* to obtain *CSI*. Then an estimate of *CI* is obtained by *deseasonalising* the data, i.e., adjusting the data for seasonal variation by dividing *CSI* by the seasonal index and thus removing *S*. A moving average (preferably of an odd order so that subsequent centering does not become necessary) of say 3, 5 or 7 months duration is then used to smoothen out the irregular component I from CI and an index of *C* is then obtained.

In many instances of business data, deseasonalized data is of relevant interest to the businessman. For instance, in perusing the data on the sale of woollen garments, a businessman need not be perturbed by a significant fall during summer months or feel unduly gleeful at what appears to be a significant rise during winter.

It is possible, for instance, that the actual data may show a rise in sales during winter, but the data adjusted for seasonal variation may indeed register a fall. The implication of such a result would

be that though sales have gone up, they have not, in fact, gone up as much as they were expected to.

When deseasonalized data in itself is also of importance, it is preferable to estimate *C* first by dividing *Y* by *S* to get *TCI* and then dividing it by *T* to obtain *CI* and finally eliminating *I*.

A third method of isolating C is to first estimate TS as a product of *T* and S and then divide Y by TS to yield CI and then get rid of I by suitable averaging. A trifling advantage of this method is that it is easier to perform one multiplication (TxS) and one division (*Y/TS*) than perform two divisions (*Y/T* and *CSI/S* or *Y/S* and *TCI/T*).

Irregular fluctuations by their very nature are unsuitable for mathematical analysis the way trend or seasonal variations are. Random variations are not caused purely by random factors but also result from sporadic factors which tend to produce extreme values. As such, I can be obtained only by dividing *Y* by *T, C* and *S*

CHAPTER 3

3 Forecasting techniques

3.1 Need for forecasting

Planning is the essence of human activity. Every individual, every firm, every economy has to plan in order to survive and grow. And planning, by definition implies choosing strategies for future action. The future, however, is shrouded in uncertainty. It is, therefore, necessary to clear this mist of uncertainty as far as possible so that present decisions can be taken with some security. The need for forecasting is essentially to minimize the uncertainty surrounding the planner's decisions. Once the future situations and events are reasonably accurately predicted, they provide a concrete basis for one's actions concerning the future.

Firms have to take decisions regarding cost, profit, production, pricing, investment, marketing etc. in order to meet competition effectively and promote its growth. An economy has to plan for its future revenues and expenditures so as to stimulate balanced and stable growth in the economy. This planning, and hence the need for forecasting, arises both at the micro and macro levels.

Forecasting in the area of economic and business activity is quite a complicated job, one of the chief reasons being that events and variables are highly interrelated. As one Greek economist Valavanis aptly stated "everything depends on everything else is the theme song of the economic and celestial spheres". Hence, to forecast one thing, it may be necessary to forecast so many others. For instance, if a firm wants to plan its future capacities, it must forecast the future demand for its product. To forecast the future demand, it will require forecasts of population growth, incomes, prices of substitute products etc. which influence its product demand.

Future, in this context, may be conceived as being of two types: *conjunctural* future and *affectable* future. Conjunctural future refers to those areas like weather, technology, political changes etc. which are not within the control of the planner or decision-maker, while affectable future is at least partly within the planner's control. For example, population growth may be to some extent sought to be checked by expenditures on family planning, price level may be sought to be controlled by monetary and fiscal policies etc. Conjunctural future is the domain of pure forecasting while affectable future involves forecasting and control.

3.2 Methods of forecasting

There are various methods of forecasting that have been devised and they all can be grouped into 4 broad categories:

a. Mechanical extrapolations;
b. Barometric techniques;
c. Sample survey approach;
d. Econometric methods;

We shall discuss the methods under each of these categories.

3.2.1 Mechanical extrapolations

The techniques falling in this category are termed "mechanical" since they are not closely integrated with economy theory and statistical data.

1 Naive (simple) models:

These models assume that the future value of a variable is related to its present value. That is, if Y_{t+1} is the forecast value of the variable for period t+1 and Y_t is the actual value in period t, these models postulate that

$\hat{y}_{t+1}=f(Y_t)$. The simplest of these simple models is the *no change model* in which the present value is assumed to continue in the next period. That is $\hat{Y}_{t+1} = Y_t$. One can also have *proportionate change models* which may be specified as:

$$Y_{t+1} = \left\{\frac{Y_t - Y_{t-1}}{Y_t}\right\} Y_t + Y_{t+1}$$

A not so naive version of the naive model is one in which the forecast of a variable for the future period is made to depend not only on its present value but also on past values. For instance, by one formulation, the forecast level of a variable in the period t+1 is given by the forecast for the current period amended by some proportion of the current forecasting error. In notation;

$$\begin{aligned}
\hat{y}_{t+1} &= \hat{y}_t + (1-\gamma)(y_t - \hat{y}_t) \\
&= \gamma\,\hat{y}_t + (1-\gamma)\,y_t \\
&= (1-\gamma)\,y_t + \gamma\,[(1-\gamma)(y_{t-1} + \gamma\hat{y}_{t+1})] \\
&= (1-\gamma)\,y_t + \gamma\,(1-\gamma)\,y_{t-1} + \gamma^2\hat{y}_{t-1} \\
&= (1-\gamma)\,y_t + \gamma\,(1-\gamma)\,y_{t-1} + \gamma^2[(1-\gamma)\,y_{t-2} + \gamma y_{t-2}] \\
&= (1-\gamma)\sum_{i=0}^{\infty}\gamma^i\,y_{t-j}
\end{aligned}$$

If γ is close to zero, γ^j will decline very rapidly and the forecast depends more on recent values than on values of remote past. If γ is close to 1, then past trends are slow to change and will be expected to continue in the future.

The advantage of naive models is that they are simple and straight-forward. However, they can be best used only for short term decisions since radical changes in the near future need not be expected. But this is not a reasonable assumption to make in the long period.

2 Time series analysis:

Of the four components of a time series T, C, S and I, S is fairly predictable and I is unpredictable but can be practically eliminated by suitable averaging. T and C are, therefore, the major preoccupations of forecasters.

Trend is usually forecast by extrapolation. That is, a trend curve is extended to past data and is projected into the future. This

projection becomes precarious the further it is extended into the future. Farther the future, greater is the probability that the forces operating in the past will change. The trend projection becomes still more questionable if it is done by moving averages. As we have already seen, moving averages trail behind the most recent data. However, if the trend is found to be rather persistent then moving averages can give a realistic portrayal of future movements for several years.

One frequently used method developed for forecasting cyclical fluctuations from the time series itself is known as *exponential smoothing*. In this method, the forecast level for period t+1 made in period t is a function of the actual value for period t and the forecast for period t made in period t-1. That is, if F_t is the forecast for period t+1, $F_t = f(Y_t, F_{t-1})$, is called the *smoothed value* for period t+1. F_t in fact is a weighted average of Y_t, and F_{t-1} with α and β being the weights given to Y_t, and F_{t-1}, where α and β are positive fractions and $\alpha + \beta = 1$. α and β called *smoothing constants.*

Exponential smoothing can be applied to any power of Y_t, and F_{t-1}. In most economic applications, simple exponential smoothing is used:

$$F_t = \alpha Y_t + \beta F_{t-1}, \alpha, \beta > 0, \alpha + \beta = 1$$

Note that this is similar to the naive model discussed earlier. If a significant trend is present, it can be adjusted for in the following way: Define

$$\Delta F_t = F_t - F_{t-1}$$

Estimate the trend as

$$T_t = \alpha \Delta F_t + \beta T_{t-1}$$

The forecast value F_t adjusted for T_t is

$$F_t' = F_t + \left(\frac{\beta}{\alpha}\right) T_t$$

If adjustment for seasonal variations also has to be done, then we have F_t adjusted for trend and seasonal variations given by

$$F_t'' = F_t'(S_{t+1})$$

where S_{t+1} is the seasonal index for period t+1.

We need to estimate the initial values F_0 and T_0 to carry out exponential smoothing. F_0 is obtained by averaging a few past observations of the series and T_0 is taken as zero or as the slope of the trend equation from past data.

The following table illustrates the process of exponential smoothing. The following values are assumed:

$$\alpha = 0.2; \beta = 0.8; F_0 = 100; T_0 = 0$$

Table 3.1 Exponential smoothing

Time period, t	Yt	Ft	Δ Ft	Tt	Forecast, Ft'
2010	105	101.00	1.00	0.20	101.80
2011	106	102.00	1.00	0.36	103.44
2012	110	103.60	1.60	0.61	106.04
2013	108	104.48	0.88	0.66	107.12
2014	115	106.58	2.10	0.95	110.38

3.2.2 Barometric techniques

While techniques falling in the category of mechanical extrapolations assume that future is an extension of the past, barometric techniques assume that future depends on present happenings.

Barometric techniques fall into two sub categories: the indicator approach and pressure indices.

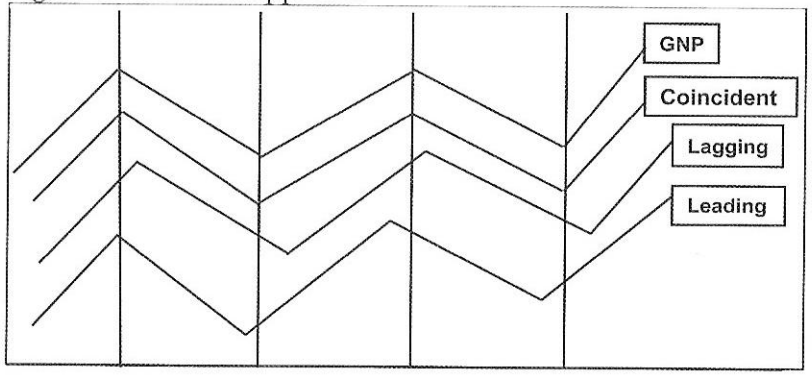
1. The indicator approach:

The Gross National Product or GNP is usually taken as a measure of aggregate economic activity. The movements of the GNP, therefore, indicate the direction in which general business activity is moving. All other economic variables are in some kind of a lead- lag relationship with the GNP. In other words, these variables move along with the GNP in different time sequences.

Some of these variables like industrial production, personal incomes, sales of retail stores etc, move in phase with the GNP. Their turning point (i.e. peaks and troughs) coincide with those of the GNP. These variables are called *coincident indicators*.

Some other variables like labour cost, book value of inventories, consumption expenditures tend to trail behind aggregate economic activity. They are called *lagging indicators*. There are yet some other variables which are observed to lead, with substantial regularity, the GNP and other measures of the current performance of business, by several months. It is the movements of these variables called *leading indicators*, which are of importance in forecasting a change in the basic performance of the entire economy. Example of these "warning signals' which move ahead are orders, raw material prices, stock prices, average workweek, etc. A diagrammatic comparison of the three types of indicators is depicted in figure 3.1.

Figure 3.1 Indicator approach



Limitations of leading indicators;

i. They are not consistent in their leading tendency. Some of them may even coincide or lag behind GNP. The reason is leading indicators are chosen not because of any causal relationship of the variables with GNP but their past behaviours. As the economic structure changes over time, their leading patterns may also change.

ii. They may be too late in signalling the reversal of trends.

iii. One can't always feel sure whether the changes they signal are significant or only a wiggle. To distinguish between genuine changes and false indications, one may have to wait and this may itself destroy the value of the forecast.

iv. They can at best indicate the direction of change but not the magnitude.

Therefore, movements of leading indicators can be used only as supplementary information. Since reliance on any one or two leading indicators is risky, *diffusion indices* are constructed as a percentage of a set of series that is moving upwards. This can range from 0 (when all are moving downward) to 100 (when all are moving upward). If it is over 50 percent, one may predict a rise in business activity and if it is less than 50 percent, one may predict business to be in a state of contraction.

2. Pressure indices: These indices are based on the recognition of the fact that amplitude differences are important in business cycles. A few examples of pressure indices are:

   i. The ratio of durable goods production to non-durable goods production. An increase in this ratio indicates prosperity.

   ii. The ratio of raw material inventories to new orders for finished goods. An increase in this ratio indicates decline in business activity.

   iii. The spread between stock yields and bond yields. As the spread becomes smaller, money flows out of the stock market into the bond market.

3.2.3 Opinion polling or sample survey approach:

As opposed to other techniques of forecasting which rest on objective method of prediction, the prediction is based on this approach looks at changes in the attitudes and expectations of

individuals who matter. For instance, in forecasting demand for a product, consumer surveys may be undertaken to know whether people intend to buy the product. One may also try to get information about consumer buying intentions indirectly or vicariously by a survey of the opinion of salesmen and dealers, retailers and wholesalers who are supposed to know consumers' needs and preferences. If the product is a new one, one may undertake test marketing, i.e. introduce the product in a sample market, and from that form an estimate of the total demand in a fully developed market.

All these methods, in fact, are forms of market research. There are several limitations of this approach to forecasting.

Firstly, interviewing and sampling are expensive affairs and hence one may be forced to cut down the sample size. But then, if the sample is small, it may not be representative of the entire market for the product in question and errors are more likely. Secondly, reliability of the opinions expressed may be low because of the over-enthusiasm of the salesmen and dealers or inconsistency of the consumers' buying intentions. One cannot often make out whether the purchase plans revealed by a consumer to an investigator are genuine or momentarily concocted by him during the interview. In other words, opinion polling approach is highly subjective in nature. For instance, if potential buyers are being personally interviewed to find out if they intend to purchase a TV set or not, a boastful buyer may respond in the affirmative even though he is not likely to buy owing to income limitations while a respondent who is suspicious of the investigator may not reveal his true buying intentions.

One of the special forms of opinion polling which seeks to reduce the subjective biases inherent in the method is the *Delphi Method* which was developed by the Rand Corporation in the U.S.A. and is used chiefly for technological forecasting.

The *Delphi method* tries to pool the opinions of experts. Each expert in a group is asked to furnish his prediction of important

events in the area in question in the form of brief statements, whose complete clarification is obtained by the investigator. After the initial predictions from all experts, each expert is then successively re-questioned in the light of the feedback supplied from other experts via the investigator until a satisfactory and consistent forecast emerges.

Opinion polling, on the whole, can give an idea of the direction of change rather than the magnitude.

3.2.4 Econometric method:

By this method, any economic activity or magnitude is forecast on the basis of mathematical equations which are supposed to express the most probable relationships among economic variables. The variables can be *endogenous* or *independent* variables. Endogenous variables are those whose values are determined from the model, while independent variables are those whose values are given from outside the model. Independent variables may be either *exogenous* (their values are assigned by the forecaster) or *predetermined* (their values are known from data of earlier periods).

There are three principal steps in econometric forecasting:

a) The *specification* on the basis of economic theory of the appropriate functional relationships among the variables which is in other words the specification of the form of the mathematical equation or equations - whether they are linear, non-linear, etc.
b) The *estimation* of the numerical values of the parameters of the equation(s), by using statistical methods;
c) The simulation or insertion of the values of the exogenous variables to get forecasts of the endogenous variables.

Econometric models may be either single equation models or simultaneous equation models.

a) Single-equation models: Suppose for instance, we need to forecast the demand for a certain product. We have the

demand function wherein demand is the endogenous variable whose value is explained in terms of its determinants which are independent variables. Usually the determinants will be very large in number and so a selection of the most important variables is made by studying, through *correlation analysis*, the closeness of the relationship between demand and its determinants. Once the independent variables have been chosen, the relationship between those variables and demand has to be specified. Suppose for our product in question, the chief demand determinants are its price, consumers' income, and weather and that the relationship is assumed linear.

We can then write, demand = a_0+ a_1 (price) +a_2 (income) + a_3 (weather).

The next step is the estimation of the values of the parameters a_0, a_1 ,a_2 and a_3 .The most frequently used technique for this estimation is *regression analysis*. After this has been done, simulation is carried out.

The forecaster assumes certain values for price, income and index of weather which are then plugged into the equations to yield forecast of product demand.

b) Simultaneous equation models: Consider the following simple model of the GNP:

$$Y = C + I + G + X \quad \ldots\ldots (1)$$

$$C = a_0 + a_1 Y + a_2 (C\text{-}1) \quad \ldots\ldots (2)$$

$$I = b_0 + b_1 r \quad \ldots\ldots (3)$$

Where,

Y = Gross National Product

C = aggregate consumption

I = aggregate investment

X = net exports

G = Government expenditures

C-1 = aggregate consumption in earlier period.

r = rate of interest

a_0, a_1 ,b_0 and b_0 are parameters

Equation (1) is an identity. Equations (2) and (3) which are consumption function and investment function respectively have been specified as being linear. *G*, *X* and *r* are exogenous variables and C-1 is a predetermined variable whose value is already known from earlier periods' data.

Once the numerical values of the parameters are estimated and specific values assumed for the exogenous variables fitted into the equations, the value of y will be obtained.

Limitations of the econometric methods;

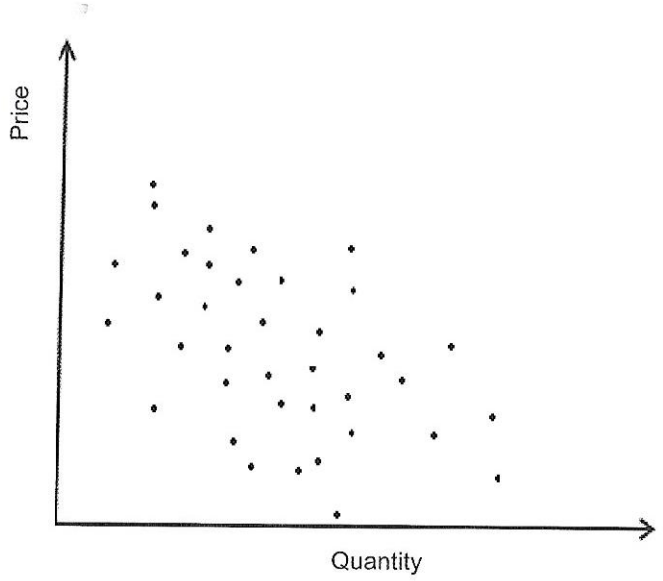
Though the econometric method is the most analytical method of forecasting based on a sound foundation of theory and data, it is nevertheless beset with a number of difficulties and drawbacks, a few of which are enumerated here.

1. When a dependent variable is determined by several independent variables, and these independent variables are themselves highly interrelated, it becomes difficult to isolate their separate influences and obtain a reasonably good estimate of the influence of each independent variable on the dependent variable. This problem is known as *multicollinearity*.
2. In simultaneous equation models, one may be faced with the *identification problem*. For instance, suppose one has a market model consisting of the following equations:
   a. $Q_t^d = \alpha + \gamma P_t$: Demand function
   b. $Q_t^s = \beta + \sigma P_t$: Supply function
   c. $Q_t^d = Q_t^s$: Market equilibrium

where Q_t^d, = demand at time t, Q_t^s = supply at time t, P_t = price at time t, and, α , γ, β and σ are parameters.

Suppose one wishes to estimate the parameters α and γ of the demand function from the price-quantity data shown in the following scatter diagram:

Figure 3.2 Demand function



Before trying to apply regression to estimate α and γ from the above diagram, what one has to make sure is whether the data really correspond to a demand function. Merely because the data show a downward trend, it f does not follow that it is demand data. Consider the following;

$$Q_t^d = Q_t^s$$

$$= \beta + \sigma P_t$$

Multiplying equation (1) by μ and equation (2) by λ, we get

$$\mu\, Q_t^d = \mu\, \alpha + \mu\, \gamma P_t$$
$$\lambda\, Q_t^s = \lambda\, \beta + \lambda\, \sigma P^t$$

Adding, we get

$$(\lambda + \mu)\, Q_t^d = \lambda\, \beta + \lambda\, \alpha + (\lambda\gamma + \mu\, \sigma) P_t$$

$$Q_t^d = \frac{\lambda\, \beta + \lambda\, \alpha}{\lambda + \mu} + \frac{\lambda\gamma}{\lambda + \mu} P_t$$

This is a "hybrid" equation obtained as a linear combination of the demand and supply equations. Since the choice of λ and μ are purely arbitrary, they can be chosen as to make + negative, and make the equation "look like" a demand equation. The scatter diagram could very well correspond to such a hybrid equation and not to the true demand function. The demand function in our model is thus not identified. If α and β were to be estimated from this data, any future forecast made will be precarious.

3. As has been noted, forecast of the endogenous variables involves, in the final stage, simulation or plugging in the values of the exogenous variables. This means forecasting of the exogenous variables will have to precede forecasting of the endogenous variables.

If the forecasts of exogenous variables have not been carefully made and involve errors, they will be carried into the model.

4. The validity of the forecasts depends on the assumption that the economic structure embedded in the econometric model will continue in the future.

3.3 Opportunistic forecasting:

A good forecasting method is one which gives high returns over cost in terms of accuracy, looks reasonable or consistent with existing knowledge, is adaptable to new circumstances and gives up-to-date results. To ensure most of these desirable characteristics of a forecast, a forecaster, in practice, does not confine himself to any one technique, since every technique has some limitations and cannot be wholly relied upon to yield the best predictions. The term *opportunistic forecasting* is used to refer to the simultaneous use of survey approach, information on leading indicators etc. together with theory to obtain consistent and reliable forecasts. The word "opportunistic" is thus not being used in this context with any pejorative slant. It only implies that the forecaster takes advantage of all opportunities.

While we have discussed the broad techniques of forecasting, specific methods or forms of these methods may be employed in particular context. For instance, for forecasting a firm's cost and profits, break-even charts are constructed.

Or in forecasting the demand for a new product, one may adopt the *evolutionary approach* which consists in extrapolating the demand for the new product as an evolution of an existing old product. For instance, the demand cell phones may be assumed to take off from where the demand for landlines left off. The substitute approach and the growth curve approach are two other approaches that may be adopted for demand projection of new products. The new product is analysed as a substitute for some existing product (e.g. ball point pen in relation to fountain pens) or its growth rate is estimated on the basis of the growth pattern for established products of the same product category as the product in question (e.g. microwave in relation to household appliance).

CHAPTER 4

4 Index numbers

4.1 Meaning and types of Index Numbers

An index number is a statistical device for studying comparative changes in a variable or a group of variables over different points of time or different geographical areas. The variables whose changes are usually studied in economics and business are prices, quantities and values.

Accordingly, we have three categories of index numbers or indices for short, namely, price indices, quantity indices and value indices. If any of these indices is computed for a single variable it is called a *single or univariate index number*, while if it is constructed for a group of variables it is called a *composite index number*. In preparing a composite index either all the variables may be treated with equal importance or each variable may be assigned a weight depending on its importance in relation to others. For instance in constructing a price index, items like food, clothing, etc. may be given more weightage in comparison to other items like fans and refrigerators. Such an index is called a *weighted index number*. Most composite indices will be weighted indices.

Since an index number is a measure of relative or comparative change, it is invariably a ratio which is generally expressed as a percentage. For instance if the price of sugar in 2010 is 4 per kilogram while it was 2.50 per kilogram in 2005, then the relative change in price in 2010 as compared to 2005 is;

$$\frac{\text{P 2010}}{\text{P 2005}} \; X\,100 = \frac{4}{2.5} \; X100 = 160$$

i.e. a 60% increase over the 2005 price. We could also compare the 2005 price with reference to the 2010 price. We would then have the index number for 2005 as;

$$\frac{\text{P 2005}}{\text{P 2010}}\ X\ 100 = \frac{2.5}{4} X100 = 62.5$$

This means that the 2005 price was 37.5% lower than the 2010 price. One could similarly prepare the 2010 sugar price index for Livingstone city in comparison to Chipata or vice versa. The period or the region in comparison to which the index number is prepared is called the *base period* or *region.*

If P_0 is the price in the base period or period 0 and P_1 is the price in period 1, then the price index is given by $\frac{P_1}{P_0}$ X 100. Similarly, the quantity index is given as by $\frac{Q_1}{Q_0}$ X 100 and the value index as $\frac{P_1\ Q_1}{P_0 Q_0}$ X 100. Often one might prepare a series of index numbers for several periods. For instance, one may prepare say the price index for a commodity from 1980 onwards. In such a case, one can prepare the price index number for every year with 1980 as the constant base and we shall then have a series of *fixed base index numbers;* or one may prepare the index for each year using the immediately preceding year as the base. We will then have a series of *chain base index numbers.* For example the index number for 1981 will be $\frac{P_{1981}}{P_{1980}}$ X 100, for 1982 $\frac{P_{1982}}{P_{1981}}$ X 100 and so on. In general, fixed base indices will be $P_{01}, P_{02}, P_{03}, P_{04}$,....., while chain indices will be $P_{01}, P_{12}, P_{23}, P_{34}$..... One can, of course convert a fixed base series into a chain base series and vice versa, as follows:

$P_{01} = \frac{P_1}{P_0}$ X 100 is the same in both types of series.

$$P_{12} = \frac{P_2}{P_1} X\ 100 = \frac{P_2}{P_0} X \frac{P_0}{P_1}\ X\ 100 = \frac{P_{02}}{P_{01}}\ X\ 100\ ;$$

$$P_{23} = \frac{P_3}{P_2} X\ 100 = \frac{P_3}{P_0} X\ \frac{P_0}{P_2}\ X\ 100 = \frac{P_{03}}{P_{02}}\ X\ 100\ ;\ \text{and so on.}$$

Conversely, we have $P_{02} = P_{01} X P_{12}$; $P_{03} = P_{02} X P_{23} = P_{01} X P_{12} X P_{23}$; and so on.

In composite index numbers, we have data on prices, quantities or values for not one commodity but several, say *n, commodities.* Thus, suppose the prices of n goods in the base period and period 1 are given as under:

Base period $\quad P_1^0\ P_2^0\ , \ldots\ldots\ldots . P_n^0$

Base period $\quad P_1^1\ P_2^1\ , \ldots\ldots\ldots . P_n^1$

Note: P_i^j is the price of commodity *I* in period *j*

To obtain a single composite index for period 1, one has to average out the *n* prices first before arriving at the index if same units are used for quantities. Theoretically, one can use any measure of average. However, in practice, the arithmetic mean is most frequently used on account of its ease in calculation, though as we shall see later, the geometric mean is the ideal average to use in index number construction. There are two ways of obtaining the composite index number. One may find the average price for all the *n* commodities for each period and then take the relevant price ratio or one may first take the ratio of the period 1 price to the base year price for each commodity and then average out the price ratios for all the *n* commodities. The index number obtained by the first method is called an *aggregate index number* and the index number obtained by the second method is called *average-of-relatives index number.* The formulae for the index numbers corresponding to each of these types are given below:

Aggregate index number using arithmetic mean:

$$P_{01} = \frac{\sum_{i=1}^{n} p_i^1 / n}{\sum_{i=1}^{n} p_i^0 / n} X100 = \sum_{i=1}^{n} \frac{P_i^1,}{P_1^0} = 100$$

Aggregate index number using geometric mean:

$$P_{01} = \frac{\sqrt[n]{P_1^1 \times P_2^1 \times P_3^1 \ldots\ldots\ldots\ldots P_n^1}}{\sqrt[n]{P_1^0 \times P_2^0 \times P_3^0 \ldots\ldots\ldots\ldots P_n^0}} X100$$

Average-of-relatives index number using arithmetic mean:

$$\mathbf{P_{01} = \frac{\sum_{i=1}^{u} \frac{P_i^1}{P_i^0}}{n} X100}$$

Average-of-relatives index number using geometric mean:

$$\sqrt[n]{\frac{P_1^1 \times P_2^1 \times P_3^1 X \times \ldots\ldots\ldots\ldots P_n^1}{p_1^0 \times P_2^0 \times P_3^0 \times \ldots\ldots\ldots\ldots P_n^0}}$$

Notice that when the geometric mean is used, the formulae for the aggregate index and the average-of-relatives index are the same whereas the two need not give the same results when the arithmetic mean is used.

The *weighted average-of-relatives index* for period 1 using the arithmetic mean is

$$\sum_{i=1}^{n} w_i (P_i^1 / P_i^0) \times w_i$$

where w is weight for i[th] commodity. It is left to the reader to obtain the formulae using geometric mean.

Indices which make use of base period values as weights are called *Laspeyre's indices* while those that make use of the given period's values as weights are called *Paasche's indices.* Thus, we have the indices as shown in table 4.1.

Table 4.1 Indices formulae

Laspeyre's Price Index Number L_P	$= \frac{\sum P_i^1 Q_1^0}{\sum P_i^0 Q_i^0} \times 100$
Laspeyre's Quantity Index Number L_q	$= \frac{\sum P_i^0 Q_1^1}{\sum P_i^0 Q_i^0} \times 100$
Paasche's Price Index number P_P	$= \frac{\sum P_i^1 Q_1^1}{\sum P_i^0 Q_i^1} \times 100$
Paasche's Quantity Index number P_q	$= \frac{\sum P_i^1 Q_1^1}{\sum P_i^1 Q_i^0} \times 100$

According to the economic law of demand, prices and quantities are generally inversely related. Hence if there has been a general rise in prices in period 1 as compared to the base period, then Q_i^1 will usually be smaller than Q_i^0. Hence Laspeyre's index tends to overestimate price changes, while Paasche's index tends to underestimate them. Hence some other formulae have been developed as a compromise between Laspeyre's and Paasche. Two of them may be noted. The *Marshall-Edgeworth Index Number* is an "arithmetic cross" between Laspeyres and Paasche and uses the arithmetic mean of both periods' values as weights. Thus the Marshall-Edgeworth index number is

$$= \frac{\sum P_i^1 (Q_1^1 + Q_1^0)}{\sum P_i^0 (Q_i^1 + Q_i^0)}$$

Fisher's index number is a geometric cross between Laspeyres and Paasche. For example, Fisher's price index number.

$$F_P = \sqrt{L_P \times P_P}$$

i.e the geometric mean of Laspeyre's and Paasche's indices. The value index number,

$$V_{01} = \frac{\sum P_i^1 Q_1^1}{\sum P_i^0 Q_i^0}$$

Consider the following time series data in table 4.1 on prices and quantities:

Table 4.2 Prices and quantities

Commodity	2010		2015	
	Price (P_i^0)	Quantity(Q_i^0)	Price(P_i^1)	Quantity(Q_i^1)
A	4	500	5	600
B	5	225	3.5	400
C	8	300	6	350
D	1	1000	2	700
E	2	600	2.5	500

Table 4.3 Computation of indices

$P_i^0 Q_i^0$	$P_i^0 Q_i^1$	$P_i^1 Q_i^0$	$P_i^1 Q_i^1$	$P_i^0(Q_i^0 + Q_i^0)$	$P_i^0(Q_i^0 + Q_i^1)$
2000	2400	2500	3000	4400	5500
1125	2000	787.5	1400	3125	2187.5
2400	2800	1800	2100	5200	3900
1000	700	2000	1400	1700	3400
1200	1000	1500	1250	2200	2750
7725	**8900**	**8587.5**	**9250**	**16625**	**17737.5**

$$\text{Laspeyre's Quantity Index Number} = \frac{8587.50}{7725.00} \times 100$$

$$= 111.2$$

$$\text{Laspeyre's Price Index Number} = \frac{8900.00}{7725.00} \times 100$$

$$= 113.1$$

$$\text{Paasche's Price Index Number} = \frac{9250}{8900} \times 100$$

$$= 103.9$$

$$\text{Paasche's Quantity Index Number} = \frac{9250}{8587.50} \times 100$$

$$= 107.7$$

$$\text{Fisher's Price Index Number} = \frac{17737.50}{16625} \times 100$$

$$= 106.1$$

Marshall-Edgeworth Price Index Number Value Index Number=

$$\frac{9250}{7725.00} \times 100$$

$$= 118.63$$

4.2 Uses of Index Numbers

Index numbers can be put to a number of uses. Firstly, they may be used to study the broad trends in the various sectors of the economy. Thus we may construct index numbers of agricultural output, industrial output, foreign trade, etc.

Index numbers depicting changes in the general price level are necessary to know the purchasing power or value of money. As we know, the value of money is determined by the quantity of goods and services that money can buy and this depends on the prices of the goods and services. The value of money varies inversely with the general price level. Thus one may construct a general price index and the reciprocal of this index will give an idea regarding changes in the value of money. For example if P_{01}

=140, then the value of money in period 1 as compared to the base period is $\frac{1}{1.4} = 71.4\%$.

In other words, it is 28.6% less. Again if $P_{01} = 90$, then 1/0.9 = 1.11. This means 11 % more purchasing power in the given period in comparison to the base period.

Index numbers are also applied in studying the standard of living of particular sections of the population such as industrial workers, farmers etc. If for instance, we find that the money wages of industrial workers show an increase during a certain period we cannot immediately conclude that the standard of living of the workers has gone up. It depends on the size and composition of their consumption basket which is a measure of their wages in real terms. If during the said period the money wages of the workers have doubled, but the cost of their consumption basket during the same period has trebled then their wages in real terms have gone down and hence their standard of living has also deteriorated. Thus real wages are more important than money wages. To know wages particularly during a period of inflation, one has to deflate the money wages by adjusting it to changes in the cost of living as measured by a *'cost of living index number.'* The formula for deflation is:

$$\text{Index of real wages} = \frac{\text{Index of money wages}}{\text{Cost of living index}}$$

Cost of living index number can be used in the fixation and periodical revision of wages and salaries and in particular minimum wages.

A more macro instance of deflation is the study of the growth in real terms of a country's GNP which is a measure of economic development. For instance, suppose we wish to know the size of the GNP in 2015 as compared to 1985. The value of the GNP in 2015 at 2015 prices as compared to the value of the GNP in 1985 at 1985 prices will be edifying to some extent. But to know the true

increase in the GNP in terms of real output, one will have to see the value of the GNP for 2015 at 1985 prices in view of the tremendous increase in the price level that has taken place during the 30-year period.

Thus we have,

$$\text{GNP at 1985 prices} = \frac{\text{GNP at 2015 prices}}{\text{Appropriate price index}}$$

Again to understand the true nature of economic growth, one has to study changes in per capita income rather than in an absolute figure like the total national income. For this one has to adjust the national income to changes in population by an index of population growth. For instance, suppose during a period the National Income of a country increased by 20%, the per capita income would have increased only by 8% if the country's population also grew at the rate of 14% during that period.

Index numbers can also be used to study the terms of exchange for a commodity or a group of commodities with respect to other commodities. For instance, suppose there are two groups of commodities X and Y, then

$$\text{Terms of exchange of X with Y} = \frac{\text{Price index of X}}{\text{Price index of Y}}$$

In international trade, one talks of the *terms of trade*. This is defined as the ratio of index of import prices to the index of export prices. This index ratio gives an idea of the amount of imports fetched by a unit export. A country is said to have favourable terms of trade if its export prices are lower than import prices.

4.3 Problems in the construction of index numbers

At each step in the construction of an index number, there are questions and problems to be resolved. Some of the major problems are discussed below:

4.3.1 Sampling

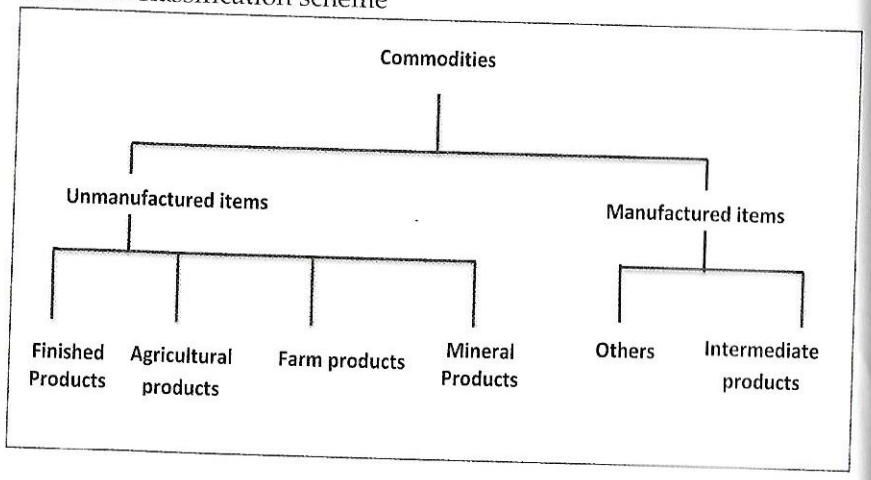
In preparing any index number it is not possible to include all the commodities or take cognizance of all the commodity prices that

prevail in all regions and times for which the index is being prepared. Hence one has to confine oneself to a sample of commodities and sample price quotations.

In case of commodities, one has to decide on the number of commodities [also which commodities are to be considered]. The number of commodities is a matter of personal discretion of the index maker and nothing more can be said than that it should be neither too large as to make the index number inconvenient nor too small to render the index unrepresentative.

The decision on the commodities to be included is governed mainly by the nature and purpose of the index. For instance, in constructing the cost of living index one is not likely to include TV's, refrigerators, etc. One may include a larger number of items in a wholesale price index than in a regional cost of living index. Generally, one classifies the commodities into various groups and subgroups and selects a few from each of them so as not to miss any important item. The classification scheme is as follows:

Table 4.4 Classification scheme



Besides *representativeness*, the products selected should be those which tend to be stable in consumption and quality.

4.3.2 Base period:

As we have already seen, index numbers can be calculated using a fixed base or a chain base. Chain-base series has the advantage of up-to-datedness. It permits the addition of new commodities, the removal of outdated or obsolete commodities, the change in the relative weights as the relative importance of the commodities changes over time and so on. However, if any computational error occurs at any point in the chain it will be carried forward cumulatively through the rest of the chain. Fixed base indices are, on the other hand, easier to calculate and errors do not accumulate. However, after some time the base period becomes remote and ceases to be of interest. Also if there are significant changes in consumption patterns warranting addition and deletion of commodities, the series will have to be recomputed. Despite the drawbacks, fixed base indices are frequently used. In this case, the choice of the base period has to be done with care. The period chosen should be quite recent and must be normal in the sense that it must not be characterized by unusual or irregular phenomena like war, families, floods, inflations, depressions etc. When the values of the variable(s) are either too high or too low index numbers for the subsequent periods will look distorted. It should also preferably be a period depicting some landmark in the country's economic or political history. For instance in the case of Zambia 1964 the year of Independence or 1991 the year of change from the second republic UNIP government to the third republic MMD government, could be chosen as a base period.

Sometimes, when it is difficult to find a proper base period, the average value of a few periods (which coincide with a full cycle in the time series) may be used as the base period value.

4.3.3 Weighting:

As already stated commodities are of unequal importance and hence have to be weighted appropriately. In preparing price index numbers, one may use quantities as weights; the quantities may be base year quantities, current year quantities, average of base and current y year quantities, average quantities of several

periods, etc. Likewise quantity indices may adopt various types of prices as weights. These are instances of explicitly weighting. But taking advantage of the fact that most commodities come in different varieties or brands, weights can also be assigned implicitly.

The number of varieties of a commodity induced in the index is determined by its importance vis-a-vis other commodities. For example, suppose weights are to be assigned to say rice and wheat in the ratio of 2:3. Then one may include two varieties of rice and three varieties of wheat.

4.3.4 Selection of average:

Though theoretically any measure of central tendency can be used, in practice, the arithmetic mean is most widely used owing to its computational simplicity. However, the geometric mean, though more difficult to calculate, has a special advantage in index number construction. Index numbers, as we have defined them are measures of comparative or relative changes and the geometric mean is an appropriate measure of relative changes. The following example will show the superiority of the geometric mean over the arithmetic mean.

Table 4.5 Arithmetic Vs Geometric mean

	Period 0		**Period 1**		**Period 2**	
	Price	Relative price	Price	Relative price	Price	Relative price
Good A	300	100	150	50	100	33.3
Good B	150	100	300	200	300	200
Index with Arithmetic mean		100		125		116.7
Index with Geometric mean		100		100		81.6

In period 1 compared to base period 0, the price of Good A has been halved and the price of Good B doubled. Hence, on the whole one should expect no change in the average price level. The Geometric index, therefore, continues to be 100 but the index using the arithmetic mean indicates a rise in the price level. Again in Period 2, the price of Good A has fallen to one-thirds its base year price while the price of Good B has doubled. One should, therefore, expect a fall in the price index which is seen in the case of index using Geometric mean. But the index based on the arithmetic average actually indicates a rise in the price level. This happens on account of the upward bias of the arithmetic mean,

4.3.5 Accuracy of index number formulae:

As we have seen, to the same price-quantity data, different formulae can be applied for obtaining the index number and they do not give the same values for the index. The question then arises as to which formula is the most satisfactory one, in the sense of giving accurate and consistent measures of changes. This is sought to be tested by applying two criteria called the Time Reversal Test and Factor Reversal Test.

The time reversal test states that the index number for period 1 calculated with period 0 as the base must be the reciprocal of each other. This is an obvious requirement. If the price level in period 1 is found to be double of the price level in period 0 by using a certain formula, then the same must reveal that the price level in period 0 is half of the price level in period 1. In other words, we expect

$$P_{01} = \frac{1}{P_{01}} \text{ or } P_{01} \times P_{10} = 1$$

The time reversal test can be extended to more than two periods. If $P_{12} = 200$ and $P_{02} = 200$ then we must have $P_{01} = 400$. This generalised version of the time reversal test is called circular test.

The factor reversal test states that the product of the Price Index and the Quantity Index must be equal to the corresponding value index. That is,

$$V_{01} = P_{01}\ X\ Q_{01}$$

Most of the index number formulae do not meet these criteria. Most of the most popular formulae which satisfies the time reversal test and the factor reversal test but not necessarily the circular test is the Fisher formula, which, for this reason has been called the *ideal index*. We shall prove below that the Fisher index satisfies both tests.

$$P_{01} = \sqrt{\frac{P_i^1 Q_i^0}{P_i^0 Q_i^0} \times \frac{P_i^1 Q_i^1}{P_i^0 Q_i^1}}$$

$$P_{10} = \sqrt{\frac{P_i^0 Q_i^1}{P_i^1 Q_i^1} \times \frac{P_i^0 Q_i^0}{P_i^1 Q_i^0}}$$

$$P_{10\times} P_{10} = \sqrt{\frac{P_i^1 Q_i^0}{P_i^0 Q_i^0} \times \frac{P_i^1 Q_i^1}{P_i^0 Q_i^1} \times \frac{P_i^0 Q_i^1}{P_i^1 Q_i^1} \times \frac{P_i^0 Q_i^0}{P_i^1 Q_i^0}}$$

$$= 1$$

Thus, the time reversal test is satisfied.

$$Q_{01} = \sqrt{\frac{P_i^0 Q_i^1}{P_i^0 Q_i^0} \times \frac{P_i^1 Q_i^1}{P_i^1 Q_i^0}}$$

$$P_{01\times} Q_{01} = \sqrt{\frac{P_i^1 Q_i^0}{P_i^0 Q_i^0} \times \frac{P_i^1 Q_i^1}{P_i^0 Q_i^1} \times \frac{P_i^0 Q_i^1}{P_i^0 Q_i^0} \times \frac{P_i^1 Q_i^1}{P_i^1 Q_i^0}}$$

$$= \sqrt{\left(\frac{P_i^1 Q_i^1}{P_i^0 Q_i^0}\right)^2}$$

$$= \frac{P_i^1 Q_i^1}{P_i^0 Q_i^0} = V_{01}$$

Thus the factor reversal test is satisfied.

4.4 Special topics

(a) Base shifting: When an index number series has been prepared for a fairly long period, the base period becomes remote and ceases to be of any interest, and hence the base needs to be shifted to a more recent period. For instance suppose indices have been constructed with 1980 as the base year. In 2015, few people, except those who are interested in history would like to compare say, current prices with 1980 prices. Hence, the base may be shifted to say 2005 or 2010. This shifting of the base is a simple matter. Suppose the base is shifted to 2005. Each index number in the series with 1980 as the base will be divided by the index for 2005 and multiplied by 100. The resulting series of index numbers will be those with 2005 as the base.

Consider the following hypothetical data.

Table 4.6 Base shifting

Year	Index (Base = 1980)	Index (Base = 2005)
1980	100	(100/408) X 100 = 14.7
1981	102	(102/408) X 100 = 25.0
.	.	
.	.	
2005	408	
2006	409	(409/408) X 100 = 100.2
.	.	
.	.	
2014	438	(438/408) X 100 = 107.4
2015	440	(440/408) X 100 = 108

(b) Splicing: Suppose we have two series of index numbers, the second series having its base period in the year in which the first series ends. The second series could have been constructed, for instance, using revised weights of the commodities. Let us say we

have one series from 2000 to 2005 and the other from 2005 onwards. 2005 may be termed the *year of overlap*. The procedure for converting the two series into one continuous series is called *splicing*. This may be done in one of the two ways. Either the new series may be made continuous with the old series or the old series may be made continuous with the new series. This virtually amounts to either computing the values of the new series with the base of the old series or shifting the base of the old series to that of the new series.

Consider the following illustration.

Table 4.7 Splicing indices

Year	Old index series	New index series
1980	100	Y_{80}
1981	102	Y_{81}
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
2005	200	100
2006	X_{06}	105
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
2015	X_{15}	190

To link the new series with the old, we have to obtain the values $X_{06}, X_{07}, \ldots\ldots X_{15}$. This can be done simply by multiplying the index for each of these years in the new series by the ratio of the old index value to the new index value in the year of overlap. For example,

$$X_{06} = 105 \times \frac{200}{100} = 210$$

To link the old series with the new, we have to obtain the values $Y_{80}, Y_{81}\ldots\ldots, Y_{04}$. These can be obtained by multiplying the index for each of these years in the old series by the ratio of the new index value to the old index value in the year of overlap. For example,

$$Y_{80} = 100 \times \frac{100}{200} = 50.$$

CHAPTER 5

5 Statistical quality control

5.1 Introduction:

Statistical quality control, popularly abbreviated as SQC, refers to the use of statistical methods in controlling the quality of manufactured products from the technical and economic points of view. The importance of SQC in recent times has been steadily increasing due to a number of factors such as high specialization of labour, increase in the precision and complexity of products, greater discernment on the part of consumers especially in monopsonistic or near-monopsonistic situations, etc.

The term "quality" is an omnibus expression for the important characteristics of any given product such as colour, density, diameter, length, breadth etc. The control of quality implies specification of the quality characteristics of the product in objective, quantitative terms and ensuring that items of the product conform to those specifications so that the product may perform its intended function. For example, if the products are component parts, they must fit properly when assembled, final products must operate satisfactorily, and so on.

Statistical control techniques are of two types:

1. *Process control* which refers to the control of the production processes by which items are manufactured, by making necessary adjustments and corrections therein, so that scrap or defective items are minimised.
2. *Acceptance sampling* which is a post-production phenomenon and refers to the control of the quality of output through inspection usually of a sample whereby

no more than a certain percentage of defective items is allowed to pass.

We shall first deal with process control techniques and then with acceptance sampling plans.

5.2 The statistical nature of production processes:

'No two human beings are exactly alike, not even twins' is a statement whose truth is quite obvious. It is not, however, equally obvious if one were to state that 'No two items of a product are exactly alike' especially when the product is completely standardized. Are not two cakes of soap that we buy in the market at the same price exactly alike? Are not two nuts or bolts produced in a factory exactly alike? If one were to be extremely fastidious in one's measurements, the probability of being able to answer these questions in the affirmative would be very close to, if not exactly, zero. Variations in a product, in other words, can never be totally eliminated. If a production process has been designed to produce screws the diameter of whose heads is 2 cm the probability that the process will turn out a screw with diameter exactly equal to 2 cm is very low. But on the other hand, the probability that the process will turn out a screw with diameter 1.8 cm or 2.2 cm is also practically nil. And indeed, if the process did produce a screw with head diameter of 2.2 cm, then one can almost surely conclude that something is wrong with the process and, therefore, needs to be rectified.

Any production process that is designed to produce a product according to certain specifications is mostly likely to produce units of the product that deviate from the given specifications. So long as the deviations are marginal or imperceptible, it does not matter as the units of the product may still perform the intended function satisfactorily. But if deviations are significant, the product will not be able to perform satisfactorily and will, therefore, be 'defective.'

The task of process control is to stipulate a tolerance range within which the measurements of the product must fall so that the good performance of the product is not inhibited.

Variations in the quality of a product (i.e. from its prefixed quantitative specifications) are due to two types of causes:

(i) Assignable causes such as men, materials, equipment and other input factors. Variations due to these causes can be detected as they will tend to be unidirectional towards poorer quality, e.g. a worker who is out of mood, a tool that is worn out, etc. An unpredictable significant variation from product specifications is proof of the presence of a non-random, assignable cause. Such causes can be traced and eliminated.

(ii) Random causes are those which cannot be traced individually but are inherently associated with the manufacturing process. They do not produce any systematic bias in the measured characteristic of the output as an assignable cause does. The variations produced by them are small and, being random, fall within predictable limits.

Suppose all assignable causes have been discovered and removed and output is being produced according to the inherent capabilities of the process. If variations continue to be outside the tolerance range, clearly the process itself is deficient and needs to be redesigned. If, however, the process design is known to be all right, then the process is said to be *in control* if no assignable cause is present. If some assignable cause is present, the process is *out of control.*

5.3 The theory of control charts

A control chart also known as the Shewart chart after Dr W.A. Shewart who developed them in 1920s is a device with a threefold purpose;

1. To describe in concrete terms what a state of control is;
2. To attain control;
3. To ascertain whether control has been attained.

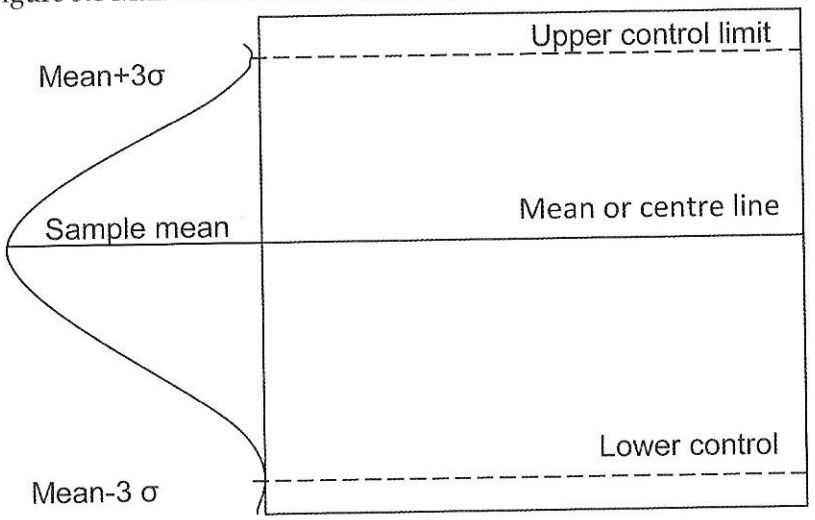
The modus operandi of setting up a control chart commences with the drawing of random samples from the production process at regular intervals. Then some statistic of interest such as the sample mean, sample range, proportion of defectives in the sample, etc. is calculated from each sample. If no assignable

causes are present and variations are purely random, then variations in the statistic will follow a statistical distribution which can be suggested by theory. For purposes of this illustration, suppose the sampling distribution of the statistic is assumed to be normal. From the mean of the samples one can estimate the mean of the distribution from the sample standard deviations, the standard deviation of the distribution. We know that three standard deviation or σ distances on either side of the mean cover over 99% of the distribution. Plus 3σ and minus 3σ from the distribution mean of centre line provide the upper control limit and the lower control limit of the control chart.

If our null hypothesis that the process is in control is true, then we can expect each of the individual sample means to fall strictly within control limits. If any sample mean falls outside the control limits, it is an indication of the presence of an assignable cause and the unacceptability of our null hypothesis.

The above theory underlying a control chart is graphically demonstrated in figure 5.1. There are different types of control charts for variables (where the quality characteristic such as length, weight etc. can be measured) and for attributes (where the quality is either not measurable or cost of measurement is too high and can therefore only say whether the quality is satisfactory or unsatisfactory).

Figure 5.1 Main features of a Control chart



5.3.1 Control chart for variables

For variables, there are two types of control charts, the X-chart for controlling the mean of the process and the R-chart for controlling the variability of the process. The two charts together give a fairly good idea of the control of the process.

5.3.1.1 $\overline{X}$ *- Chart*

This chart shows the fluctuations in the means of the samples about the mean of the process. The process mean is obtained by averaging the sample means. That is, process mean;

$$\overline{\overline{X}} = \frac{\overline{X}1 + \overline{X}2 + ...\overline{X}K}{K}$$

where K is the number of samples. K is usually about 25 and the samples are drawn when no assignable cause is present.

$$UCL = \overline{\overline{X}} + 3\hat{\sigma}(\bar{x})$$

$$LCL = \overline{\overline{X}} - 3\hat{\sigma}(\bar{x})$$

Where $\hat{\sigma}(\bar{x}) = \frac{\hat{\sigma}}{\sqrt{n}}$ and $\hat{\sigma} = \sqrt{\bar{S}^2 \frac{n}{n-1}}$ $\bar{S}^2$ being average of all sample variances or $\hat{\sigma} = \frac{\bar{S}}{C_2}$ where $C_2 = \sqrt{\frac{2}{n}} \frac{\left(\frac{n-2}{2}\right)}{\left(\frac{n-3}{3}\right)}$

We then have

$$UCL = \bar{\bar{X}} + 3\frac{\bar{S}/C_2}{\sqrt{n}} = \bar{\bar{X}} + \frac{3\bar{S}}{C_2\sqrt{n}}$$

$$LCL = \bar{\bar{X}} - \frac{3\bar{S}}{C_2\sqrt{n}}$$

Letting $A_1 = \frac{3\bar{S}}{C_2\sqrt{n}}$

$$UCL = \bar{\bar{X}} + A_1\bar{S}$$

$$LCL = \bar{\bar{X}} - A_1\bar{S}$$

Since A_1 depend only on n, the values of A_1 have been tabulated for different values of n, the sample size.

In summary,

$$\boxed{\text{Central line} = \bar{\bar{X}} = \Sigma \bar{X}/K}$$

Where K = number of samples

$$\bar{X} = \Sigma X/n$$

Where X= individual observation and n = sample size

$$\boxed{\text{Control limit} = \bar{\bar{X}} \pm A_1\bar{S}}$$

Where A_1 = tabulated values

$$\bar{S} = \Sigma S/K$$

and $S = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}}$

$K, X, n, \bar{X},$ have the previous meanings.

The number of control charts required in any SQC programme is usually large and it is therefore desirable to save time in the computations. Hence in setting up the control limit for $\bar{X}-Chart$, estimate of the $\hat{\sigma}$ may be obtained on the basis of sample ranges which are much easier to calculate than sample standard deviations. The relation between the standard deviation of the population and sample ranges is given by

$$\hat{\sigma} = \frac{\bar{R}}{d_2}$$

Where $\bar{R}$ is the mean of the sample ranges and the values of d_2 are tabulated for various values of sample size, n.

It must, of course, be noted that this estimate of $\hat{\sigma}$ based on $\bar{R}$ will not be as good as the estimate based on $\bar{S}$ in the sense that, on the average, $\frac{\bar{S}}{C_2}$ will lie nearer to $\hat{\sigma}$ than $\frac{\bar{R}}{d_2}$

Now, the upper control limit, $UCL = \bar{\bar{X}} + \frac{3\frac{\bar{R}}{d_2}}{\sqrt{n}}$

$$= \bar{\bar{X}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$

Letting $A_2 = \frac{3\bar{R}}{d_2\sqrt{n}}$

$$UCL = \bar{\bar{X}} + A_2\bar{R}$$

Where A_2 values which depend only on n are found in tables.

In summary,

Centre line $= \bar{\bar{X}} = (\sum \bar{X})/K$
Control limits $= \bar{\bar{X}} \pm A_2 \bar{R}$

Where A_2 values are listed in the tables.

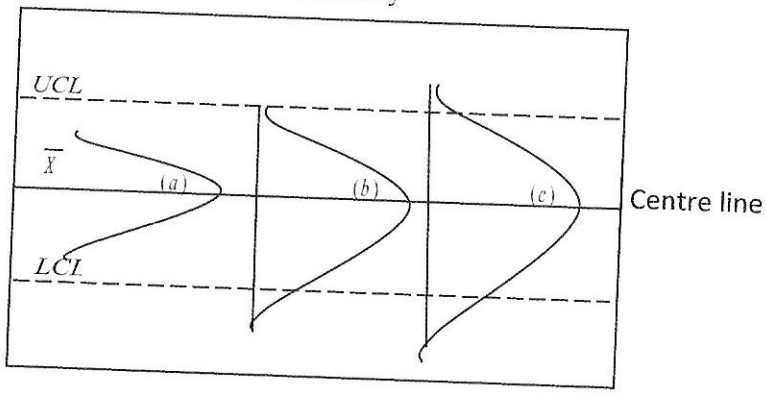
$$\bar{R} = \sum R / K$$

And R = difference between the largest and smallest observations.

5.3.1.2 R-Charts

The control chart for the process mean assumes that the process variation remains constant. But it is possible that the variability in the process may change while the mean remains the same. This is illustrated in figure 5.2:

Figure 5.2 $\bar{X}$ - Chart with Variability



In the above diagram, only samples that are drawn from distribution (c) can give the value of X outside the control limits, but this is not the efficient method to discover shifts in the process variability. It is thus clear that a process can go out of control without the mean going out of control. In such cases, a control chart that reveals process variations is a better warning signal

than the $\bar{X} - Chart$. The R-Chart is one such chart. It shows the fluctuations of ranges of samples about the mean range $\bar{R}$. The control limits are given by $3\hat{\sigma}R$ where $\hat{\sigma}R$ is the population standard deviation of the range.

The sampling distribution of the range is, however, not normal. The lower control limits for many samples could, therefore, be negative. But since the range, by its definition can never be negative, the LCL is set at zero in such situations. There is in other words, virtually no LCL in such cases.

$$UCL = \bar{R} + 3\hat{\sigma}R$$

$$LCL = \bar{R} - 3\hat{\sigma}R$$

An estimate of $\hat{\sigma}R$ is given by taking

$$\hat{\sigma}R = d_2\hat{\sigma} = d_3\frac{\bar{R}}{d_3}$$

The values of d's for various values of sample sizes are given in tables. Then

$$UCL = \bar{R} + 3d_3\frac{\bar{R}}{d_3} = (1 + 3\frac{d_3}{d_2})\bar{R}$$

Letting $D_4 = 1 + 3\frac{d3}{d2}$, we get $UCL = D4\,\bar{R}$

And likewise letting, $LCL = \bar{R} - 3\,d3\frac{\bar{R}}{d2} = (1 - 3\frac{d3}{d2})\bar{R}$

And also letting $D3 = 1 - 3\frac{d3}{d2}$, we get $UCL = D3\,\bar{R}$

The values of D_3 and D_4 which depend only on n are listed in the tables.

In summary,

Centre line $\bar{R} = \sum R/K$
Control limits : $UCL = D4\,\bar{R}$

$$LCL = D3\,\bar{R} \text{ if } D3\,\bar{R} \geq 0$$
$$= 0 \text{ if } D_3\,\bar{R} < 0$$

Consider the following table:

Table 5.1 Computation of $\bar{X}$ and R Statistics

Sample No.	Observation No.					Statistics	
	1	2	3	4	5	$\bar{X}$	R
1	45	30	43	35	32	37	15
2	31	34	34	35	35	33.8	4
3	33	32	14	26	24	25.8	19
4	18	36	30	28	21	26.6	18
5	45	35	20	38	39	35.4	25
6	47	26	37	34	40	30.8	37
7	0.29	46	25	22	37	31.8	24
8	33	29	33	28	13	27.2	20
9	42	18	30	11	23	24.8	31
10	18	37	34	26	34	29.8	19
11	31	29	52	39	11	32.4	41
12	18	24	20	29	30	24.2	12
13	19	1	27	27	28	20.4	27
14	32	39	17	25	29	28.8	22
15	24	21	37	30	32	28.8	16
16	35	16	22	24	25	24.4	19
17	28	21	23	39	41	30.4	20
18	22	13	11	45	31	24.4	34
19	42	43	33	29	47	38.8	18
20	25	44	32	24	36	32.2	20

N.B.: The numbers in the sample observations correspond to the last two digits in the measurement of a product whose desired dimension is 5136. e.g. 32 is to be read as 5132.

For the $\bar{X}$ - Chart,

Centre line $= \bar{\bar{X}} = (\Sigma \bar{X})/K = \dfrac{587.8}{20} = 29.39, \quad UCL = \bar{\bar{X}} + A_2\bar{R}$

and $LCL = \bar{\bar{X}} - A_2\bar{R}$.

From the tables, the value of A_2 for n=5 is found to be 0.577

$$\bar{R} = \Sigma R/K = \frac{441}{20} = 22.05$$
$$A_2\bar{R} = (0.577)(22.05) = 12.72$$

Therefore,

$$UCL(\bar{X}) = 29.39 + 12.72 = 42.11$$
$$LCL(\bar{X}) = 29.39 - 12.72 = 16.67$$

For the R-chart,
Centre line= R =22.05

$$UCL(R) = D_4\bar{R}$$
$$LCL(R) = D_3\bar{R}$$

From the tables, for n=5, we get

$$D_4\bar{R} = 2.115, \; D_3\bar{R} = 0$$
$$UCL(R) = (2.115)(22.05) = 46.64$$
$$LCL(R) = (0)(22.05) = 0$$

From the following 5.3 it is clear that the process is in control.

Figure 5.3 Process in control

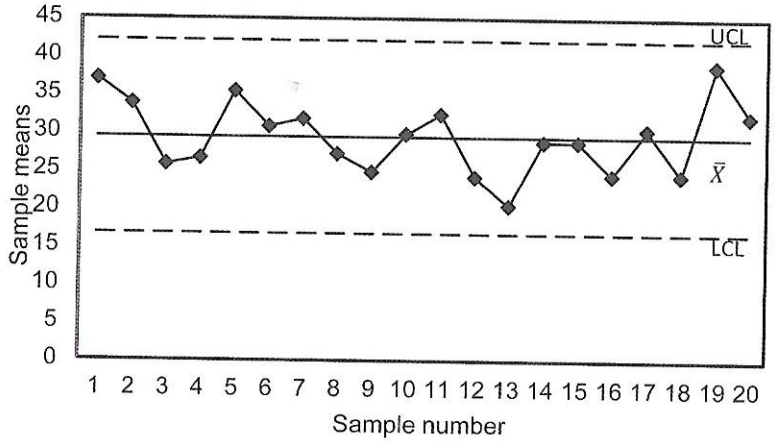
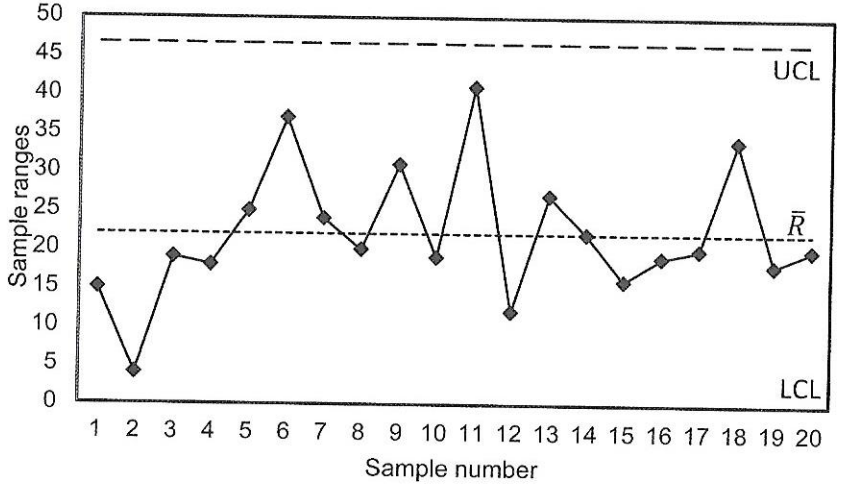


Figure 5.4 Process in control



Consider now the following data for which an X-chart and R-chart are to be prepared. Each sample consisted of five observations.

Sample No.	1	2	3	4	5	6	7	8	9	10
$\bar{X}$	25	38	13	37	24	28	45	33	37	20
R	12	40	21	39	18	16	15	14	2	23

For the $\bar{X}$ - Chart ,

$$\text{Centre line} = \bar{\bar{X}} = \Sigma \bar{X}/K = \frac{300}{10} = 30$$

$$UCL = \bar{\bar{X}} + A_2 \bar{R}$$

$$LCL = \bar{\bar{X}} - A_2 \bar{R}$$

$$\bar{R} = \Sigma R/K = \frac{200}{10} = 20$$

$$A_2 \bar{R} = (0.577)(20) = 11.54$$

$$UCL(\bar{X}) = 30 + 11.54 = 41.54$$

$$LCL(\bar{X}) = 30 - 11.54 = 18.46$$

For the R-Chart,

$$\text{Centre line} = \bar{R} = 20$$

$$UCL(R) = D_4 \bar{R} = (2.115)(20) = 42.3$$

$$LCL(R) = D_3 \bar{R} = 0$$

The $\bar{X}$- and R charts are drawn in figure 5.5:

Figure 5.5 Process out of control

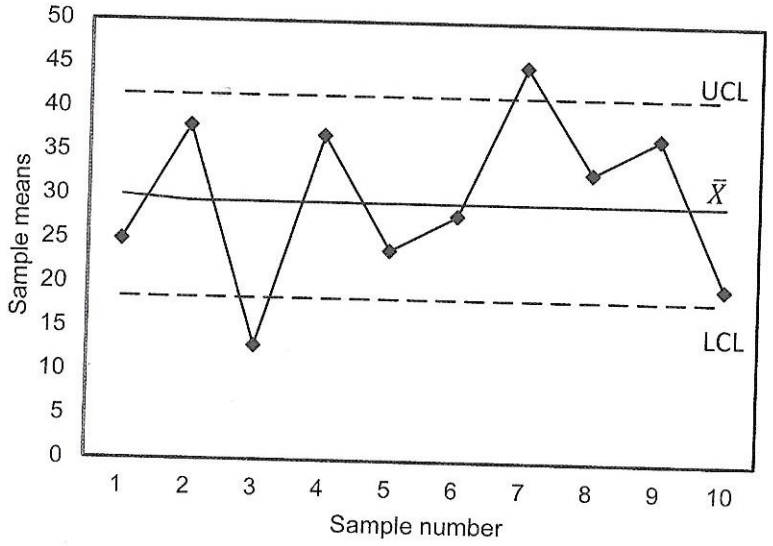
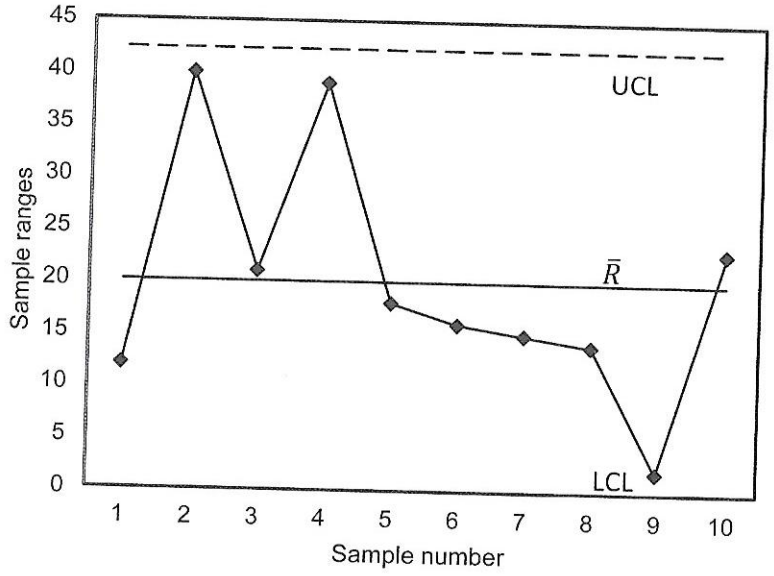


Figure 5.6 Process out of control



It can be seen that while the R chart shows the process variation

to be in control, the $\bar{X}$-Chart shows that the process mean is out of control since one of the sample means lies above the $UCL(\bar{X})$ and another falls below the $LCL(\bar{X})$. We have therefore to eliminate these two sample points and re-compute the control limits.

$$\text{Centre line} = \bar{\bar{X}} = \Sigma \bar{X}/K = \frac{242}{8} = 30.25$$

$$UCL = \bar{\bar{X}} + A_2\bar{R} = 30.25 + 11.54 = 41.9$$

$$LCL = \bar{\bar{X}} - A_2\bar{R} = 30.25 - 11.54 = 18.74$$

None of the remaining points now fall outside these control limits and so the process may be assumed to be in control.

5.3.2 Control charts for attributes

As already stated, in many cases, the quality characteristic cannot be quantitatively measured and can only be expressed as being defective or non-defective or as meeting or not meeting specifications. In such cases, the sample output is described by the binomial distribution dependent only on the number or proportion of defectives and the sample size.

a) The d-chart or np-chart

The probability of any number of defectives, d, in a sample of size n, is

$p(d) = {}^{n}C_d\,(p)^d\,(q)^{n-d}$ where $q = 1-p$ and p is the process average fraction defective and estimated by taking it equal to $\frac{d}{n}$.

We then test the hypothesis that p is a random sample drawn from a population for which the fraction defective is some expected value p'. This chart has as the centre line, $\bar{d}$ (or $\overline{np}$), the average number of defectives in all the samples. The control limits are given by $\bar{d} + 3\hat{\sigma}R$. Since d has a binomial distribution, its

standard deviation is $\sqrt{npq}$. However, in practice when n is large, the binomial is approximated by the normal and the standard deviation of d is taken as $\sqrt{n\bar{p}\bar{q}}$ where $\bar{p}=\frac{d}{n}$ and $\bar{q}=1-\bar{p}$. In summary,

Centre line= $\bar{d}=\sum d/K$

Where d = number of defectives per sample and K =number of samples.

Control limits= $\bar{d}\pm 3\sqrt{n\bar{p}\bar{q}}$

where n=sample size, and $\bar{p}=\frac{d}{n}$, $\bar{q}=1-\bar{p}$.

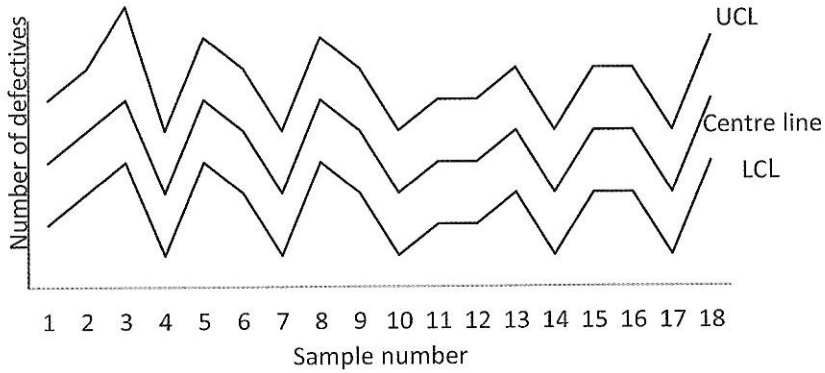
Very often in problems dealing with fraction defective, the sample size varies. But with unequal sized samples, the d-chart becomes unwieldy as the centre line d will not be a straight line but varies with the sample size.

We will have $\bar{d}=n\bar{p}$, where $\bar{p}=\frac{\sum d}{n}$

Control limits will be $=n\bar{p}\pm 3\sqrt{n\bar{p}\bar{q}}$

The d-chart will then appear something like the following figure:

Figure 5.7 The d-chart



(b) The p-chart

The unwieldiness of the d-chart when sample sizes vary is greatly mitigated by the p-chart which is based on the proportion rather than the number of defectives and hence is also more meaningful. If the samples are of equal sizes, the centre line of the p-chart is given by

$\bar{p}=\frac{\sum d}{Kn}$ and to obtain the control limits, 3 times the standard error of proportions, $\sigma_{\bar{p}}$ is used.

$$\sigma_{\bar{p}}=\frac{\sigma_{(p)}}{n}\text{ , where }\sigma_{(p)}=\frac{\sqrt{np'q'}}{n}=\sqrt{\frac{p'q'}{n}}$$

Again for large samples, we substitute $\bar{p}$ and $\bar{q}$ with p' and q'

In summary,

Centre line= $\bar{p}=\sum d/Kn$

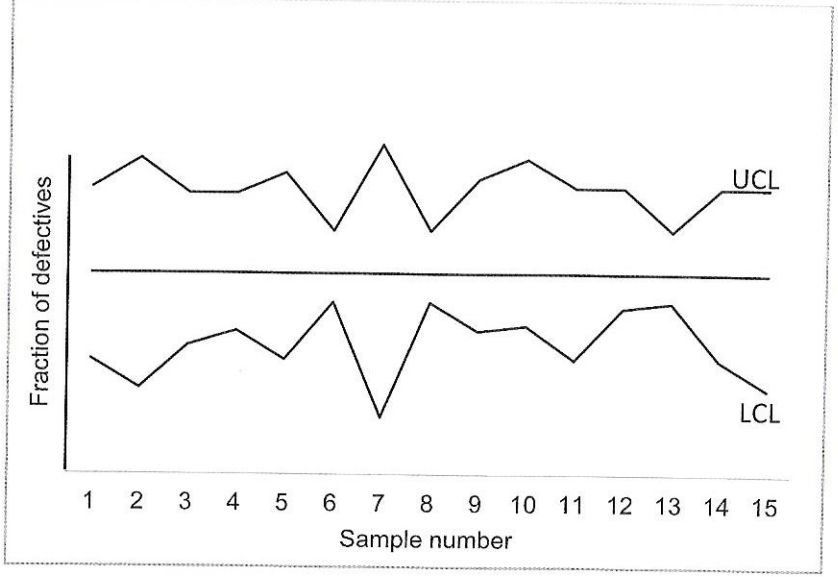
Control limits = $\bar{p}\pm 3\sqrt{\frac{n\bar{p}\bar{q}}{n}}$

If sample sizes vary, p is taken as $\frac{\sum d}{\sum n}$. The centre line thus remains constant. The control limits, however, will vary with the sample size as they are given by $\bar{p} \pm 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$.

N.B. Since $3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 3\sqrt{\bar{p}\bar{q}}/\sqrt{n}$, the numerator is constant for all samples hence needs to be computed just once.

The figure of a p-chart with varying sample sizes will look like the following

Figure 5.8 The P-Chart



(c)The distinction between defect and defective and the C-chart:

'Defective' is a term applied to items which fail to conform to specifications in any of the characteristics. 'Defect' refers to a characteristic which does not conform to specifications. An item is defective if it contains at least one defect. E.g. a sheet of glass

which contains air bubbles is defective; a piece of cloth which has imperfections is defective; and so on. The air bubbles, the imperfections are the defects. In such situations, the sample size is unspecified, though presumably large, and one cannot count the number of defectives in the usual manner. To gauge the extent to which the product is defective, one has to arbitrarily divide it into a number of units and see how many units contain any defect. For instance one may think of a piece of cloth as consisting of so many squares of equal dimension, each square being checked to see if it is perfect or imperfect. The number of defects or imperfections is assumed to follow a Poisson distribution.

The C-chart is constructed for defects per unit. The centre line of the C- chart is $\bar{C}$, which is the mean number of defects in several units (usually 25 or more) i.e. $\bar{C} = \frac{\sum C}{Kn}$ where C is the number of defects per unit and n is the number of units. The control limits are given by

$$\bar{C} \pm 3\sqrt{\bar{C}}$$

Examples

(a) The following table shows the results of production and inspection of 100 induction coils per day for about a fortnight.

Table 5.2 Defective coils from production process

Day	No of defectives	Fraction of defective
1	Nil	0.00
2	6	0.06
3	2	0.02
4	Nil	0.00
5	1	0.01
6	2	0.02
7	Nil	0.00
8	2	0.02
9	5	0.05
10	2	0.02
11	Nil	0.00
12	2	0.02
13	3	0.03
14	4	0.04
15	1	0.01
Total	**30**	

For the p-chart,

$$\text{the centre line } \bar{p} = \Sigma d / K\, n = \frac{30}{1500} = 0.02$$

The control limits are given by $\bar{p} \pm 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$

$$\sqrt{\frac{\bar{p}\bar{q}}{n}} = \sqrt{\frac{(0.02)(0.98)}{100}} = \frac{\sqrt{0.0196}}{10} = \frac{0.14}{10} = 0.014$$

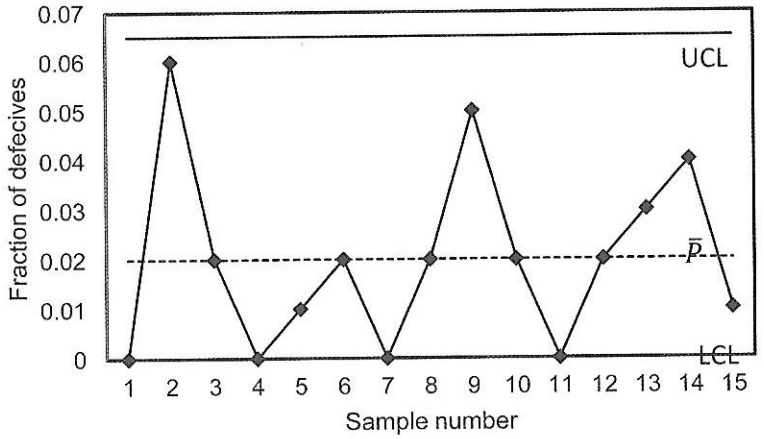
$$3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 3(0.014) = 0.042$$

$$UCL = 0.02 + 0.042 = 0.062$$

$LCL = 0.02 - 0.042$ which is taken as zero.

The following figure shows all the sample points within the control and hence the process is in control.

Figure 5.9 Sample points within a process control



(b) The following data shows the number of defects in 20 prices of cloth each of 100 metres

Sample No.	No of defects
1	1
2	3
3	3
4	1
5	6
6	4
7	3
8	7
9	10
10	2
11	2
12	6
13	4
14	3
15	2
16	7
17	1
18	5
19	6
20	4

The centre line of the C-chart is $\bar{C} = \sum d / K\, n = \frac{80}{20} = 40$

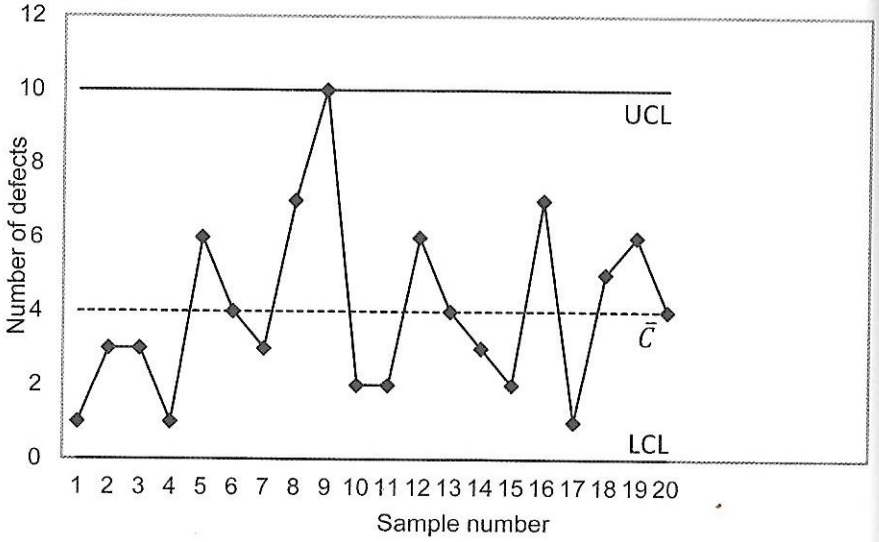
The control limits are given by $\bar{C} \pm 3\sqrt{\bar{C}}$

$$UCL = 10$$

$$LCL = 4 - 6 \text{ which is taken as zero}$$

The control chart drawn below, shows that the process is in control.

Figure 5.10 Process control



5.4 Acceptance Sampling:

As already mentioned at the outset, acceptance sampling, besides process control is the other major area of SQC. In this section, we shall present only the elementary ideas of acceptance sampling.

The purpose of acceptance sampling is quality assurance. It seeks to prescribe a method whereby a company which receives a consignment of goods can decide either to accept it as conforming to standards or to reject it as being below standard. In making this decision, it faces a risk of accepting lots of low quality or rejecting

lots of acceptable quality. Acceptance sampling enables it to specify the risk of accepting lots of any given quality,

Though the purpose of acceptance is primarily quality assurance and not quality estimation or quality control, it indirectly but effectively influences quality. A high rate of acceptance encourages good quality and a high rate of rejection strongly indicates bad quality. A supplier whose goods are being rejected at a high rate is bound to take immediate steps to step up quality. Often, the company itself may send experts to guide the supplier in solving problems of quality.

Acceptance sampling can either be in the nature of an 'attribute inspection ' which merely classifies a lot as being good or bad or it may be based on quantitative measurements. We shall discuss acceptance sampling by attributes only.

The inspection involved in acceptance sampling may range from no inspection to inspection through sampling to 100% inspection. The extent of inspection depends very much on economic considerations. If the cost of inspection is high or the loss involved in accepting a defective item is quite small or inspection is destructive there will be little or no inspection. In the opposite situations, inspection will be high. For example, if the consignment consists of say trucks or tanks or computers or some such products where every unit is highly expensive, one cannot take the risk of accepting even one defective unit and hence 100% inspection will be warranted.

A consumer in order to decide whether to accept or not a certain lot, inspects a sample of n units. He then selects a certain number N called the *acceptance number*. The null hypothesis H_0 is that the lot is acceptable. The decision rule then is:

Accept H_0 if $d \le N$, for the given sample size n.

Do not accept H_0 if $d > N$ for the given sample size n.

(d =number of defective units in the sample).

If H_0 is not accepted, there are two alternative courses of action available:

I. reject the lot—this is an *acceptance-rejection plan;*
II. remove the defective ones and charge the cost of additional inspection to the supplier—*this is an acceptance rectification plan.*

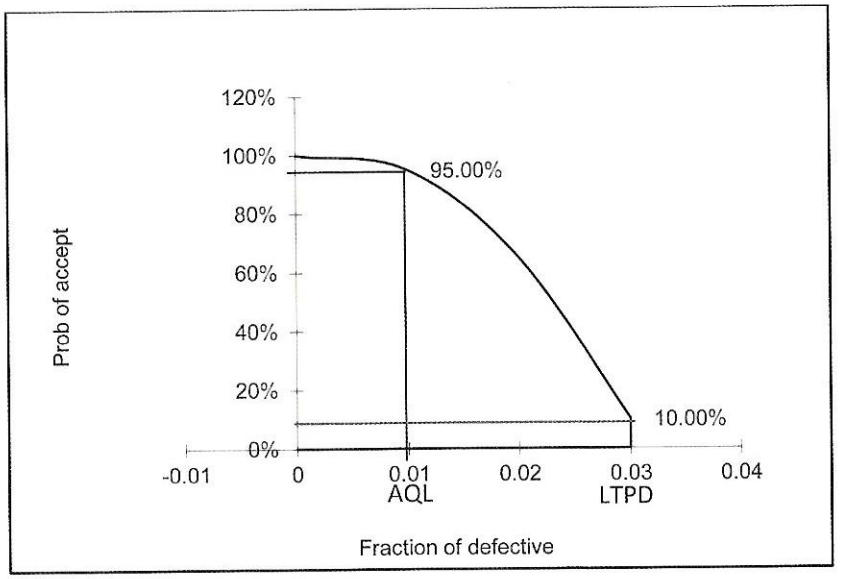
There are four bases for specifying a given sampling plan:

a. AQL or Acceptable Quality Level. Lots of this level are regarded as good.
b. α or Producer's risk. This is the probability that lots of quality level AQL will not be accepted.
c. LTPD or Lot Tolerance Per cent defective. This is the dividing line between good and bad lots. Lots at this level are of poor quality.
d. β or consumer's risk. This is the probability that lots of quality level LTPD will be accepted.
A good sampling would be one which gives a high probability of acceptance of AQL and other relatively good lots and a low probability of acceptance of LTPD and other relatively poor lots.

For specific values of AQL, LTPD, α and β, the probabilities of acceptance of lots with varying proportions of defectives are depicted by the operating characteristic or OC curve. The OC curve is drawn for a given sample size n and given acceptance number N. The probabilities used for constructing the OC curve are binomial or Poisson.

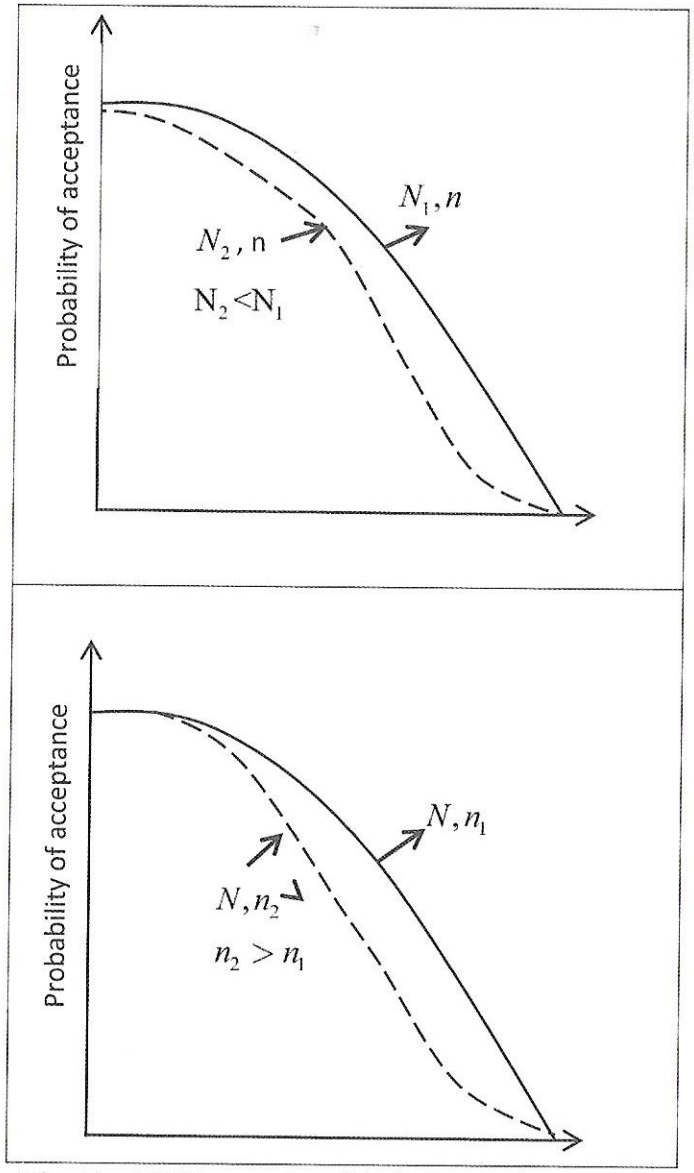
For example, suppose AQL=0.01, α =0.05, LTPD=0.03 and β =0.1. The OC curve will look like as under.

Figure 5.11 The OC curve



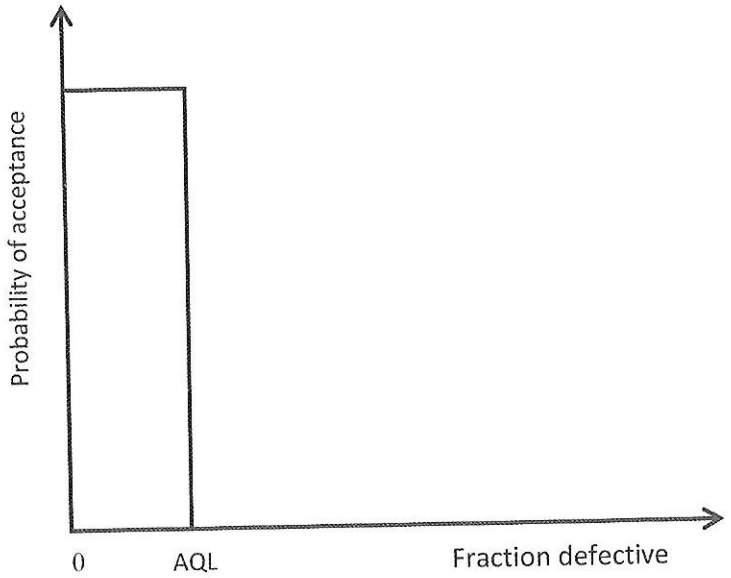
For a given N, larger value of n would enable a better discrimination between good and bad lots. For a given n, a smaller value of N will alter the level of the OC curve. These are depicted in the figures below.

Figure 5.12 OC Curve with varying n and N



The ideal OC curve would obviously be the following:

Figure 5.13 Ideal OC curve



CHAPTER 6

6 Zambian Statistics

6.1 The cardinal relevance of statistical data

It was the 19th century English physicist Lord Kelvin who stressed the need for objective data as a requirement for conducting any scientific enquiry. He wrote:

"When you can measure what you are speaking about, and express it in numbers, you know something about it; when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science..."

Since economics is a science, the above statement is very much applicable to it. It was the first Nobel prize winning economist Jan Tinbergen whose research programme in the 1930s was characterized by three components: selection of data, processing of data, and estimation of relations in the data. This facilitated the bridging of theory and facts, as represented in statistics. It marked the incipient stage of the rapid development of econometrics.

Tinbergen himself published a book in 1937 entitled "An Econometric Approach to Business Cycle Problems". Tinbergen showed three uses of this approach: *interpreting* the movements in variables; *extrapolating* the movements which involve calculating the movements if no exterior changes took place; and variation in movements when external conditions change. This is finding out the consequences of a *given policy*. One may also want to

determine the optimum variation, in which case one is interested in determining the *best policy*.

Since the time of Tinbergen, successive generations of development economists have stressed the importance of statistical data for development planning. It is for this reason that most countries the world over, including the emerging countries in Africa, have set up national centers for the production and dissemination of a wide range of statistical data. Such centers go by a variety of names. Examples of some of these in Africa are:

a. Bureau of Statistics (Gambia, Lesotho, Uganda)
b. Central Bureau of Statistics (Namibia)
c. National Bureau of Statistics (Kenya, Seychelles, Tanzania)
d. Central Statistical Office (Botswana, Mauritius, Swaziland, Zambia, Zimbabwe)
e. National Statistical Institute (Benin, Burkina Faso, Cape Verde, Cote D'Ivoire)
f. National Institute of Statistics (Cameroon, Madagascar, Mozambique, Rwanda)
g. National Institute for Statistics (Guinea, Guinea-Bissau)
h. National Statistical Office (Malawi, Mali, Mauritania, Niger)

The following mandate given to the Central Statistical Office of Botswana encapsulates the *raison d'être* for any national government statistical agency: "CSO is the principal data collecting, processing and disseminating agency responsible for coordinating, monitoring and supervising the National Statistical System. It thus has the statutory mandate to produce and provide government, the private sector, the parastatal organizations, international organizations, the civil society, and the general public with statistical information for evidence-based decision-making, policy formulation and planning as well as investment purposes. The statutory mandate also includes the responsibility

of providing advisory and technical service to all users on statistical matters."

6.2 Central Statistical Office, Zambia

Zambia's CSO is entrusted with the following mission: "To coordinate and provide timely, quality and credible official statistics for use by stakeholders and clients for sustainable development". Its goal statement is: To achieve an effective, efficient and coordinated National Statistical System (NSS) that will ensure sustainable production and disseminating of demand driven official statistics for national development".

The CSO produces statistics in four main areas – Economic, Social, Agricultural and Environmental. Its key partners are: Ministry of Finance, Ministry of Health, Ministry of Agriculture and Livestock, Bank of Zambia and Zambia Revenue Authority.

The CSO produces a variety of reports. Some of the major ones are:

- Census of Population and Housing Reports once every decade; the latest one to date is for 2010.
- Living Conditions Monitoring Survey; latest survey: 2010.
- Demographic and Health Survey; latest survey: 2013-14.
- Labor Force Survey; latest survey: 2012.
- Crop Forecast Survey; latest survey: 2012-2013.

In addition to the above reports, the CSO also brings out reports on specific themes from time to time. Examples of these are:

- Gender Statistics Report 2010;
- Micro Level Estimates of Poverty, 2007;
- National Accounts Bulletin 2005;
- Energy Statistics, 2007.

In addition the CSO also produces a Monthly Bulletin on inflation and changes in the CPI.

6.3 Utility of survey reports for policy and planning

6.3.1 Zambia Census of Population and Housing (ZCPH)

Zambia's first ZCPH was conducted in 1969. Since then it has been conducted at ten-year intervals in 1980, 1990, 2000 and 2010.

The census report provides detailed information on the demographic, social, education, economic and fertility characteristics of the Zambian population. It also contains data on mortality, language and ethnicity, and disability.

Data obtained from the Census can be used for planning in a number of ways. The following table shows some of the possible plan-related uses of data on population and housing:

Table 6.1 Census information and its uses

Census information	Potential uses
Total population size	When two or more census counts are compared for the same province/district/region, planners can determine if its population is increasing or decreasing in size
Age and sex	Useful for identifying segments of the population that require different types of services
Sex	Sex ratios can be calculated by 5-year age groups to crudely observe migration, especially among the working age cohorts
Marital status	Provides information on family formation and housing needs
Household composition and size	Useful to determine housing needs for related and unrelated households
Educational attainment and literacy	Provides information on the educational skills of the work force
Location of residence and place of prior residence	Helps assess changes in rural and urban areas. Place of prior residence helps to identify communities that are experiencing in-and out-migration
Occupation and labor force participation	Helps to provide insights into the labour force of a given area. The information can be used to develop economic development strategies
Living quarter characteristics	Helps planners determine housing and community facility needs

Source: MEASURE Evaluation – M&E Learning Center: Lesson 4: Use of Census and Related Population Information.

6.3.2 Living Conditions Monitoring Survey (LCMS)

This is a very comprehensive survey. It is a survey with a large sample of households drawn from both rural and urban areas in all the districts of Zambia. Its objectives are to monitor the effects of various government policies on the wellbeing of households and individuals, monitor levels and intensity of poverty, highlight the vulnerable groups in society, and to monitor change in the living conditions of the population over time.

The survey provides data on a large number of variables relating to demography and migration, orphan-hood, health and nutrition, economic activities, household income and expenditure, household amenities and housing conditions, household access to facilities, agricultural production, community development issues, death in households, self-assessed poverty, shocks to household welfare and household coping strategies.

The Living Conditions Monitoring Survey has been conducted since 1992 and there have been six reports until 2010. The large volume of data thrown up by the LCMS allows for both time series and cross section analyses of changes in living conditions over time and across provinces, districts and socioeconomic groups.

6.3.3 Zambia Demographic and Health Survey (ZDHS)

Since 1992, Zambia has been implementing a number of health reforms with a view to improving the health status of the

population. The series of Zambia Demographic and Health Survey undertaken since 1992 to date provide data to assess the achievements of the health reforms. In particular, there are several Millennium Development Goals relating to health that Zambia was expected to achieve by end 2015. The ZDHS data help to monitor the progress in the attainment of the MDH health targets – targets such as the infant and under-5 mortality rates, the maternal mortality ratio and reversal in the trends in diseases notably HIV and AIDS.

6.3.4 Labor Force Survey

This survey provides detailed information on the employment situation in the country. It is today widely recognized that GDP or economic growth can translate into meaningful development only if it is accompanied by a commensurate growth in employment characterized by decent conditions of work and pay. Jobless growth cannot translate into human development. Thus employment statistics becomes one of the crucial indicators of a country's economic performance.

The labor force survey in Zambia takes cognizance of the fact that more than formal employment, informal employment is highly predominant. A series of surveys have shown that in Zambia about one-tenth of the employed population is in formal employment while the remaining 90% are in informal employment. In the rural sector, informal employment constitutes 96% of the total employment, leaving only 4% in the formal sector.

6.3.5 Crop Forecast Survey

A Crop Forecast Survey obtains estimates from agricultural holdings on the area under major crops as well as production and sales estimates during the agricultural season. This information is useful for assessing the expected food security situation in the country and for producing the National Food Balance Sheet (NFBS). This NFBS in turn helps to determine the food surplus or

deficit situation in the country with respect to the major cereals and tubers produced in the country. This information is vital to all stakeholders interested in the country's agricultural development as it serves as a guide to their strategic planning and decision-making. Zambia's Crop Forecast Survey uses an internationally recognized methodology promoted by the Food and Agricultural Organization.

Practice Questions

Problem Set 1

Q1. Define and briefly explain/illustrate the following:

a. Slutsky-Yule effect;
b. Circular test for index numbers;
c. Operating characteristic curve;
d. Bootstrapping;
e. Delphi method

Q2. A company wants to select a sample of 1200 customers for a marketing survey in four districts. The following data are available:

Marketing region	Population frequency	Variability (standard deviation	Cost per unit (Zambian kwacha)
District 1	18000	4.3	18
District 2	600	6.4	10
District 3	12000	9.4	39
District 4	24000	7.1	24

a. How many persons should be selected from each district:

I. In proportion to the population frequency in each district;
II. In proportion to the population frequency and variability in each district;
III. In proportion to the population frequency, variability and cost per unit in each district?

b. Do you think proportional sampling is appropriate for this survey? Give reasons for your answer.

Q3. A researcher wants to study the living conditions of both male- and female-headed households in three provinces of

Zambia: Eastern, North-Western and Western. For this purpose, he wants to select a total sample of 100,000 households from the three provinces.

Data are provided by the Living Conditions Monitoring Survey 2010 as shown below:

Total population of households in the three provinces:
Eastern: 342,000; North-Western: 138,000; Western: 205,000

Percentage of female-headed households in the three provinces:
Eastern: 23.5%; North-Western: 23.8%; Western: 35.3%

How many female-headed- and male-headed households should be included in the sample?

Q4. Apply simple exponential smoothing to the series given below, using:

$\alpha = .4$; $\beta = .6$; $F_0 = 40$; $T_0 = 1$

Year	Y_t
2009	44.5
2010	41.8
2011	40.5
2012	44.0
2013	46.1
2014	48.0

Q5. The length of a certain product is specified to be 30 ± 1 units.

Sample averages and ranges are given below (sample size =5).

Sample No	Average	Range
1	30.7	1.1
2	30.2	0.9
3	30.4	0.5
4	30.3	1.0
5	30.1	0.2
6	30.8	0.7
7	30.2	0.3
8	30.5	1.2
9	30.0	0.5
10	30.6	0.5

Plot $\overline{X}$ and R charts and check for the control. Estimate the per cent defective of the process when in control at the level indicated by the above data. Can the per cent defective be reduced by changing the averages?(For n=5, $A_2 = 5.77, D_3 = 0, D_4 = 2.115$)

Problem Set 2

Q1. Suppose a systematic sample of 10 units is to be chosen from a population of 100 units. What sample would you obtain if:

a. The first randomly selected unit is 7;
b. The first randomly selected unit is 77?

Q2. In each of the following cases, what is the appropriate mix of sampling techniques that you would recommend?

a. A health researcher wants to study malaria control in Zambia. He intends to conduct the study in an area where the incidence of malaria is the highest. District Commissioners and other leaders in five districts in the area will be asked to select individuals for participation in focus groups. The focus group results will then be used to conduct a survey of 1400 households in three geographical zones.
b. A researcher wants to study the differences between government-run schools and private schools in Zambia in respect of students' participation in sports and other extra-curricular activities. Lusaka is chosen as the province of study. A list of government-run schools is obtained from the Ministry of Education and a list of private schools is obtained from the Independent Schools of Zambia, ISAZ. A total of 30 schools is chosen and five teachers and 10 students are selected randomly from each selected school to know their views on the school climate that determines participation in sports and other extra-curricular activities.
c. A researcher wants to understand how persons infected with hepatitis C make decisions regarding alcohol consumption. The study intends to use semi-structured interviews with hepatitis C patients who are being treated

in the University Teaching Hospital. In selecting the patients, one has to keep in mind variability across three variables: gender, race/ethnicity and level of alcohol use.

d. In the United States, a study was conducted of youth smoking cessation programs. Since most of the programs operate at the county level, the sample design started by looking at all the counties in the US. It was then decided to eliminate counties with populations under 10,000. This left 2500 counties available for the study. In selecting the sampling units, factors such as urbanization, socioeconomic status, youth smoking prevalence and tobacco control expenditures had to be considered. The research then initially used key informants to know about actual smoking cessation programs. These informants were then asked to provide contact information regarding others involved in youth tobacco cessation programs. This process was continued until no new names appeared on the lists. In this way, 1300 informants and 756 smoking cessation programs were obtained for the study.

Q3. Choose the right answer in each of the following statements:

a. Given a multiplicative time series model, calculating a moving average, and then dividing the model by the moving average, will isolate which components of the series?

A. T and C
B. T, C, and I
C. S and I
D. T, C, and S
E. None of the above

b. Suppose that the equation for a trend line to predict sales of new houses in Ibex Hill in Lusaka, where the time period is in years, is given by: $Y_t = 0.92t + 4.1$. Assuming

that year 1 is 2006, predict the sales for 2013. (Sales are in $100, 000s).

A. $1,146, 000
B. $1,054, 000
C. $1, 238,000
D. $ 962, 000

Q4. (a) An electronics company is selling portable CD players and its actual demand for three periods is: 54, 57 and 44 respectively. Using the exponential smoothing method, obtain the unadjusted and trend-adjusted forecasts for periods 2 & 3.

You are given:
Smoothing constant $\alpha = 0.2$; trend adjustment factor $\beta = 0.7$;
Unadjusted forecast for period 1 = 50; initial trend = 0

(b). A quality control inspector of a soft drink company has taken 25 samples of four observations each of the volume of bottles filled. The sum of the means of the 25 samples is found to be 398.75 and the standard deviation of the bottling operation is 0.14 ounces. Use this information to develop control limits for the X-bar chart of the bottling operation.

Problem Set 3

Q1. Suppose you have the following numbers: 225, 50, 350, 175, 200. Show that the arithmetic mean of the sampling distribution of means of all possible samples of size 3 drawn from this population of five numbers will be equal to the population mean. To what do we owe this result?

Q2. Fill in the blanks in the following sentences:

A. Demand for seats in a university is at its highest in the beginning of the academic year; demand also tends to grow and fall off every 20 years. In time series forecasting, the former demand characteristic would be called _____ and the latter would be called _____.

B. The method of forecasting where the value of the variable in the next time period is assumed to be equal to the present value of the variable is called _____ forecasting.

C. When a moving average series introduces cycles that are not present in the original data, it is called _____.

D. The method used to reduce the subjective biases in opinion polling is called _____.

Q3. The seasonal indices for the 12 months January to December for a company whose annual sales have remained more or less constant at ZMW 100 million are: 76, 93, 112, 83, 105, 108, 81, 88, 98, 130, 108 and 118 respectively. Estimate the company's monthly sales for February, March, July, August, November and December next year.

Q4. In the production of pig iron, percent of silicon is an important property. The following results were obtained on a series of consecutive cast in subgroups of 4.

Sample No	*Percent silicon*			
1	1.13	1.0	0.96	0.67
2	0.77	0.65	0.83	0.92
3	0.80	0.94	0.96	0.76
4	0.80	0.60	0.77	0.73
5	0.85	0.62	0.60	0.66
6	0.84	1.00	0.90	0.96
7	0.60	0.82	0.87	0.80
8	0.70	0.91	1.20	1.00
9	0.90	1.00	1.08	0.85
10	0.71	0.58	0.94	0.96

Construct $\overline{X}$ and R charts on the data. (For $n=4, A_2 = 0.73, D_3 = 0, D_4 = 2.28$).

Q5(a). At the end of a school year, a state wants to administer a reading test to 36 third graders. The school system has 20,000 third graders, half boys and half girls. The results of last year's test are as shown below.

Stratum	Mean score	Standard deviation
Boys	70	10.27
Girls	80	6.66

This year, the researchers plan to use a stratified sample, with one stratum consisting of boys and the other, girls. Use the results of last year to decide how many sampled students

should be boys and how many should be girls in order to maximize precision.

(b). What is the role of Zambia's Central Statistical Office? List some of its major publications and their utility for policy analysis and development planning.

Problem Set 4

Q1. You have learnt that that there are four components in a time series: trend, cycle, seasonal variations and irregular movements. In each of the following cases, identify the component(s) that will mainly influence the time series.

a. Daily fluctuations in stock price index;
b. Level of construction activity;
c. Hotel occupancy rates in Zambia's game parks;
d. Full-time employment levels;
e. Part-time employment levels;
f. Electricity consumption;
g. Monthly sales in a gift shop;
h. Changes in demographic patterns in a country;
i. Airline ticket sales;
j. Global warming;
k. Weather forecasts;
l. Demand for swimming pools;
m. Sub-prime mortgage lending;
n. Changes in cigarette consumption;
o. The effects of Zambia hosting the UNWTO General Assembly meeting in 2013 on Zambia's trade, transport and tourism.

Q2. State whether the following statements are true (T) or false (F):

I. When using the simple exponential smoothing method, all observations are given the same weight
II. A moving average of a time series is the value around which a series moves

Q3 Consider the following data:

Year	Production (in million tons)
2004	30
2005	33
2006	28
2007	35
2008	38
2009	40
2010	44
2011	47
2012	46
2013	52
2014	55

Fit a linear trend to the data by:

a. the method of moving averages, using a three-year period;
b. the method of semi-averages.
c. Suppose that the equation for a trend line to predict sales of new houses in Ibex Hill in Lusaka, where the time period is in years, is given by: $Y_t = 0.92t + 4.1$. Assuming that year 1 is 2006, predict the sales for 2013. (Sales are in $100, 000s).

E. $1,146, 000
F. $1,054, 000
G. $1, 238,000

$ 962, 000

Q4 Calculate the following:

a. Fisher's Price Index Number from the data given below:
$\Sigma P_0Q_0 = 1340$; $\Sigma P_0Q_1 = 1700$; $\Sigma P_1Q_0 = 1990$; $\Sigma P_1Q_1 = 2555$

b. The forecast revenue from the export of durable goods for 2015 by fitting a straight line trend by the method of semi-averages to the following data:

Year	Revenue (in $ million)
2006	394.9
2007	466.2
2008	481.2
2009	503.6
2010	569.2
2011	522.2
2012	491.2
2013	499.8
2014	556.1
2015	609.7

Problem Set 5

Q1 For each of the following, identify the sampling method(s) utilised:

a. All teachers with 25 years or more teaching experience are selected by the University of Zambia for the award of a long-service bonus.
b. The Southern African Institute for Policy Research wants to bring out a research document consisting of ten chapters. Ten experts are chosen to prepare the ten chapters and then each of those experts is asked to identify three other experts to edit the whole document.
c. A researcher studies his own children for literacy development.
d. Thirty percent of the students who graduated with a masters degree in economics from UNZA in the past ten years are randomly chosen for a tracer study.
e. Five Grade 7 classes are randomly chosen from all the primary schools in Chongwe district.
f. From the list of students registered for the course ECN 2342, every 10th student is selected.
g. Twenty-five percent of the teachers each from the Natural Sciences, Engineering and Agriculture schools in the University of Zambia are chosen randomly for a survey.
h. All the students in ECN 2342 were identified by number. A table of random numbers was then used to select 30 of these students for a sample.
i. 100 primary, 100 secondary and 100 higher secondary school teachers were selected by assigning I.D. numbers and using a table of random numbers.
j. A group of differently-abled children were selected to study their learning behavior.

Q2. Calculate the price index number using the Marshall-Edgeworth formula for the data given below (Base period = 2004):

Commodity	Price (per kg)		Quantity (1000' kg)	
	2004	2008	2004	2008
Wheat	14	22	40	60
Rice	12	18	25	35
Cotton	8	11	60	55

Q3(a). An industrial process fills containers with breakfast oats. The mean fill for the process is 510 grams and the standard deviation of fills is known to equal 5 grams. Four containers are selected every hour and the mean weight of the subgroup of four weights is used to monitor the process for special causes and to help to keep the process in statistical control. Find the lower and upper control limits for the X-bar control chart.

b. An experienced operator has operated a machine over several days. Twenty-five random samples of n = 6 were taken during this time. The sum of the ranges $\Sigma R = 4.270$ cm. Compute the control limits for the R chart.

You are given: for n = 6, $D_4 = 2.00$; $D_3 = 0.00$

Q4. Which of the following would you use to forecast the direction of change in business activity?

A. Number of new building permits issued;
B. Change in manufacturers' unfulfilled orders (durable goods industries);
C. Index of consumer expectations;
D. Change in labor cost per unit of output;
E. Unemployment rate;
F. Money supply;
G. Length of the average work week;
H. Industrial production;
I. Commercial loans;
J. Manufacturing inventories.

Subject Index

Q3 Consider the following data:

Year	Production (in million tons)
2004	30
2005	33
2006	28
2007	35
2008	38
2009	40
2010	44
2011	47
2012	46
2013	52
2014	55

Fit a linear trend to the data by:

a. the method of moving averages, using a three-year period;
b. the method of semi-averages.
c. Suppose that the equation for a trend line to predict sales of new houses in Ibex Hill in Lusaka, where the time period is in years, is given by: $Y_t = 0.92t + 4.1$. Assuming that year 1 is 2006, predict the sales for 2013. (Sales are in $100, 000s).

E. $1,146, 000
F. $1,054, 000
G. $1, 238,000

$ 962, 000

Q4 Calculate the following:

a. Fisher's Price Index Number from the data given below:
$\Sigma P_0Q_0 = 1340$; $\Sigma P_0Q_1 = 1700$; $\Sigma P_1Q_0 = 1990$; $\Sigma P_1Q_1 = 2555$

b. The forecast revenue from the export of durable goods for 2015 by fitting a straight line trend by the method of semi-averages to the following data:

Year	Revenue (in $ million)
2006	394.9
2007	466.2
2008	481.2
2009	503.6
2010	569.2
2011	522.2
2012	491.2
2013	499.8
2014	556.1
2015	609.7