Course on
Data Collection, Analysis and Validation

January 19-21, 2021

STATISTICAL INDIAN INSTITUTE

भिन्नेष्वैक्यस्य दर्शनम्

UNITY IN DIVERSITY

Sponsored by

cpcb

Indian Statistical Institute,

7, S. J. S Sansanwal Marg, New Delhi-110 016

Preface

This course is an introduction to *Statistical Methods for Data Collection, Analysis and Validation* with hands on training & case study and it aims to provide you to carry out statistical analysis of simple data sets and an exposure to use commonly available computer software packages. The development of such packages has made it much easier for everyone to apply statistics to their observations of the world.

The theory presented in the lectures will be supported by a series of tutorials/computer practical and case studies. These will put into practice what you have learned in the lectures. You shall be given a set of exercises each day which will be completed outside of the practical hour when necessary. Most of the exercises will be based on calculations that can be done without the aid of a computer; the answers to these problems will be discussed in the tutorial sessions.

After attending this course, we hope that the participants will be able to:

- do some basic statistical analysis of data.
- have the confidence to apply the techniques studied and perhaps slightly more advanced techniques.

January 19, 2021

S K Neogy
Professor & Course Coordinator

CONTENTS

CHAPTER – 1

INTRODUCTION TO ENVIRONMENTAL STATISTICS

Environmental statistics is a branch of statistics, which has developed rapidly over the past 10-15 years, in response to an increasing concern among individuals, organizations and governments for protecting the environment. It differs from other applications topics (e.g. industrial statistics, medical statistics) in the very wide range of emphases, models and methods needed to encompass such broad fields as conservation, pollution, evaluation and control, monitoring of ecosystems, management of resources, climate change, the greenhouse effect, forests, fisheries, agriculture and food. It is also placing demands on the statisticians to develop new approaches or new methods (e.g. for sampling when observations are expensive or elusive or when we have specific information to take into account) as well as to adapt the whole range of existing statistical methodology to the challenges of the new environmental fields of application.

Environmental statistics is indeed becoming a major, high-profile, identified theme in most of the countries where statistical analysis and research are constantly advancing our understanding of the world we live in. Its growing prominence is evident in a wide range of relevant emphases throughout the world.

Other expressions of concern for environmental statistics are found in the growing involvement of national statistical societies, such as the Royal Statistical Society in the UK and the American Statistical Association, in featuring the subject in their journals and in their organizational structure. Other nations also express commitment to the quantitative study of the environmental change networks and with governmental controls and standards on environmental emissions and effects. Many universities throughout the world are identifying environmental statistics within their portfolios of applications in statistical research, education and training.

Of course, concern for quantitative study of environmental issues is not a new thrust. This is evidenced by the many individuals and organizations that have for a long time been involved in all (including the statistical) aspects of monitoring, investigating and proposing policy in this area. These include health and safety organizations; standards bodies; research institutes; water and river authorities, meteorological organizations; fisheries protection agencies; risk, pollution, regulation and control concerns, and so on.

Such bodies are demanding more and more provision of sound statistical data, knowledge and methods at all levels (from basic data collection and sampling to specific methodological and analytic procedures). The statistician is of course ideally placed to represent the issues of uncertainty and variation inevitably found in all environmental problems. An interesting case in point was in relation to the representation of uncertainty and variation in the setting of environmental pollution standards.

Environmental statistics is thus taking its place besides other directed specialties; medical statistics, econometrics, industrial statistics, psychometrics, etc. It is identifying clear fields of application, such as pollution, utilities, quality of life, radiation hazard, climate change, resource management, and standards. All areas of statistical modeling and methodology arise in environmental studies, but particular challenges exist in certain areas such as official statistics, spatial and temporal modeling and sampling. Environmentally concerned statisticians must be pleased to note the growing public and political acceptance of their role in the environmental debate.

Many areas of statistical methodology and modeling find application in environmental problems. Particular modern sampling methods have special relevance and potential in many fields of environmental study ; they are important in monitoring and in standard setting. For example, ranked-set sampling aims for high efficiency inference, where observational data are expensive, by exploiting associated (concomitant, often ‘expert-opinion’) information to spread sample coverage. Composite sampling seeks to identify rare conditions and from related inferences again where sampling is costly and where sensitivity issues arise, whilst adaptive sampling for elusive outcomes and rare events modifies the sampling scheme progressively as the sample is collected.

Other topics such as size-biasing, transect sampling and capture-recapture also find wide application in environmental studies.

Time-series methods have been widely applied and developed for environmental problems but more research is needed on non-stationary and multivariate structures, on outliers and on non-parametric approaches. We will start our study of environmental statistics by considering briefly some practical examples, from different fields. In this five days programme we will be examining how statistical principles and methods can be used to study environmental problems. Our concern will be directed to :

- Data collection, monitoring and representation;
- Drawing inferences about important characteristics of the problem;
- Using statistical methods to analyze data and to aid policy and action.
- Probabilistic and statistical models;

The principles and methods will be applicable to the complete range of environmental issues (including pollution, conservation, management and control, standards, sampling and monitoring) across all fields of interest and concern (including air and water quality, forestry, radiation, climate, food, noise, soil condition, fisheries, and environmental standards).

Any models or methods applicable to situations involving uncertainty and variability will be relevant in one guise or another to the study and interpretation of environmental problems and will thus be part of the armoury of environmental statistics or environmetrics. Environmental statistics is a vast subject. In an article in the journal Environmetrics, Hunter (1994) remarked: Measuring the environment is an awesome challenge, there are so many things to measure, and at so many times and places. But, however awesome, it must be faced! The recently published four-volume Encyclopedia of Environmetrics (El-Shaarawi and Piegorsch, 2002) bears witness to the vast coverage of our theme and to its widespread following.

As we enter the new millennium the world is in crisis – in so many respects we are placing our environment at risk and not reacting urgently enough to reverse the effects.

- The average European deposits in a lifetime a monument of waste amounting to about 1000 times body weight, the average North American achieves four times this.

- Sea-floor sediment deposits around the UK average 2000 items of plastic debris per square metre.

- Over their lifetime, each person in the Western world is responsible for carbon dioxide emissions with carbon content on average 3500 times the person's body weight.

The problem of acid rain, accumulation of greenhouse gases, climate change, deforestation, disposal of nuclear waste products, nitrate leaching, particulate emissions from diesel, fuel, polluted streams and rivers, etc., have long been crying out for attention. Ecological concerns and commercial imperatives sometimes clash when we try to deal with the serious environmental issues. Different countries show different degrees of resolve to bring matters under control; carbon emission is a case in point, with acclaimed wide differences of attitude and practical between, for example, the United States and the European Union. Environmental scientists, and specialists from a wide range of disciplines, are immersed in efforts to try to understand and resolve the many environmental problems we face.

Playing a major role in these matters are the statisticians, who are uniquely placed to represent the issues of uncertainty and variation inevitably found in all environmental issues. This is vital to the formulation of models and to the development of specific statistical methods for understanding and handling such problems.

CHAPTER – 2

ENVIRONMENTAL DATA QUALITY MANAGEMENT

INTRODUCTION

This chapter discusses evolution of the environmental data quality model by evaluating the relationship between data quality and decision quality, and by distinguishing analytical quality from data quality. A ‘ ‘next-generation’ ’ data quality model can create the framework needed for explicitly managing both data and decision uncertainties using new strategies to produce greater decision confidence (‘ ‘better’ ’), while simultaneously shortening project lifetimes (‘ ‘faster’ ’) and cutting overall project costs (‘ ‘cheaper’ ’) more than ever before possible .

“QUALITY’’ AS A POLICY GOAL”

Exhortations for ‘ ‘sound science’ ’ and ‘ ‘better quality data’ ’ within the context of regulatory environmental decision making are increasingly popular. Is the current data quality model sufficient to achieve sound science? Is ‘ ‘data quality’ ’ really the key issue, or is there something more fundamental at stake?

‘ ‘Data quality’ ’ is too often viewed as some independent standard established by outside arbiters independent of how the data will actually be used. Project managers tend to follow a checklist of ‘ ‘approved’ ’ analytical methods as the primary means of achieving ‘ ‘data quality.’ ’ Yet, striving for ‘ ‘high quality data’ ’ under the current model has proven to be an expensive and sometimes counterproductive exercise.

In contrast to checklist approaches to ‘ ‘data quality,’ ’ sound science in regulatory and project decision-making is achieved by acknowledging and managing decision uncertainty. Correspondingly, acceptable data quality is achieved by managing all aspects of data uncertainty to the degree needed to support the decisions for which the data are intended. Managing uncertainty, either of decisions or of data, requires careful planning using relevant expertise and technical skills. Calls for ‘ ‘sound science’ ’ and

‘‘better data quality’’ are meaningless without a simultaneous commitment to include scientifically qualified staff when planning science-based programs and projects. Environmental programs exist because there is work that must be done at the project level. Policy-makers that desire to see sound science in environmental decisions need to provide a coherent vision that will steer the
development of program infrastructure that focuses on managing decision quality at the project level.

It is a mistake to assume that scientific data are (or can be) the only basis for regulatory decision-making. Science may be able to provide information about the nature and likelihood of consequences stemming from an action, but the decision to pursue or reject that action (i.e., accept or reject the risk of consequences) based on scientific information is within the province of values, not science. Even the choice of how much uncertainty is tolerable in statistical hypothesis testing lies in the realm of values. Thus, it is appropriate that many nonscientific considerations feed into a regulatory decision-making process. This does not invalidate a foundation of ‘‘sound science’’ as long as the various roles of science and values are differentiated, and any underlying assumptions and other uncertainties in both data and decision making are openly declared with an understanding of how decision making could be affected if the assumptions were erroneous.

DECISION QUALITY AS DEFENSIBILITY

The term ‘‘decision quality’’ implies that decisions are defensible (in the
broadest scientific and legal sense). Ideally, decision quality would be equivalent to the correctness of a decision, but in the environmental field, decision correctness is often unknown (and perhaps unknowable) at the time of decision-making. When knowledge is limited, decision quality hinges on whether the decision can be defended against reasonable challenge in whatever venue it is contested, be it scientific, legal, or otherwise. Scientific defensibility requires that conclusions drawn from scientific data do not extrapolate beyond the available evidence. If scientific evidence is insufficient or conflicting and cannot be resolved in the allotted time frame, decision defensibility will have to rest on other considerations, such as economic concerns or political sensitivities. No matter what considerations are actually used to arrive at a decision, decision quality (i.e., defensibility)

implies there is honest and open acknowledgment and accountability for the full range of decision inputs and associated uncertainties impacting the decision making process.

FIRST-GENERATION STEPPING-STONES THAT BECAME STUMBLING BLOCKS

When immediate action is desired, but knowledge and expertise are not yet sufficient to plot the smartest plan of attack, a reasonable tactic is to initially create a consistent, process-driven strategy based on the best available information so everyone can ‘‘sing from the same sheet of music’’ while experience and knowledge are being accumulated. Certainly this made sense for the emerging cleanup programs. To be consistent with sound science, however, such a process-driven approach should be openly acknowledged by all participants as the first approximation that it is, with the understanding that one-size-fits-all oversimplifications will be discarded in favor of more scientifically sound information as it becomes available. Although science may be comfortable viewing first approximations as short-lived stepping-stones subject to continual improvement and revision, this view is less welcome when economic and litigious forces intersect with broader societal goals in a regulatory crucible. This is one of the fundamental conflicts faced by policy makers seeking ‘‘sound science’’ as a basis for regulation.

EVOLVING A SECOND-GENERATION DATA QUALITY MODEL

To set the stage for an updated data quality model, we must clarify the term ‘‘data quality.’’ Data quality is ‘‘the totality of features and characteristics of data that bear on its ability to meet the stated or implied needs and expectations of the user/customer’’ What data users ‘‘need,’’ ultimately, is to make the correct decisions. Therefore, data quality cannot be viewed according to some arbitrary standard, but must be judged according to its ability to supply information that is representative of the particular decision that the data user intends to make. Said in a different way, anything that compromises data representativeness compromises data quality, and data quality should not be assessed except in relation to the intended decision. The assumptions of the current data generation model and routine application of this model to environmental decision-making for site cleanup

are inadequate to ensure that data are representative of the site decisions being made. The root cause of data non-representativeness is the fact that environmental data are generated from environmental samples (i.e., specimens) taken from highly variable and complex parent matrices (such as soils, waste piles, sludges, sediments, groundwater, surface water, waste waters, soil gas, fugitive airborne emissions, etc.). This fact has several repercussions:

1. The concept of representativeness demands that the scale (spatial, temporal, chemical speciation, bioavailability, etc.) of the supporting data be the same (within tolerable uncertainty bounds) as the scale needed to make the intended decisions (does unacceptable risk exist or not; how much contamination to remove or treat; what treatment system to select; what environmental matrix to monitor; what analytes to monitor for; where and how to sample; etc.).
2. The concept of representativeness can be coarsely broken into sample representativeness and analytical representativeness, both of which are critical to managing data uncertainties:

- Sample representativeness includes procedures related to specimen selection, collection (i.e., extraction from the parent matrix), preservation, and subsampling (although this is often included with ‘‘analytical’’ since it typically takes place in the lab). All are crucial to data quality, but the representativeness of specimens is difficult to ensure without sufficient sampling density to understand the scale and characteristics of matrix heterogeneities. Even perfectly accurate analysis is no guarantee of good data quality if the sample were not representative of the properties of concern to the decision-maker. Since many environmental matrices are highly heterogeneous on many different scales that affect contaminant concentration and behavior in analytical and biotic systems, most of the uncertainty in most of today’s site data stems from the sampling side, although inaccurate analysis certainly can (and do) occur.
- Analytical representativeness involves selecting an analytical method that produces test results that are representative of the decision. Causes of analytical non-representativeness include selecting the wrong method or erroneously interpreting method results Analytical

representativeness is compromised when matrix interferences degrade method performance to the point where erroneous decisions would be made if the data were not recognized as suspect. If interferences are found, sound science demands that method modification or an alternate method be used to compensate.

In contrast to the assumptions that underlie the current data quality model, a second-generation data quality model for the environmental field will explicitly recognize that:

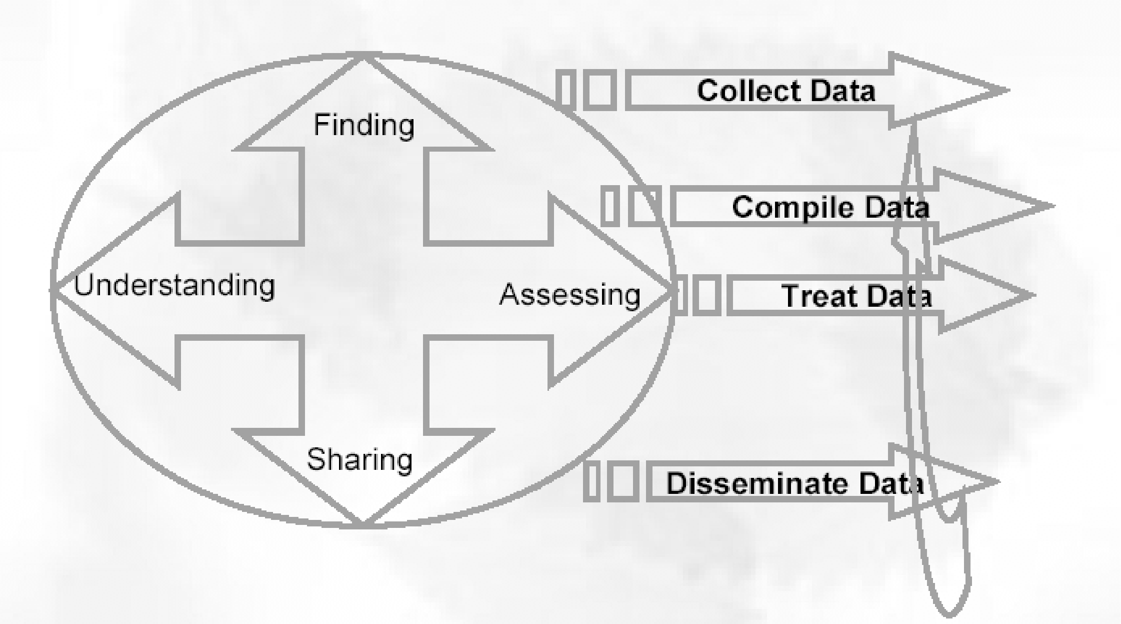
- Data quality is an emergent property arising from the interaction between the attributes of the analytical data (such as its bias, precision, detection and quantitation limits, and other characteristics that together contribute to data uncertainty) and the intended use of the data (which is to assist managing decision uncertainty).
- Data uncertainty is comprised of both sampling and analytical uncertainties.
- Analytical uncertainty in a test result arises from both the analytical uncertainty of the measurement method itself and from interaction between the sample matrix and the analytical process. The analytical uncertainty arising from the method itself is only a fraction (and often a negligibly small fraction) of the overall data uncertainty.
- Sampling uncertainty accounts for the majority (and in some situations, nearly all) of the data uncertainty. This uncertainty can be managed by increasing the sampling density and/or by targeting sample collection designs to yield the most valuable information (i.e., gather more data where decisions are more uncertain, such as boundaries between ‘‘clean’’ and ‘‘dirty’’ areas, and less data where there decisions are more certain, such as obviously ‘‘clean’’ or obviously ‘‘dirty’’ areas). Sample representativeness requires that all aspects of sampling design be matched to the scale of decision-making.
- Procedures to estimate and report data uncertainties (e.g., uncertainty intervals) to the data user need to be developed for the environmental field.
- Investment in properly educated and experienced technical staff is a necessary and cost-effective means to achieve data quality and good science where numerous complex and interacting variables must be evaluated and balanced.

CHAPTER – 3

DATA COLLECTION, SUMMARISATION AND PRESENTATION OF ENVIRONMENTAL DATA

Introduction

Many time, we are tempted to exercise control or take decisions on the basis of our experience, impression or intuition. If we are lucky we might be rewarded with expected results - otherwise we fail. The risk involved in such exercises can be controlled/minimized only through systematic collection of data and their analysis where data is a numerical expression of the characteristics of an activity or process. The environmental data components are presented in the following diagram.



Environmental Data Components

Type of Environmental Data:

Data are of two categories : (i) routine ; (ii) special

Routine data are collected for monitoring of air quality, monitoring the benzene level, evaluating the water quality status of Yamuna river, assessing hazardous wastes etc. Special data might be collected for investigating chronic problems or for studies involving experimentation for improvement purpose. Whatever be the type of data, we must follow some cardinal principles while collecting data.

Attribute and Variable Data

There are two types of data – the Attribute type and the Variable type. In the attribute type the characteristic considered is not measurable. It only gives a comparative view of the characteristic of interest. For example good, bad or worse, whereas the variable type of characteristic is measurable. For example if temperature, the characteristic then the importance of the characteristic is given by a numerical value say 50^0C at one time point and say 52^0C at another time point.

1. Have a clearly defined objective

Remember that data are meant for ACTION. Before collecting data, it is important to determine which charateristic is to be considered. Next, what we are going to do with the data. Both short term and long term objectives are to be kept in mind. It is no use collecting data which are not utilized at all. Collecting of data costs money. A balance has to be struck between the cost and the worth of the information for action.

2. Collecting data to suit the purpose

Once the objective for collecting data is defined, the types of comparison which need to be made are also determined, and this in turn identifies the type of data, time interval which should be collected. Suppose we are interested in assessing the variability in certain characteristic. If only one observation is collected per day, it is impossible to determine the variability which occurs in a day.

3. Ensure reliability of measurements

Data are input for decisions made by an organization. In view of the impact it makes, it is absolutely essential that the reliability of data is ensured. Unreliability of data might result from deliberate attempts to conceal true facts or ignorance. In one chemical process industry, the inprocess wastage figures were being under-reported only to show the management that targets are being met. In another case for acceptance sampling purposes, an inspector was reporting the average of measurements based on varying sample size at his discretion and using such results for determining acceptability of the product supplied by a vendor. This was a violation of the stipulated procedure to follow the sampling plan.

4. Decide the subsequent treatment of data

Once some data are collected, it is necessary to decide in advance how to present or summarize the data and what kind of statistical analysis to perform so that meaningful inferences can be drawn for action. The responsibility at all the stages of data collection, presentation, analysis, reporting and action is also to be simultaneously fixed. The paper work system needs proper planning to make it effective.

5. Find right ways to record data

While collecting data it is necessary to arrange the data neatly so that subsequent processing is facilitated. Relevant details such as day of the week, hour of the day, inspector, measuring instrument used etc. needed to be recorded properly. The frequency of data can be decided keeping in view the purpose, cost of data collection, data-processing facility and the availability of relevant resources. It is important for the data to be capable of being collected in a simple way and in a form which is easy to use. Appropriate data-sheet or proforma or check-sheet has to be designed. So that data recording becomes easy and the data are arranged automatically so that they can be easily taken on for further processing.

Tools and Techniques for looking at data :

The following techniques are commonly used for preliminary analysis.

1. Check Sheet
2. Pareto Analysis
3. Brain storming
4. Ishikawa Diagram or Cause and Effect Diagram

5. Scatter Diagram and Regression Analysis
6. Stratification
7. Histogram

Some of these techniques have been discussed in this chapter and some others have been discussed in the following chapters.

Check Sheet

To collect data check sheets are used. This facilitates easy collection, summarization and analysis of data. It can be used in the following functions: (a) vital items check (c) problem location checks (d) problem cause checks

Pareto Analysis

Whenever any problem related to pollution or cost is taken up to investigate the ways of improvement, the first task usually is to narrow down the problem area so that the problem becomes easier to handle and the root causes of the maladies are identified quickly. Fortunately there is a natural law which almost always ensures that a relatively few of the contributors account for the bulk of the problem. This is segregation of ``vital few and trivial many" contributors is known as PARETO ANALYSIS, a term coined by Dr. J.M. Juran.

Procedure for making Pareto Chart

Step 1 Decide what items are to be investigated and how to collect data.

(i) Decide what kind of data you want to investigate. For example, the problem of hazardous waste in Noida or Ghaziabad area, amount of loss in monetary terms, number of customer complaints etc.

(ii) Decide how to classify the data. For example, by the type of defect, location, process, machine, worker, method etc.; if the necessary record items appearing infrequently under the heading ``others".

(iii) Determine the method of collecting the data and the period for which it is to be collected.

Step 2 Design a data tally sheet listing the items with space to record their totals.

Step 3 Fill in the tally sheet and calculate the totals.

Step 4 Make a Pareto chart data sheet listing the items, their individual totals, cumulative totals, percentages of overall total, and cumulative percentage.

Step 5 Arrange the items in order of quantity and fill in the data sheet. The item ``others" should be placed in the last row, no matter how large it is because it is composed of a group of items each of which is smaller then the smallest item listed individually.

Step 6 Draw two vertical axes and a horizontal axis. Mark the left hand vertical axis with a scale from 0% to 100%. The horizontal axis is to be divided into a number of equal intervals, equal to the number of items investigated.

Step 7 Construct a bar chart with bars over the intervals corresponding to each item. The height of each bar is proportional to the corresponding frequency.

Step 8 Draw the cumulative curve (Pareto Curve) by marking the cumulative percentage points above the right hand intervals of each item, and connecting the points by a solid line.

Step 9 Write other relevant information on the chart so that it becomes self-explanatory.

Pareto Analysis of the number of persons affected due to different types of hazardous waste in a city is shown next.

Exercise: Make Pareto Analysis of types of hazardous waste in a city.

Table : Data Tally Sheet

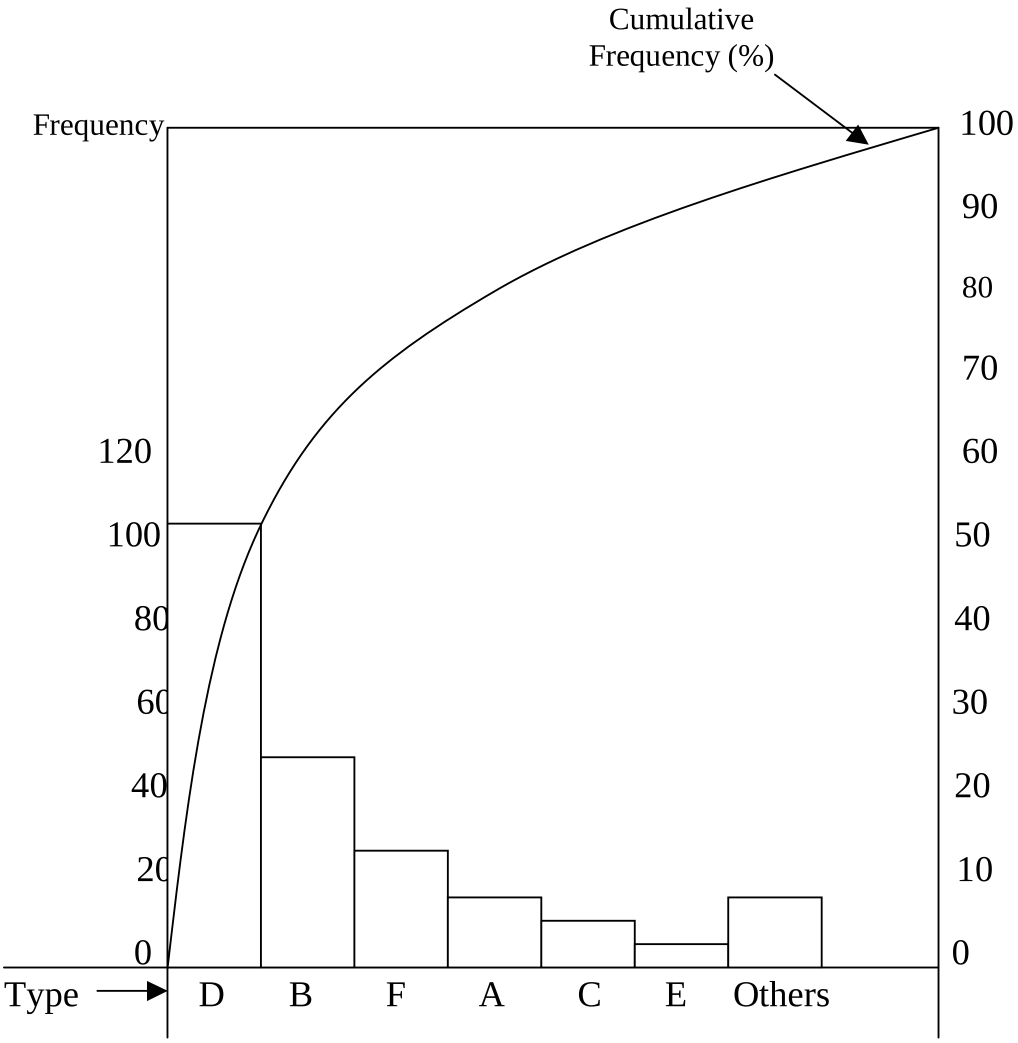
Sl.No.	**Type of hazardous waste**	**Tally**	**Frequency (no. of person in thousands)**
1	A	𝍸 𝍸	10
2	B	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||	42
3	C	𝍸 |	
4	D	𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸	
		𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸𝍸 ||||	104
5	E	||||	4
6	F	𝍸𝍸𝍸𝍸	20
7	Others	𝍸𝍸 ||||	14
		TOTAL	200

Table : Data Sheet for Pareto Chart

Sl. No.	**Type of hazardous waste**	**Frequency**	**Cumulative Frequency**	**% Contribution by Type of waste**	**Cumulative Percentage**
1	D	104	104	52	52
2	B	42	146	21	73
3	F	20	166	10	83
4	A	10	176	5	88
5	C	6	182	3	91
6	E	4	185	2	93
7	Others	14	200	7	100
	Total	200		100	

The chart brings out very clearly that if the hazardous waste is to be brought down, we should first concentrate on type of waste D which contributes more than 50\% to the total. Such information is obviously very useful in directing the priorities of the study.

The information obtained through Pareto analysis when presented in the form of a chart is known as PARETO CHART. Dr. Juran was the first to use the concept introduced by the Italian economist V. Pareto who showed that the largest share of income or wealth is held by a much smaller number of people.



Pareto Chart by Effect

This is a chart concerning poor performance and is used to find out what the major problem is. This performance may be related to :

i. Quality : faults, failures, complaints, returned items, repairs, recovery etc.
ii. Cost : amount of loss, expenses
iii. Delivery : Stock, Shortage, defaults in payments, delays in delivery.
iv. Safety : Accidents, Mistakes, Breakdowns

Brain Storming And Ishikawa Diagram

Any problem we take up for study normally involves a large number of factors originating from different departments of the organization. One or a few persons may not have complete knowledge about all the possible causes or factors which contribute to the problem. As such it is necessary to conduct a group exercise wherein all concerned and knowledgeable people must sit together and discuss. Such an exercise is known as brain storming. This will help us to prepare a complete list of the factors involved in any experiment.

The list of factors can be presented in tabular form. However, a most comprehensive way of presentation is a pictorial or diagrammatic form known as the Ishikawa Diagram.

Cause & Effect Diagram (Ishikawa Diagram)

In order to achieve the goal of ``making right the first time". It is necessary that we understand the root cause which create the problems. The cause of poor air quality or water quality may be attributed to a number of factors depending upon the complexity of the problem, and a cause-and-effect relationship can be found among those factors. Approached individually, different people might offer different explanations as to the root causes. But jointly we can determine structure of a multiple cause and effect relation by observing it systematically. It is difficult to solve complicated problems without considering this structure which consists of a chain of causes and effects, and a CAUSE & EFFECT DIAGRAM is a method of expressing it simply and easily.

In 1953 Prof. K. Ishikawa of Tokyo University summarized the opinions of engineers at a plant in the form of a {Cause & Effect Diagram} as they discussed a quality problem. This was the first time, this approach was used. Since then it proved to be quite useful in visual display of the relation between the characteristics to be investigated and the factors so that systematically the theories could be tested and remedies developed.

The diagram is a fish-bone like structure where the quality characteristic having problem is indicated by the horizontal main arrow (backbone) and the major factors such as Materials, Methods, Man, and Machines etc. which contribute primarily to the causes of the problems are represented in the form of slanting arrows meeting the main arrow from the top or bottom. The secondary causes of each primary causes are indicated through horizontal small arrows meeting the arrows for primary causes. The causes are listed through brain-storming sessions attended by all concerned.

Procedure for making and using cause & effect diagram

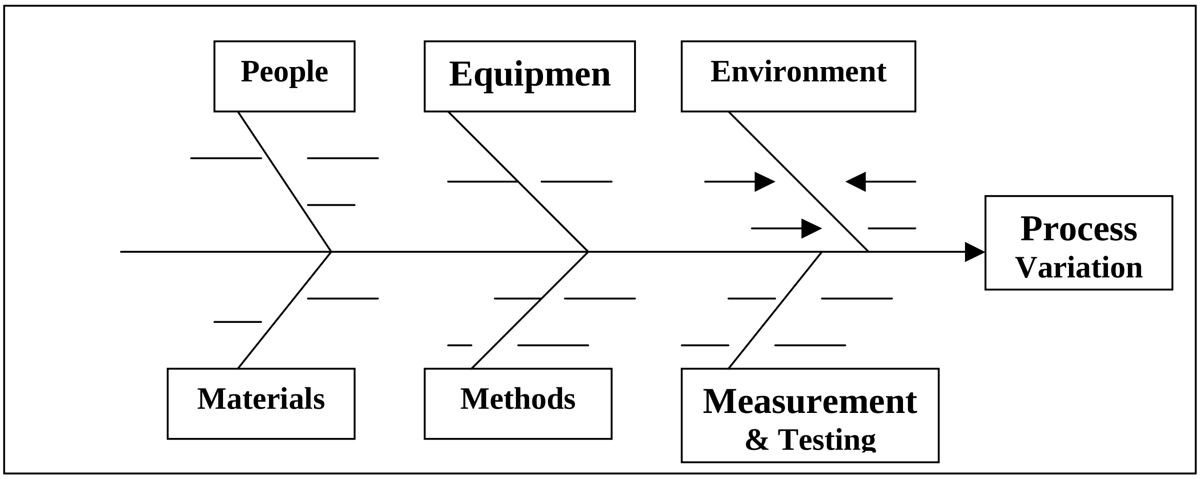
- Decide on the quality characteristic having problem.
- Find as many causes as possible which are considered to affect the quality characteristic.
- Sort out the relations among the causes and make a cause & effect diagram consisting of arrows which represent the primary and secondary causes.
- Determine priorities of the causes for verification with data already available or to be collected specially.
- Assign importance or significance to each factor objectively on the basis of data and device appropriate measures to get rid of the problems.

Checkpoints for preparing cause effect diagram

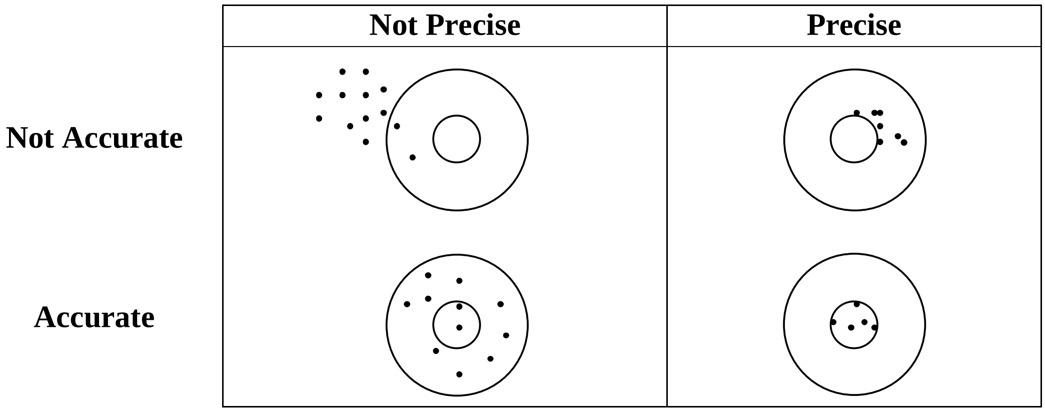
- State the objective very clearly. Is it ``increasing or decreasing average level of some variable characteristic or decreasing the variability or decreasing the occurrence of some undesirable events."
- Secure participation from all concerned. The participants should express their viewpoints honestly and fearlessly. Even very odd ideas might click subsequent for solving a problem.
- Express the factors as concretely as possible. Factors expressed in an abstract manner only result in a cause & effect diagram based on generalities which will not help in solving the problem.
- Choose measurable characteristics and factors so that they are amenable to statistical verification with data.

- ❖ Discover factors amenable to action. If the cause you have identified can not be acted upon the problem will not be solved. If improvements are to be affected, the cause should be broken down to the level at which they can be acted upon, otherwise identifying them will become a meaningless exercise.

See Figure for cause and effect diagram.



The measurement method must produce accurate and precise results over time



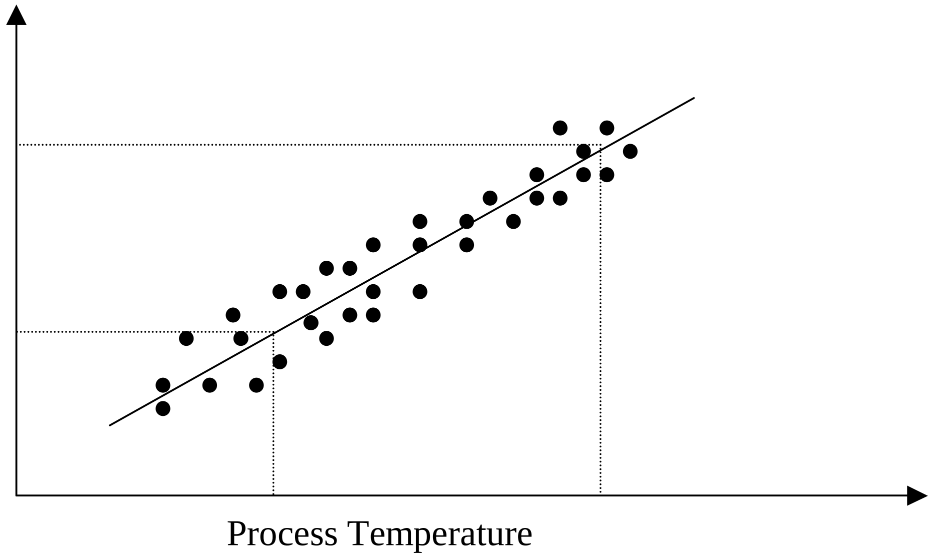
Stratification

This is the sample selection method used when the whole population, or lot, is made up of a complex set of different characteristics, e.g. region, income, age, race, sex, education, operators, shift, days etc. In these cases the sample must be very carefully drawing in proportions which represent the makeup of the population.

Stratification involves simply collecting or dividing the set of data into meaningful groups or strata and depicting the data in stratified form so as to bring out if the different groups are significantly different. Groups which are worse than the others are singled out and appropriate actions are taken to bring them at par with the others thereby effecting significant improvement in the overall performance.

Scatter Diagram

Scatter diagram are used to examine the relationship between two factors to see if they are related. If they are, then by controlling the independent factor, the dependent factor will also be controlled. For example, if the temperature of the process and the purity of a chemical product are related then by controlling temperature, the purity of the product is determined. Figure illustrates use of Scatter Diagram in different situations.



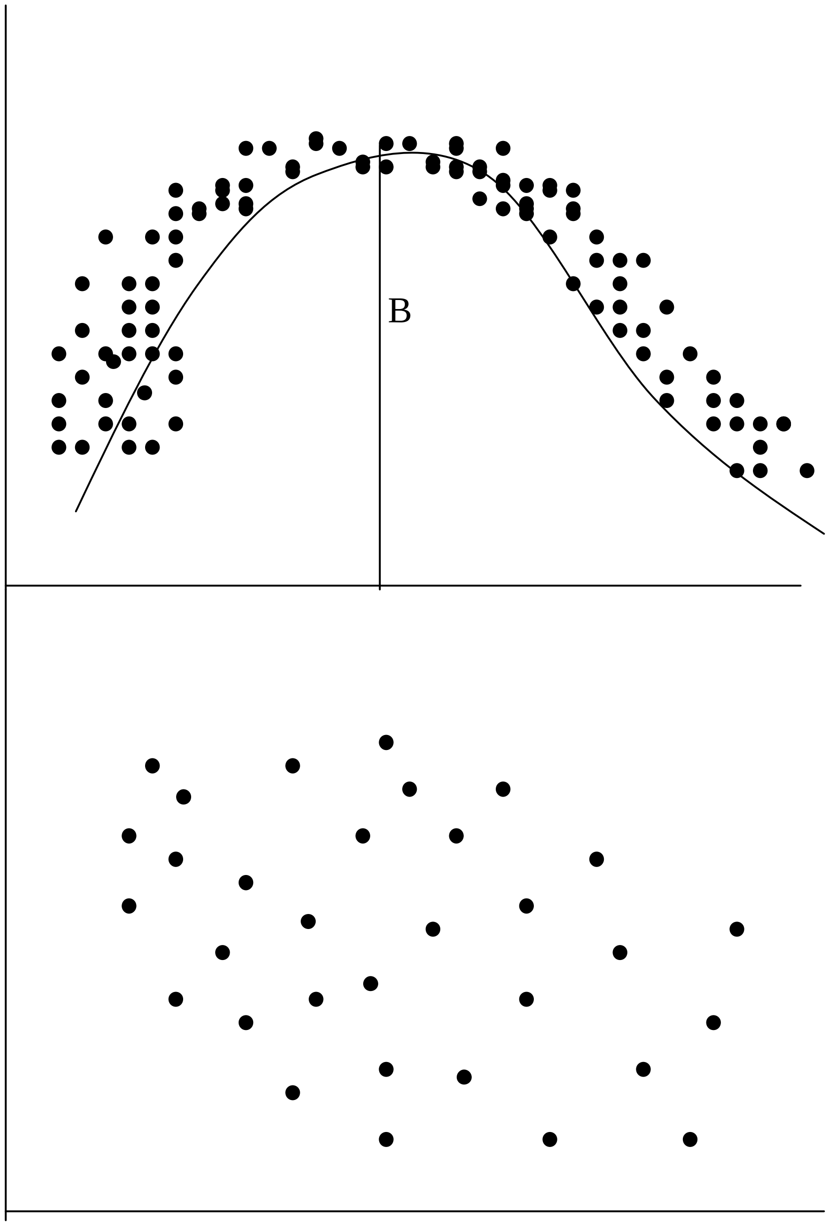


Figure : Scatter Diagram

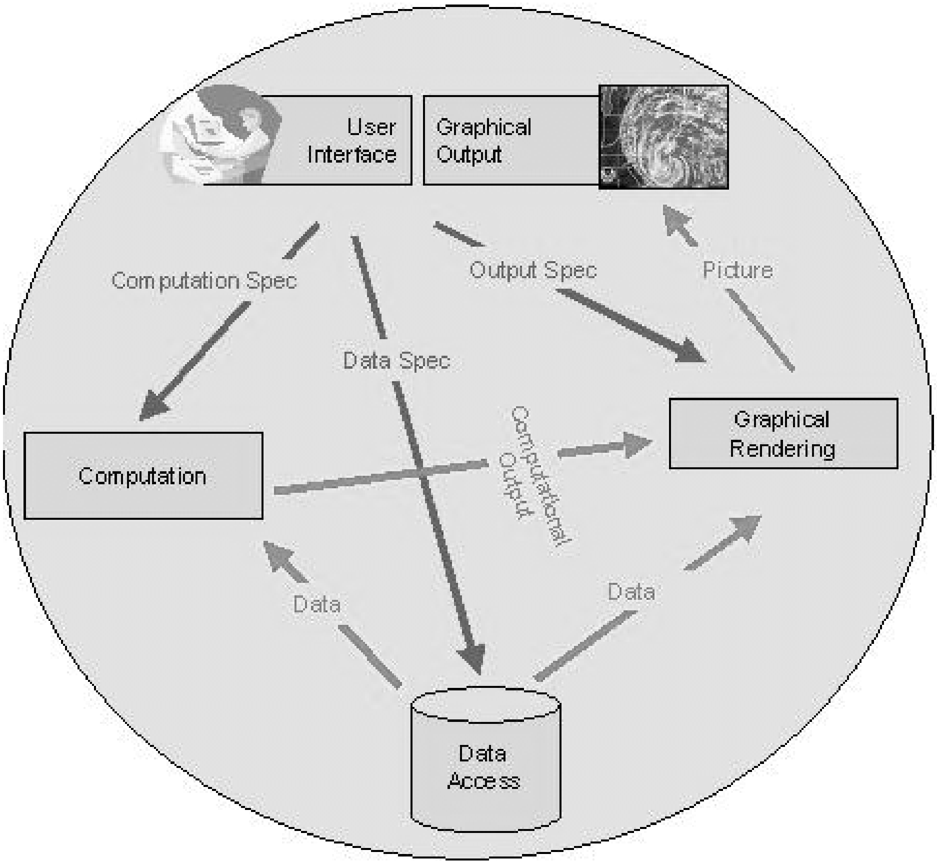
Motivation:

A quantitative approach to environmental problem solving can be broken down into a number of steps. We assume that you start with a qualitative problem much like the acid rain question in the first chapter.

1. Formulate quantitative hypotheses and questions that will help you address your general question or issue.
2. Determine the variables to observe.
3. Collect and record the data observations.
4. Study graphics and summaries of the collected data to discover and remove mistakes and to reveal relationships between variables.

5. Choose a model describing the important relationships seen or hypothesized in the data.
6. Fit the model using the appropriate modeling technique.
7. Examine the fit using model summaries and diagnostic plots.
8. Repeat steps 5-7 until you are satisfied with the model.
9. Interpret the model results in light of your general question.

Data Analysis and Display System Components



CHAPTER - 4

DESCRIPTIVE STATISTICS FOR ENVIRONMENTAL DATA

What is Statistics?

The word ‘statistics’ is used in two senses – data and the science. The science of statistics deals with:

1. Collection of data
2. Summarisation of data
3. Analysis of data, and
4. Drawing valid inference from data which are usually subject to variation.

A layman usually considers statistics in the sense of ‘ Data ‘ only. As was the case with many other sciences, Statistics has also been much abused knowing or unknowingly, by people involved in public dealings. All these ultimately led to the comment ‘There are lies, damned lies, and statistics’.

Need for the Science Statistics:

Other than the people engaged in professional statistics activities, it is scientists, engineers and managers at different levels of manufacturing, laboratories, or service organizations, who handle maximum amount of data and interpret them for decision making and action. The efficiency of an organization depends upon the quality of decision making to a large extent. There are many situations, where common sense is a poor guide when it comes to interpretation of data. The quality of decision making on the basis of data can improve only with the help of the science of statistics. Of course, sometimes the basic problems remains, namely, not talking with facts but talking on the basis of opinions, impressions etc. which make the decision making highly subjective. Typically we are interested in a population - a well defined groups of cases.

Population: Collection of all elements under consideration and about which we are trying to draw conclusions.

Population elements may be :

- Objects;
- Entities;
- Units;
- People; .. etc
- A batch of material

Generally each case has one or more characteristics (attributes) of interest. When a particular characteristic is measured we obtain a value which varies from case to case – hence each characteristic is termed a variable. Recording the value of a variable for each case amounts to collecting data.

Sample: A subset of the elements selected from the population with a view to draw inference about the population characteristics. Thus a sample is part of population. The objective of statistical inference is to draw conclusions about the population using a sample data from that population.

Data Summarisation Methods:

- Graphical Methods
- Tabular Summarisation
- Numerical Indices

Graphical Methods:

Graphic displays provide better insight that often is not possible with words or numbers.

Graphical Tools:

- Bar Chart
- Pie Chart
- Run Chart
- Histogram
- Frequency Curve
- Scatter Diagram
- Control Charts
- Box Plots
- Youden Plots

Tabular Methods: Summarises data in the form of a table.

Table 1 : Concentration of benzene in 100 air samples (units in $\mu\, g/m^3$)

3.37	3.34	3.38	3.32	3.33	3.28	3.34	3.31	3.33	3.34
3.29	3.36	3.30	3.31	3.33	3.34	3.34	3.36	3.39	3.34
3.35	3.36	3.30	3.32	3.33	3.35	3.35	3.34	3.32	3.38
3.32	3.37	3.34	3.38	3.36	3.37	3.36	3.31	3.33	3.30
3.35	3.33	3.38	3.37	3.44	3.32	3.36	3.32	3.29	3.35
3.38	3.39	3.34	3.32	3.30	3.39	3.36	3.40	3.32	3.33
3.29	3.41	3.27	3.36	3.41	3.37	3.36	3.37	3.33	3.36
3.31	3.33	3.35	3.34	3.35	3.34	3.31	3.36	3.37	3.35
3.40	3.35	3.37	3.35	3.32	3.36	3.38	3.35	3.31	3.34
3.35	3.36	3.39	3.31	3.31	3.30	3.35	3.33	3.35	3.31

Procedure for Constructing a Frequency Distribution :

1. Decide on the number of cells.
2. Calculate the approximate cell interval. The cell interval equals the largest observation minus the smallest observation divided by the number of cells. Round this results to some convenient number.
3. Construct the cell by listing cell boundaries.
4. Tally each observation into the appropriate cell.
5. List the total frequency of each cell.

Number of cells in frequency distribution:

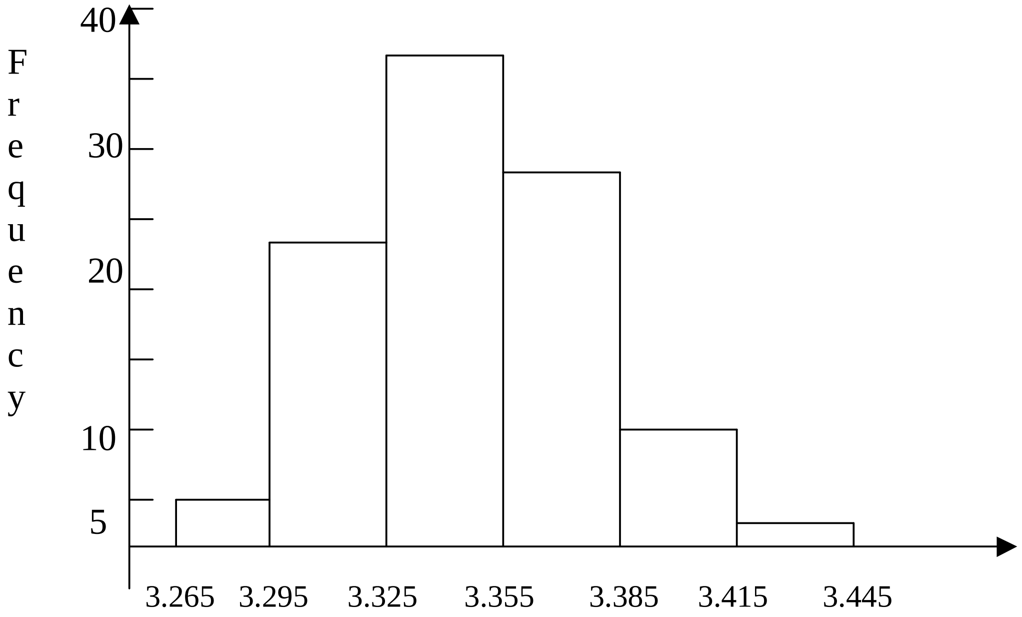
Number of observations	Recommended number of cells
20-50	6
51-100	7
101-200	8
201-500	9
501-1000	10
Over 1000	11-20

Table 2: Frequency Table of Concentration of benzene in 100 samples :

Diameter	Tally Mark	Frequency	Cumulative Frequency
3.265 – 3.295	𝍸	**5**	**5**
3.295 – 3.325	𝍸 𝍸 𝍸 𝍸 ///	**23**	**28**
3.325 – 3.355	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 /	**36**	**64**
3.355 – 3.385	𝍸 𝍸 𝍸 𝍸 𝍸 //	**27**	**91**
3.385 – 3.415	𝍸 ///	**8**	**99**
3.415 – 3.445	/	**1**	**100**
Total		100	

Histogram : It is bar chart of a frequency distribution. It highlights the center and amount of variation in the sample of data. The simplicity of construction and interpretation of the histogram makes it an effective tool in the elementary analysis of data. Many problems in quality control have been solved with this one elementary tool alone. Figure - 1 gives the histogram of data given in Table – 1. The following steps are used to construct histogram :

1. Mark the Y – axis with frequency scale.
2. Mark the X- axis with class boundaries using a suitable scale.
3. Draw rectangles on X – axis with base equal to the width of the class interval and height equal to class frequency.



Histogram illustrates how variable data provides much more information than do attribute data. Centering of histogram, width of the histogram and the shape of the histogram reflect the ability of the process to meet specification limits and presence of assignable causes of variation in the process. Figure – 2 gives typical histograms encountered in practice.

Frequency Polygon :

It is the line graph of class frequency against midpoint of class interval.

Numerical Indices : Data can be summarized using

- Measures of central tendency
- Measure of dispersion

Measures of Central Tendency: A value which is representative of the set of data as most of the data is centered around this value. Important measures of central tendency are Mean, Mode and Median.

Mean : Total of all the observations divided by the number of observations.

$$\text{Mean } \left(\overline{\mathrm{X}}\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

X	Frequency (f)	Xf
x_1	f_1	$x_1\ f_1$
x_2	f_2	$x_2 x_2$
$\vdots$	$\vdots$	$\vdots$
x_k	f_k	$\vdots$
Total	$\sum f_i$	$\sum x_i\ f_i$

$$\text{Mean }\left(\overline{X}\right)=\frac{x_1f_1+x_2f_2+\cdots+x_kf_k}{f_1+f_2+\cdots+f_k}=\frac{\sum x_if_i}{\sum f_i}$$

Example:

Benzene Conc. $\left(\mu\ g/m^3\right)$	No of days (f)	Xf
25	2	50
26	3	78
27	4	108
28	3	84
29	1	29
30	2	60
Total	15	409

$$\text{Average temp. }\left(\overline{\boldsymbol{X}}\right)=\frac{409}{15}=27.266$$

Mode : That value for which frequency is maximum.

Median_: It is the middle most central value when all values are arranged by order of magnitude. Half the values lie above this value and the other half lie below it. That is median divides the data into two equal parts.

Steps to compute the Median:

1. Arrange all values in order of size. From smallest to largest.

2. If the number of values (n) is odd, the median is center value in the ordered list. The location of median is obtained by counting $\frac{(n+1)}{2}$ observations from the bottom of the list.

   Consider the data set : 490, 400, 450, 420 and 430 to find the median of this data, we first arrange the data from smallest to largest value.

e.g. 400, 420, 430, 450, 490. The median is in the position

$\frac{(n+1)}{2}$ = $\frac{5+1}{2}=3$. Therefore the median is 430.

3. If the number of observations (n) is even, the median M is given by the average of the two center observations in the ordered list. i.e for example 70, 75, 77, 82, 88, 100, 105, 108, the median is the average of the 4th and 5th value i.e, $\frac{82+88}{2}=85$

The median has several advantages over the mean. The most important is that extreme values do not affect the median as strongly as they do the mean. That is the mean is much more sensitive to outlier values as compared to the median.

Percentile: The p^{th} percentile of the data is the value such that p percent of the observations fall at or below it.

The median is the 50th percentile the first quartile is 25th percentile and the third quartile is the 75th percentile.

Dispersion: Variation is a fact of nature and in industrial life too. No two items produced by same process are exactly the same. Tests done on the same samples may vary from chemist to chemist or from laboratory to laboratory. This is true whether the test equipment involved is automatic or manually operated. Variation can be because of lack of complete homogeneity of chemicals used in test, variation in test environmental conditions or due to difference in the skill of chemists or testing equipments etc. Variation in the test results add to the uncertainty of decisions and hence it is important to measure variation and control it.

In summarizing data, the variability in the values is often an important feature of interest. Major measures of dispersion are :

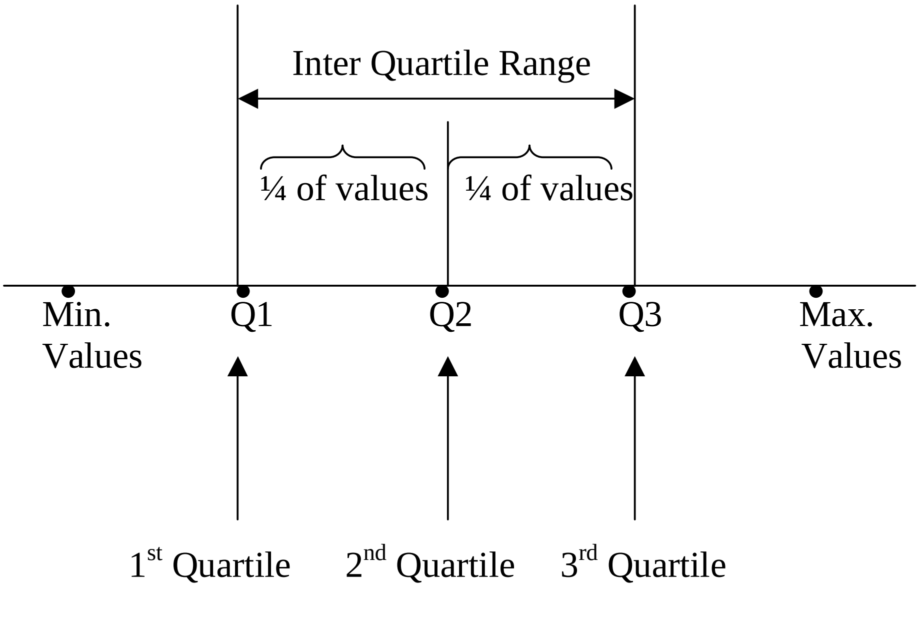
1. **Range (R) :** The range is the difference between the largest and smallest value in a data set. That is Range (R) = Largest Value - Smallest Value.

2. **Quartile:** Quartiles divide the data into four equal parts. Each part contains 25% of the values. Q_1 is call the first or lower quartile and Q_3 is called the third quartile or higher quartile. Q_2 is the median.

Interquartile Range (IQR): It is the difference between the third and the first quartiles of a set of values. That is Interquartile Range (IQR) = Q_3 - Q_1

Interquartile range is a simple measure of speed that gives the range covered by the middle half of the data. It reflects the variability of the middle 50 percent of the data.

The quartiles and the IQR are unaffected by extreme values.



Calculation of Quartiles:

1. Arrange the data in the increasing order and locate the median.

2. The first quartile in the median of the observation below the location of the median.

3. The third quartile in the median of the observations above the median of all observations.

Example: Data below gives the daily presence of sulphur oxides in a city.

15.8, 26.4, 17.3, 11.2, 23.9, 24.8, 16.2, 12.8, 22.7, 28.8, 7.7, 13.5, 18.1,, 17.9, 23.5,

Determine the quartiles and inter-quartile range.

1. Arrange the data in the increasing order i.e

7.7, 11.2, 12.8, 13.5, 15.8, 16.2, 17.3, 17.9, 18.1, 22.7, 23.5, 23.9, 24.8, 26.4, 28.8

2. Q_2 = Median = Middle value i.e 8th value = 17.9
   Q_1 = 13.5 and Q_3 = 23.9.

3. Interquartile range (IQR) = Q_3 - Q_1 = 23.9 - 13.5 = 10.4

Standard deviation and Variance: The Most commonly used measure of dispersion is called the standard deviation. It takes into account all the values in a set of data.

Suppose the test result values are : $x_1, x_2, \cdots, x_N$

Population Standard Deviation: It is denoted by the Greek symbol σ and is given by root mean squared deviation from the mean μ. That is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Where μ is the population mean $(\mu) = \frac{x_1 + x_2 + \cdots + x_N}{N}$

Sample Standard deviation (s): If the sample result values are $x_1, x_2, \cdots, x_n$. It is given by

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where $\bar{X} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

Variance : Population variance (σ^2) and sample variance (s^2) are given by

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}, \qquad s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Example : Standard deviation of the sample test values:

X	$X - \bar{X}$	$(X - \bar{X})^2$	
15	-5	25	$\bar{X} = \frac{100}{5} = 20$ $s = \sqrt{\frac{1}{n-1}\sum(X - \bar{X})^2}$ $= \sqrt{\frac{66}{4}} = 4.062$
18	-2	4	
20	0	0	
21	1	1	
26	6	36	
100	0	66	

Sample Standard deviation (s) = 4.062 and sample variance = $\frac{66}{4} = 16.5$

Benzene Concentarions	Frequency	Mid-Point	$f \, x M$	$\left(M - \overline{M}\right)^2 f$
	(f)	(M)		
3.265 - 3.295	5	3.28	16.40	0.02042
3.295 - 3.325	23	3.31	76.13	0.02643
3.325 - 3.355	36	3.34	120.24	0.00005
3.355 - 3.385	27	3.37	90.99	0.01839
3.385 - 3.415	8	3.40	27.20	0.02518
3.415 - 3.445	1	3.43	3.43	0.00074
	100		334.39	0.09121

$$\overline{\mathrm{M}} = \frac{334.39}{100} = 3.3439, \text{ and s} = \sqrt{\frac{0.09121}{99}} = 0.03035$$

$$\text{Note : s} = \sqrt{\frac{1}{n-1}\left[X^2 f - n\left(\overline{X}\right)^2\right]}$$

Coefficients of Variation: The Standard deviation is an absolute measure of dispersion that expresses variation in the same units as the original data. It cannot be a sole basis for comparing two distributions especially if the data are measured on different scales or if larger mean has larger variation. In such cases, we use coefficient of variation. It is a relative measure of variation. It relates the standard deviation and the mean and expresses standard deviation a percentage of mean. The formula for coefficient of variation is

$$\text{Coefficient of variation (CV)} = \frac{\text{S tan dard deviation } (\sigma)}{\text{Mean } (\mu)} \text{ x } 100.$$

Example: Laboratory 1 can compete on an average 40 analyses per day with a standard deviation of 5. Whereas laboratory 2 can complete 162 analyses per day with a standard deviation of 15. Which laboratory shows more consistency.

Solution: At first glance, it appears that laboratory B has three times more variation in the output as compared to Laboratory A. But Laboratory B has more output per day. Considering all this, we need to compute the coefficient of variation.

Lab 1: Coefficient of Variation = $\frac{5}{40} \text{ x } 100 = 12.5\%$

Lab 2: Coefficient of Variation = $\frac{15}{160} \text{ x } 100 = 9.4\%$

Laboratory B has less relative variation.

BOX PLOT

Box-and-whisker plot (box Plot) is a powerful graphical summary of distributional characteristics of data. The box plot captures main features of location, spread and shape of a distribution. It provides an informative, transparent data display for decision making.

A box plot consist of a box, whiskers and outliers.

The box plot as drawn is shown in Figure 1

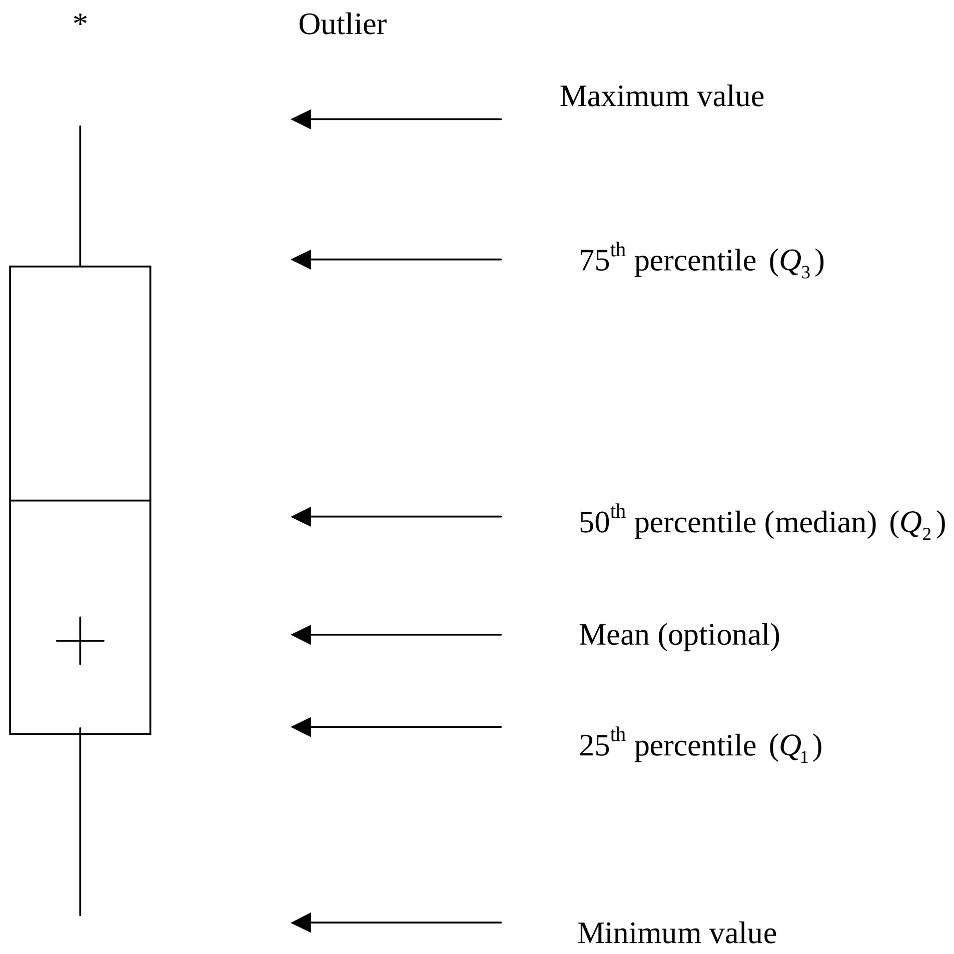


Figure 1 : Box Plot

The box contains a middle 50 percent of the data, the bottom of the box is at the first quartile (Q_1) and the top is at the third quartile (Q_3) value. The median, the mid point of the data set is shown as a line across the box. Therefore $\frac{1}{4}$ of the distribution is between this line and the top of the box, and $\frac{1}{4}$ of the distribution is between this line and the bottom of the box. Thus the median line divides the box into two smaller boxes which represent the upper and the lower central quarter of the data.

The outer parts of the data set or the ``tails'' are the whiskers. The whisker are plotted by lines that extend from the top and the bottom of the box to the extreme data values (maximum and minimum) that are not taken to be outliers.

The mean of the data set is indicated by plus sign (+).

Interpretation of the box plot

- Box covers middle half of the data
- Whiskers show range of the data
- Symmetry is indicated by the box, the whiskers and the location of the mean. Closer the mean is to the median, the more symmetrical the distribution. In case of skewed data the box plot is not symmetric
- Position of the box and median gives the location of the data
- The length of the box is proportional to the inter quartite range (IQR) gives the dispersion. Thus larger boxes have larger dispersion
- Outliers are points outside the lower and upper limit and are plotted with asterisks.

  The limit used for identifying the outliers are:

  Lower Limit: $Q_1 - 1.5(Q_3 - Q_1)$

  Upper Limit: $Q_3 + 1.5(Q_3 - Q_1)$

  The length of the whiskers should not exceed 1.5 $(Q_3 - Q_1)$
- Different data sets from two or more groups can be compared by constructing box plots side-by-side. In this case width of the boxes can be drawn proportional to the sample size of the data sets.

Exercise : Consider the data on presence of methyl tertiary butyl ether in water from two different sources.

Source A: 509, 509, 516, 518, 510, 514, 511, 504, 523, 501, 511,
503, 510, (MTBE in Water) 495, 506, 511, 533, 512, 509, 507

Source B: 504, 524, 515, 508, 513, 520, 536, 529, 521, 510, 502,
528, 536, (MTBE in Water) 528, 516, 511, 519, 512, 524, 523

CHAPTER – 5

PROBABILITY AND STATISTICAL DISTRIBUTIONS

Introduction:

In our everyday language, when we use the phrases ' most likely ; 'highly probable', 'less likely' etc. what we are really doing, almost unconsciously, is that we are expressing our degree of belief in the occurrence of certain special events. In probability theory, these concepts are formalized and rules are developed to obtain quantitative estimates of probability of events so that the estimation procedure is freed from the shackles of subjective judgment.

Experiments And Events :

An experiment is some well-understood procedure or process governed by a set of rules, whose outcome can be observed. A random experiment is an experiment whose outcome is not uniquely determined by any theory; but the set of possible outcomes is determined. The characteristic feature of an experiment is that it can be repeated infinitely. The set of all possible outcomes is called the sample space. An event is any subset of the sample space.

Examples of deterministic experiments are :

(i) Observing the distance traveled by a car running with an average speed of 45 Km. for three hours.

(ii) Measuring exactly the amount of heat generated in an electrical circuit having resistance R ohms and carrying current I for t seconds,

In these experiments the outcomes can be predicted with great accuracy using the laws of physics and elementary mathematics.

On the other hands a random experiment is characterized by a set of possible outcomes, the set having more than one number.

Examples of random experiment are :

(i) Tossing a coin and observing whether it falls head or tail.
(ii) Observing whether it rains or not the next day.
(iii) Noting down number of absentees exceeding a specified number in a shift.
(iv) Whether an electrical/electronic equipment passes all tests in final inspection.
(v) Observing the number of batches of items accepted by the customer on sampling inspection.
(vi) Number of blow holes in a casting.
(vii) Test result value obtained after testing the sample

An event (or subset of the sample space) of a random experiments may be called a random event.

Probability Theory provided the foundation for use of Statistical methods in solution of problems involving random events.

Probability And its Measure:

Before we evaluate the probabilities of random events in numerical terms, we must choose a unit of measurement. Such a unit is called the probability of a sure event. An event is called a *sure* event . An event is called a *sure event* if it will certainly occur in the given experiment. For instance, the appearance of either head or tail on tossing a coin is a sure event. The probability of a sure event is assumed equal to one, and zero probability is assigned to an *impossible event* i.e the event which in the given experiment, cannot occur at all (e.g. the appearance of ten number of spots on the face of a six-faced die). We can also state intuitively that the probability of the occurrence of head in tossing a coin is 0.5. In most of the practical cases of a random events, the probability will be having values other than 0, 1 and 0.5. However, probability of any random event will always be in the interval between zero and one. Now the question is "How do we measure the probability of an event ".

Classical Method :

The classical method of measuring the probability of an event A arising out of an experiment, denoted by P(A) is

$P(A) = \frac{f}{n}$ where f = number of outcomes favourable to A.

and n = Total number of equally likely, mutually exclusively and collectively exhaustive outcomes of the experiment.

Example - 1 :

Tests of water samples declares water fit (F) or unfit (U) for drinking. What is the probability of getting [Unfit, Unfit] drinking water samples in two test-tubes ?
The possible outcomes are (U,U), (U,F), (F,U), and (F,F).
Thus the number of outcomes favourable for the event = 1.
The number of equally likely, mutually exclusive and collectively exhaustive outcomes = 4
So, the required probability $= \frac{1}{4} = 0.25$

Example

What is the probability that the sum of scores will be 9 while throwing two dice together having six faces each?

Here the numerator can be realized from four outcomes viz . (3,6), (4, 5), (5,4) and (6, 3) and the total number of all possible outcomes = 36.

So, the required probability = $\frac{4}{36} = \frac{1}{9}$

The Classical Method is inapplicable in situations where we can not assume that the outcomes of experiments are equally likely. For example, classical method cannot help us in finding the probability of the following events :

1. Occurrence of head in tossing a biased coin.

2. Failure of a tube/equipment/device after working for 1000 hours.

To help us in such situations, we make use of the Relative Frequency method of assessing probability :

Relative Frequency Method :

Relative frequency of an event A

$$= \frac{\text{Number of times event A is observed}}{\text{Total number of trials conducted or observations made}} = \frac{f}{n}$$

Probability can be estimated by the relative frequency where *n* is infinitely large. Symbolically,

Probability of an event

$$A = \lim_{n \to \infty} \frac{f}{n}$$

What we really mean is that the relative frequency of an event tends to its probability. To be more specific, the implication is that if the number of independent trials is sufficiently large, then with a practical confidence the relative frequency will be as close to the probability as desired.

The above method is valid due to ' statistical stability ' in the occurrence of random events and provides us with the method of estimating probability as long term proportion of occurrences.

Next we discuss a few indirect methods which make it possible to calculate the possibilities of composite events in terms of probabilities of simpler events. They are the Addition Rule and Multiplication Rule of probability.

Addition Rule of Probability

The probability that one of two mutually exclusive events (it does not matter which of them) occurs is equal to the sum of the probabilities of these events.

This rule is expressed by the formula.

$$P(A \cup B) = P(A) + P(B).$$

Example :

The LSL and USL of ambient air of an area of good category has AQI in the range of 20 to 40.

Let us define the elementary events as follows :

A : the event that sample has AQI below LSL

B: the event that sample has AQI above USL

The probabilities of these events was estimated from the past inspection records from a particular area by making use of relative frequency method. Accordingly it is found that P(A) = .05 and P(B) = .03. What is the probability that a random sample taken from the same area will be outside the specification?

So, the required probability can be estimated as

P(A U B) = (A) = .05 + .03 = .08

The practical interpretation of this probability is that we expect 8% of the sample will be outside the specification as long as no major changes in the air quality takes place.

Multiplication Rule of Probability

If A and B are two independent events, the probability of their joint occurrence is equal to the product of the probabilities of the two events .
This rule is expressed by the formula

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

If the two events occur in successive trials then the order occurrence should also be taken into account.

Example :

a. In the previous example of AQI of ambient air, what is the probability that in two successive samples taken, one sample is below LSL due to low AQI and the other sample is above USL due to high AQI ?.

D_L : Sample is below LSL due to Low AQI
D_H : Sample is above USL due to high AQI

A : The event that first sample has low AQI and second sample has high AQI
B : The event that first sample has high AQI and second sample has Low AQI.

$$P(A) = P(D_L)P(D_H) = .05 \text{ x } .03 = .0015$$
$$P(B) = P(D_H)\ P(D_L) = .03 \text{ x } .05 = .0015$$

So, the required probability = P(A) + P(B) = 0.0015 + 0.0015 = 0.003

Statistical Distributions:

Binomial Probability Distribution

The conditions for occurrence of Binomial probability distribution as follows :

1. Outcome of a trial is classified as ' success' or 'failure'.
2. Probability of success ' p' remains constant from trial to trial .
3. *n* independent trials are made.
4. Random variable of interest is the number of success (x) in n trials.

The probability of getting x number of successes in n trials is given by

$$P(X=x)=p(x)=\binom{n}{x}p^{x}(1-p)^{n-x} \text{ for } x=0,1,2\cdots,n$$

Also note that $E(X) = np$ and $V(X) = npq$ where q = 1 – p

Example : For n = 18 and p = 0.1

x :	0	1	2	3	4	5	6	7
P(x) :	0.150	0.300	0.284	0.168	0.070	0.022	0.001	0.000

Average (X) = *np*

Variance (X) = *npq* and standard deviation of (X) = $\sqrt{npq}$

Poisson Distribution:

Examples of random variables having Poisson distribution are

i. number of breakdowns in equipments in fixed time intervals
ii. number of defects in slides of same type
iii. number of pathogens in drops of water of same point.

Thus the random variables, of interest can occur either in time or in space. The conditions given rise to Poisson distribution are

1. Occurrence of event in short interval of time or space is proportional to time or space interval.
2. Probability of two or more occurrence in short interval of time or space is negligible and can be considered zero.,
3. occurrences are independent

Random variable : Discrete

Range of $X = 0, 1, 2, 3, \cdots$

$$P(X = x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

Where λ = average number of occurrences in fixed time interval or fixed size in space.

Average (X) = λ &

Variance (X) = λ &

Standard Deviation = $\sqrt{\lambda}$

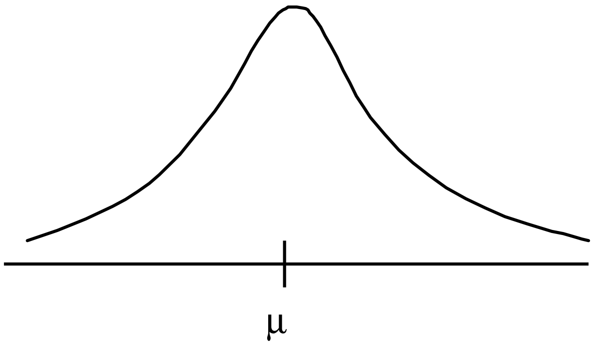
Normal distribution:

The normal distribution is the most important continuous probability distribution. It has been useful in countless applications involving every conceivable discipline. The usefulness is due in part to the fact that the distribution has a number of properties that make it easy to deal with mathematically. More importantly, however, the distribution happens to describe quite accurately the random variables associated with a wide variety of experiment.

The normal distribution is completely specified by two parameters viz mean (μ) and standard deviation (σ). The probability density; function of normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}} \quad for \;\; -\infty < x < +\infty$$

where (μ) = mean and (σ) = standard deviation



Graph of normal distribution

Examples of random variables are measurable characteristics like length and diameter of components, chemical properties (say S_i % in Grey Iron), electrical characteristics like resistance etc.

The theoretical justification for the occurrence of Normal distribution is provided by the CENTRAL LIMIT THEOREM which states that the sum of a number of independent and identically distributed random variables each with a finite mean and variance will be closer to a Normal distribution as the number of random variables increases.

Thus, when a random variable represents the total effect of a large number of independent small causes, the Central Limit Theorem leads us to expect the distribution of that variable to be Normal.

Properties of a Normal distribution

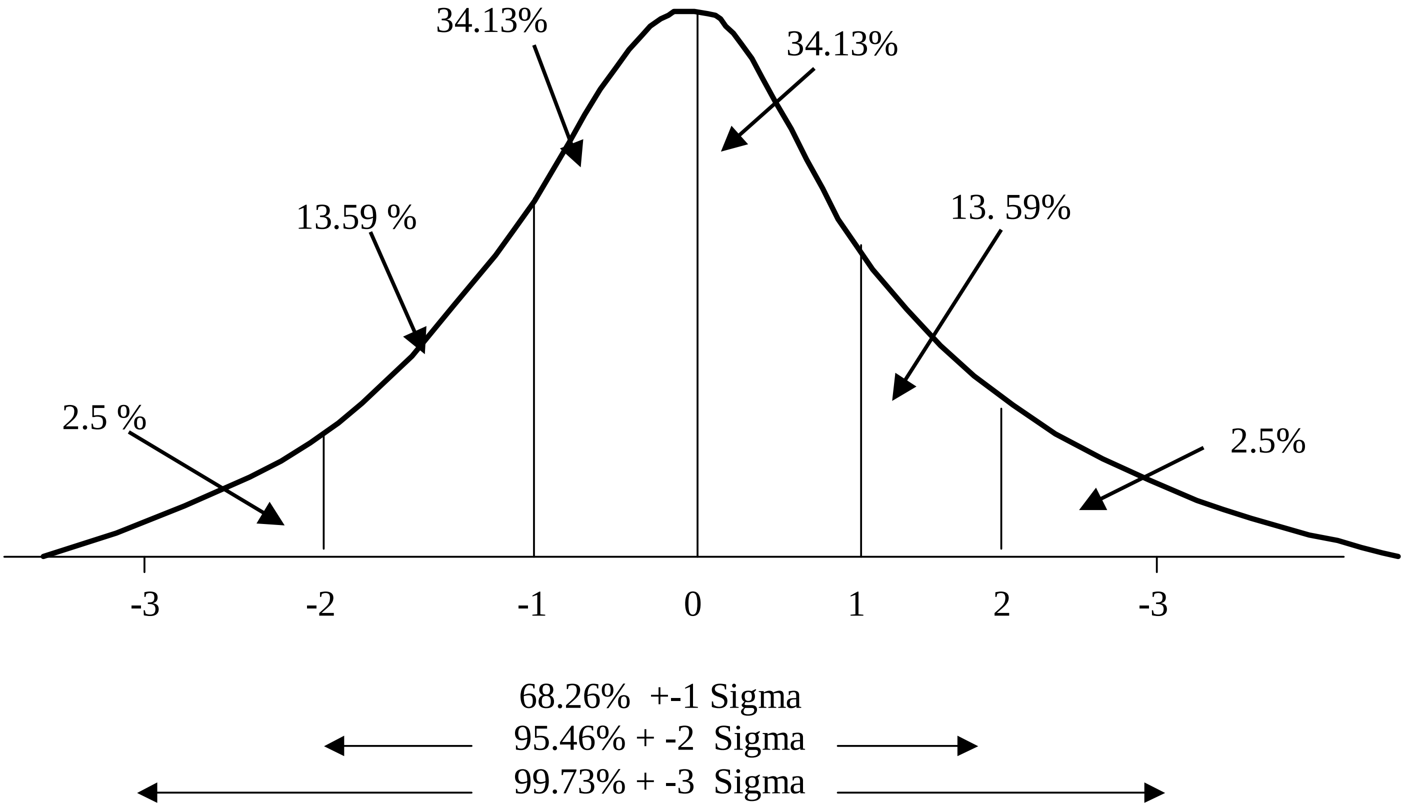
1. It has a bell shape.
2. It is single peaked and thus unimodal.
3. It is symmetric about the mean and Mean = Median = Mode. All located at the center of the curve.
4. The two tails of the extend indefinitely and never touch the horizontal axis.
5. The location and shape of the curve is determined completely by mean (μ) and standard deviation (σ).

Area under the Normal Curve

The total area under the normal curve is one. The area under the curve is interpreted as the probability that is for any normal distribution with mean (μ) and standard (σ), the area under the curve for selected interval between mean $\pm k\sigma$ that is between Mean $-k\sigma$ and Mean $+k\sigma$ is tabulated below

K	**Area in %**
1	68.26
2	95.46
3	99.73
1.96	95.00
2.58	99.00
3.09	99.90
4	99.99366
5	99.99994266
6	99.99999980

For k = 2, Area under the curve between the limits Mean 2σ and mean $+2\sigma$ is given by 0.9546. That is
$\Pr ob\,[\mu - 2\sigma < x < \mu + 2\sigma\,] = 0.9546$ *and so on*



Standard Normal Distribution

A normal distribution with mean zero and standard deviation equal to one is called a standard normal distribution. Thus the standard normal distribution has $\mu = 0$ and $\sigma = 1$. If a random variable X has a normal with mean μ and standard deviation σ, the random variable defined as

$$Z = \frac{X - \mu}{\sigma} = \frac{X - \text{Mean}}{\text{S.D.}}$$

has mean zero and standard deviation one. Z is called the Standard normal variable. Normal probability table is provided for the standard normal variable (Z). It may be noted that

$$P[a < x < b] = P\left[\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right]$$

This implies that

$$P[\mu - \sigma < X < \mu + \sigma] = P[-1 < Z < 1] = 0.6826 \textit{ and}$$
$$P[\mu - 2\sigma < X < \mu + 2\sigma] = P[-2 < Z < 2] = 0.9546$$

Example :

A softdrink machine is regulated so that it discharges on average 300 ml per cup. If the amount of drink is normally distributed with a standard deviation of 20 ml.

(a) What fraction of cups will contain more than 325 ml.

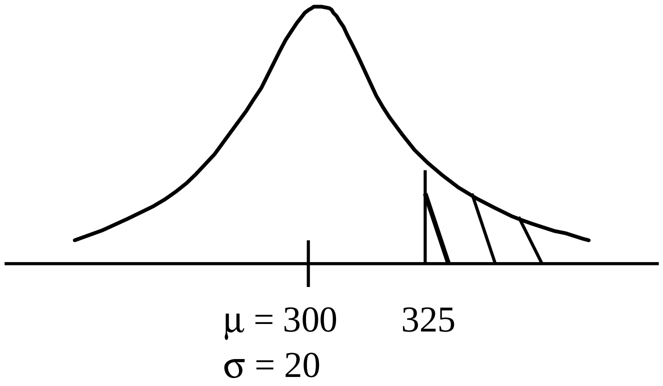
(b) What is the probability that a cup contains between 285 ml and 335 ml.

(c) What is the probability that cup will contain exactly 300 ml.

(d) How many cups will overflow on an average if cups of size 340 ml are used for the next 2000 drinks.

Solution : Let X denoted the quantity of soft drink per cup. Here Mean (μ) = 300 ml and $\sigma = 20\,\text{ml}$ and X follows normal distribution. We are required to find :

P[X > 325 ml] or

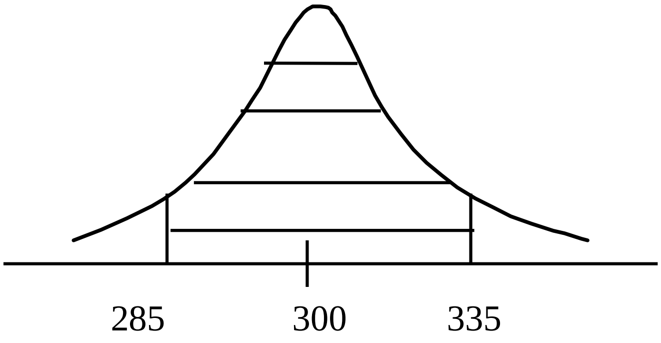


Area under the normal curve above 325 ml. Value of Z corresponding to X = 325 is

$Z = \frac{325 - 300}{20} = \frac{25}{20} = 1.25$. From tables we get this area equal to

1 - P[Z < 1.25] = 1 - 0.8944 = 0.1056 = 10.56%. Hence 10.56% of the cups are expected to contain more than 325 ml.

(b) In this case, we need to find the area between the limits 285 ml and 335 ml.



Z value corresponding to 335 is

$$Z = \frac{335 - 300}{20} = \frac{35}{20} = 1.75$$

Z value corresponding to 285

$$Z = \frac{285 - 300}{20} = \frac{-15}{20} = -0.75$$

Required probability is = P [Z < 1.75] - P [Z < - 0.75] = 0.7333

Hence 73.33% of cups are likely to have soft drink between 285 ml and 335 ml.

(c) The probability is zero.

(d) In this case, we first find the chances that X will be more than 340 ml.

Value of Z corresponding to 340 is

$$Z = \frac{340 - 300}{20} = \frac{40}{20} = 2.$$

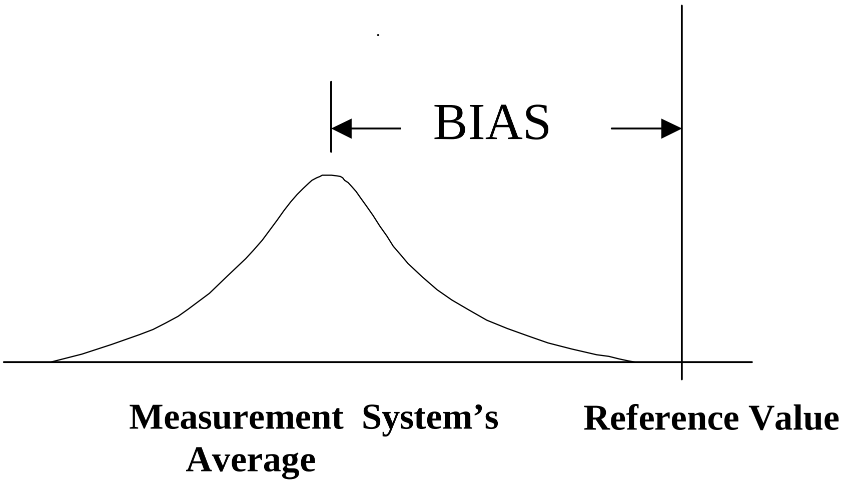
Hence, P [Z > 2] = 1 - [Z ≤ 2] = 1 – 0.9772 = 0.0228.

Number of cups likely to overflow if cups of size 340 ml are used for the next 2000 drinks = 2000 x 0.0228 = 45.6 cups.

CHAPTER – 6

REPEATABILITY & REPRODUCIBILITY

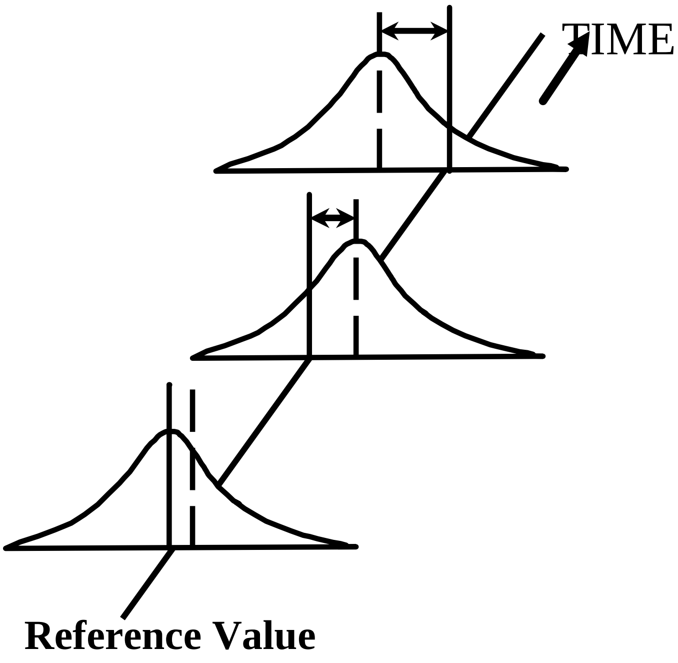
Bias : It is the difference between the observed average value of a characteristic and the true value or accepted reference value. Bias is the measure of the total systematic error as compared to random error. There may be one or more systematic error components contributing to the bias. A large systematic difference from the accepted reference value is reflected by a larger bias value.



Laboratory Bias: The difference between the observed average of the test results from a particular laboratory and an accepted reference value.

Bias of Measurement Method: The difference between the observed average value of test results obtained from all laboratories using that method and an accepted reference value.

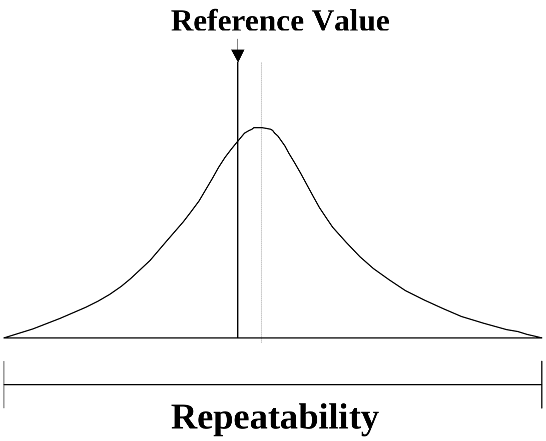
Stability: Stability is the total variation in the measurements obtained with a measurement system on the same samples when measuring a single characteristic over an extended time period. That is, stability is the change in bias over time.



Precision: The closeness of agreement between independent test results obtained under stipulated conditions. It describes the net effect of discrimination, sensitivity and repeatability over the operating range: Size and time of the test or measurement system. ASTM defines precision to include the variation from different reading, instruments, people or conditions.

The measure is usually expressed in terms of imprecision and computed as a standard derivation of test result. Higher standard derivation reflects less precision.

Repeatability : Variability in independent test results obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short interval of time. It is the inherent variation in a within a system when conditions of testing are fixed and defined, fixed samples, parts, instrument, standard, method, operator and environment and assumptions.



Measures commonly used are : Repeatability standard derivation, repeatability variance, repeatability coefficient of variance.

Possible Causes for poor repeatability include :

- Within sample or parts: form, position, consistency
- Within-instrument : repair, wear, equipment or fixture failure, poor quality or maintenance.
- Within-standard: Quality, class, wear
- Within-method: Variation in set up, technique, zeroing, holding, clamping, point density.
- Within-appraiser: Technique, position, lack of experience, manipulation skill or training, feel, fatigue.

- Within-environment: Short cycle fluctuations in temperature, humidity, vibration, lighting, cleanliness.

- Violation of an assumption: stable, proper operation

- Instrument design or method lacks robustness, poor uniformity

- Wrong gauge for the application

- Application: part size, position, observation, error (readability, parallax)

Repeatability: Variability or precision in test results obtained with the same method on identical samples in different laboratories with different operators using different equipment.

Reproducibility is typically defined as the variation in the average values of the measurements or test values obtained by different laboratories.

Potential sources of reproducibility error include:

- **Between samples:** average difference when measuring types of parts, A, B, C etc, using the same instrument, operators and method

- **Between-instruments**: average difference using instruments A, B, C etc., for the same parts, operators and environment.

- **Between-Standards:** average influence of different setting standards in the measurement process.

- **Between-methods:** average difference caused by changing point densities, manual versus automated systems, zeroing, holding or clamping methods, etc.

- **Between-appraisers (Lab Assistants):** average difference, between A, B, C, etc., caused by training, technique, skill and experience.

- **Between-environment:** average difference in the measurements over time caused by environmental cycles, this is the most common study for highly automated systems in product and process qualifications.

- Violation of an assumption in the study.

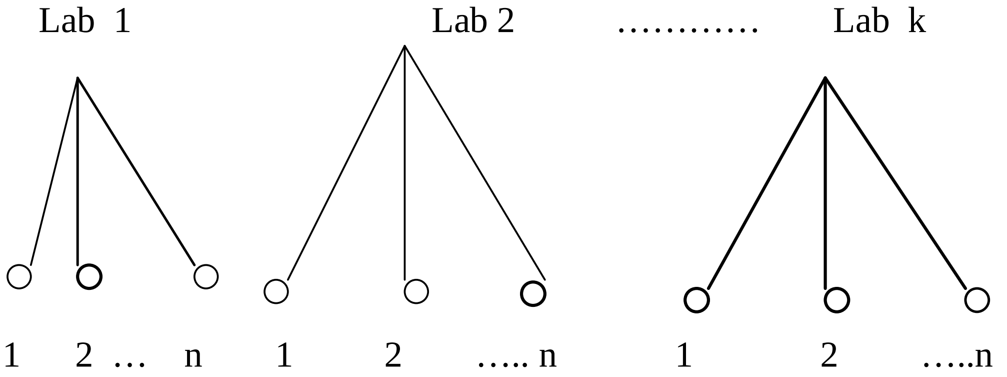
- Instrument design or method lacks robustness.

- Lab Assistants training effectiveness

- Application – part size, position, observation error (readability, parallax).

Estimation of Repeatability and Reproducibility :

In an inter-laboratory programme, a large number of laboratories carry out repeats test on the same sample of homogeneous material. The scheme can be depicted as shown below.



If the sample is tested only once in different laboratories, the variation present in the test results value will reflect combined variability arising from with-in laboratory variation and of the variability arising from between-laboratory variation. In such cases it is not possible to estimate repeatability and reproducibility of the measurement system.

Description of the Model: The repeatability and reproducibility of the measurement system can be estimated from the analysis of the data from a group of laboratories selected from a population of laboratories using the same method model

y = m + B + e

where

m : is the mean of the results
B : is the laboratory components of the bias under repeatability condition
e : is the random variability occurring during any measurement under Repeatability condition.

Let σ_r^2 denote within-laboratory variance. It is known as a repeatability variance.

σ_L^2: the variance of B it is the between Laboratory variance, σ_L^2

The sum of the between laboratory variance and the within-laboratory variability is known as reproducibility variance (σ_R^2)

$$\text{Reproducibility variance} = \sigma_R^2 = \sigma_L^2 + \sigma_r^2$$

Example:

Suppose there are four participating laboratories and each laboratory carry out three repeat test on the same sample.

				Mean	Variance
Laboratory 1	25	27	26	26	1
Laboratory 2	26	22	24	24	4
Laboratory 3	21	24	24	23	3
Laboratory 4	25	24	26	25	1

Estimate of Repeatability

$$S_r^2 = \frac{1 + 4 + 3 + 1}{4} = \frac{9}{4} = 2.25$$

Variance in laboratory means

$= 1.67$ is an estimate of $\frac{\sigma_r^2}{3} + \sigma_L^2$

Hence $1.67 = \frac{2.25}{3} + \sigma_L^2$

$S_L^2 =$ estimate of $\sigma_L^2 = 1.67 - 0.75 = 0.92$

Estimate of reproducibility $\left(S_R^2\right) = S_r^2 + S_L^2 = 2.25 + 0.92 = 3.17$

1. Repeatability limit = $2.8 \text{ x } \sqrt{2.25} = 4.20$
2. Reproducibility limit = $2.8 \text{ x } \sqrt{3.17} = 4.98$

CHAPTER - 7

ESTIMATION AND CONFIDENCE INTERVALS

The objective of statistical inference is to draw conclusions about population characteristic or true value using sample test result values. Statistical estimation makes considerable use of quantities computed from the observations in the sample. We define a statistic as any function of the sample test results. For example, sample mean and sample standard deviations are both statistics.

Types of Estimation:

We can make two types of estimation about a population characteristic or true value using the test results obtained.

A Point Estimation : A point estimate is a single value that is used to estimate an unknown population parameter or true value.

An interval Estimate: An interval estimate is a range of values used to estimate a population parameter or true value.

Point Estimation : The objective of statistical point estimation is to make an estimate of the population or true characteristics with the help of sample statistic. For example we might estimate the true mean octane rating of a particular type of fuel with the help of sample mean $(\overline{X})$. When the size of population is very large or infinite, we never know the true value of the population mean. We can only make estimate of the population mean. The estimate is bound to vary depending upon the random samples selected and the precision of testing method used. The sample mean rarely coincides with the population mean. Some related concepts and definitions are

Estimator: The function of the observation chosen to estimate the population parameter, e.g. sample mean is an estimator of population mean μ. Two important desirable properties of an estimator are: (i) unbiasedness (ii) minimum variance.

Estimate: The particular value of the estimator in given situation

Unbiased estimator: Whose expected or average value taken over an infinite number of similar samples or all possible samples equals the population parameter being estimated.

Standard error : The standard derivation of the estimator. For example, the standard error of sample mean $(\overline{X})$ is given by $\frac{\sigma}{\sqrt{n}}$

Population Parameter	**Point Estimate**	**Standard error**
Population Proportion (P)	Sample Proportion $(\overline{P})$	$\sqrt{\frac{p(1-p)}{n}}$
Population Mean (μ)	Sample Mean $(\overline{X})$	$\frac{\sigma}{\sqrt{n}}$
Population Variance (σ^2)	$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$	

The main disadvantages of point estimate is that it provides us only with a single value as the estimate of unknown population parameter. It does not provide information about the precision of the estimate i.e. about the magnitude of error due to sampling. The sampling distribution of the estimator and the sample size will determine the extent of closeness of the estimate to the true value. In many practical situations it will be desirable not only to provide an estimate but also to establish an interval within which we can expect with a given degree of probabilistic confidence, that the unknown parameter would like. The procedure is known as confidence interval estimation. The width of the interval provides.

Confidence Interval: Any confidence interval has two aspects

(i) An interval computed from the data.
(ii) The confidence level attached to the interval.

It gives the probability that the interval include the parameter or true value. User can choose the confidence level. In most cases 95% confidence or higher is taken. The confidence level is usually written in the $1-\alpha$. For example, 95% confidence level corresponds to $1-\alpha = 0.95$ or $\alpha = 0.05$.

A $1-\alpha$ confidence interval for the parameter θ is given by two statistics U and L such that $P[L \leq \theta \leq U] = 1-\alpha$. L is called the lower confidence limit.
U is called the Upper confidence Limit.

Confidence Interval of Population mean (μ)

(1) When population standard deviation (σ) is known

The $(1-\alpha)$ percent confidence interval for population mean (μ) is given by :

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $Z_{\alpha/2}$ is the value of the standard normal variate exceeded with probability $\alpha/2$.

(2) When population standard deviation σ is unknown. In this case the confidence interval is given by

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $\overline{X}$ is the sample mean and S denotes the sample standard deviation,

given by $s = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$

and $t_{\alpha/2}$ is the value obtained from t distribution with (n-1) degrees of freedom such that it is exceeded by probability $\frac{\alpha}{2}$

Example: 16 observations made on the SO_2 content in the air and the values are given below :

97.99, 96.25, 97.51, 93.63, 92.63, 92.51, 95.44, 94.80
99.21, 89.19, 89.50, 93.73, 97.34, 93.64, 87.25, 96.11.

(i) Obtain a point estimate of the population mean SO_2 content.
(ii) Establish 95% confidence interval for the true mean SO_2 content.

Solution :

We have sample mean ($\overline{X}$) = 94.98 and sample standard deviation = 3.73.

(i) Point estimate of true mean SO_2 content is $\overline{X}$ = 97.98
(ii) 95% confidence level for true mean SO_2 content is given by $\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

We have t - distribution (table) with degrees of freedom 16 -1 = 15, the value of $t_{\alpha/2} = 2.13$. Hence the required confidence interval is given by

$$94.98 \pm 2.13 x \frac{3.73}{\sqrt{16}} = 94.98 \pm 1.97.$$

That is the true mean SO_2 content lies between 93.01 and 96.95.

Example : Out of 1000 people treated with air of particular composition, 200 showed allergic reaction. With 99% confidence level, estimate the proportion of the population that would show an allergic reaction to the air of particular composition.

Solution : Here sample proportion showing allergic reaction is $\overline{P} = \frac{200}{1000} = 0.2$. This gives us a point estimate . That is 20% of the people are likely to show allergic reaction.

99 % confidence level is given by

$$\overline{P} \pm z_{\alpha/2} \sqrt{\frac{\overline{P}\left(1-\overline{P}\right)}{n}}$$

$$0.20 \pm 2.58 \sqrt{\frac{0.2 \, x \, 0.8}{1000}} = 0.20 \pm 0.0326$$

that is 99% confidence interval for the proportion of population likely to show an allergic reaction to the particular composition of air is 16.74% to 23.26%

CHAPTER - 8

TEST OF HYPOTHESIS FOR DECISION MAKING

Often, it is desired to test on the basis of sample data whether the population mean or proportion differs from a specified standard or historical value. Hence we are concerned with drawing conclusion about population mean (μ) or proportion based on sample data.

Hypothesis testing begins with an assumptions, called a Hypothesis, that we make about population parameter. Then we collect sample data, produce sample statistics, and use this information to decide how likely it is that our hypothesized population value is correct.

Basic Concepts :

'Null Hypothesis' (denoted by *H_0*), it asserts that there is no difference between the population from which the sample has been selected and the population whose parameter is specified under the hypothesis. Null hypothesis is formulated with the hope of rejecting it. Simultaneously we must stipulate the alternative hypothesis. **Alternative Hypothesis (denoted by *H_1*)** which is formulated with the hope of provisionally accepting it.

Type I and Type II Error :

Whenever we test a statistical hypothesis with sample data, we shall have one of the four possible results along with their probabilities in parenthesis summarised in the table :

Reality → ↓ Decision	H_0 True	H_0 False
Accept H_0	Correct Decision $(1 - \alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Correct Decision $(1 - \beta)$

Type I Error : Rejecting a null hypothesis (H_0) when it is true.

Type II Error : Accepting a null hypothesis (H_0) when it is false.

Level of significance (α). The probability of committing Type I error is designated by α called the level of significance.
The probability of committing Type II error is designated by β.

The decision criteria for rejecting the Null Hypotheses can also be stipulated in terms of 'critical region' of the *test statistic* which is a function of sample observations and whose sampling distribution is known under the assumption of null hypothesis.

Critical region is that range of values for test statistic whose probability of belonging to that range is equal to the level of significance when the null hypothesis is true.

One Sided Test

It may be noted that a test of any statistical hypothesis where the alternative is one sided such as

$$H_0 : \mu = \mu_0 \; ; \; H_1 : \mu < \mu_0 \quad \text{or perhaps}$$

$$H_0 : \mu = \mu_0 ; \; H_1 : \mu > \mu_0$$

is called a one-tailed test. The critical region for the $\mu > \mu_0$ lies entirely to the right tail of the distribution of the test statistic concerned, while the critical region for the alternative hypothesis $\mu > \mu_0$ lies entirely to the left tail. A test of any statistical hypothesis where the alternative is two sided such as

Both Sided Test

$$H_0 : \mu = \mu_0 ; \quad H_1 : \mu \neq \mu_0$$

is called a two-tailed test. The alternative hypothesis states that either $\mu < \mu_0$ or $\mu > \mu_0$. Values in both tails of the distribution constitute the critical region.

Whether one sets up a one-sided or a two sided alternative hypothesis will depend on the conclusion to be drawn if H_0 is rejected. The location of the critical region can be determined only after H_1 has been stated.

Steps for Test of Hypothesis

Now, for give sample size, we can outline the steps in tests of hypothesis as follows :

1. State the null hypothesis (H_0) and alternative hypothesis (H_1).
2. Choose the level of significance α .
3. Select a statistic whose sampling distribution is known if (H_0) is true and certain other assumptions are satisfied.
4. Find the critical region for the statistic, which depends on α and the probability distribution of the statistic.
5. Compute the statistic.
6. Draw the conclusions. If the statistic falls in the critical region, reject (H_0) . Otherwise accept it provisionally till further evidence is accumulated and tested again.

Test for Specified Proportion :

Here we have to test the null hypothesis $H_0 : P = P_0$ on the basis of sample proportion p = r / n where r is the number of items falling into the category of interest out of n randomly selected items.

For $H_0 : P = P_0$ the appropriate test statistic may be 'number of items falling into the category of interest' (R) which follows Binomial Distribution.

When either $P \leq 0.5$ and $np > 5$ or $P > 0.5$ and $n(1 - P) > 5$, we may use Normal approximation to Binomial distribution so that the test statistics becomes

$$Z = \frac{r - n\,P_0}{\sqrt{n\,P_0\,(1 - P_0)}}$$

Z follows standard Normal distribution.

The test statistic can also be expressed in terms of sample proportion p = r/n. In this case

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0\,(1 - P_0)}{n}}}$$

Here also Z follows standard Normal Distribution.

Example :

A company claimed that in a particular region 55% of the consumer use products made by them. In a random sample of 1000 consumers 510 consumers agreed that they actually use the product made by this company. Can be conclude at 5% level of significance that the claim of the company is correct ?

H_0 : (P = .55) against H_1 : (P < 0.55)

$$\text{Test Statistic } Z = \frac{r - nP_0}{\sqrt{nP_0\,(1 - P_0)}} = \frac{510 - 550}{\sqrt{1000 \times .55 \times .45}}$$

$$= \frac{-40}{\sqrt{247.5}} = \frac{-40}{15.732} = -2.54$$

Critical region will be Z = < - 1.64.

Conclusion :

Since the observed value of the test statistic falls in the critical region, hence we reject H_0 and conclude that the company claim is not correct.

Testing for a Mean (μ) :

A random sample of 8 cigarette of a certain brand has an average tar content of 18.6 milligram and sample standard deviation of 2.9 mg. Is this in line with manufacturer's claim that average tar content does not exceed 17.5 mg. Take $\alpha = 1\%$.

$$H_0 : (\mu = 17.5) \;\; VS \;\; H_1 : (\mu > 17.5)$$

$$T = \frac{\overline{X} - \mu}{s / \sqrt{n}} = \frac{(18.6 - 17.5)}{2.9 / \sqrt{8}} = \frac{1.1}{1.025} = 1.073$$

Critical Region : From t – distribution with d.f. = 8 – 1 = 7 lies in the upper tail (2.99 + ∞). Do not reject *H_0* and hence accept *H_0* .

CHAPTER – 9

REGRESSION ANALYSIS

INTRODUCTION

In many fields like business, administration, transport, education and in industry, we are required to establish the relationships among variables of interest. For example, the relationship between price and demand, the number of units produced and production costs, absenteeism rates and overtime costs, input and output. The nature of the relationship helps us to make predictions or forecasts, provide detailed understanding of processes, exercise better control, and to optimise our processes and systems. One way to find the relationship is by means of regression analysis.

Regression Analysis

Regression analysis provides quantitative techniques for establishing the relationship as a formula between the variables being considered. Regression analysis enables us to determine and utilize a relation between a variable of interest, called the dependent variable, and one or more independent variables or predictor variables. Y denotes the dependent variable whose value we want to predict. X denotes the independent variable or predictor variable. After we have estimated the relationship, we use correlation analysis to determine the strength of the linear relationship. The correlation analysis tells us how well a formula/equation actually describes the linear relationship.

Relationship between two variables

It is important to understand the difference between mathematical and statistical relationships.

Mathematical Relationships

When the mathematical relationship between X and Y is exact, the value of Y is exactly determined once the value of X is specified. For example

$$Y = 100 + 50\,X$$

where X denotes the number of persons attending a dinner party and Y the cost of the dinner party. Overhead cost for the party is Rs. 100 and the dinner cost is Rs. 50 per head. In this case, once we specified the value of the variable X, the value taken by the variable Y is completely known to us. If 10 people attend the party, that is X = 10, the cost of dinner is

$$Y = 100 + 50 \times 10 \text{ or } Y = 600.$$

Statistical Relationship

In this case, the value of the dependent variable Y is not completely determined by the value of the independent variable X. For example, we may find that two families have the same income but their expenditures on food items are not same. This difference may be due other factors like, age or a difference in food habits that we are not considering.

Similarly, consider the relationship between fuel consumption and the speed of a vehicle. This relationship will not be exact as the fuel consumption depends on other factors apart from the speed of the vehicle. These factors include driving habits, road conditions and the age of the vehicle. In regression analysis, we must know or assume the functional form of the relationship between the variables. This is done by setting Y equal to some function which depends on X and on some parameters. We may arrive at the desired function by one of two methods.

(a) From analytical or theoretical considerations.
(b) By examining the scatter diagram obtained by plotting the data on a X by Y plane. The values of Y are represented on the vertical scale and the X values on the horizontal scale. The pattern of points in the scatter diagram reveals what function form may be used for the purpose of analysis.

FITTING A STRAIGHT LINE

We now consider a basic regression model where the relationship is linear, i.e. for any value of X the mean of Y is given by $\beta_0 + \beta_1 X$. Since, in general, an observed value of Y will be different from this mean value, we denote the difference by ε and write the statistical relation in the form

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where β_0 and β_1 are unknown parameters. β_0 represent the Y intercept and β_1 represents the slope of the line. ε represent the random deviation of the observed Y from the mean value $\beta_0 + \beta_1 X$. ε is called the random component. The value of ε for any observation depends on the possible error of measurement and on the values of the variables other than X. We make the following assumptions for carrying out regression analysis:

i) $E(\varepsilon) = 0$, variance $(\varepsilon) = \sigma^2$ and the error component follows a normal distribution.

ii) The values of X are known i.e., there is no randomness involved in the value of X.

Least squares criterion

The regression parameters β_0 and β_1 are unknown and must be estimated from sample data. Point estimates of the parameters are commonly obtained by a method of estimation called the method of least squares. As per this method we choose the parameter such that the sum of the squares of error is minimum. That is, parameters are obtained such that

$$S = \sum \varepsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

is minimum. It can be shown that the least square estimators for the linear regression are obtained by solving the following two equations:

$$\sum_{=1}^{n} Y_i \quad = \quad n\ \beta_0 \ + \ \beta_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i \quad = \quad \beta_0 \sum_{i=1} X_i \ + \ \beta_1 \sum_{i=1} X_i^2$$

These equations are called normal equations, the solution of which is given by

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1\ \overline{X}$$

where b_0 and b_1 denote the estimates of β_0 and β_1 respectively. $\overline{Y}$ and $\overline{X}$ are the mean values of Y and X respectively. In the least squares method, we minimize the sum of the squares of the vertical distances of the points from the line.

Prediction of mean value of Y

The prediction of E(Y), the mean value of Y for a given value of X, is denoted by $\hat{Y}$, and is given by

$$\hat{Y} = b_0 + b_1 X.$$

$\hat{Y}$ is also called the fitted value of Y. The difference between the observed value Y and the fitted value $\hat{Y}$ is called the residual. Thus the residual for the ith observation is given by $\varepsilon_i = Y_i - \hat{Y}_i$.

The sum of the residual is zero, i.e. $\sum_{i=1}^{n} \varepsilon_i = 0$. The estimate of error variable σ^2 is given by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2$$

Analysis of Variance Approach

Analysis of variance (ANOVA) is highly useful technique for regression analysis.

Partitioning of Total Sum of squares

The uncertainty associated with a prediction is related to the variability of the Y observations as given by the deviations $Y_i - \overline{Y}$. The greater the variability in the data, the larger will be the deviations $Y_i - \overline{Y}$. The measure of the variability of the observations is expressed in terms of the sum of the squares of the deviations $Y_i - \overline{Y}$, it is denoted by TSS, Total Sum of Squares and equals

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

It is also called the sum of squares about the mean. This total sum of squares can be expressed as $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$

$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is called sum of squares about the regression or the sum of squares due to error (SSE). $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ is called the sum of squares due to regression (SSR). SSR may be viewed as a measure of the effect of the regression in reducing the variability of Y_i 's. If all the observations fall on the fitted regression line, all deviations will be zero. We have

TSS = SSE + SSR

Partitioning of Degrees of Freedom

Corresponding to the partitioning of the total sum of squares TSS into components SSR and SSE, we have the partitioning of the degrees of freedom.

TSS = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ has $n-1$ degrees of freedom associated with it. SSE has $n-2$ degrees of freedom and finally SSR has one degree of freedom.

Mean Square

A sum of squares divided by its associated degrees of freedom is called a mean square. The regression mean square, denoted by MSR is given by $MSR = \frac{SSR}{1}$. The error mean square is denoted by MSE and $MSE = \frac{SSE}{n-2}$.

ANOVA TABLE

It is helpful to summarise the information on sum of squares (SS), degrees of freedom (df) and mean square (MS) in the form of a table called ANOVA Table. The table below provides a basic format for the ANOVA table for regression analysis.

ANOVA TABLE

Source of Variation	**Sum of Squares (SS)**	**d.f.**	**M.S.**
Regression	SSR = $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	1	$MSR = \frac{SSR}{1}$
Residual	SSE = $\sum_{i=1}^{n}(Y_i - \hat{Y})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	TSS = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	n-1	

$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$ provides an estimate of σ^2 and r^2 = coefficient of determination is given by = $\frac{SSR}{TSS} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}(Y - \overline{Y})^2}$ and from this we can, obtain the correlation coefficient (r).

EXAMINING THE FITTED STRAIGHT LINE

Standard error of intercept (b$_0$) and the slope (b$_1$).

We note that $b_0 = \overline{Y} - b_1 \overline{X}$, and $V(b_0)$ is given by

$$V(b_0) = \frac{\sum_{i=1}^{n} X_i^2 / n}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \sigma^2$$

and standard error s.e. of b$_0$ is the square root of v(b$_0$).

When σ is unknown, the estimate of s.e. (b$_0$) is obtained by replacing σ by s, i.e.

$$s.e.(b_0) = \left(\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X - \overline{X})^2}} \right) s$$

the standard error of slope (b$_1$) is given by

$$s.e.\,(b_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

and is estimated by $\dfrac{s}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$

Confidence Interval for b$_1$

$A\,(1-\alpha)\,100\%$ confidence interval for the parameter β_1 is given by

$$b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

where $t_{\alpha/2}$ is the value of the t-distribution with (n – 2) degrees of freedom.

To test the Null Hypothesis:

$H_0 : \beta_1 = \beta$, we use the test statistics

$$t = \frac{b_1 - \beta}{s / \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

The critical regions are

H_1 :	Reject H_0 if
$\beta_1 < \beta$	$t < t_\alpha$
$\beta_1 > \beta$	$t > t_\alpha$
$\beta_1 \neq \beta$	$t < -t_{\alpha/2}$ or
	$t > t_{\alpha/2}$

Confidence Interval for Mean Response:

$A\ (1 - \alpha)$ 100% confidence interval for the mean response at $X = X_0$ is given by

$$\hat{Y} \pm t_{\alpha/2}\ s \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}$$

where $t_{\alpha/2}$ is a value of the $t-$ distribution with $\mathrm{n}-2$ degrees of freedom

Example

Air quality Index (AQI) is calculated hourly for a certain area (town). The SO_2 content and corresponding AQI given in the following table for 9 hours. Establish a linear relationship between the two variables.

SO_2 $\mu g / m^3$	Air quality index.
34	59
39	69
30	50
33	56
36	64
38	66
45	77
41	73
48	83

Sl. No.	Y	X	XY	X^2	Y^2
1	34	59	2006	3481	1156
2	39	69	2691	4761	1521
3	30	50	1500	2500	900
4	33	56	1848	3136	1089
5	36	64	2304	4096	1296
6	38	66	2508	4356	1444
7	45	77	3465	5929	2025
8	41	73	2993	5329	1681
9	48	83	3984	6889	2304
	344	597	23299	40477	13416

$$\sum Y = 344, \sum X = 597, \sum XY = 23299, \sum X^2 = 40477, \sum Y^2 = 13416$$

$$b_1 = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n(\overline{X})^2} = \frac{23299 - 9 \times (597/9)\,(344/9)}{40477 - 9 \times (597/9)^2}$$

$$= \frac{480.3337}{876} = 0.548326 \quad \text{and}$$

$$b_0 = \overline{Y} - b_1\,\overline{X}$$

$$b_0 = 38.22 - 36.37229 = 1.8499. \quad \textit{We get}$$

$$\hat{Y} = 1.8499 + 0.548326\,X$$

Table of predicted values and residuals

Y	X	$\hat{Y}$	$Y - \hat{Y}$
34	59	34.20117	-0.20117
39	69	39.68442	-0.68442
30	50	29.26624	0.733765
33	56	32.55619	0.44381
36	64	36.9428	-0.9428
38	66	38.03945	-0.03945
45	77	44.07103	0.92897
41	73	41.87773	-0.87773
48	83	47.36098	0.639016

$$s^2 \;=\; \text{estimate of } \sigma^2 \;=\; \frac{\sum (Y-\hat{Y})^2}{9-2}$$

$$= \frac{4.176433}{7} \;=\; 0.5966$$

Total sum of squares (TSS)

$$= \sum_{i=1}^{9}(Y_i - \overline{Y})^2 = \sum_{i=1}^{9} Y_i^2 - n(\overline{Y})^2$$

$$= 13416 - 9 \times (344/9)^2 = 267.5556$$

Sum of squares due to residual

$$= \sum_{i=1}^{9}(Y_i - \hat{Y}_i)^2 = 4.176433$$

Sum of square due to regression

$$= \sum_{i=1}^{9}(\hat{Y}_i - \overline{Y})^2 = 263.3791$$

ANOVA TABLE

Source of variation	SS	d.f.	MS	F – ratio
Regression Residual	263.3791 4.176433	1 7	263.3791 0.5966	441.4422
Total	267.5556	9		

$$R^2 \;=\; \text{coefficient of determination}$$

$$= \frac{S.S.\ due\ to\ regression}{Total\ \ sum\ of\ \ squares} = \frac{263.3791}{267.556} \;=\; .98439.$$

This suggests that 98.439% of the variation in Y is due to the linear relationship.

Standard error of $b_1 = \dfrac{s}{\sqrt{\sum_{i=1}^{9}(X_i - \overline{X})^2}} = \sqrt{\dfrac{0.5966}{876}} = 0.026097$

Standard error of b$_0$ $= s\sqrt{\dfrac{\sum_{i=1}^{9} X_i^2}{n\sum_{i=1}^{9}(X_i - \overline{X})^2}} = \sqrt{\dfrac{40477 \times 0.5966}{9 \times 876}} = 1.75$

95% confidence interval for β_1 is given :

$$b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{\sum_{i=1}^{9}(X_i - \overline{X})^2}}$$

$= \quad 0.548326 \quad + \quad 2.365 \times 0.026097$

$= \quad 0.548326 \quad \pm \quad 0.610019$

Multiple linear Regression

In most regression analysis problems, more than one independent variables are needed. For example the demand for the product may depend on price of the product, the disposable income and price of the substitute product. In such a case, three independent variables will be needed. The model with k independent variables is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \in$$

The unknown parameters $\beta_0, \beta_1, \beta_2, \beta_k$ are called the regression coefficient. $\in$ is the error component with $E(\in) = 0$ and $V(\in) = \sigma^2$.

The method of least square is used for estimating the parameters. In this case we will have $k+1$ normal equations to solve. We generally use a statistical computer package for solving the multiple regression problem. We can also use such model for fitting polynomial models. Consider the quadratic model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

We can redefine the variables as X$_1$ = X and X$_2$ = X^2 and get $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ and solve it as a multiple regression problem. If we let $\hat{Y}$ denote the predicted value of Y, the estimate of error variance (σ^2) is given by $s^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}$ where k is the number of independent variables. It has $n-k-1$ degrees of freedom. In this case, we measure the strength of the relationship in terms of the multiple correlation coefficient (R) or the coefficient of multiple determination (R^2). (R^2) gives the fraction of the variation in Y that is explained by multiple regression. R^2 is computed as follows

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Example The following table gives the data on Air Quality Index *(Y)* and corresponding carton monoxide in mg / m^3 *(X_1)* and NO_2 in $\mu g / m^3$ *(X_2).*

Y	X_1	X_2	$\hat{Y}$
38	1	5	47.2375
40	2	5	55.0625
85	3	5	62.8875
59	4	5	70.7125
40	1	10	38.4625
60	2	10	46.2875
68	3	10	54.1125
53	4	10	61.9375
31	1	15	29.6875
35	2	15	37.5125
42	3	15	45.3375
59	4	15	53.1625
18	1	20	20.9125
34	2	20	28.7375
29	3	20	36.5625
42	4	20	44.3875

(a) Calculate the least-square equation to predict the AQI from SO_2 and NO_2.

(b) Predict AQI if SO_2 is 3 and NO_2 is 6.

We assume that: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
To estimate the parameters β_0, β_1 and β_2, we use the software package Excel (MS Office)
It gives the following output:

Multiple correlation coefficient (R) = 0.8001194 coefficient of multiple determination (R^2) = 0.640191

Anova Table

Source of variance	S.S.	d.f	M.S F
Regression Residual	2764.63 1553.81	2 13	1382.3125 119.52404
Total	4318.44	15	

$\hat{Y} = 48.1875 + 7.825\ X_1 - 1.755\, X_2$

Predicted value of Y for $X_1 = 3$ and $X_2 = 6$

$\hat{Y} = 48.1875 + 7.825 \times 3 - 1.755 \times 6 = 61.1325$

Activity A

Distinguish between the mathematical relationship and the statistical relationship

Activity B

Identify at least three independent variable to predict the weight of an individual and write down the multiple regression model.

Activity C Specify the type of relationship that is expected between fuel consumption of a car and its speed. Write down the model and suggest how you can estimate the parameters.

The following data pertain to the demand for a product in thousand of units and price charged in rupees in eight different locations

Price	Demand
14	150
12	180
15	112
13	140
18	86
15	124
9	223

a) Obtain the line of best fit.
b) Determine the coefficient of correlation
c) Interpret your results
d) Predict the demand if the price is Rs. 10.

CHAPTER – 10

INTRODUCTION TO DESIGN OF EXPERIMENTS

An experiment is the planning and collection of measurements or observations according to a prearranged plan under controlled conditions for the purpose of obtaining factual evidence supporting or not supporting a stated theory or hypothesis.

In a statistically designed experiment the layout for conducting the individual trials is decided on statistical basis to facilitate subsequent analysis.

Role of Statistically designed experiments:

The main reason for designing an experiment statistically is to obtain unambiguous results at a minimum cost. Obtaining valid results from test programme calls for sound statistical design. In fact, a proper experimental design is more important than sophisticated statistical analysis. Results of a well planned experiment are often evident from simple graphical analysis. However, the world's best statistical analysis cannot rescue a poorly planned experimental programme.

The need to learn about interaction among variables and to measure experimental error are some of the added reasons for statistically designing experiments. The designing of an experiment is essentially the determination of the pattern of observations to be collected. A good experimental design is one that answers efficiently and unambiguously those questions which it is intended to resolve and furnishes the required information with a minimum of experimental effort. To do this, the problem must first be posed as succinctly as possible; as also a list of questions to be answered by the experiment must be correctly formulated. Any experiment must be set up to answer a specific question or questions. Precise information of the question (or questions) to be answered enables the experimenter to state his hypothesis more effectively. The major and minor variables which are supposed to have influence on the process have to be identified and the ranges within which these are to be tested will have to be determined. It is also to be ascertained before the commencement of the experiment whether the factors are independent of functions of other factors. The experimenter should also know before hand what extraneous or disturbing factors must be controlled balanced or minimized and the kind of control that is desirable. To obtain the best and most economical design for an experiment some prior estimate of the experimental error would be required and that must be estimated. The experimenter must also set a clear goal as to what improvements are acceptable. In other words to be of technical and practical importance, he should specify the acceptable degree of difference between the effects and consequences. He should also set failure risks and consequences. For instance, the acceptable risks of failing to find an improvement of size noted above and the risk of claiming an improvement, when none exists, must be specified before hand. When the experimenter has to deal with effects which are large compared with random errors, intuitive judgement may be satisfactory but when the errors are appreciable such a procedure may be misleading. Apparent effects, attributed by the experimenter to such factors as he has varied may in reality arise solely through the accidental fluctuations due to the errors. It is difficult to decide whether a particular result is genuine or due to error. Statistical methods alone in such situations, offer sound and logical means of treatment of data and there can be no alternative to rigid statistical tests. These methods should therefore be regarded as part of the technique which industrial scientists should learn in order to deal with their problems effectively. Statistical tests of significance are often required to establish the significance and extent of each of the regulating variable with the lowest number of trials. In such tests it is usual to postulate that the effect sought does not exist, and to see whether on this hypothesis the observed difference can be attributed to chance.

In summary, planned experimentation on Statistical basis is necessary under the following situations:

i. Even with strict adherence to the process specification evolved over a period of time or borrowed from foreign collaboration, the product quality or productivity remain unsatisfactory.

ii. To confirm the desired results for an alternative design/process assembly methods with minimum data collection.

iii. To distinguish between critical factors (which need to be controlled within narrow limit) and non-critical factors (which do not require to be closely controlled) to prevent or minimize the occurrence of defective product.

iv. To determine the optimum process conditions.

v. To locate source of variability.

vi. To correlate process variables with product characteristics

vii. To compare different products, processes, machines, materials and methods.

viii. To evaluate process capability.

ix. To test different hypothesis and theories.

Advantages of Statistically Designed Experiments:

1. Evaluate the experimental error.
2. Isolate the effect of factors in a quantitative manner.
3. Evaluate the interrelationship or interaction between factors.
4. Reduce uncertainty from conclusions.
5. Extract maximum information from given data.
6. Predict the extent of improvement possible over the existing performance.
7. Obtain answers to questions with optimum cost at known risks.

BASIC PRINCIPLES OF EXPERIMENTATION:

Any experiment is required to establish or disprove some theory formulated about a process. Verification of the theory cannot be absolute and if only it can be shown that the observations are compatible with the theory within reasonable limits of error to which the observations are subjected, we can assume that the hypothesis made are correct.

Experimental Error: It is well known that the results of no two experiments will be in complete agreement despite every effort to maintain the same conditions. This is due to a large number of factors beyond economic control. These differences known as experimental errors introduce a degree of uncertainty into any conclusions that may be drawn from the results of the experiment. The experimental errors can be kept within check by following three cardinal principles of experimentation .

Randomisation, Replication and Local Control:

Randomisation: It consists of scheduling the experiments with the different treatments in a random manner. For example, if two methods of processing are to be compared and two machines are available for the trials, the comparisons may be biased by machine differences, if any. This bias is removed by allocating machines to the methods randomly, thus giving validity to the conclusions drawn on the basis of the results. Randomisation is described as an insurance against extraneous factors. It is to balance the effects of unknown variations in materials, equipments, time etc over which we have no control, to prevent any factor being unduly favoured or handicapped . Randomisation assures validity of statistical tests. Random Number Tables can be used for this purpose. Treatments and experimental units are numbered and then allotment of one to other is made using Random Number tables.

Replication: It is repetition of experiments. The replication of observations helps in estimating the experimental error. This in turn aids in deciding whether the observed differences in responses are due to the treatment effect or due to chance. Also such replication increases the sensitivity of the experiment i.e., the power of detecting true differences between treatments. Number of replications will depend on the magnitude of experimental error and real effect of the factors, desired to be detected.

Local Control or Blocking: To obtain maximum sensitivity it is necessary that different trials are subjected to the same background conditions to the extent feasible. In practice, it may be difficult to ensure such uniformity due to natural variability of material, environmental conditions etc. However, it may be possible to split up a set of treatment within small groups where such variations are less. This is known as Local Control. One method of introducing Local Control is to see that all trials are repeated the same number of times under different conditions. This is known as Balancing. It helps in taking control of heterogeneous experimental conditions. Thus local control is the technique of balancing the effect of known disturbing factors and thereby reducing the error. It ensures uniformity in the background conditions of comparison by isolating known disturbing factors. Local control is ensured by dividing the experimental units into smaller groups, within which the variations are likely to be less than that of the set as a whole. Local control makes an experiment more sensitive, thus avoiding need for a large number of repetitions or replications.

TERMINOLOGY

FACTOR: A variable or an attribute which influences or is suspected of influencing the characteristics or response being investigated e.g., speed, feed, temperature, operator, material etc.

Types of Factors: Qualitative and Quantitative

Qualitative Factors: The level of qualitative factors are limited in number and have no intrinsic order for example operators, machine, type of material etc.

Quantitative Factor: Is the one that can take continuum of possible values e.g., temperature, speed

LEVEL: The values of a factor being examined in an experiment for example, three levels of temperature may be:

Level 1:	800^0 C
Level 2:	850^0C
Level 3:	900^0C

The levels may be chosen at fixed values or they may be chosen from all possible levels by a random process. Unless otherwise mentioned, we shall consider only fixed levels.

TREATMENT: One set of combination of levels (one from each factor) employed in a given experimental trial e.g., an experiment conducted using temperature 800^0C. Furnace F_1 and operator B would constitute one treatment combination. To investigate the effect of different factors we conduct trials with different treatments.

EXPERIMENTAL UNIT: Basic units subjected to trial. The experimental units are allocated to different treatment combinations to facilitate conducting trials with different treatments that is material, equipment and other facilities provided for conducting each trial is an experimental unit.

RESPONSE: Numerical (or attribute) result of a trial with given treatment e.g., output/yield per shift/day, dimension, strength, hardness, success/failure, defect rate, rejection rate etc.,

EFFECT: Change in response due to changes in levels of the factors.

MAIN EFFECT: Estimate of the effect of a single factor obtained independently of the other factors.

INTERACTION: If the effect of one factor is not the same at different levels of another factor, interaction between the two factors exist.

EXPERIMENTAL ERROR: It is the variation in response caused by conditions not controlled in experiment due to either ignorance or inability when the same treatment is repeated.

STATISTICAL APPROACH TO EXPERIMENTATION

Statistical approach to designing and analyzing an experiment requires that the experimenter must have a clear idea in advance of what needs to be studied , what and how the data is to be collected and type of analysis to be done. This calls for a detailed planning of the experimental study process. The various steps involved in designing and conducting experiment are given below:

1. **Defining Purpose and Scope:** It is important that we develop a clear and a specific statement of the problem to be studied. It must be accepted by the team members. It is necessary to list all objectives of the experiment including hypothesis to be tested and questions to be answered. The scope of the experiment in terms of products, markets, customers, processes etc to be covered must also be clearly identified and stated. An unambiguous statement of the problem often helps in better understanding of the process and arriving at the final solution.

2. **Process Analysis:** The purpose of the process analysis is to study the related processes, inputs, outputs and measurement involved to have a clear understanding of the process functioning and control mechanisms used. Considerations should also be given to customer feedback and complaints received. It helps in gathering all information about the systems relevant to designing of experiment. Use of process flow chart is often made for carrying out process analysis. At this stage the experimenter will have rediscovered which factors are important and worth investigating. It also provides some

understanding of which responses need to be considered in the objective of the study and whether the runs of the process need to be grouped or not.

3. **Choosing Factors to Study:** Here, we must decide on the factors of investigation. Making a cause and effect or fish bone diagram of the problem will help identify possible factors. In the initial study, we wish to include as many factors in the design as possible. Such initial studies are called screening experiments. Screening experiments help in the elimination of many factors from further consideration because of their minimal effect on the response variables. We use the factors which various studies or experiences have shown to be influential. Often we consider more levels of a smaller number of factors to better characterize the relationship between the responses and the remaining factors.

After having selected the factors to be studied, we fix the range within which each factor can be experimented. The factors in an experiment may be either quantitative or qualitative. In case of quantitative factors, we must consider how these factors are to be controlled at the designed values and measured. We must also fix the number of levels or values of the factors to be used in the experiment. The levels may be chosen specially or selected at random from the set of all possible factor levels. In case of linear effect of a factor, two levels of a factor are adequate. If we wish to study quadratic or non linear effect of a factor, we require minimum three levels of a factor.

4. Choosing the Response: Response(s) is the observed system output in an experiment or the dependent variable(s) of the experiment. Response selected must be relevant to the objective of the study. Taken together responses represent all the aspects of quality, productivity and performance we wish to study. Response of an experiment could be measured on :

Continuous scale: It should also be decided how each response is to be measured, the type of instrument to be used, accuracy and least count of the instrument needed. The capability of the measurement process needs to be assessed.

OR

Binary: The response is classified as good or bad. Binary response drastically reduces the power of the experimental design to detect the effect of change.

OR

Subjective Rating: Categorise the response in categories with intrinsic ordering of the categories. For example response may be classified as good, normal, or bad. Response can also be obtained in a 10 point scale, say 1 to 10 (bad to good).

In case of binary and subjective rating we need to standardize the inspection process in order to reduce and minimize inspection errors.

5. Choice of Experimental Design: In this step, we decide on the statistical design to be used for conducting the experiment. The number of trials to be made, the composition of each trial and the number of replications needed for each trial. We must also determine the order in which data will be collected and the method of randomization to be used. It also involves writing down the statistical model and finding out what statistical methods to be based for carrying out data analysis. In fixing the size of the experiment, we must strike a balance between the statistical efficiency and cost of experiment.

6. Conducting the experiment: This step involves conducting the experiment and actual data as per plan prepared in the previous steps. It is important to monitor the progress of the experiment and to train and involve concerned persons for better conduct of the experiment. Proper care should be taken while fixing levels of the factors, and taking measurements on the response variables. It is important to maintain uniform environment conditions and not allowing factors which do not form part of the experiment to vary. All non experimental factors must be maintained at constant levels throughout the experiment.

7. Data Analysis: Once the experiment has been conducted and data collected, we go into analysis of data. Statistical methods should be used for analyzing the data. We first verify the model assumption made about the data using appropriate techniques. We commonly use methods of testing of hypothesis, analysis of variance, regression analysis etc for analyzing data. Graphical methods are also frequently used and play an important role in analysis.

8. Conclusion and Recommendation: Once the data analysis is over, we draw conclusions about the results and factors. The statistical conclusion must be physically interpreted and practical significance evaluated. Based on this, recommendations of the findings are made. The use of graph and chart is a very effective way to make presentation of the results and conclusions to the management. If necessary, further experiment may be planned.

COMPLETELY RANDOMISED DESIGN

Description:

These are single factor experiments with no restriction of randomization. The different levels (quantitative, qualitative) of the factor are allotted at random to different experimental units. The number of units for each level of the factor is determined from cost consideration and the power of the test. It is required that the experiment is performed in a random order so that environment in which the treatments are used is as uniform as possible. Such experiments are called Completely Randomised Design

When Used:

i. The experimental units are homogenous, or
ii. The pattern of heterogeneity in the experimental material is not known and hence it is not possible to group them into smaller homogenous blocks.

Data Layout: Here we have k different treatments to compare and **ith** treatment is repeated **n_i** times. Let y_{ij} denote the response on the j^{th} experimental unit $(j = 1, 2, \cdots, n)$ corresponding to the i^{th} level $(i=1,2,\cdots,k)$ of the experimental factor or treatment. The complete data for k treatments are as follows:

	Treatments (levels)				
	1	2	$\cdots$	j	k
Observation	y_{11}	y_{21}	$\cdots$	y_{j1}	y_{k1}
	.	.	$\cdots$	.	.
	.	.	$\cdots$	.	.
	y_{1n_1}	y_{2n_2}	$\cdots$	y_{jn_j}	y_{kn_k}
Total	T_1	T_2	$\cdots$	T_j	T_k
Number	n_1	n_2	$\cdots$	n_j	n_k

The layout of data is also known as one-way classification because only one factor is being investigated.

Model: The statistical model describing the observations is

$$y_{ij} \;=\; \mu \;+\; \tau_i \;+\; e_{ij} \quad for \;\; i \;=\; 1,\; 2,\; \cdots, k \;\; and \;\; j = 1,\, 2,\; \cdots \;, \;\; n_i$$

Where y_{ij} is the j^{th} observation on the i^{th} treatment, μ is the common effect for the whole experiment, τ_i represents the effect of the i^{th} treatment and e_{ij} represents the random error present in the j^{th} observation on the i^{th} treatment.

The error e_{ij} is usually considered a normally and independently distributed (NID) random effect whose mean value is zero and whose variance $\left(\sigma^2\right)$ is the same for all levels. μ is always a fixed parameter, and $\tau_1, \tau_2, \cdots, \tau_k$ are considered to be fixed parameters, if the levels of treatment are fixed. It is also assumed that $\sum_{i=1}^{k} \tau_i = 0$.

If the k levels of treatments are chosen at random, the τ_i's are assumed NID $\left(0, \sigma_\tau^2\right)$. Whether the levels are fixed or random depends upon how these levels are chosen in a given experiment.

Hypothesis and Assumptions: The analysis of a single factor completely randomized experiment usually consists of a one-way analysis of variance (ANOVA) test where the hypothesis $H_0 : \tau_i = 0$ for all i is tested. If this hypothesis is not rejected, then no treatment effects exist and each observation y_{ij} is made up of its population mean μ and a random error e_{ij}. If the null hypothesis is rejected, we shall be interested in grouping or ranking the τ_i's through multiple comparisons.

In applying, the ANOVA techniques, the basic assumptions are :

1. The process is in control i.e., it is repeatable.
2. The distribution of population being sampled is normal.
3. The variance of the errors within all k levels of the factors are homogeneous.

The lack of normality in the dependent variable Y does not seriously affect the analysis when the number of observations per treatment is the same for all treatments.

Rationale for Analysis of Variance: We use the following notation

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij} = Total\ response\ of\ i^{th}\ treatment$$

$$y_{..} = \sum_{i=1}^{k} y_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

$$= Total\ response\ for\ all\ treatments$$

It can be shown that

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2$$

$$= \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2$$

where $\bar{y}_i. = i^{th}$ treatment mean and $\bar{y}_{..}$ = the grand mean. This may be referred as the fundamental equation of analysis of variance. It shows that the total sum of squares of deviation from the grand mean is equal to the sum of squares of deviations between treatment means and the grand mean plus the sum of squares of deviations within treatments i.e.,

$$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{error}}$$

Where SS_{total} is total sum of squares. $SS_{\text{treatment}}$ is called the sum of squares due to treatment and SS_{error} is called the sum of squares due to error (i.e, within treatment). Since $\sum_{i=1}^{k} n_i = N$ in all SS_{total} has N - 1 degrees of freedom. $SS_{\text{treatment}}$ has k - 1 degrees of freedom as the experiment has k level of a factor or k treatments and SS_{error} has N - k degrees of freedom. Each of sum of squares divided by the corresponding degrees of freedoms is called mean square. Mean Square (M.S.) due to treatment $= \dfrac{SS_{treatment}}{(k-1)}$ *and Mean Square due to error equal* $\dfrac{SS_{error}}{(N-k)}$

Mean square due to error provides an estimate of error variance $\left(\sigma^2\right)$.

Further it can be shown that if each of the terms (sum of squares in the above equation is divided by its appropriate degrees of freedom, it will yield two independent chi-square distributed unbiased estimates of same σ^2 when

H_0 is true, their ratio will be distributed as F distribution with k - 1 and $\sum_{i=1}^{k} n_i - k$ degrees of freedom i.e.,

$$\left[\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}_{..}\right)^2 / (k-1)\right] / \left[\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2 / \left(\sum_{i=1}^{k} n_i - k\right)\right]$$

follows $F_{k-1}\ ,\ \sum n_i - k$ distribution.

The critical region is normally taken as the upper tail of the F distribution (Table A) rejecting H_0, if F > F_α where α is the area above F_α.

ANOVA Table : The actual computation will be much easier if we use the following relations.

$$\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{N} = Sum\,of\ squares\ due\ to\,treatment$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{T^2}{N} = Total\,sum\ of\ squares$$

where $T_i = \sum_{j=1}^{n_i} y_{ij}$ and $T = \sum T_i\ \ and\ \ N = n_1 + n_2 + \cdots + n_k$

The term $\frac{T^2}{N}$ is called the Correction Factor (C.F.)

The error SS (or within treatment SS) can be obtained by subtraction . Then, the ANOVA table may be set up as follows.

Source of variance	d.f.	S.S.	M.S.	F
Between Treatments	k - 1	$\sum_{i=1}^{K} \frac{T_i^2}{n_i} - \frac{T^2}{N}$	$\frac{SS_{treatment}}{k-1}$	$\frac{MS_{treatment}}{MS_{error}}$
Within treatment/Error	N - k	*	$\frac{SS_{error}}{(N-k)}$	
Total	$N-1$	$\sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij} - \frac{T^2}{N}$		

* obtained by subtraction

Paired Comparison of Means:

When ANOVA indicates significant differences between treatment means, we shall be interested in making ordered groups of treatments such that they may be considered homogeneous within a group. No unique best method exists but one useful method is Duncan's Multiple Range Test. The steps are as follows:

1. Arrange the treatment means in ascending order.

2. Find the value of the last significant studentised range $r_\alpha(p, f)$ from Table - B for each p = 1, 2,……, k where α is the significance level, p is the number of means lying within and including two means being compared and f is the number of degrees of freedom associated with MS_E , the error mean square.

3. For each p, find the least significant range as $R_p = r_\alpha(p, f)\sqrt{MS_E/n}$

   where n is the sample size for each treatment. For unequal sample sizes, the least significant range should be calculated as $R_p = r_\alpha(p, f)\sqrt{MS_E}$.

4. Consider any subset of p adjacent sample means. Let $\bar{y}_{i.} - \bar{y}_{j.}$ denote the range of the means in this subgroup. The population means μ_i and μ_j are considered to be different if

   $\bar{y}_i - \bar{y}_j > R_p$ for equal sample size

or $(\bar{y}_i - \bar{y}_j)\sqrt{\frac{2n_i n_j}{n_i + n_j}} > R_p$ for unequal sample size.

5. If, within a subgroup, the most extreme pair of means is found to be not significantly different then all means within the subgroup are assumed to be equal with no further testing required. First we compare the largest mean with the smallest mean. If these are found to be significantly different then compare the largest and the second smallest. These comparisons are continued until all means have been compared

with the largest mean. The same process of comparison is repeated for the second largest mean and is continued until all possible pairs of means have been compared.

6.Summarise the results by underlying any subset of adjacent sample means that are not considered to be significantly different at the chosen α level.

7. If the sample sizes are all the same, then there is no question as to the validity of the groupings obtained. However, one must be careful when sample sizes are unequal. In such a situation, all possible paired comparisons should be made.

Example : Four different air-injection systems are being investigated for their efficiency. It is desired to test if there is significant difference between them. Five items of each system are taken and the efficiency of injection in each of them measured. The results are as follows:

Efficiency in System

	A	B	C	D	Total
	35	39	39	23	
	34	37	27	28	
	46	17	35	21	
	30	31	29	17	
	40	21	20	21	
T_i	185	135	150	110	580
$\bar{y}$	37	27	30	22	29.0
$\sum y^2$	6997	3901	4716	2484	18098

Correction Factor $= \frac{T^2}{N} = \frac{(580)^2}{20} = 16820$

SS $_{\text{treatment}}$ $= \frac{1}{5}\left[185^2 + 135^2 + 150^2 + 110^2\right] - 16820 = 590$

Total SS $= \left[35^2 + 34^2 + \cdots + 21^2\right] - 16820 = 1278$

The ANOVA table can now be set up as

ANOVA Table

Source of variance	d.f.	S.S.	M.S.	F
Between systems	3	590	196.67	4.57
Within Systems	16	688	43.0	
Total	19	1278		

$F_{.05\,;\,3,\ 16} = 3.24 \quad and \quad F_{.01\ \ 3,16} = 5.29$

There appears to be significant differences among the systems. We now proceed to group the systems on the basis of their average as per the Duncan's Multiple Range Test Method.

The sample means in ascending order are

$\bar{y}_D$	$\bar{y}_B$	$\bar{y}_C$	$\bar{y}_A$
22	27	30	37

For $\alpha = 0.05,\ f = 16,\ n = 5\ and\ MS_E = 43.0,$ we get from Table - B

p	2	3	4
$r_\alpha(p, f)$	2.998	3.144	3.235
R_p	8.79	9.22	9.49

The results of comparisons of the treatment means is as follows

Treatment pair	P	R_P	Range of Treatment means	Reject $\mu_i = \mu_j$?
A - D	4	9.49	15	Yes
A - B	3	9.22	10	Yes
A - C	2	8.79	7	No

The first group is therefore

D B $\underline{\text{C} \qquad \text{A}}$

C - D 3 9.22 8 No

The second grouping is

D B $\underline{\text{C} \quad \text{A}}$

C - B 2 Not Needed No

So the groupings of the treatment means are

First Group : A, C Second Group :C, B, D

FACTORIAL EXPERIMENTS

In Industrial applications frequently we know that several factors may affect the characteristics in which we are interested and we wish to estimate the effects of each of the factors and how the effect of one factor varies over the level of the other factors. For example quality of weld joints may be affected by type of electrode used, current voltage and gap etc. We are often tempted to test each of the factors separately holding all other factors constant in a given experiment but with a little thought it might be clear that such an experiment might not give the information required. The logical procedure would be to vary all factors simultaneously within the framework of the same experiment. When we do so, we have what is now widely known as a **factorial experiment.**

The factorial experiments are particularly useful in those situations which require the study of the effects of varying two or more factors. In a full factorial experiment all combinations of the different factor levels must be examined in order to elucidate the effect of each factor and their interactions.

Advantages of a Factorial Design:

1. It increases the scope of the experiment and gives information not only on the main factors but on their interactions also.

2. The various levels of one factor constitute replications of other factors and increase the amount of information obtained on all factors.

3. When there are no interactions, the factorial design gives the maximum efficiency in the estimate of the effects.

4. When interactions exist, their nature being unknown, a factorial design is necessary to avoid misleading conclusions.

5. In the factorial design the effect of a factor is estimated at several levels of other factors and the conclusions hold over a wide range of conditions.

Factor : A variable which is believed to affect the outcome or response of the experiment.

Level : Various values of a factor examined in an experiment.

Treatment Combination: A combination of levels of the factors in the experiment.

Experimental Units : Items used in an experiment are referred as experimental units. Examples are machines, patients, cars, plots, engines etc.

Response: A response is the numerical result observed for a particular treatment combination.

Effect : The effect of a factor is the change in response produced by a change in the level of the factor.

Main Effect : Average effect of a factor.

Interaction Effect: If the effect of one factor is different at different levels of the second factor then the two factors are said to interact.

Experiment with all Factors At Two Levels : (2^k Series)
In this case each factor is at 2 levels and there being k factors in all. These levels may be quantitative or qualitative, such as two machines, two operators, the high and low level of a factor. A complete replication of such a design requires $2x2x\cdots x2=2^k$ observations and is called 2^k factorial design.

Notation:

Let A , B , C denote the factors; the levels of A,B,C … are denoted by (1), a; (1), b; (1), c; ……respectively. As a convention, the lower case letters a,b,c ….denote the higher level of the factors. The lower level is signified by the absence of the corresponding letter. Thus the treatment combination bc, in a 2^3 factor experiment, represent the experiment in which factor A is at low level and factor B and C are at high level. The treatment combination which consists of low level of all factors is represented by (1). We shall extend this notation by letting (1) , (a), (b), (ab), (c) ,… be the treatment total corresponding to experimental conditions (1), a, ab, c, respectively.

Main effects and interactions : (2^2 Design)

Consider an experiment involving two factors: Reaction time (A) and amount of catalyst (B) each at two levels - low or high. The effect of these two factors on the chemical yield is to be studied. The results are as follows:

Table 1

		Reaction Time (A)	
		Low (1)	High (a)
Catalyst (B)	Low (1)	40	50
	High (b)	60	72

Effect of reaction time at low level of catalyst (B) = 50 - 40 = 10

Effect of reaction time (A) at high level of catalyst (B) = 72 - 60 = 12

Main effect of reaction time, $A = \frac{10 + 12}{2} = 11$

Alternatively, it can be thought as difference between the average response at high level of A and the average response at the low level of A.

Main effect of A = $\frac{50+72}{2} - \frac{40 + 60}{2} = 11$

That is, increasing reaction time (A)from low level to high level causes an increase of 11 units in the yield. Similarly

Main effect of B = $\frac{60 + 72}{2} - \frac{40 + 50}{2} = 21$

In some experiments, we may find that the difference in response between the levels of one factor is not the same at all levels of the other factors. When this occurs, there is interaction between the factors. For example consider the response data in the Table Shown below:

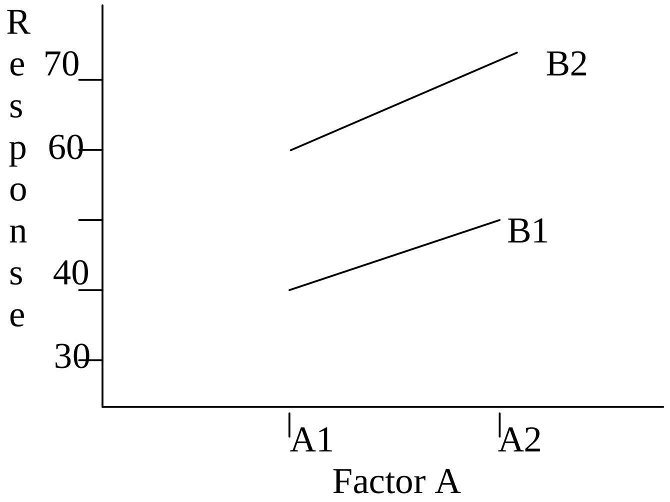
Table 2

		Reaction Time (A)	
		Low (1)	High (a)
Catalyst (B)	Low (1)	40	60
	High (b)	70	32

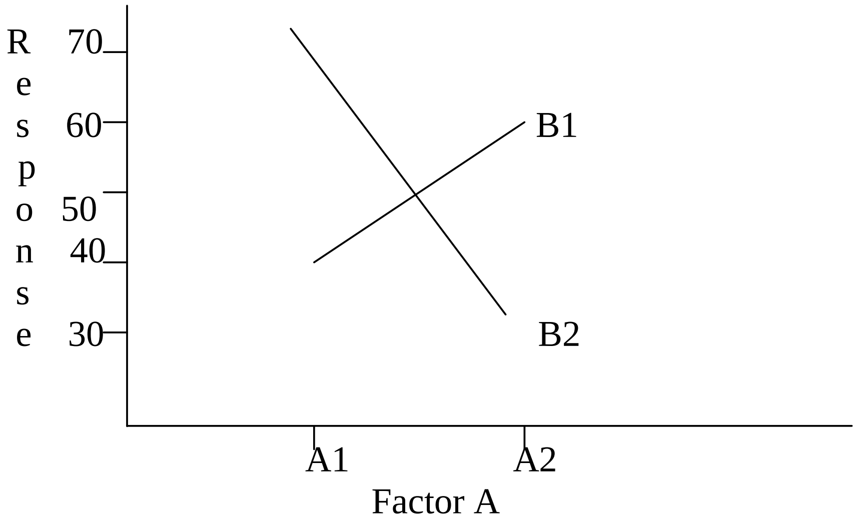
Effect of factor A at low level of B = 60 - 40 = 20
Effect of factor A at high level of B = 32 - 70 = - 38

Since the effect of A depends on the level chosen for factor B, we say that there is interaction between A and B. These ideas may be illustrated graphically.

Factorial Effect Without Interaction



Factorial Effect with Interaction



It is often convenient to write down the treatment combination in the order (1) , a, b, ab. This is referred to as standard order.

Main Effects and Interactions expressed in terms of treatment total for 2^2

Factorial Effect	(1)	(a)	(b)	(ab)	Divisor
M	+	+	+	+	4r
A	-	+	-	+	2r
B	-	-	+	+	2r
AB	+	-	-	+	2r

Where r is the number of replication. Sum of Squares for any effect = $\frac{T^2}{4r}$ where T is factorial effect total.

Example: Consider the experimental data given in table I

$$Main\ effect\ A = \frac{(a) + (ab) - (1) - (b)}{2r}$$

$$= \frac{50 + 72 - 40 - 60}{2} = 11$$

Main effect of B $= \frac{(b) + (ab) - (1) - (a)}{2r}$

$$= \frac{60 + 72 - 50 - 40}{2} = 21$$

$$Interaction\ \ AB = \frac{(1) + (ab) - (a) - (b)}{2r}$$

$$= \frac{40 + 72 - 50 - 60}{2} = 1$$

Sum of squares due to main effect of A

$$= \frac{[(a)+(ab)-(1)-(b)]^2}{4r} = \frac{(22)^2}{4} = 121$$

Sum of squares due to main effect of B

$$= \frac{(42)^2}{4} = 441$$

Sum of squares due to AB

$$= \frac{(2)^2}{4} = 1$$

2^3 Factorial Design:

Main Effects and Interactions expressed in terms of treatment totals

Factorial Effect	(1)	(a)	(b)	(ab)	(c)	(ac)	(bc)	(abc)	Divisor
M	+	+	+	+	+	+	+	+	8r
A	-	+	-	+	-	+	-	+	4r
B	-	-	+	+	-	-	+	+	4r
AB	+	-	-	+	+	-	-	+	4r
C	-	-	-	-	+	+	+	+	4r
AC	+	-	+	-	-	+	-	+	4r
BC	+	+	-	-	-	-	+	+	4r
ABC	-	+	+	-	+	-	-	+	4r

Where r is the number of replications. Each factorial effect has same variance $\frac{2\sigma^2}{r}$. Sums of squares of any factorial effect is $\frac{T^2}{8r}$ where T is factorial effect total as given by the above table.

Yates' Method of Computing Factorial Effects:

Yates has developed a systematic tabular method for computing factorial effects. The steps in the computations are as follows:

1. Arrange the treatment combination in standard order. That is, for one factor we simply write (1) , a. For two factors add b, ab derived by multiplying the first two by the additional letter b. For three factors add c, ac, bc, abc, derived by multiplying the first four by the additional letter c and so on.
2. Place the corresponding treatment totals in the next column.
3. Derive the top half of the column (1), by adding the response in pairs
4. Obtain the lower half of the next (column 1) by taking the differences of the first member of each pair from the second in each case.
5. Repeat the process k times until we reach column k where k is the number of factors involved in the experiment. Column k gives factorial effect totals.
6. Obtain the factorial effect by dividing the factorial affect by $r\,x\,2^{k-1}$ where r is the number of replicates.
7. Sum of squares due to factorial effects is obtained by dividing the squares of factorial effects total by $2^k r$

Example: Consider the following 2^3 factorial experiment, designed to determine the effects of certain variables on the reliability of a rotary stepping switch. The factors studied were:

Code	Factor	Low Level	High Level
A	Lubrication	dry	lubricated
B	Spark suppression	no	yes
C	Current	0	0.5 amp

Each switch was operated continuously until a malfunction occured, and the number of hours of operations was recorded.

The whole experiment performed twice, with following results:

Hours of Operation

Experimental	Rep.1	Rep.2	Total
(1)	828	797	1625
a	997	948	1945
b	994	949	1943
ab	1069	1094	2163
c	593	813	1406
ac	773	1026	1799
bc	748	970	1718
abc	1202	1182	2384
Total	7204	7779	14983

Treatment Combination	Treatment Total	(1)	(2)	(3)	Effect	S.S.	F - Ratio
(1)	1625	3570	7676	14983	936.44		
a	1945	4106	7307	1599	199.88	159800.06	15.21
b	1943	3205	540	1433	179.12	128343.06	12.21
ab	2163	4102	1059	173	21.62	1870.56	0.18
c	1406	320	536	- 369	-46.12	8510.06	0.81
ac	1799	220	897	519	64.88	16835.06	1.60
bc	1718	393	- 100	361	45.12	8145.06	0.75
abc	2384	666	273	373	46.62	8695.56	0.83

Sum of Squares due to total

$$= 14446895 - \frac{(14983)^2}{16} \qquad = 416251.9375$$

Sum of Squares due to treatment

$$= \frac{28725685}{2} - 14362842.5 = 332199.44$$

Sum of Squares due to error = 84052.50

Error Variance $= \dfrac{84052}{8} = 10506.563$

From Table A, $F_{0.05,1,\ 8}$ value at 5% level of significance is given by 5.32. Hence main effects of factor A and B have significant effect and remaining effects are insignificant.

CHAPTER – 11

ENVIRONMENTAL SAMPLING

This chapter discusses means of obtaining data for environmental studies. Either the data will come from a planned experiment in the lab or from sampling done in the field. This chapter discusses several methodologies for obtaining data in a scientifically valid way via sampling.

One of the key points to understand is that a valid sampling plan is needed in order to obtain useful data. If the scientist simply goes out into the field and picks sites to sample with no plan ahead of time, then biases and other problems can lead to poor or worthless data.

Example: Estimate the number of trees in a forest with a particular disease. How can we do this? One idea is to divide the forest into plots of size 1 acre say and then obtain a random sample of these acres. Count the number of diseased trees in each sampled acre. From this sample, we can use statistical principals to estimate the number of trees in the forest with the disease.

Some of the most well-known sampling designs used in practice and discussed here are as follows:

- Simple Random Sampling
- Stratified Random Sampling
- Systematic Sampling
- Double Sampling
- Multistage Sampling

Introduction

First, we introduce some terminology and basic ideas.

Census: This occurs when one samples the entire population of interest.
The United States government tries to do this every 10 years. However, in practical problems, a true census is almost never possible.

In most practical problems, instead of obtaining a census, a sample is obtained by observing the population of interest, hopefully without disturbing the population. The sample will generally be a very tiny fraction of the whole population.

One must of course determine the population of interest - this is not always an easy problem. Also, the variable(s) of interest need to be decided upon.

Element : an object on which a measurement is taken.

Sampling Units : non-overlapping (usually) collections of elements from the population.

In some situations, it is easy to determine the sampling units (households, hospitals, etc.) and in others there may not be well-defined sampling units (acre plots in a forest for example).

Example. Suppose we want to determine the concentration of a chemical in the soil at a site of interest. One way to do this is to subdivide the region into a grid. The sampling units then consist of the points making up the grid. The obvious question then becomes - how to determine grid size. One can think of the actual chemical concentration in the soil at the site varying over continuous spatial coordinates. Any grid that is used will provide a discrete approximation to the true soil contamination. Therefore, the finer the grid, the better the approximation to the truth.

Frame: A list of the sampling units.

Sample: A collection of sampling units from the frame.

Notation:

N Number of Units in the Population
n Sample size (number of units sampled)
y Variable of interest.

Two Types of Errors.

- Sampling Errors - these result from the fact that we generally do not sample the entire population. For example, the sample mean will not equal the population mean. This statistical error is fine and expected. Statistical theory can be used to ascertain the degree of this error by way of standard error estimates.

- Non-Sampling Errors - this is a catchall phrase that corresponds to all errors other than sampling errors such as non-response and clerical errors. Sampling errors cannot be avoided (unless a census is taken). However, every effort should be made to avoid non-sampling errors by properly training those who do the sampling and carefully entering the data into a database etc.

Simple Random Sampling (SRS)
One of the simplest sampling designs available is the simple random sample. Simple Random Sample : is the design where each subset of n units selected from the population of size N has the same chance (i.e. probability) of being selected.

Note: It is possible to have a sampling plan where each of the possible samples considered have the same probability of selection *but* the sampling plan is not a SRS.

Example: Suppose the frame for the population consists of sampling units labeled A, B, C, and D. Thus, $N = 4$ and we wish to obtain a sample of size $n = 2$. Then there are 6 possible random samples of size 2:

AB, AC, AD, BC, BD, CD

A simple random sample then requires that each of these 6 possible samples have an equal chance of being selected. In other words, the probability of obtaining anyone of these 6 samples is 1⁄6.

Now, if we only considered two possible samples: AB or CD, each with probability 1⁄2, then each sampling unit has a probability of 1/2 of being selected. But this is not a simple random sample.

Therefore, a simple random sample guarantees that each sampling unit has the same chance of being selected. On the other hand, a sampling plan where each unit has the same chance of being selected is not necessarily a simple random sample.

Question: How do we obtain a simple random sample? The answer is easy – simply label all the sampling units in the population as 1, 2,…,*N* and then pick at random from this list a set of *n* numbers. This sampling is generally done *without replacement*. This is akin to putting the numbers 1 through *N* on a slip of paper, putting them in a hat and then random picking *n* slips of paper from the hat. Of course, actually writing numbers on a slip of paper and picking from a hat is quite tedious, especially if *N* is large. Instead, what is done in practice is to have a statistical or mathematical software package generate a random sample automatically. Many books make use of a table of random digits but these tables are rather archaic and it is suggested to simply use a computer for the task of choosing random samples.

Estimating the Population Mean

Let

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \text{and} \quad \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2$$

denote the population mean and variance respectively. These population parameters are estimated by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2,$$

the sample mean and variance respectively. Using combinatorial counting techniques, it can be shown that the sample mean $\bar{y}$ is unbiased for μ. That is, the average value of $\bar{y}$ over all possible samples of size *n* is exactly equal to μ. Additionally, the sample variance S^2 is unbiased for σ^2*:*

Furthermore, using counting techniques, it also follows that

$$\mathrm{var}(\bar{y}) = \{\sigma^2/n\}(1 - n/N).$$

The factor *(1-n/N*) is called the finite population correction factor which is approximately equal to 1 when *n* is a tiny fraction of *N*. The square-root of the variance of $\bar{y}$ is the Standard error of the sample mean. This is usually estimated by Estimated Standard Error of the mean:

$$= \frac{s}{\sqrt{n}}\sqrt{1 - n/N}.$$

Example: Consider two populations of sizes *N*1 = 1,000,000 and *N*2 = 1000. Suppose the variance of a variable *y* is the same for both populations. What will give a more accurate estimate of the mean of the population: a SRS of size 1000 from the first population or a SRS of size 30 from the second population? In the first case, 1000 out of a million is 1/1000th of the population. In the second case, 30/1000 is 3% of the population. Surprisingly, the sample from the larger population is more accurate.

Confidence Intervals. A $(1-\alpha)100\%$ confidence interval for the population mean can be formed using the following formula:

$$\bar{y} \pm t_{\alpha/2,n-1}\widehat{SE}(\bar{y}) = \bar{y} \pm t_{\alpha/2,n-1}(s/\sqrt{n})\sqrt{1 - n/N},$$

where $t_{(\alpha/2,n-1)}$ is the $\alpha/2$ critical value of the *t*-distribution on *n*-1 degrees of freedom. This confidence interval is justified by applying a finite population version of the central limit theorem to the sample mean obtained from random sampling.

Estimating a Population Total

Often, interest lies in estimating the population total, call it T_y. For instance, in the diseased tree example, one may be interested in knowing how many trees have the disease. If the sampling unit is a square acre and the forest has *N* = 1000 acres, then $T_y = N_\mu = 1000\mu$ Since μ is estimated by $\bar{y}$, we can estimate the population total by

$$t_y = N\bar{y} \qquad (1)$$

and the variance of this estimator is

$$\mathrm{var}(t_y) = \mathrm{var}(N\bar{y}) = N^2\mathrm{var}(\bar{y}) = N^2(1 - n/N)\sigma^2/n.$$

Confidence Interval for Population Total. A $(1-\alpha)100\%$ confidence interval for the population total T_y is given by

$$t_y \pm t_{\alpha/2,n-1}(s/\sqrt{n})\sqrt{N(N-n)}.$$

Sample Size Requirements.

When using a confidence interval to estimate μ or T_y, the total, we may require that our estimate lies within *d* units from the true population parameter. How large a sample size is required so that the half-width of the confidence interval is *d*? The following two formulas give the (approximate) sample size required for the population mean and total:

For the mean $\mu\colon\ n \geq \dfrac{N\sigma^2 z^2_{\alpha/2}}{\sigma^2 z^2_{\alpha/2} + Nd^2},$

and

For the total $T_y\colon\ n \geq \dfrac{N^2\sigma^2 z^2_{\alpha/2}}{N\sigma^2 z^2_{\alpha/2} + d^2},$

where $z_{\alpha/2}$ is the standard normal critical value (for instance, if α = 0.05, the $z_{0.025}$ = 1.96). These two formulas are easily derived algebraically solving for *n* in the confidence interval formulas.

Note that these formulas require that we plug a value in for σ^2 which is unknown in practice. To overcome this problem, one can use an estimate of σ^2 from a previous study or a pilot study. Alternatively, one can use a reasonable range of values for the variable of interest to get an estimate of $\sigma^2 : \sigma \approx$ Range/6.

Example. Suppose a study is done to estimate the number of ash trees in a state forest consisting of N = 3000 acres. A sample of n = 100 one-acre plots are selected at random and the number of ash trees per selected acre are counted. Suppose the average number of trees per acre was found to be $\bar{y}$ = 5.6 with standard deviation s = 3.2. Find a 95% confidence interval for the total number of ash trees in the state forest.

The estimated total l is $t_y = N\,\bar{y} = 3000(5.6)$= 16800 ash trees in the forest. The 95% confidence interval is

$$16800 \pm 1.96(3.2/\sqrt{100})\sqrt{3000(3000-100)} = 16800 \pm 1849.97.$$

A Note of Caution. The confidence interval formulas given above for the mean and total will be approximately valid if the sampling distribution of the sample mean and total are approximately normal. However, the approximate normality may not hold if the sample size is too small and/or if the distribution of the variable is strongly skewed. To illustrate the problem, consider the following illustration. Suppose a survey is to be conducted to estimate the total number of students in Ohio public schools suffering from asthma. Let us take each county as a sampling unit. Then N = 88 for the eighty eight counties in Ohio.

For the sake of illustration, suppose we know the number of students in each county suffering from asthma and that the data is given in the following table:

1 Adams 359	15 Columb 1221	29 Greene 1550	43 Lake 2499
2 Allen 1296	16 Coshocton 415	30 Guerns 464	44 Lawren 822
3 Ashlan 520	17 Crawford 522	31 Hamilton 8250	45 Lickin 1979
4 Ashtab 1274	18 Cuyahoga 14570	32 Hancock 888	46 Logan 558
5 Athens 580	19 Darke 637	33 Hardin 448	47 Lorain 3618
6 Auglaize 558	20 Defian 447	34 Harris 209	48 Lucas 4632
7 Belmont 638	21 Delaware 1448	35 Henry 346	49 Madison 517
8 Brown 679	22 Erie 1012	36 Highland 601	50 Mahoni 2608
9 Butler 3980	23 Fairfield 1710	37 Hockin 264	51 Marion 824
10 Carrol 249	24 Fayett 373	38 Holmes 380	52 Medina 2250
11 Champaign 549	25 Frankl 13440	39 Huron 867	53 Meigs 264
12 Clark 1748	26 Fulton 658	40 Jackson 383	54 Mercer 602
13 Clermo 2083	27 Gallia 389	41 Jefferson 778	55 Miami 1192
14 Clinton 586	28 Geauga 941	42 Knox 613	56 Monroe 185

57 Montgo 5459	68 Preble 572	79 Tuscararawas 1117
58 Morgan 178	69 Putnam 435	80 Union 572
59 Morrow 413	70 Richla 1473	81 VanWert 289
60 Muskin 1206	71 Ross 893	82 Vinton 179
61 Noble 181	72 Sandus 713	83 Warren 2404
62 Ottawa 436	73 Scioto 849	84 Washington 784
63 Pauldi 267	74 Seneca 601	85 Wayne 1279
64 Perry 440	75 Shelby 684	86 Willia 499
65 Pickaw 699	76 Stark 4576	87 Wood 1363
66 Pike 406	77 Summit 6205	88 Wyando 247
67 Portage 1812	78 Trumbu 2556	

Figure 1 shows the actual distribution of students with asthma for the $N = 88$ counties and we see a very strongly skewed distribution. The reason for the skewness is that most counties are rural with small populations and hence relatively small numbers of children with asthma. Counties encompassing urban areas have very large populations and hence large numbers of students with asthma.

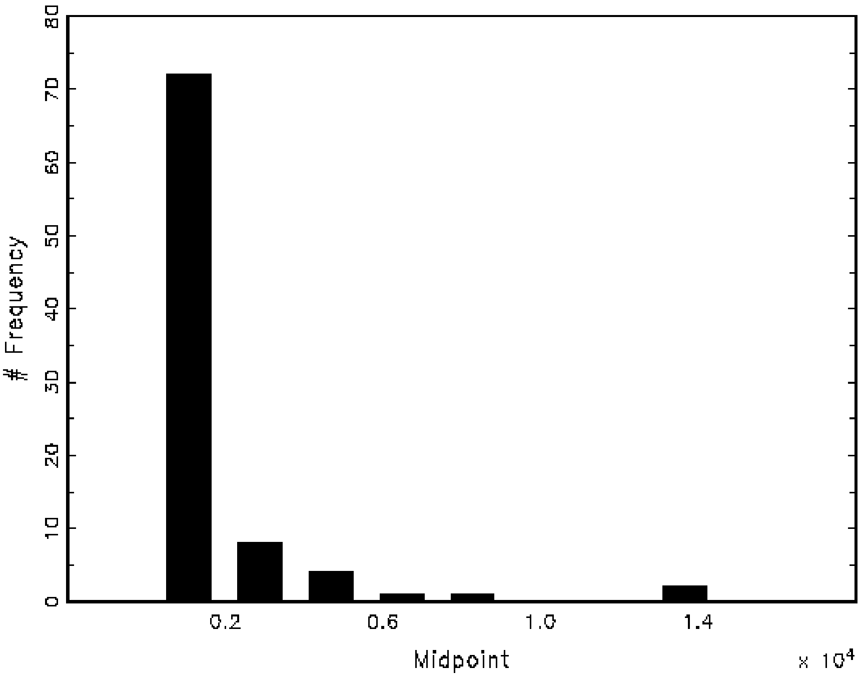


Figure 1: Actual distribution of student totals per county. Note that the distributionis very strongly skewed to the right.

To illustrate the sampling distribution of the estimated total t_y where

$$t_y = N\,\bar{y},$$

10,000 samples of size *n* were obtained and for each sample, the total was estimated. The histograms show the sampling distribution for t_y for sample sizes of *n* = 5, 25, and 50. The long vertical line denotes the true total of *T* = 131, 260.

Clearly the sampling distribution of *ty*, the estimated total, is not nearly normal for *n* = 5. We see a bimodal distribution which results due to the presence of lightly populated and heavily populated counties.

Cochran (1977) gives the following rule of thumb for populations with positive skewness: the normal approximation will be reasonable provided the sample size *n* satisfies

$$n \geq 25G_1^2,$$

where *G*1 is the population skewness,

$$G_1 = \sum_{i=1}^{N} (y_i - \mu)^3/(N\sigma^3).$$

For this particular example, we find

$$25G^2 = 357$$

which is much bigger than the entire number of sampling units (counties)!

In order to get an idea of how well the 95% confidence interval procedure works for this data, we performed the sampling 10,000 times for various sample sizes and

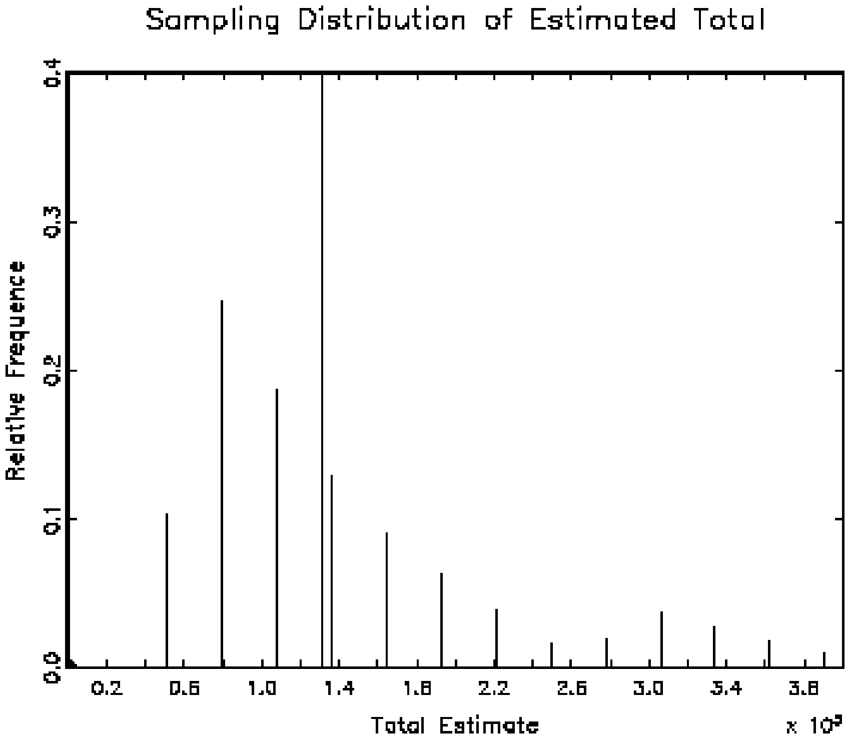


Figure 2: SRS of $n = 5$ for estimating the total number of students.

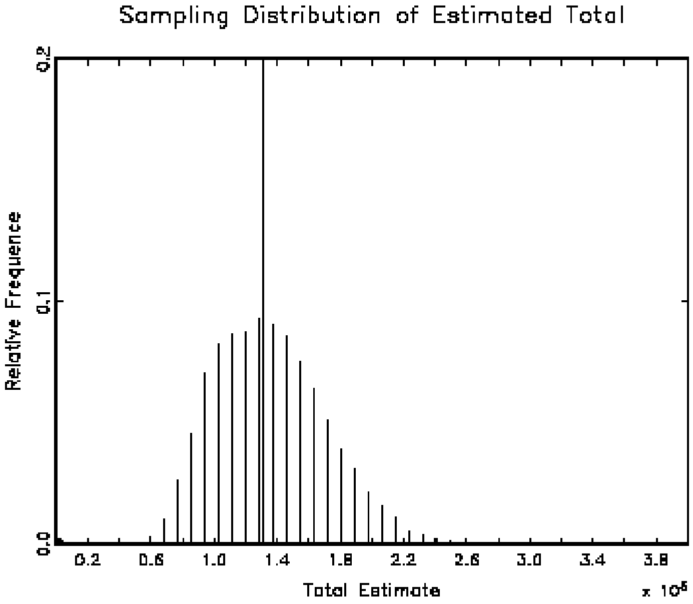


Figure 3: SRS of $n = 25$ for estimating the total number of students.

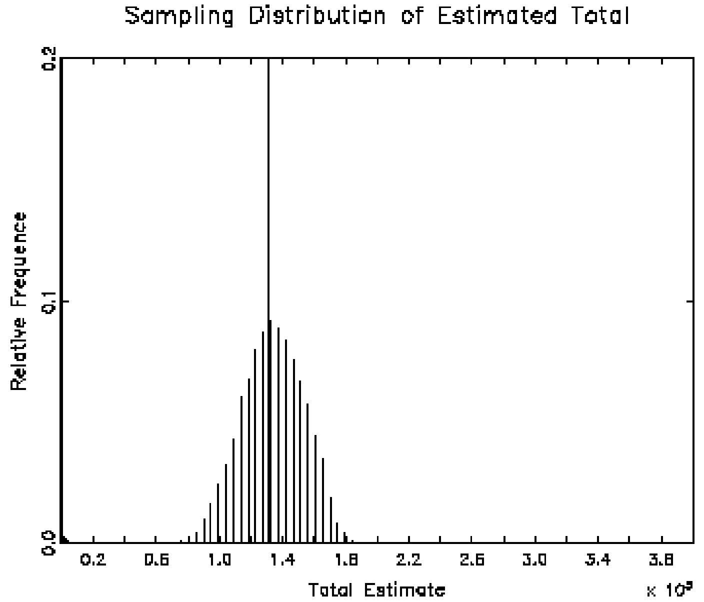


Figure 4: SRS of $n = 50$ for estimating the total number of students.

computed the percentage of intervals that contained the true population total. If the confidence procedure works correctly, the percentage of intervals containing the true population total should be approximately 95%. The results are given in the follow table:

Sample Size	Percentage
5	70%
10	74%
25	83%
50	89%

The simulation indicates that the true confidence level is quite a bit lower than the stated confidence level of 95%. For *n* = 5, only 70% of the 10,000 intervals contained the true population total.

Thus, this example illustrates that for a strongly non-normal population and relatively small sample sizes, the sample mean (and hence estimated total) will not be approximately normal and the confidence interval formulas given above are not valid.

Estimating a Population Proportion

Consider a situation where for each sampling unit we record a zero or a one indicating whether or not the sampling unit is of a particular type or not. A very common instance of this type of sampling is with opinion polls - do you or do you not support candidate X? Suppose you take a survey of plants and you note whether or not each plant has a particular disease. Interest in such a case focuses on the proportion of plants that have the disease. In this section we look at how to estimate the population proportion.

If we obtain a sample of size *n* from a population of size *N*, and each unit in the population either has or does not have a particular attribute of interest (e.g. disease or no disease), then the number of items in the sample that have the attribute is a random variable having a hypergeometric distribution. If *N* is considerably larger than *n*, then the hypergeometric distribution is approximated by the binomial distribution. We omit the details of these two probability distributions.

The data for experiments such as these looks like $y_{1,}, y_2, ..., y_n$, where

$$y_i = \begin{cases} 1 & \text{if the } i\text{th unit has the attribute} \\ 0 & \text{if the } i\text{th unit does not have the attribute.} \end{cases}$$

The population proportion is denoted by *p* and is given by

$$p = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

We can estimate *p* using the sample proportion $\hat{p}$ given by

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Note that in statistics, it is common to denote the estimator of a parameter such as *p* by $\hat{p}(\text{"}p\text{"}-hat)$. This goes for other parameters as well.

Using simple random sampling, one can show that

$$\text{var}(\hat{p}) = (\frac{N-n}{N-1})\frac{p(1-p)}{n}.$$

This variance can be estimated by replacing *p* by $\hat{p}$ in the above formula.

An approximate $(1-\alpha)100\%$ confidence interval for the population proportion is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{(N-n)\hat{p}(1-\hat{p})}{N(n-1)}}.$$

This confidence interval is justified by assuming that the sample proportion behaves like a normal random variable which follows from the central limit theorem. The approximation is better when the true value of p is near 1/2. If p is close to zero or one, the distribution of $\hat{p}$ tends to be skewed quite strongly unless the sample size is very large.

The sample size required to estimate p with confidence level $(1-\alpha)$ with half-width d is given by

$$n \geq \frac{z^2_{\alpha/2}p(1-p)N}{z^2_{\alpha/2}p(1-p)+d^2(N-1)}.$$

Note that this formula requires knowing p which is what we are trying to estimate! There are a couple ways around this problem. (1) Plug in p = 1/2 for p in the formula. This will guarantee a larger than necessary sample size. (2) Use a guess for p, perhaps based on a previous study.

Stratified Random Sampling.

Data is often expensive and time consuming to collect. Statistical ideas can be used to determine efficient sampling plans that will provide the same level of accuracy for estimating parameters with smaller sample sizes. The simple random sample works just fine, but we can often do better in terms of efficiency. There are numerous sampling designs that do a better job than simple random sampling. In this section we look at perhaps the most popular alternative to simple random sampling: Stratified Random Sampling.

The idea is to partition the population into K different *strata*. Often the units within a strata will be more homogeneous. For stratified random sampling, one simply obtains a simple random sample in each strata. Of course, the problem arises as to how many observations to allocate to each strata. Another issue is how to define the strata in the first place.

There are three advantages to stratifying:

1. Parameter estimation can be more precise with stratification.
2. Sometimes stratifying reduces sampling cost, particularly if the strata are based on geographical considerations.
3. We can obtain separate estimates of parameters in each of the strata which may be of interest in of itself.

Examples.

- Estimate the mean PCB level in a particular species of fish. We could stratify the population of fish based on sex and also on the lakes the fish are living.
- Estimate the proportion of farms in Ohio that use a particular pesticide. We could stratify on the basis of the size of the farm (small, medium, large) and/or on geographical location etc.

These two examples illustrate a couple of points about stratification. Sometimes the units fall naturally into different stratum and sometimes they do not.

Notation. Let N_i denote the size of the i^{th} stratum for i = 1, 2, ….. K, where K is the number of strata. Then the overall population size is

$$N = \sum_{i=1}^{K} N_i.$$

If we obtain a random of size n_i from the i^{th} stratum, we can estimate the mean of the i^{th} stratum, $\bar{y}_i$ by simply averaging the data in the i^{th} stratum. The estimated variance of $\bar{y}_i$ is

$$(s_i^2/n_i)(1 - n_i/N_i),$$

where s_i^2 is the sample variance at the i^{th} stratum.
The population mean is given by

$$\mu = \sum_{i=1}^{K} N_i \mu_i / N,$$

which can be estimated by

$$\bar{y}_s = \sum_{i=1}^{K} N_i \bar{y}_i / N,$$

with an estimated variance given by

$$\hat{\sigma}^2_{\bar{y}_s} = \sum_{i=1}^{K} (\frac{N_i}{N})^2 (s_i^2/n_i)(1 - n_i/N_i).$$

The estimated standard error of $\bar{y}_s = \hat{S}E(\bar{y}_s)$ is the square root of this quantity.

The population total $T = N\mu$ can be estimated using

$$t_s = N\bar{y}_s$$

with estimated standard error

$$\widehat{SE}(t_s) = N \cdot \widehat{SE}(\bar{y}_s)$$

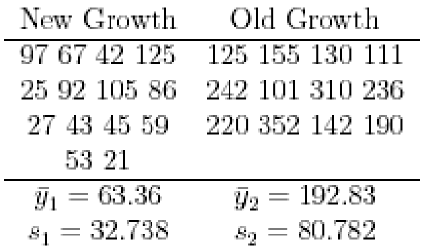
Approximate $(1-\alpha)100\%$ confidence intervals for the population mean and total using stratified random sampling are given by

$$\text{Population Mean: } \bar{y}_s \pm z_{\alpha/2}\widehat{SE}(\bar{y}_s),$$

and

$$\text{Population Total: } t_s \pm z_{\alpha/2}\widehat{SE}(t_s).$$

Example. A survey was done to estimate the average number of invasive honeysuckle plants per acre in a forest. The forest is partitioned into 158 acre plots. N_1= 86 acres of the forest are new growth and N_2= 72 acres are old growth. A sample of n_1= 14 acres of new growth and n_2= 12 acres of old growth forest were obtained yielding the following data:



New Growth	Old Growth
97 67 42 125	125 155 130 111
25 92 105 86	242 101 310 236
27 43 45 59	220 352 142 190
53 21	
$\bar{y}_1 = 63.36$	$\bar{y}_2 = 192.83$
$s_1 = 32.738$	$s_2 = 80.782$

The average number of plants per acre using the two-strata sampling is estimated to be:

$$\bar{y}_s = N_1\bar{y}_1/N + N_2\bar{y}_2/N = 86(63.36)/158 + 72(192.83)/158 = 122.36.$$

The standard error of this estimate is given by

$$\begin{aligned}\widehat{\text{SE}}(\bar{y}_s) &= \sqrt{(N_1/N)^2 s_1^2/n_1(1-n_1/N_1) + (N_1/N)^2 s_2^2/n_2(1-n_2/N_2)} \\ &= \sqrt{(86/158)^2(32.738)^2/14(1-14/86) + (72/158)^2(80.782)^2/12(1-12/72)} \\ &= 10.635.\end{aligned}$$

Thus, with 95% confidence, we estimate that the average number of honeysuckle per acre in the forest is

$$122.36 \pm 2(10.635) = 122.36 \pm 21.270 \text{ plants.}$$

It is interesting to note what would have happened if we had ignored the stratification and simply treated this as a simple random sample of size $n = n_1 + n_2 = 14+12 = 26$. The sample mean of all $n = 26$ acres is $\bar{y} = 123.12$ which is very close to the estimated mean found using the stratification formulas. The standard deviation for the $n = 26$ measurements is $s = 88.100$. The standard error of the mean using the simple random sampling formula is

$$\widehat{\text{SE}}(\bar{y}) = s/n(1-n/N) = 88.100/26(1-26/158) = 15.792.$$

Thus, using a stratified sampling plan led to a much smaller standard error of the mean (10.635 compared to 15.792) than if we had just treated the data as a simple random sample. That is, the stratified design leads to a much more precise estimator of the mean. In addition, the stratification design allows us to obtain separate estimates of honeysuckle abundance in new and old growth parts of the forest.

Post-Stratification

Sometimes the stratum to which a unit belongs is unknown until after the data is collected. For example, values such as age or sex which could be used to form stratum, but these values may not be known until individual units are sampled. The idea of post-stratification is to take a simple random sample first and then stratify the observations into strata after. Once this is done, the data can be treated as if it were a stratified random sample. One difference however is that in a post-stratification setting, the sample sizes at each stratum are not fixed ahead of time but are instead random quantities. This will cause a slight increase in the variability of the estimated mean (or total).

Allocation in Stratified Random Sampling

If a stratified sample of size *n* is to be obtained, the question arises as to how to allocate the sample to the different strata. In deciding the allocation, three factors need to be considered:

1. Total number of elements in each stratum.
2. Variability in each strata, and
3. The cost of obtaining an observation from each stratum.

Intuitively, we would expect to allocate larger sample sizes to larger stratum and/or stratum with high variability. Surveys are often restricted by cost, so the cost may need to be considered. In some situations, the cost of sampling units at different strata could vary for various reasons (distance, terrain, etc.). The optimal allocation of the total sample *n* to the *i*th stratum is to chose n_i proportional to

$$n_i \propto \frac{N_i \sigma_i}{\sqrt{c_i}},$$

where c_i is the cost for sampling a single unit from the *i*th stratum. Therefore, the i stratum will be allocated a larger sample size if its relative size or variance is big or its cost is low. If the costs are the same per stratum, then the optimal allocation is given by

$$n_i \propto N_i \sigma_i,$$

which is known as *Neyman Allocation.*

A simple allocation formula is to use *proportional allocation* where the sample size allocated to each stratum is proportional to the size of the stratum. This will be nearly optimal if the cost and variance at each stratum are nearly equal.

Stratification for Estimating Proportions.

A population proportion can be thought of as a population mean where the variable of interest takes only the values zero or one. Stratification can be used to estimate a proportion, just as it can be used to estimate a mean. The formula for the stratified estimate of a population proportion is given by

$$\hat{p}_s = \frac{1}{N}\sum_{i=1}^{K} N_i \hat{p}_i,$$

and the estimated variance of this estimator is given by

$$\widehat{\text{var}}(\hat{p}_s) = \frac{1}{N^2}\sum_{i=1}^{K} N_i(N_i - n_i)\hat{p}_i(1-\hat{p}_i)/(n_i - 1).$$

Systematic Sampling.

Another sampling design that is often easy to implement is a systematic sample. The idea is to randomly choose a unit from the first *k* elements of the frame and then sample every *k*th unit thereafter. This is called a *one-in-k systematic sample*. A systematic sample is typically spread more evenly over the population of interest. This can be beneficial in some situations. In addition, a systematic sample may yield more precise estimators when the correlation between pairs of observations in the systematic sample is negative. However, if this correlation is positive, then the simple random sample will be more precise. We can use the same formulas for estimating the population mean and total as were used for a simple random sample. These estimators will be approximately unbiased for the population mean and variance. If the order of the units in the population are assumed to be arranged in a random order, then the variance of the sample mean from a systematic sample is the same of the variance from a simple random sample on average. In this case, the variance of $\bar{y}$ from a systematic sample can be estimated using the same formula as for a simple random sample: $(N-n)s^2\ (Nn)$.

An alternative to estimating the variability is to consider the order of the observations in the systematic sample: $y_1, y_2, \ldots., y_n$ and then note that for consecutive neighboring points y_i and y_{i-1}, we have $E\left[(y_i - y_{i-1})^2\right] = 2\sigma^2$ assuming that neighboring points are independent. From this, it follows that

$$s_L^2 = 0.5 \sum_{i=2}^{n} (y_i - y_{i-1})^2/(n-1)$$

can be used to estimate the variance and therefore the standard error of the mean $\bar{y}$ can be estimated using

$$\widehat{\mathrm{SE}}(\bar{y}) = s_L/\sqrt{n}.$$

If the population has some periodic variation, then the systematic sampling approach may lead to poor estimates. Suppose you decide to use a systematic sample to monitor river water and you plan on obtaining samples every seventh day (a 1-in-7 systematic sample). Then this sampling plan reduces to taking a sample of water on the same day of the week for a number of weeks. If a plant upstream discharges waste on a particular day of the week, then the systematic sample may very likely produce a poor estimate of a population mean.

Systematic sampling can be used to estimate proportions as well as means and totals.

Systematic sampling can be used in conjunction with stratified random sampling. The idea is to stratify the population based on some criterion and then obtain a systematic sample within each stratum.

Other Design Strategies

There are many different sampling designs used in practice and the choice will often be dictated by the type of survey that is required. We have discussed simple random sampling, stratified random sampling and systematic sampling. Now we briefly discuss a few other well-known sampling methodologies.

Cluster Sampling.

The situation for cluster sampling is that the population consists of groups of units that are close in some sense (clusters). These groups are known as *primary units*. The idea of cluster sampling is to obtain a simple random sample of primary units and then to sample *every* unit within the cluster.

For example, suppose a survey of schools in the state is to be conducted to study the prevalence of lead paint. One could obtain a simple random sample of schools throughout the state. But this could lead to high costs due to a lot of travel. Instead, one could treat school districts as clusters and obtain a simple random sample of school districts. Once an investigator is in a particular school district, she could sample every school in the district.

A rule of thumb for determining appropriate clusters is that the number of elements in a cluster should be small (e.g. schools per district) relative to the population size and the number of clusters should be large. Note that one of the difficulties in sampling is obtaining a frame. Cluster sampling often makes this task much easier since it if often easy to compile a list of the primary sampling units (e.g. school districts).

Cluster sampling is often less efficient than simple random sampling because units within a cluster often tend to be similar. Thus, if we sample every unit within a cluster, we are in a sense obtaining redundant information. However, if the cost of sampling an entire cluster is not too high, then cluster sampling becomes appealing for the sake of convenience. Note that we can increase the efficiency of cluster sampling by increasing the variability within clusters. That is, when deciding on how to form clusters, say over a spatial region, one could choose clusters that are long and thin as opposed to square or circular so that there will be more variability within each cluster.

Estimation and standard error formulas for cluster sampling can be found in most textbooks on sampling (e.g. Scheaffer, Mendenhall, and Ott 1996).

Notation.

$$\begin{array}{rcl}
N & = & \text{The number of clusters} \\
n & = & \text{Number of clusters selected in a simple random sample} \\
m_i & = & \text{Number of elements in cluster } i \\
M & = & \sum_{i=1}^{N} m_i = \text{Total number of elements in the population} \\
y_i & = & \text{The total of all observations in the } i\text{th cluster}
\end{array}$$

The population mean μ is estimated by

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}.$$

This estimator is a special case of a *ratio estimator* which we shall introduce a bit later. The estimated variance of $\bar{y}$ is given by

$$\widehat{\text{var}}(\bar{y}) = \{(N-n)/(Nn\bar{M}^2)\}s_r^2,$$

where

$$s_r^2 = \sum_{i=1}^{n}(y_i - \bar{y}m_i)^2/(n-1),$$

and

$$\bar{M} = M/N,$$

the average size of a cluster for the population. Note that often in practice *M* and hence $\bar{M}$ are unknown in which case $\bar{M}$ can be estimated by

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i.$$

Estimating the Population Total in Cluster Sampling. An estimate of the population total in cluster sampling can be obtained in much the same way it was obtained in simple random sampling:

$$t_y = M\bar{y}.$$

The estimated variance of t_y is simply $M^2\widehat{\text{var}}(\bar{y})$. What is wrong with using this estimator of the population total? The problem is that it requires that we know *M* which is often unknown.

Alternatively, if we do not know *M*, we could estimate the population total using

$$N\bar{y}_t,$$

where

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

is the average of the cluster totals for the samples clusters. The estimated variance of $N\bar{y}_t$ is

$$\widehat{\text{var}}(N\bar{y}_t) = N(N-n)s_t^2/n,$$

where

$$s_t^2 = \sum_{i=1}^{n}(y_i - \bar{y}_t)^2/(n-1).$$

$N\bar{y}_t$ is an unbiased estimator of the population total, but because it does not use the information on the cluster sizes (e.g. the m_i's), the variance of $N\bar{y}_t$ tends to be bigger than the variance of t_y.

Example. Roberts et al (2004) used a cluster sampling approach to estimate the number of additional deaths in Iraq that resulted due to the Iraq war that started in 2003. From this article, it was widely reported that the number of Iraqi's killed from the war (so far) is 100,000. Their estimate of Iraqi deaths due to the war was 98,000 (not including Falluja which had a very high number of deaths). A 95% confidence interval for this total was given as (8000, 194000). 33 clusters were sampled based on Governorates and 30 households were interviewed in each cluster. The 33 clusters were sampled using a systematic sampling approach. Additional details can be found in the article.

Question: How is a cluster sample different from a stratified sample?

Multistage Sampling

Multistage sampling is similar to cluster sampling. The idea is to determine a set of clusters (i.e. primary units). The first stage is to obtain a simple random sample of these clusters. The second stage is to obtain a simple random sample of units from each of the selected clusters. In cluster sampling, one would sample every unit within the cluster. However, for multistage sampling, only a sample of units within the selected clusters is obtained. In the school lead sampling, if the number of schools in districts is large, then multistage sampling may be preferred over cluster sampling. Multistage sampling differs from stratified sampling in that only a sample of clusters are obtained. In stratified sampling, every cluster would be sampled.

Of course, multistage sampling can be generalized to any number of stages. Suppose you want to survey lakes in the country. The first stage may be to randomly select a sample of states. In the second stage, select a sample of counties from each of the selected states. Finally, sample lakes in each county.

Composite sampling - mixing samples that were obtained near each other to save on the cost of analyzing the sample. For example, consider the problem of testing blood to determine the proportion of people with syphilis. Initially, take one drop from each blood sample, mix these drops, and test the mixture for syphilis. If the test is negative, then syphilis is not present in any of the blood samples. However, if the test is positive, then the individual samples need to be tested. On average, the expected number of tests using composite sampling is much less than the number of samples present.

Ranked set sampling - used to save time and money for analyzing samples. The following example will help illustrate the procedure.

Ranked set sampling example. The goal is to estimate the average amount of spray deposit on apple tree leaves. The sampling units are the leaves of the tree. Accurately computing the deposit density from the spray is time consuming: it requires an image analysis of the leaf to obtain a total pixel grey-scale value which is then divided by the leaf area. Suppose a sample of size $n = 5$ is to be obtained. The basic idea of ranked set sampling is to obtain a random sample of five leaves and *rank* them from highest to lowest spray deposit density. Pick the leaf with the highest spray concentration and accurately measure this concentration. Ranked set sampling requires that the observations can be quickly ranked. In this example, ranking the observations can be done if leaves are sprayed with a fluorescent dye and examining them visually under ultraviolet light. Next, randomly pick five more leaves, rank them and then measure the spray density on the *second* highest leaf. Again, randomly pick five leaves, rank them and perform the measurement on the third highest leaf. Repeat this to get the fourth and fifth measurements. We can think of the data in the following illustration - each row corresponds to five sampled leaves. In the first row, the largest value is denoted by $x_{1(1)}$ and in the second row, the second largest value is denoted by $x_{2(2)}$; and so on.

$$\begin{array}{ccccccc}
x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & : & x_{1(1)} \\
x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & : & x_{2(2)} \\
x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & : & x_{3(3)} \\
x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & : & x_{4(4)} \\
x_{51} & x_{52} & x_{53} & x_{54} & x_{55} & : & x_{5(5)}
\end{array}$$

An unbiased estimator of the mean is given by the ranked set mean estimator:

$$\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_{i(i)}.$$

It can be shown that the ranked set sample mean is more efficient than the simple random sample mean, i.e. the variance of $\overset{=}{x}$ is less than the variance of the sample mean from an ordinary simple random sample. In fact, the increased efficiency of ranked set sampling can be quite substantial. Of course if errors are likely when ranking the observations in each row above, then the efficiency of the ranked set sampling will decrease.

Ratio Estimation.

It is quite common that we will obtain auxiliary information on the units in our sample. In such cases, it makes good sense to use the information in this auxiliary information to improve the estimates of the parameters of interest, particularly if the auxiliary information provides information on the variable of interest.

Suppose *x* is the variable of interest and for each unit, there is another (auxiliary) variable *u* available. If *u* is correlated with *x*, then measurements on *u* provide information on *x*. Typically in practice, measurements on the variable *u* will be easier and/or less expensive to obtain and then we can use this information to get a more precise estimator for the mean or total of *x*. For instance, suppose we want to estimate the mean number of European corn bore egg masses on corn stalks. It is time consuming to inspect each and every leaf of the plant for corn borers. We could do this on a sample of plants. However, it is relatively easy to count the number of leaves on each given stalk of corn. It seems plausible that the number of egg masses on a plant will be correlated with the number of leaves on the plant.

A common use of ratio estimation is in situations where u is an earlier measurement taken on the population and x represents the current measurement. In these situations, we can use information from the previous measurements to help in the estimation of the current mean or total.

Suppose we obtain a sample of pairs (u_1, x_1), ….. , (u_n, x_n). We can compute the means of the two variables $\bar{x}$ and $\bar{u}$ and form their ratio:

$$r = \frac{\bar{x}}{\bar{u}}.$$

Letting μ_x and μ_u denote the population means of x and u respectively, then we would expect that

$$\frac{\mu_x}{\mu_u} \approx \frac{\bar{x}}{\bar{u}},$$

in which case

$$\mu_x \approx r\mu_u.$$

Using this relationship, we can define the ratio estimator of mean μ_x as

$$\bar{x}_{\text{ratio}} = r\mu_u,$$

and if N is the total population size, then the ratio estimator of the total τ is

$$t_x = rN\mu_u.$$

What is the intuition behind the ratio estimator? If the estimated ratio remains fairly constant regardless of the sample obtained, then there will be little variability in the estimated ratio and hence little variability in the estimated mean using the ratio estimator for the mean (or total).

Another way of thinking of the ratio estimator is as follows: suppose one obtains a sample and estimates μ_x using $\bar{x}$ and for this particular sample, $\bar{x}$ underestimates the true mean μ_x. Then the corresponding mean of u will also tend to underestimate μ_u for this sample if x and u are positively correlated. In other words, $\mu_u / \bar{u}$ will be greater than one. The ratio estimator of μ_x is

$$\bar{x}_{\text{ratio}} = r\mu_u = \bar{x}\left(\frac{\mu_u}{\bar{u}}\right).$$

From this relationship, we see that the ratio estimator takes the usual estimator $\bar{x}$ and scales it upwards by a factor of $\mu_u / \bar{u}$ which will help correct the under-estimation of $\bar{x}$.

There is a problem with the ratio estimator: it is biased. In other words, the ratio estimator of μ_x does not come out to μ_x on average. One can show that

$$E[\bar{x}_{\text{ratio}}] = \mu_x - \text{cov}(r, \bar{x}).$$

However, the variability of the ratio estimator often tends to be smaller than the variability of the usual estimator of $\bar{x}$ indicating that it may still be preferable.

An estimate of the variance of the ratio estimator $\bar{x}_{ratio}$ is given by the following formula:

$$\widehat{\text{var}}(\bar{x}_{\text{ratio}}) = (1 - n/N)\sum_{i=1}^{n} (x_i - ru_i)^2/[n(n-1)]. \tag{2}$$

By the central limit theorem applied to the ratio estimator, $\bar{x}_{ratio}$ follows an approximate normal distribution for large sample sizes. In order to guarantee a good approximation, a rule of thumb in practice is to have $n \geq 30$ and the coefficient of variation $\sigma_x / \mu_x < 0.10$. If the coefficient of variation is large, then the variability of ratio estimator tends to be large as well.

An approximate confidence interval for the population mean using the ratio estimator is

$$\bar{x}_{\text{ratio}} \pm z_{\alpha/2}\widehat{se}(\bar{x}_{\text{ratio}}),$$

where $\widehat{se}(\bar{x}_{\text{ratio}})$ is the square-root of the estimated variance of the ratio estimator in (2).

An approximate confidence interval for the population total using the ratio estimator is given by

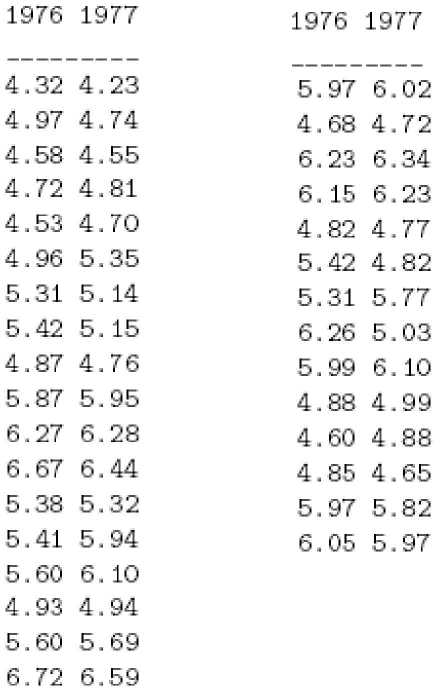
$$t_x \pm z_{\alpha/2}\widehat{se}(t_x),$$

where

$$\widehat{se}(t_x) = N\widehat{se}(\bar{x}_{\text{ratio}}).$$

When estimating the mean or total of a population when an auxiliary variable is available, one needs to decide between using the usual estimator $\bar{x}$ or the ratio estimator. If the correlation between *x* and *u* is substantial, then it seems that using the ratio estimator should be preferred. A rough rule of thumb in this regard is to use the ratio estimator when the correlation between *x* and *u* exceeds 0.5. There is a theoretical justification for this given in Cochran (1977, page 157) based on assuming the coefficient of variation for *x* and *u* are approximately equal.

Example. A study of acid rain was undertaken by examining samples of water in 32 lakes in 1977. In 1976, the pH was measured in the population of all *N* = 68 lakes which gave a mean value of μ_u = 5.715 in 1976. Figure 5 shows a scatterplot of the pH values from the sample of *n* = 32 lakes in 1977. The goal is to estimate the mean pH level μ_x for all *N* = 68 lakes for 1977. The data for the *n* = 32 lakes are given in the following table:



1976	1977
4.32	4.23
4.97	4.74
4.58	4.55
4.72	4.81
4.53	4.70
4.96	5.35
5.31	5.14
5.42	5.15
4.87	4.76
5.87	5.95
6.27	6.28
6.67	6.44
5.38	5.32
5.41	5.94
5.60	6.10
4.93	4.94
5.60	5.69
6.72	6.59
5.97	6.02
4.68	4.72
6.23	6.34
6.15	6.23
4.82	4.77
5.42	4.82
5.31	5.77
6.26	5.03
5.99	6.10
4.88	4.99
4.60	4.88
4.85	4.65
5.97	5.82
6.05	5.97

The sample means for the n = 32 lakes are

$$\bar{x} = 5.3997 \text{ and } \bar{u} = 5.4159,$$

which gives an estimated ratio of

$$r = \frac{\bar{x}}{\bar{u}} = \frac{5.3997}{5.4159} = 0.9970.$$

The ratio estimator of μ_x, the average pH in the 68 lakes is

$$\bar{x}_{\text{ratio}} = r\mu_u = (0.9970)(5.715) = 5.6979,$$

which is higher than the simple estimate of $\bar{x}$ = 5.3997. Therefore, the ratio estimate takes the usual estimate of 5.3997 and scales it up by a factor of $\mu_U / \bar{u}$=5.715=5.4159 = 1.0552. The sample correlation between pH in 1976 and 1977 for the 32 lakes is 0.883 which indicates that the ratio estimator will be more efficient than the usual simple random sample estimator of the mean. The estimated coefficient of variation for 1976 and 1977 are respectively 0.1234 and 0.1244. Although the coefficient of variation for 1977 exceeds our rule of thumb value of 0.10, it does not exceed it by much.

The estimated variance for the ratio estimator can be computed as

$$\widehat{\text{var}}(\bar{x}_{\text{ratio}}) = (1 - n/N)\sum_{i=1}^{32}(x_i - 0.9970u_i)^2/[32(31)] = (1-32/68)(3.2473)/[32(31)] = 0.0017.$$

The standard error of $\bar{x}_{ratio}$ is obtained by taking the square root of this quantity which gives $\hat{se}\left(\bar{x}_{ratio}\right) = \sqrt{0.0017} = 0.0412$. A 95% confidence interval for μ_x is

$$5.6979 \pm 1.96(0.0412) = 5.6979 \pm 0.0808.$$

Note that if we had just used the sample mean to estimate the population mean (obtaining $\bar{x}$ = 5.3997), the associated standard error would be

$$\widehat{se}(\bar{x}) = (s/\sqrt{n})\sqrt{1 - n/N} = (0.6716/\sqrt{32})\sqrt{1 - 32/68} = 0.0864$$

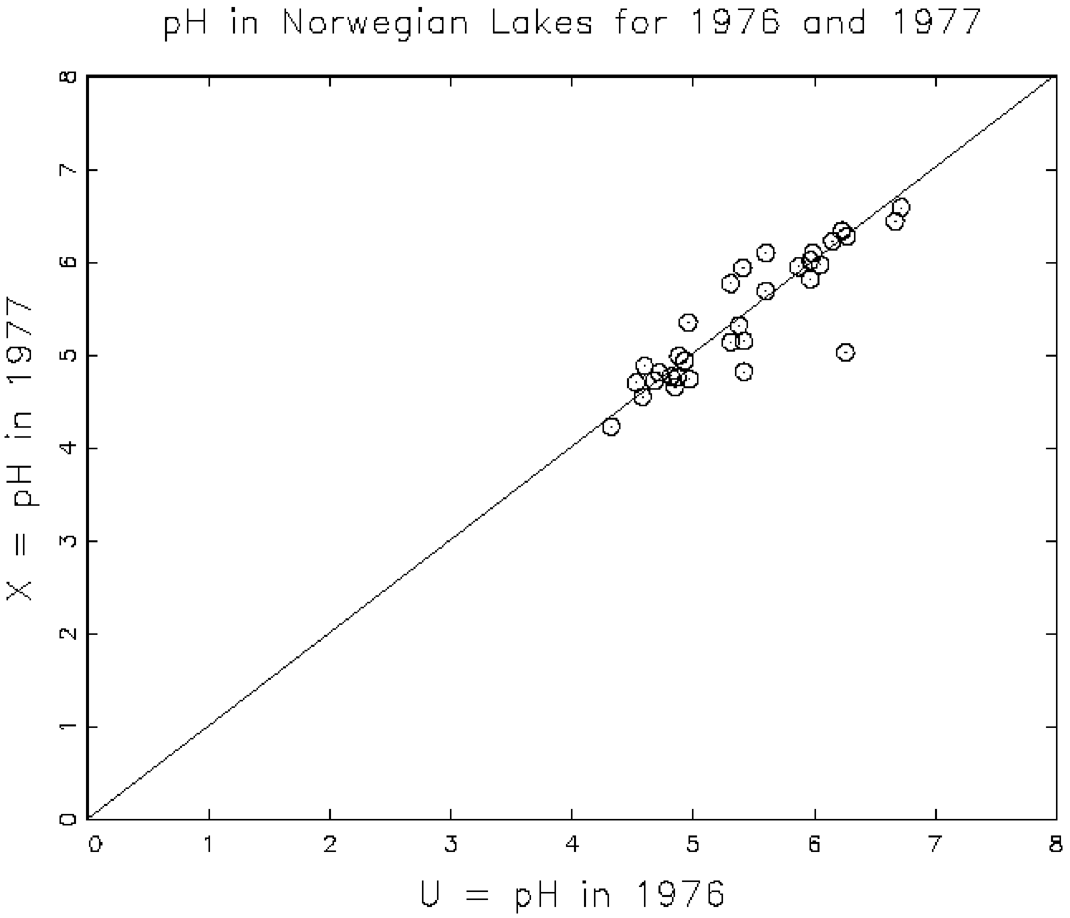


Figure 5: Scatterplot of pH in 1977 versus 1976 at 32 Norwegian lakes. These 32 lakes are a subset of all 68 lakes.

which is more than twice the standard error of the ratio estimator. This indicates that the ratio estimator is a more efficient estimator of the population mean.

There exist sample size formulas for estimating means and totals using a ratio estimator which can be found in most textbooks on sampling. Note that if ratio estimation is more efficient than the usual simple random sample estimate, then smaller sample sizes will be required for the same level of precision.

Regression Estimation

Note that the line in Figure 5 appears to go through the origin which stands to reason if the relationship $x = ru$ is approximately valid. There exist other examples where an auxiliary variable is available and the relationship between *x* and *u* is linear, but the line does not necessarily go through the origin. In these situations, it makes sense to utilize the information in the auxiliary variable using a simple linear regression relation between *x* and *u*:

$$x = \beta_0 + \beta_1 u + \epsilon,$$

where β_0 and β_1 are the intercept and slope of the line and ε is a random error to account for the fact that the sample points will not all lie exactly on a line.

Let $\hat{\beta}_1$ denote the usual least-squares estimator of the slope. Then the estimated regression line is given by

$$\hat{x} = \bar{x} + \hat{\beta}_1(u - \bar{u}).$$

Additionally, the least-squares regression line always passes through the mean $(\bar{u}, \bar{x})$. This suggest the following least-square regression estimator of the mean of *x*, denoted $\hat{\mu} L$:

$$\hat{\mu}_L = \bar{x} + \hat{\beta}_1(\mu_u - \bar{u}).$$

Thus, the regression estimator takes the usual estimator $\bar{x}$ of the mean and adjusts it by adding $\hat{\beta}_1(\mu_u - \bar{u})$.

- Typically the ratio estimator is preferred over the regression estimator for smaller sample sizes.
- Ratio and regression estimation can be used in conjunction with other types of sampling such as stratified sampling.

Double Sampling

Double sampling (also known as 2-phase sampling) is similar to ratio estimation in that it uses information from an auxiliary variable. For ratio estimation, it was assumed that the population mean μ_u was known for the auxiliary variable, but this may not always be the case.

The basic idea of double sampling is to first take a large preliminary sample and measure the auxiliary variable. It is assumed that the auxiliary variable will be easy and/or inexpensive to measure and that it will be correlated with the variable of interest. Then another sample (often a sub-sample of the first sample) is obtained where the variable *x* of interest is measured.

Some examples of easy-to-measure auxiliary variables are

- Examine aerial photographs of sampling units to get rough counts of trees, animals etc.
- Published data from past surveys.
- A quick computer search of files using a keyword for example.

In order to perform a double sampling, one first obtains a preliminary sample of size *n'* say and measures the variable *u*. From this preliminary sample, we can get an estimate of μ_u using

$$\hat{\mu}'_u = \sum_{i=1}^{n'} u'_i/n'.$$

Then one obtains the usual sample of size *n*, perhaps as a sub-sample of the preliminary sampled units. From this sample, we can compute the ratio as in a ratio sample:

$$r = \frac{\bar{x}}{\bar{u}}.$$

Then, the population total for *x* can be estimated using

$$t_x = r\hat{\mu}'_u.$$

The variance for the estimated total using double sampling is more complicated than the variance of the ratio estimator because we have an extra source of variability with double sampling - namely the variability associated with the preliminary sample. The estimated variance of the double sampling total estimator is given by

$$\widehat{var}(t_x) = N(N-n')s^2/n' + \frac{N^2(n'-n)}{nn'}s_r^2,$$

where

$$s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - ru_i)^2.$$

Notice that if *n'* = *N*, that is if the preliminary sample is of the entire population (i.e. a census), then the first term in this variance formula becomes zero and we end up with the same formula as the ratio estimator variance.

Unequal Probability Sampling

The sampling procedures discussed up to this point involve simple random sampling of sampling units in which case each unit has the same chance of being selected for the sample. Even with sampling designs more complicated than simple random sampling, such as stratified random sampling, a simple random sample was obtained in each stratum. In many situations, a simple random sample is either not possible or not preferable.

In *line-intercept* sampling for example, a line is more likely to intercept larger units than smaller units. If we divide an area into plots of sampling units, the plots may not all have the same size. In these cases, the probability of the unit to be selected into the sample will depend on the size of the unit. This is sometimes known as *probability proportional to size* estimation.

Let p_i denote the probability that the *i*th unit will be selected.

Hansen-Hurwitz Estimator: Suppose sampling is done with replacement. Recall that when using simple random sampling, the population total is estimated by $t_y = N\bar{y}$. We can rewrite this as

$$t_y = \frac{1}{n}\sum_{i=1}^{n} y_i/(1/N).$$

If we are sampling with replacement when each unit has the same chance of being selected, then the probability that a unit is selected at any given draw is 1/*N*. For the Hansen-Hurwitz estimator, we simply replace the 1/*N* by p_i for the *i*th unit:

$$t_{HH} = \frac{1}{n}\sum_{i=1}^{n} y_i/p_i \quad \text{(Hansen-Hurwitz estimation of total)}$$

Horvitz-Thompson Estimator: Sampling with replacement is not done often in

practice as in the case of the Hansen-Hurwitz estimator. With the Horvitz-Thompson estimator, the sampling can be done either with or without replacement. We shall consider the case when the sampling is done without replacement. Let π_i denote the probability the i^{th} sampling unit is selected in the sample. (Note that if all units have the same chance of being selected and we sample without replacement, then $\pi_i = n/N$: Can you explain why?)
The estimator of the population

total is given by

$$t_{HT} = \sum_{i=1}^{n} y_i/\pi_i \quad \text{(Horvitz-Thompson Estimator).}$$

The population mean can be estimated using

$$\hat{\mu}_{HT} = t_{HT}/N.$$

assuming the *n* units selected are all distinct (this will not necessarily be the case when sampling with replacement). The variance formula for the Horvitz-Thompson estimator is quite complicated and involves probabilities of the form π_{ij} which denotes the probability that units *i* and *j* are both selected. Recent research into simpler variance formulas that do not require knowing the π_{ij} has been published, see for example Berger (2004). If sampling is done proportional to size and size of units vary, then the π_{ij} will vary in value as well.

Detectability

In some sampling cases, the elements may be difficult to detect within the sampling units. This may be the case in certain wildlife populations (e.g. fish, birds, etc.). If one is obtaining a simple random sample from a population of *N* units, then whether or not an animal in the unit is detected may not be certain, but instead a probability is associated with the chance the animal is detected. A non-animal example could occur when soil samples are assessed for a particular contaminant, some of the material may be missed due to sparsity of the contaminant.

Definition. The probability that an object in a selected unit is observed is termed its *detectability*.

For the sake of discussion, we shall refer to the objects as “animals." The following is some notation:

$$\begin{aligned} y &= \#\text{ of animals observed} \\ \tau &= \text{total } \#\text{ of animals} \\ p &= \text{probability an animal is observed.} \end{aligned}$$

If we assume independence between observations and a constant detectability probability *p* throughout a region, then

$$Y \sim \text{Binomial}(\tau, p),$$

that is, *Y* , the number of animals observed follows a binomial distribution on τ trials and success probability *p*. Therefore, the expected value of *Y* is

$$E[Y] = \tau p,$$

which indicates that we can estimate the total number of animals by solving for τ and using an estimate for the mean:

$$\hat{\tau} = y/p.$$

The variance of the binomial random variable *Y* is $\tau\, p$(1- *p*) and thus

$$\text{var}(\hat{\tau}) = \frac{\tau p(1-p)}{p^2} = \frac{\tau(1-p)}{p},$$

which can be estimated by substituting $\hat{\tau}$ for τ to get

$$\widehat{\text{var}}(\hat{\tau}) = \frac{y(1-p)}{p^2}.$$

Notice that if the probability *p* of detection is small, then this variance becomes large. If the area of the region of interest is *A*, then we can define the animal *density* as

$$D = \tau/A,$$

the number of animals per unit area. An estimate for the density then is

$$\hat{D} = \frac{y}{pA},$$

which has an estimated variance of

$$\widehat{\text{var}}(\hat{D}) = \frac{y}{A^2}(\frac{1-p}{p^2}).$$

These formulas require that we know the value of p but this is typically not the case in practice.

The question arises as to how to estimate p. Methods such as double sampling, capture-recapture or line transects can be used to estimate p. One way to estimate p is to select n sampling units and let x_i denote the number of animals detected in the i^{th} unit using the standard sampling technique. Then do an intensive search of each of these sampling units and let y_i denote the actual number of animals at the i^{th} unit. Then an estimate of p is obtained by computing

$$\hat{p} = \frac{\bar{x}}{\bar{y}}.$$

The variance of this estimator can be estimated using ideas from ratio estimation.

If p has to be estimated, then the previous estimate of the population total τ can now be given as

$$\hat{\tau} = \frac{y}{\hat{p}}.$$

Since we now have the random $\hat{p}$ in the denominator instead of a fixed p, the variance of the estimated total increases by an extra term. An approximate formula for the variance of this estimated total can be derived using a Taylor series approximation to the ratio

$$\text{var}(\hat{\tau}) = \tau(\frac{1-p}{p}) + \frac{\tau^2}{p^2}\text{var}(\hat{p}).$$

In the formulas above, we have let y denote the number of animals observed from our sample. The value of y obtained depends on the sampling design used. For instance, if a simple random sample was used, then the estimate of the total was found to be $N\bar{y}$ assuming all animals could be detected. If *p* is the probability of detection, then the estimate of the total becomes

$$\hat{\tau} = N\bar{y}/p.$$

We can replace *p* by $\hat{p}$ in this formula when *p* needs to be estimated. The variance formula approximations become quite complicated in this case (e.g. see Thompson 1992).

Line Transect Method

In this section we give a brief introduction to some of the basic ideas of line transect sampling. The basic idea of the line transect method of sampling is for the observer to move along a selected line in the area of interest and note the location of animals (or plants) along the line and the distance from the line. The goal of the line transect method is to estimate the animal density *D* = (# of animal/unit area)*:* Then the total number of animals can be found by computing

$$\tau = DA,$$

where *A* is the area of the region of interest. The observer will obtain a random sample of line transects. Let y_i denote the number of animals detected along the *i*th transect.

The Narrow Strip Method: Choose a strip of length *L* and let w_0 denote the distance to the left and right of the line where the observer will observe the animals - w_0 is called the half-width. A simple estimate of the density along the strip is

$$\frac{\text{Number of animals in the strip}}{\text{Area of the strip}} = \frac{y}{2w_0L}.$$

The narrow strip method assumes that animals anywhere in the strip are just as likely to be observed as anywhere else in the strip. However, a more realistic scenario is that the detectability decreases with the distance from the transect.

Instead of using the narrow strip method then, the data can be used to estimate a detectability function where the probability of detection drops off with the distance from the line transect. A couple popular parametric choices for the detectability functions are given by the exponential function and the half-normal function:

$$\begin{aligned} g(x) &= e^{-x/w} \text{ Exponential Function} \\ g(x) &= e^{-\pi x^2/(4w^2)} \text{ Half-Normal Function,} \end{aligned}$$

where w is a parameter typically estimated using maximum likelihood and x is the distance from the line. Instead of specifying a parametric form for the detection function (e.g. exponential and half-normal), nonparametric detection functions can be estimated using *kernel* methods.

For line transect sampling, more than one transect is obtained. One can obtain a simple random sample of transects. This is usually accomplished by drawing a line along one edge of the region and then selecting n points at random along this line. Then the transects are perpendicular lines extending from this baseline into the region at the n points. Note that biases can occur for transects that occur near the boundary of the region (e.g. there may be few animals along the boundary - there are ways of dealing with this that we will not go into here). If the region has an irregular shape, then the lengths L_i of the n transects will have varying lengths and therefore the lengths are random variables.

Instead of taking a simple random sample of transects, one could instead obtain a systematic sample of transects. This will help guarantee a more even coverage of the region.

Also, transect lines can also be selected with probability proportional to the length of the transect. The probability proportional to length selection can be accomplished by selected n points at random from the entire two-dimensional region and then select transects based on perpendicular lines that go through these selected points from the baseline.

The Data Quality Objectives Process

The collection of data can be time consuming and expensive. Therefore, it is very important to plan matters very carefully before undertaking a survey or experiment. If too small a sample size is used, then there may not be enough information to make the resulting statistical analysis useful. For instance, confidence intervals may be too wide to be of any use or a statistical test may yield insignificant results even if there is a real effect. On the other hand, one does not want to unnecessarily expend too much money and resources obtaining more data than what is necessary in order to make a decision.

The steps of the DPO can be summarized as following:

1. State the problem: describe the problem, review prior work, and understand important factors.

2. Identify the decision: what questions need to be answered?

3. Identify the inputs to the decision: determine what data is needed to answer questions.

4. Define the boundaries of the study: time periods and spatial areas to which the decisions will apply. Determine when and where data is to be gathered.

5. Develop a decision rule: define the parameter(s) of interest, specify action limits,

6. Specify tolerable limits on decision errors: this often involves issues of type I and type II probabilities in hypothesis testing.

7. Optimize the design for obtaining data: consider a variety of designs and attempt to determine which design will be the most resource-efficient.

This process may very well end up being an iterative process. Not only will later steps depend on the earlier steps but the later steps may make it necessary to rethink earlier steps as the process evolves. For instance, one may initially set unrealistic error bounds (type I and/or II) and then come to realize that these constraints would make the project go way over budget.

References

Berger, Y. G. (2004), “A Simple Variance Estimator for Unequal Probability Sampling without Replacement,” *Journal of Applied Statistics*, 31, 305-315.

Cochran, W. G. (1977), *Sampling Techniques*, 3rd edition, Wiley, New York.
Roberts, L., Lafta, R., Garfield, R., Khudhairi, J., Burnham, G., (2004), \Mortality
before and after the 2003 invasion of Iraq: cluster sample survey," *The Lancet*, 364, 1857-1864.

Scheaffer, R., Mendenhall, W. and Ott, R. (1996), *Elementary Survey Sampling*, 5th edition, New York: Duxbury Press.

Thompson, S. K. (1992), *Sampling*, New York: Wiley.

CHAPTER – 12

CONTROL CHARTS

Introduction

What is a Control Chart

Control Chart is a chart on which the values of the quality characteristic being controlled are plotted in sequence. The chart consists of a central line (corresponding to the desired average level) and two statistical limit lines called Upper Control Limit (UCL) and Lower Control Limit (LCL) which indicate the limits of *Natural variation* (not the specified variation) for the sample `statistic' (like average, range, % defective, no. of defective items per sample, no. of defects per item etc.) being plotted.

The control limits are supposed to strike a balance between two kinds of errors, viz., (1) looking for trouble that does not exist and (2) failing to look for trouble that does exit. Neither of these kinds of errors should be unduly large, yet neither should be reduced to such an extent that it unduly increases the other.

Sample constituting *rational subgroup* are taken at regular intervals of production and suitable `statistic' computed from the sample measurements are plotted on the control chart. Suitable technical action is called for whenever the statistic violates the control limits or some *abnormal* pattern is developed in the chart.

Histogram has certain limitations. Even if the original data were collected in sequence of time and the identification of time-sequence was kept for each observation, that information is totally lost when we make frequency distribution and histogram. If there has been a gradual drift or occasional changes in the process level during the period of data collection, histogram does not reveal these aspects which might be vital inputs for necessary corrective action. In such situations, we might sometimes wrongly conclude that the process is under statistical control. Therefore, it is necessary that we examine the behaviour of the process over sequence of time wherever feasible. This is done through RUN CHARTS.

Run Chart

Run chart is a simple chart where the quality characteristic is plotted in sequence of time for consecutive items produced. The chart contains the specification limits and also the mid-specification line where the process average is supposed to be centered.

Advantages of a Run Chart

- Very easy and simple to plot.
- Needs little statistical training for interpreting the chart.
- Provides good feedback on approximate average level and variability for prompt corrective action.

Disadvantages of a Run Chart

- One has to wait for a long time to detect a small change in average or variability level.
- Objective and precise decision criteria as to when to take action and when not to take action are not provided. These disadvantages are taken care of through CONTROL CHARTS.

Rational Subgroup

Product streams can usually be divided into homogeneous groups (or lots) with reference to time or other characteristics, ensuring that products in the group have been made under conditions of statistical control. Rational subgroup is a sample which represents a homogeneous group. Assignable causes, if they exist, cause variation between groups. The objective of the control chart technique is to check whether the variation between groups measured by subgroup difference is in conformity with the variation within subgroup. Appropriate statistic from these rational subgroups are plotted on the control chart. Process standards are evolved after excluding those subgroups where assignable causes are suspected to have operated.

Control limits

The control limits are usually placed at Mean ± 3 standard deviation standard deviation of the statistic. The standard deviation represents the variability within a rational subgroup.

Procedure for installing and operating control charts

1. Decide on the characteristic (measurable or attribute) to be controlled.
2. Define the groups or lots which will provide rational subgroups.
3. Decide subgroup size. It really depends on the amount of shift to be detected quickly in the process level. For measurable characteristic, for shifts of as much as 2σ, sample size 4 or 5 is usually used whereas for small shifts, say 1σ, sample sizes 15-20 are suitable.
4. To develop process standards obtain data for 20-25 subgroups. Through appropriate statistical procedure, `homogenise' the data, evolve process standard and calculate the standard deviation of the sample statistic to be plotted on the chart.
5. Obtain the control limits as Average ± 3 standard deviation of the sample statistic.
6. Draw the central line and the control limits on a graph, continue to obtain rational subgroup measurements and plot the sample statistic on the chart.
7. As soon as a point violates the limits or there is `abnormal pattern' of points, infer that some assignable cause of variation has disturbed the process. Accordingly investigate and take corrective action.

Types of Control Charts

Type of control charts depends on the nature of quality characteristic being controlled. The Charts are broadly classified as Attribute and Variable.

Control charts for attributes

These charts are used when we are interested in controlling percentage or proportion of occurrences of some event. The typical example is when quality data are generated in the form of attribute data like `good' and `bad' and the quality characteristic of concern is the `proportion defective' (p) or number of defectives in a sample of constant size.

In order that we are able to calculate the standard error of the statistic ``number of defectives in a sample of constant size" we must know the long term pattern of variation (i.e. probability distribution) of the concerned statistic (a random variable) under stable process conditions producing proportion of defective.

The long run average (μ), number of defectives and the standard deviation $(\sigma)$of the number of defectives in a sample of size n will be given by

$$\mu = np$$
$$\sigma = \sqrt{np\,(1-p)}$$

If the characteristic chosen is p i.e. the sample proportion defective, then the mean and s.d. are given by

$$\mu = p$$

$$\sigma = \sqrt{p(1-p)/n}$$

Where n is the sample size.

These results help us in calculating the control limits for the relevant characteristics.

np Chart

When the subgroup size (n) for inspection remains constant in each subgroup, we use np chart to examine the state of control with respect to number of defectives in each subgroup

The formula for the control limits are

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

when $n\bar{p}$ is the homogenised average number of defectives per sample of size n.

C-Chart

This chart is used for those characteristics which indicate only the number of occurrences of some rare events like occurrence of defects, breakdown, accidents, absenteeism, etc. during fixed time intervals or length, area and volume space. The concerned random variable follows what is known as POISSON DISTRIBUTION.

In control chart operations the variable denoting no. of occurrence is named as c. The long run average and standard deviation are given by

$$\mu = \bar{c}$$
$$\sigma = \sqrt{\bar{c}}$$

This result helps us in obtaining the control limits for the desired characteristic.

$\bar{X} - R$ Control charts

In order to ensure that the production of defectives (not conforming to specification) is minimised, we have to exercise control over average level as well as variability. For controlling the average level, the appropriate control chart is $\bar{X}$ chart where the subgroup averages $\bar{X}$ are plotted in time sequence. Similarly, for controlling the variability, the chart to be used is the Range or R-chart where the statistic plotted are the sample ranges (R). A sample $\bar{X}$ - R-chart is given in the next page.

The following data relate to CO content in air in a particular locality.

Table : Data on CO content

Sample No.	1	2	3	$\overline{X}$	R
1	2.12	2.13	2.12	2.123	0.01
2	2.13	2.10	2.13	2.120	0.03
3	2.15	2.13	2.13	2.137	0.02
4	2.10	2.12	2.14	2.120	0.04
5	2.11	2.14	2.16	2.137	0.05
6	2.07	2.13	2.15	2.117	0.08
7	2.12	2.12	2.14	2.127	0.02
8	2.10	2.12	2.15	2.123	0.05
9	2.15	2.13	2.11	2.130	0.04
10	2.11	2.12	2.12	2.117	0.01

Assuming that the measurable characteristic follows Normal distribution, the formulae for the control limits in the $\overline{X}$ and R charts are given as follows:

Table : Formulae for Control Limits

Chart for	Central Line	Upper Control	Lower Control Limit
Average $\overline{X}$	$\overline{X}$	$\overline{X} + A_2\ \overline{R}$	$\overline{X} - A_2\ \overline{R}$
Range, R	$\overline{R}$	$D_4\ \overline{R}$	

Where
$\overline{X}$ = Average of the sample average .
$\overline{R}$ = Homogenised average range

For routine operation of control chart, $\overline{X}$ is to be replaced by target value A2, D3, D4 are constants depending on a subgroup size (Refer TABLE-A). For the given data, we have sub group size

n	=	3
A2	=	1.023
D3	=	0
D4	=	2.574

Homogenisation of Ranges

First we shall evaluate the process standards for Range. We shall detect the abnormally high ranges which are likely to have arisen due to some assignable cause of variation. If no assignable cause has disturbed the process, the individual `range' values will all fall within the control limits for range. If any range value violates the limits, it is an indication that it does not belong to the set of remaining range values. In such a case, we eliminate that `range' and reexamine the control aspect for the remaining `ranges'. This procedure known as "Homogenisation of ranges" is continued till we are left with a set of ranges which falls within the latest revised control limits. The average of the ranges remaining at last within the control limits is called the ``standard (homogenised) average range"(R).

One note of caution. If in the process of homogenisation more than 20 % of points are to be discarded, do not use the data for evolving process standards for future control because such a situation indicates that the process is very much disturbed and so it should not be considered for evolving the standards.

For our data

$$\overline{R} = 0.35/10 = 0.035$$

$$UCL_R = 2.574 \text{ X } 0.035 = 0.090$$

$$LCL_R = D_3 \ \overline{R} = 0$$

All ranges are within limits.

Incidently the long run average value of $\overline{R} \ / \ \sigma$ where σ is the population standard deviation stabilises at a constant value denoted by d_2 which again depends on subgroup or sample size. So, an estimate of standard deviation is provided by $\overline{R} \ / \ d_2$

For subgroup size 3, d_2= 1.693
The estimate of standard deviation = $\hat{\sigma}$ = 0.035/1.693 = 0.027

This standard deviation is a measure of variability.

Setting Limit for Averages

1. Since specification is given between 2.10 ± 0.05 for future control the target is kept at 2.10 (Lower Specification Limit + Upper Specification Limit/2) and control limits are calculated as
$\pm\ A_2\ \overline{R}$ (homogenised) from target
where A_2= 1.023 X 0.035 = 0.036
Upper Control Limit = 2.10 + 0.036 = 2.136
Lower Control Limit = 2.10 - 0.036 = 2.064

After installing charts fixed number (here 3) of subgroup observations are to be collected and from each `rational' sub group, $\bar{x}$ and R are to be calculated and plotted on the respective charts. Successive points can be joined by straight lines. So long as the plotted points exhibit natural patterns of variation within the control charts, no action is called for since the process is in control. But as soon as any abnormal pattern of variation is noticed, we must hunt for the trouble-maker and not rest till we catch the culprit.

Natural pattern of variation in control charts

So long as the process conditions are quite stable i.e. the process is governed only by chance causes, the behaviour of the plotted points should satisfy all the following conditions:

1. Most of the points are centered around the central line.
2. A few of the points are spread out and approach the control limits.
3. None of the points (or at most only a very rare and occasional point) exceeds the control limits.

Unnatural pattern of variation in control charts

The various indications and conclusions are presented in the following list:

Table : Unnatural patterns and conclusions from control chart

Sl.No.	Pattern of points	Conclusions
1.	Point violating control limits	Change in level
2.	Run of points on same side of central line but within control limits - 7 Successive - 10 out of 11 - 12 out of 14 - 14 out of 17 - 16 out of 20	Sustained shift in level
3.	Trend of points	Gradual change in level
4.	Points mostly near UCL as well as LCL	Two or more overlapping distribution of characteristic under observation
5.	Appearance of cycles	Some factor influencing the monitoring characteristic periodically
6.	Points too close to central line	In correct rational subgrouping
7.	Correlation between $\bar{X}$ and R Charts	Skewness in underlying distribution

Table-A

Factors for $\overline{X} - R$ Charts

Factors for estimating s from R

No. of Observations in a sample	A_2	D_3	D_4	For the Estimate from $\overline{R}\ (d_2)$
2	1.880	0	3.268	1.128
3	1.023	0	2.574	1.693
4	0.729	0	2.282	2.059
5	0.577	0	2.114	2.326
6	0.483	0	2.004	2.534
7	0.419	0.076	1.924	2.704
8	0.373	0.136	1.864	2.847
9	0.337	0.184	1.816	2.970
10	0.308	0.223	1.777	3.078
11	0.285	0.256	1.744	3.173
12	0.266	0.284	1.717	3.258
13	0.249	0.308	1.692	3.336
14	0.235	0.329	1.671	3.407
15	0.223	0.348	1.652	3.472

Control Charts Based On Weighted Averages

The Moving-Average Control Chart

Shewhart $\bar{x}$ control chart is relatively insensitive to small shifts in the process mean. Various modifications and supplemental criteria have been suggested to improve its ability to detect small shifts. Control charts based on the moving average are also very effective in detecting small process shifts.

Suppose that samples of size *n* have been collected, and let $\bar{x}_1, \bar{x}_2, \ldots.., \bar{x}_t, \ldots.$ denote the corresponding sample means. The moving average of span *w* at time *t* is defined as

$$M_t = \frac{\bar{x}_t + \bar{x}_{t-1} + \cdots + \bar{x}_{t-w+1}}{w}$$

That is, at time period *t*, the oldest sample mean is dropped and the newest one added to the set. The variance of the moving average M_t is

$$V(M_t) = \frac{1}{w^2} \sum_{i=t-w+1}^{t} V(\bar{x}_i) = \frac{1}{w^2} \sum_{i=t-w+1}^{t} \frac{\sigma^2}{n} = \frac{\sigma^2}{nw}$$

Therefore, if $\bar{\bar{x}}$ denotes the center line of the control chart, then the 3-sigma control limits for M_t are

$$UCL = \bar{\bar{x}} + \frac{3\sigma}{\sqrt{nw}}$$

and

$$LCL = \bar{\bar{x}} - \frac{3\sigma}{\sqrt{nw}}$$

The control procedure would consist of calculating the new moving average M_t as each sample mean $\bar{x}_t$ becomes available, plotting M_t on a control with upper and lower control limits and concluding that the process is out of control if M_t exceeds the control limits. In general, the magnitude of the shift of interest and *w* are inversely related; smaller shifts should be guarded against more effectively by longer moving averages.

Example

An $\bar{x}$ control chart with center line $\mu = 10.0$ and upper and lower 3-sigma control limits at 16.0 and 4.0 is shown in Figure 1. Values of the sample statistic $\bar{x}_i$ plotted on the chart for periods 1,2, …. ,t are listed in table – 1. The statistic plotted on this chart will be

$$M_t = \frac{\bar{x}_t + \bar{x}_{t-1} + \cdots + \bar{x}_{t-7}}{8}$$

for period $t \geq 8$. For time periods $1 \leq t < 8$ the average of the observations for periods 1,2, … t is plotted. The values of these moving averages are shown in Table 1.

The control limits for the moving average control chart may be easily obtained. Since for the $\bar{x}$ chart we have $3\sigma_{\bar{x}} = 6.0$ then $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 2.0.$ Consequently, we find the upper and lower control limits for the moving-average control chart as

$$UCL = \overset{=}{x} + \frac{3\sigma}{\sqrt{nw}} = \overset{=}{x} + \frac{3\sigma_{\bar{x}}}{\sqrt{w}} = 10.0 + \frac{(3)(2.0)}{\sqrt{8}} = 12.12$$

and

$$LCL = \overset{=}{x} - \frac{3\sigma}{\sqrt{nw}} = \overset{=}{x} - \frac{3\sigma_{\bar{x}}}{\sqrt{w}} = 10.0 - \frac{(3)(2.0)}{\sqrt{8}} = 7.88$$

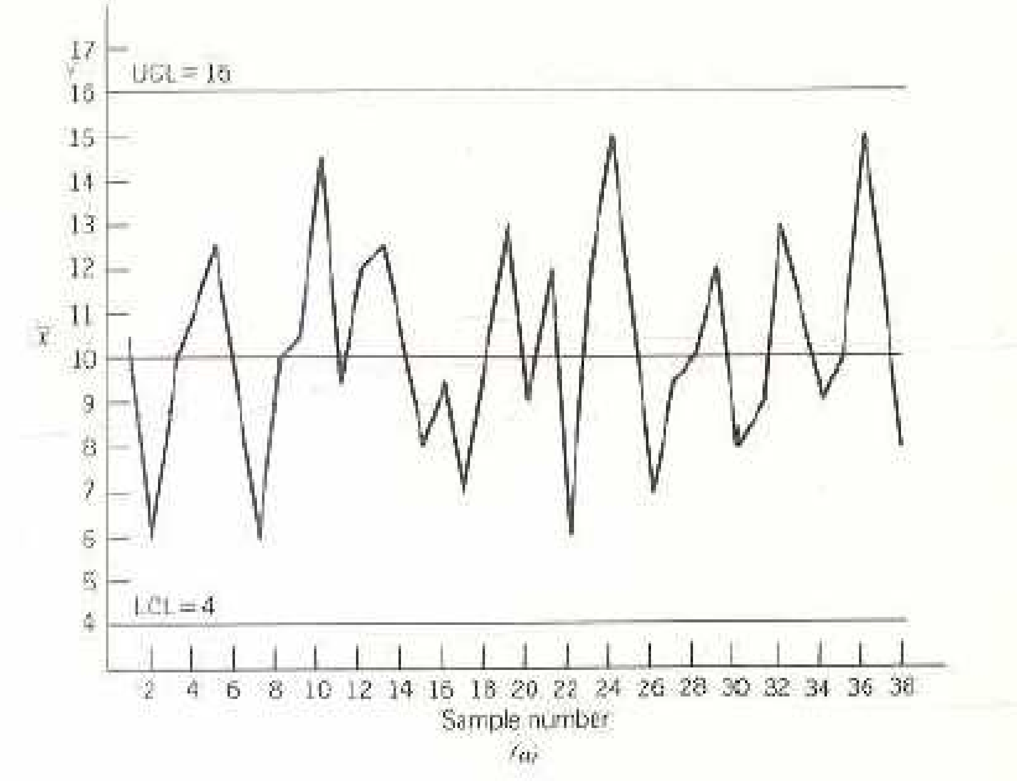


Figure – 1

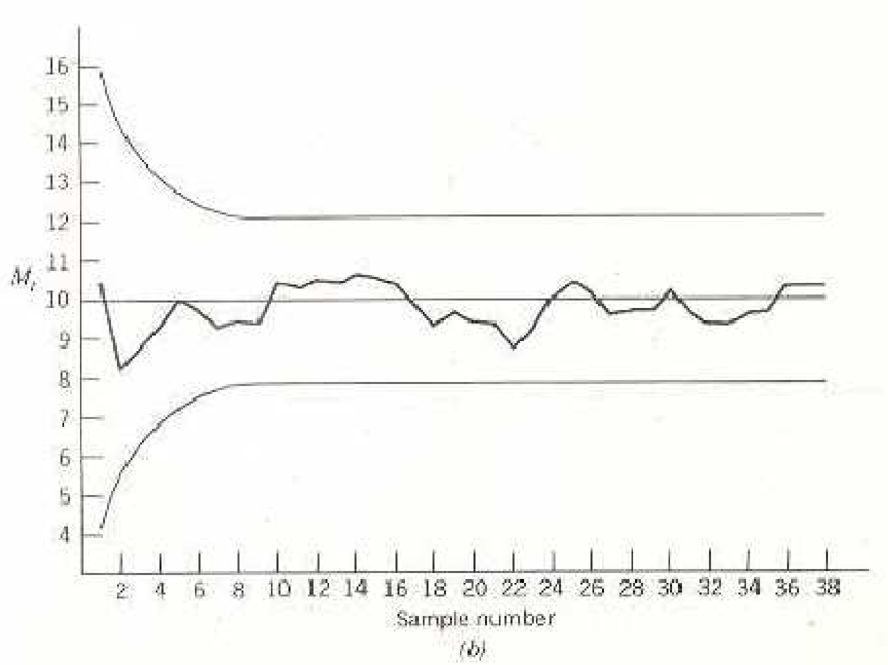


Figure – 2

Table 1 : Data and Calculations for the Moving-Average Control Chart

Sample, t	$\bar{x}_t$	M_t	Control Limits for M_t	
			LCL	UCL
1	10.5	10.5	4.00	16.00
2	6.0	8.25	5.76	14.24
3	10.0	8.83	6.54	13.46
4	11.0	9.38	7.00	13.00
5	12.5	10.00	7.32	12.68
6	9.5	9.92	7.55	12.45
7	6.0	9.36	7.73	12.27
8	10.0	9.44	7.88	12.12
9	10.5	9.44	7.88	12.12
10	14.5	10.50	⋮	⋮
11	9.5	10.44		
12	12.0	10.57		
13	12.5	10.57		
14	10.5	10.70		
15	8.0	10.50		
16	9.5	10.44		
17	7.0	10.00		
18	10.0	9.44		
19	13.0	9.88		
20	9.0	9.51		

Sample, t	$\bar{x}_t$	M_t	Control Limits for M_t	
			LCL	UCL
21	12.0	9.45		
22	6.0	8.89		
23	12.0	9.39		
24	15.0	10.08		
25	11.0	10.58		
26	7.0	10.21		
27	9.5	9.77		
28	10.0	9.90		
29	12.0	9.90		
30	8.0	10.15		
31	9.0	9.78		
32	13.0	9.53		
33	11.0	9.53		
34	9.0	9.78		
35	10.0	9.84		
36	15.0	10.47		
37	12.0	10.47		
38	8.0	10.47		

The control limits for M_t apply for periods $t \geq 8$. For period 0< t <8, the control limits are given by $\overline{\overline{x}} \pm 3\sigma / \sqrt{nt}$. These control limits are shown in Table. An alternative procedure that avoids using special control limits for periods t < w is to use an ordinary $\bar{x}$ chart until at least *w* sample means have been obtained.

The moving-average control chart is shown in Figure 2. No points exceed the control limits. Note that for the initial periods t < w the control limits are wider than their final steady-state value. Moving averages that are less than *w* periods apart are highly correlated, and this often complicates interpreting patterns on the control chart. This is easily seen by examining.

The moving-average control chart is more effective than the usual $\bar{x}$ chart in detecting small process shifts. Using both the moving-average and $\bar{x}$ control charts simultaneously can also yield good results. If the two charts are used simultaneously, the process is considered to be out of control if either M_t *or* $\bar{x}_t$ (or both) plot outside their respective control limits. It is also helpful to plot the points M_t on a standard $\bar{x}_t$ chart, so that a single chart could be used to record the data.

Moving-average control charts can also be used in cases where each sample consists of a single observation. This situation occurs frequently when production of a single unit of product requires a very long time, and where automatic measurement and test procedures are used.

CHAPTER – 13

FORECASTING AND TIME SERIES

INTRODUCTION

Forecasting helps managers respond quickly and accurately to market changes and customer needs. If various activities, operations and processes are properly planned and organized, the control is easier and smoother. Forecasting helps reducing failures and cost in making unnecessary changes in the processes and systems. For example, if the demand for the product can be estimated accurately, the operational efficiency of the organization goes up.

Forecasting deals with what we think will happen in the future. Planning deals with what we think should happen in the future. Through adequate planning, we attempt to change and control future events and forecasting helps us to predict those future events.

Good planning uses forecasts as a valuable input for planning the design and operations of an organization. Forecasts are necessary for planning, scheduling and controlling the system to facilitate effective and efficient output of goods and services.

Marketing uses forecasts to plan products, pricing, positioning and promotion purposes. Finance uses forecasting for financial management and for allocation of funds. Operations managers use forecasts for the procurement of raw material, fixing targets, scheduling of jobs and equipments. Top management uses forecasts for planning expansions, diversification and for making strategic decisions. Thus, forecasting plays a vital role in the decision making process of a manager.

Planning decisions may be classified as long, medium and short term. Long term decisions involve the development of new products and markets, setting up new plants, expansions and diversifications. Long term may mean about 2 years or more into the future. Such decisions generally lack quantitative information and historical data on which to base our forecasts.

The collective wisdom of experts in the field plays a significant role in the development of forecast for long term planning.

Medium term decisions involve issues such as fixing production and sales targets, determining manpower requirements, etc.
Medium range may be taken to mean from 6 months and up to about 2 years, which is the normal time frame for aggregate planning, budgeting and resource acquisition and allocation decisions.

Short range refers to less than 6 months. Examples of areas where such a time frame is appropriate are the procurement of materials, scheduling of jobs and activities. Similarly, managers need forecasts to make decisions about controlling inventory, production, labour and costs. Accurate forecasts are also needed for immediate future hours, days, and weeks ahead.

Some distinguish between forecasts and predictions. A forecast is seen as an estimate of a future event based on scientific methodology that uses past data. Forecasting requires past data, statistical techniques and managerial skills.

On the other hand, a prediction is an estimate of future events obtained through subjective factors like, hunch, experience and intuition. The various methods used for forecasting can be classified as follows:

(i) Qualitative methods
(ii) Quantitative methods based on averages, moving averages and exponential smoothing
(iii) Regression methods
(iv) Econometric models
(v) Auto Regressive and Moving Average (ARMA) models

FORECASTING FOR LONG TERM DECISIONS

Long term decisions cover areas such as capacity expansion, plant or facility location, mergers and acquisitions, and product development over a longer time span. These decisions require forecasts for many years into the future. For making long term decisions, past data may not be a reliable indicator of future events. Under such conditions, we mainly rely on qualitative forecasting methods. Qualitative methods depend upon managerial judgement and experience and not on any specific model. Thus, different individuals may use the same qualitative technique and arrive at different forecasts. Quantitative methods for dealing with these opinions and judgements are best suited for long term forecasting. Through such methods, we can obtain reasonable forecasts in the face of a great deal of uncertainty and lack of data. Four such techniques are the Delphi Method, the Nominal Group Technique, Survey Methods and Life Cycle Analogy approaches.

Delphi Method

This method relies on the subjective opinions of experts and aims at minimizing bias and error of judgement. A panel of experts provides written responses on the questions being considered. The co-ordinator edits and summarises the responses. On the basis of summary, the panel is then asked to reconsider the individual responses and respond again to the set of questions prepared. The answers are provided in writing. The responses of the second round are again summarized and fed back to the experts. This process is repeated three to five times until sufficient convergence is achieved. In this method, direct interpersonal relations are avoided and personalities do not conflict nor some members can dominate the group.

THE DELPHI TECHNIQUE HAS BEEN APPLIED IN THE FOLLOWING AREAS :

a) Forecasting
b) Evaluating possible budget allocations,
c) Setting corporate goals and objectives
d) Generating and evaluating strategies
e) Exploring urban and regional planning options, and
f) Planning health care systems.

GUIDELINES FOR CONDUCTING A DELPHI STUDY :

The following guidelines should be followed while conducting a Delphi study.

a) All members should agree to serve on the panel.
b) The procedure for conducting the study should be explained to the panelists in detail.
c) Every panel member should be assigned a code number.
d) Two copies of each questionnaire should be sent to the panelists in each round so that he can retain a copy for his own record.
e) The questionnaires should be easy to understand.
f) It should not contain too many statements. A practical limit is suggested as 25.
g) Contradictory forecasts should be included to initiate debate.
h) Injection of moderator's opinion should be avoided because it has been found to substantially bias the results.
i) When editing the respondent's comments for clarity, the intent for the originator should not be lost. Similarly, when editing from round to round, meaning of a statement should not be changed.
j) The questionnaire should be pre-tested on any willing guinea pigs outside the respondent group.

Nominal Group Technique

This technique is similar to the Delphi technique. However this method provides an opportunity for interaction and encourages discussions among the experts and permit creativity. At the end of discussions, the experts who arrive at a consensus rank the ideas.

Survey Methods

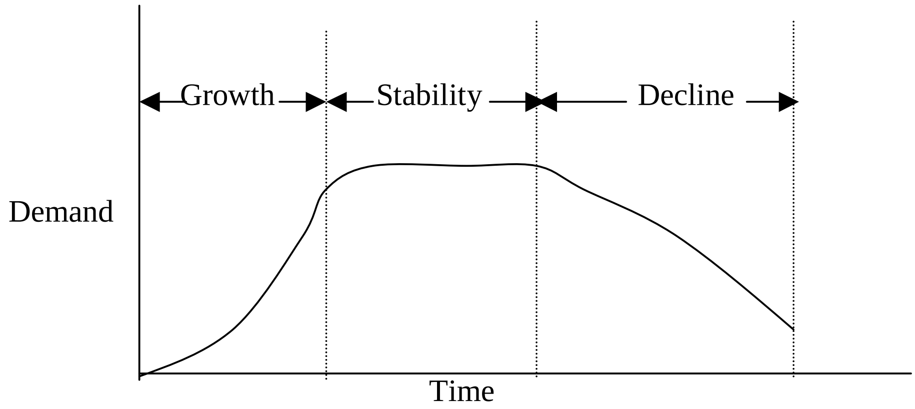
Surveys are designed and conducted to gather relevant information. Surveys generally use structured questionnaires. The responses to the questionnaires are obtained through various means:

(a) personal interviews
(b) telephone interviews
(c) mail/fax mode
(d) Internet communications

Life Cycle Analogy

In this case, predictions are based on the patterns related to the introduction, growth and saturation phases of similar products.

Product Life Cycle



The demand for a product generally tends to follow a predictable pattern called the product life cycle. When a new product is introduced, it has a low demand during market development phase. Followed up by a rapid growth phase and high demand and finally the demand declines. The time span for various phases from birth to death may vary considerably from product to product. Using various forecasting methods and characteristics of the product cycle of similar products we can make prediction for product variety, volumes and capacity needed.

FORECASTING FOR MEDIUM AND SHORT TERM DECISIONS

Medium and short terms forecasts are commonly used for production planning, scheduling, procurement and financial planning decisions in an organisation. The methods are better structured and are data based as compared to long term forecasting. Descriptions of common methodologies follow.

Time Series Forecasting

Time series analysis methods are used to study past data and to identify the patterns that are present. These patterns are then projected into the future. A time series can be decomposed into component such as average level, trend, seasonality, cycle and error. The magnitude and form of the component are estimated from the available data and projected forwarded into the future to make forecasts. Methods used are moving averages and exponential smoothing.

Moving Averages

The moving averages method is used to estimate the average of a time series and thereby remove the effect of random fluctuations. It is most useful when the time series has no pronounced trend or seasonal influence.

This technique involves calculating the average of the n most recent observation of a time series and using it as the basis for forecasts for the next time period. Large values of n should be used for a time series that is stable and small values of n if it is susceptible to change in the average value.

Exponential Smoothing

This is the most frequently used method for smoothing data. It is a weighted moving average method that gives recent observations more weight than earlier observations.
It requires three items of data:

1. Estimated average of the series for the last period (A_{t-1}) i.e. forecast for the next period.
2. Demand for the period (D_t).
3. A smoothing parameter, (α), $0 < \alpha < 1$.

Forecast for period *t* is given by
$A_t = \alpha$ (Demand for this period) $+ (1 - \alpha)$ (Average calculated for the previous period)

$$= \alpha D_t + (1 - \alpha) A_{t-1}$$

$$= A_{t-1} + \alpha (D_t - A_{t-1})$$

Such a forecast for the next period equals the forecast for the current period plus a proportion of the forecast error for the current period.

A_t is a weighted average of all past observations. It can be written as the linear combination of past data and weights decay exponentially.

For example if $\alpha = 0.20$

$$A_t = 0.2\ D_t + 0.80\ A_{t-1}$$
$$= 0.20\ D_t + 0.16\ D_{t-1} + 0.128\ D_t + 0.1024\ D_{t-2}$$

Weights given are:

$$\alpha, \alpha\ (1-\alpha),\ \alpha(1-\alpha)^2, \ldots$$

Example

Week	Demand
1	400
2	380
3	411

If the actual Demand for the week 4 is 415. What is the forecast for week 5?

1. *Using moving average*:

   The moving average at the end of week 3 is $A_3 = \dfrac{411+380+400}{3} = 397$

   Thus forecast for week 4 is 397.

   Forecast for week 5, $A_4 = \dfrac{415+411+380}{3} = 402.$

2. *Using Exponential smoothing*:

   Forecast for the week 4 using $\alpha = 0.1.$

   $A_3 = 0.10\ (411) + 0.90\ (390) = 392.1$

If the actual demand for week 4 is 415, the average for week 4 would be:

$A_4 = 0.1 \times 415 + 0.9 \times 392.1 = 394.4$

Causal Forecasting Methods

In this case, we develop a cause and effect relationship model between the variable of interest and its causal factors. For example, the demand for tyres of a particular type may be related to the population of existing vehicles, road mileage of existing vehicles, wear out rate of tyres and road conditions. We collect relevant data on these variables and use regression analysis techniques to identify the nature of the statistical relationship. The regression model fitted may be linear, polynomial or non-linear. Once the regression relationship is established, we make a prediction based on it. Causal forecasting methods also include econometric models, simulation model and input and output models.

Box and Jenkins have proposed a sophisticated technique for stochastic model building and forecasting using time series data.

Auto-Regressive Model, AR(p)

In this case, the current value is expressed as a linear combination of p-previous values of time series and random component.

Moving Average Model, MA(q)

In this model, the current value is made linearly dependent on q previous error terms.

Mixed Auto Regressive-Moving Average (ARMA) Model

Sometimes, it is advantageous to include both autoregressive and moving averages in the model. This leads to the mixed auto regressive (ARMA) model.

Activity A

Describe the role of forecasting in planning

____

____

____

Activity B

Describe the method to be used for forecasting the monthly demand for a newspaper.

____

____

____

Activity C

Describe the advantages of quantitative methods of forecasting over qualitative methods

____

____

____

Activity D

What methods you think are used for weather forecasting ?

____

____

____

7. Using a three period moving average predict the demand for the next period

Period	1	2	3	4	5
Demand	15	24	34	21	35

8. Using exponential smoothing and with alpha $(\alpha) = 0.2$, predict the demand for week 7

Week	1	2	3	4	5	6
Demand	21	32	43	43	30	45

GENERAL STEPS IN THE FORECASTING PROCESS

1. Identify the General Need.
2. Select the Period (Time Horizon) of Forecast
3. Select the Indicators Relevant to the Need:
   (i) Industry Sales
   (ii) Competitors (collective) present and projected capacity.
   (iii) Population projection (in case product is directly related to the population).
   (iv) Income levels.
   (v) Economic development etc.
4. Select the Forecast Model to be Used : For this, knowledge of various forecasting models, in which situations these are applicable, how reliable each one of them is; what type of data is required. On these considerations; one or more models can be selected.
5. Data Collection : With reference to various indicators identified-collect data from various appropriate sources-data which is compatible with the model(s) selected in step (4). Data should also go back that much in past, which meets the requirements of the model.

6. Prepare Forecast : Apply the model using the data collected and calculate the value of the forecast.
7. Evaluate.

TIME SERIES COMPONENT

A sequence of observations on a variable of interest at equally spaced points in time

Variable of Interests :

- Sales
- Demand
- Population
- Inventory
- Power Consumption
- Production
- No. of Accidents
- No. of Tourists visiting
- Traffic Intensity

Equally Spaced Points in Time

Days

Weeks

Months

Quarters

Years

Example of Time Series

TIME	1985	1986	1987	1988	1989	1990	1991
SALES **(Rs. Crore)**	2.2	2.1	2.4	2.6	2.7	2.9	2.8

Time Series Analysis :

We use quantitative techniques to study past behaviour of a Time Series to

- Identify the pattern present
- Detect changes in pattern
- Use changes and pattern to predict the future behavior of Time Series
- Forecasts provide valuable input for decision making

Components or Variations in Time Series

- Trend
- Cyclical Fluctuation
- Seasonal Variation
- Irregular or Random Variation

TREND : Long Term direction in which the Series is moving
The value of the variable tends to increase or decrease over a long period of time.

Example

- Steady increase in cost of living
- Increase in Population
- No. of tourist visiting a particular place

Trend Reflects the Net Effect of Factors :

- Change in population
- Demographic Characteristics
- Technological Improvements
- Economic Development
- Gradual Shift in Habit and Attitude

These factors tend to operate fairly gradually and in one direction or other over a long period.

We describe the trend component by a smooth continuous curve types of trends

- Linear trend
- Non-linear trend
- Linear trend or straight line trend

$$Y = a + bt + e$$

Y : Value of dependent variable

t : Value of time variable

a : The Y – intercept (value of Y when $t = 0$)

b : Slope of the trend line

e : Random component

- Non – linear trend : Quadratic trend :

$$Y = a + bt + ct^2 + e$$

Example of Quadratic trend :

T (Year)	**1992**	**1993**	**1994**	**1995**	**1996**
Y (Sales in Millions)	13	24	39	65	106

$$\hat{Y} = 39.3 + 158.9t + 248.4t^2$$

Where t = T – 1994. Predict sales for the year, 1997.

- Exponential trend :

$$Y = k\ a^t \text{ for } k > c, a > o$$

For a > 1 : Growth curve

For 0 < a < 1 Decay curve

Example : Population, Money invested, Depreciation, GNP

- Cyclical variation :

Component of a time series that trends to oscillate above and below the trend line for periods longer than one year

Factors leading to cyclical variation

- Buildups and depletion of inventories
- Shifts in rates of capital expenditure
- Year to year variation in harvests
- Change in Government monetary and fiscal policy.

We first eliminate the effect of seasonal component by using time series consisting of annual data.

Measures of cyclical variation

- Percent of trend
- Relative cyclical residual

Seasonal variation :

- Means a periodic movement in a time series where period is not longer than one year.
- A periodic movement is one which repeats at regular interval of time.
- It is repetitive and predictable.

Main Causes :

Climatic changes of different seasons

Customs and habits which people follow at different times

We can project past pattern into future and eliminate its effect from time series to get deseasonalized time series.

Ratio to moving average method :

- Develop an index to describe the degree of seasonal variation index is based on a mean of 100
- Periodic fluctuation are eliminated by taking moving average of period equal to the period of fluctuation
- Moving averages are centred against the time which is the mid-point of the time points included in the calculation of moving averages.
- When period is odd, moving averages correspond to time point given in time series
- When period is even, we calculate a subsequent 2 – item moving averages.
- Calculate percentage of actual value to moving value for each time point.
- Collect these percentage for some period and find average by deleting extreme values.
- Adjust the modified means.

Year	**Sales per quarter (x $ 10,000)**			
	I	**II**	**III**	**IV**
1988	16	21	9	18
1989	15	20	10	18
1990	17	24	13	22
1991	17	25	11	21
1992	18	26	14	25

1. Deseasonalizing the time series
2. Developing the trend line

3. Finding the cyclical variation around the trend line

Year (1)	Quarter (2)	Actual Sales (3)	Step 1:4 Quarter Moving Total (4)	Step 2:4 Quarter Moving Average (5) = $\frac{(4)}{4}$	Step 3:4 Quarter Centered Moving Average (6)	Step 4: Percentage of Actual to Moving Average (7) = $\frac{(3)}{6}\times100$
1988	I	16				
	II	21				
	III	9	64	16.00	15.825	56.7
	IV	18	63	15.75	15.625	115.2
1989	I	15	62	15.50	15.625	96.0
	II	20	63	15.75	15.750	127.0
	III	10	63	15.75	16.000	62.0
	IV	18	65	16.25	16.750	107.5
1990	I	17	69	17.25	17.625	96.5
	II	24	72	18.00	18.500	129.7
	III	13	76	19.00	19.00	68.4
	IV	22	76	19.00	19.125	115.0
1991	I	17	77	19.25	19.000	89.5
	II	25	75	18.75	18.625	134.2
	III	11	74	18.50	18.625	59.1
	IV	21	75	18.75	18.875	111.3
1992	I	18	76	19.00	19.375	92.9
	II	26	79	19.75	20.250	128.4
	III	14	83	20.75		
	IV	25				

Year	Step 5			
	I	II	III	IV
1988	--	--	56.7	115.2
1989	96.0	127.0	62.5	107.5
1990	96.5	129.7	68.4	115.0
1991	89.5	134.2	59.1	111.3
1992	92.9	128.4	--	--
	-------	---------	-------	--------
Modified sum	188.9	258.1	121.6	226.3

Modified mean : Quarter I $\dfrac{188.9}{2} = 94.45$

II $\dfrac{258.1}{2} = 129.05$

III $\dfrac{121.6}{2} = 60.80$

IV $\dfrac{226.3}{2} = 113.15$

397.45

Step 6

Adjusting factor = $\dfrac{400}{397.45} = 1.0064$

Quarter	Indices	×	Adjusting Factor	=	Seasonal
			Indices		
I	94.45	×	1.0064	=	95.1
II	129.05	×	1.0064	=	129.9
III	60.80	×	1.0064	=	61.2
IV	113.15	×	1.0064	=	113.9
			Sum of seasonal indices	=	**400.1**

Year (1)	Quarter (2)	Actual Sales (3)	Seasonal Index / 100 (4)	Deseasonalized Sales (5) = (3) + (4)
1988	I	16	0.951	16.8
	II	21	1.299	16.2
	III	9	0.612	14.7
	IV	18	1.139	15.8
1989	I	15	0.951	15.8
	II	20	1.299	15.4
	III	10	0.612	16.3
	IV	18	1.139	15.8
1990	I	17	0.951	17.9
	II	24	1.299	18.5
	III	13	0.612	21.2
	IV	22	1.139	19.3
1991	I	17	0.951	17.9
	II	25	1.299	19.2
	III	11	0.612	18.0
	IV	21	1.139	18.4
1992	I	18	0.951	18.9
	II	26	1.299	20.0
	III	14	0.612	22.9
	IV	25	1.139	21.9

CHAPTER – 14

SIX SIGMA CONCEPT IN ENVIRONMENTAL MANAGEMENT

What is "Six Sigma"?

Six Sigma is a well-structured, data-driven methodology for eliminating defects, waste, or quality control problems of all kinds in manufacturing, service delivery, management, and other business activities. Six Sigma methodology is based on the combination of well-established statistical quality control techniques, simple and advanced data analysis methods, and the systematic training of all personnel at every level in the organization involved in the activity or process targeted by Six Sigma.

Why is Six Sigma so popular?

Six Sigma methodology has recently gained wide popularity because it has proven to be successful not only at improving quality but also at producing large cost savings along with those improvements. Some spectacular Six Sigma "success stories" at large corporations have been widely publicized and they captured the imagination of many business leaders.

For example, Jack Welch, a CEO of General Electric (one of the largest manufacturing businesses in the world) said *"Six Sigma is the most important initiative GE has ever undertaken--it is part of the genetic code of our future leadership."* and he credits Six Sigma with cost savings at GE in the range of billions of dollars.

Technically Speaking...

The term Six Sigma (a trademark of Motorola, where it originated over 12 years ago) reflects the statistical objective of the approach, namely striving to achieve a negligible number of defects, corresponding to the probability associated with a six sigma value for the normal curve: Applying the normal curve, Six Sigma attempts to relegate defects and quality problems to the very tails of the distribution, making such problems literally rare exceptions in a process that operates almost without defects. To achieve this "Six Sigma objective," a process must not produce more than 3.4 defects per million opportunities to produce such defects (where a "defect" is defined as any kind of unacceptable outcome produced by the process under scrutiny). Note that the 3.4 defects-per-million criterion actually corresponds to a normal z value of 4.5 because the Six Sigma approach allows for 1.5 times sigma worth of so-called "drift" or process "slop" (termed by Motorola the "Long-Term Dynamic Mean Variation"). Hence, the most basic statistical tool for the Six Sigma effort is the Six Sigma calculator that will compute the number of defects given the respective one, two, .., six sigma process. In addition, a wide variety of much more complex analytic techniques are recommended by the Six Sigma approach and need to be used at the consecutive stages of the Six Sigma project, depending on the nature of the process.

Key Concepts of Six Sigma

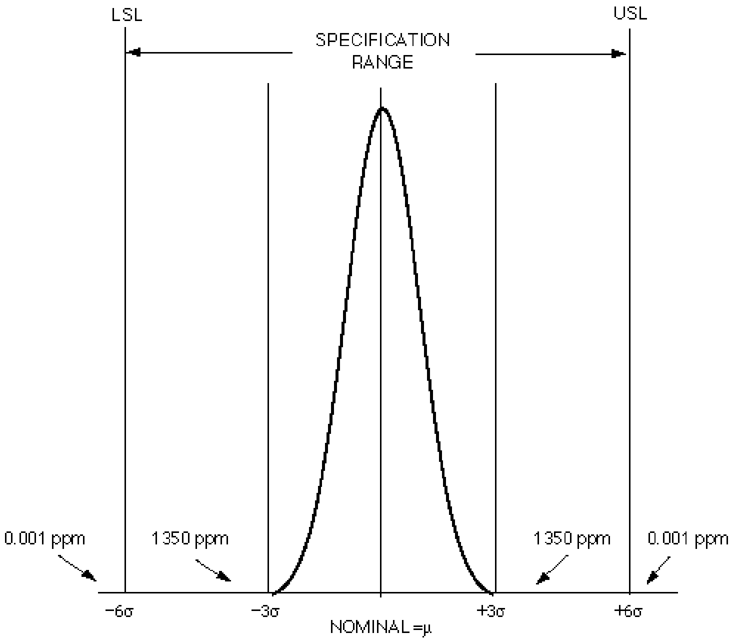
At its core, Six Sigma revolves around a few key concepts.

Critical to Quality:	Attributes most important to the human being
Defect:	Failing to deliver what the human being wants
Process Capability:	What your process can deliver
Variation:	What the human being sees and feels
Stable Operations:	Ensuring consistent, predictable processes to improve what the human being sees and feels
Design for Six Sigma:	Designing to meet human being needs and process capability

Human Beings Feel the Variance, Not the Mean

Often, our inside-out view of the business is based on average or mean-based measures of our recent past. Customers don't judge us on averages, they feel the variance in each transaction, each product we ship. Six Sigma focuses first on reducing process variation and then on improving the process capability.

Customers value consistent, predictable business processes that deliver world-class levels of quality. This is what Six Sigma strives to produce.



How does it work?

The power of Six Sigma lies in its "empirical," data-driven approach (and its focus on using quantitative measures of how the system is performing) to achieve the goal of the process improvement and variation reduction. That is done through the application of so-called "Six Sigma improvement projects" which, in turn, follow the "Six Sigma DMAIC" sequence of steps (Define, Measure, Analyze, Improve, and Control). Specifically:

- **Define.** The ***Define*** phase is concerned with the definition of project goals and boundaries, and the identification of issues that need to be addressed to achieve the higher (better) sigma level.
- **Measure.** The goal of the ***Measure*** phase of the Six Sigma strategy is to gather information about the current situation, to obtain baseline data on current process performance, and to identify problem areas.
- **Analyze.** The goal of the ***Analyze*** phase of the Six Sigma quality effort is to identify the root cause(s) of quality problems, and to confirm those causes using the appropriate data analysis tools.
- **Improve.** The goal of the ***Improve*** phase is to implement solutions that address the problems (root causes) identified during the previous (***Analyze***) phase.

- **Control.** The goal of the ***Control*** phase is to evaluate and monitor the results of the previous phase (***Improve***).

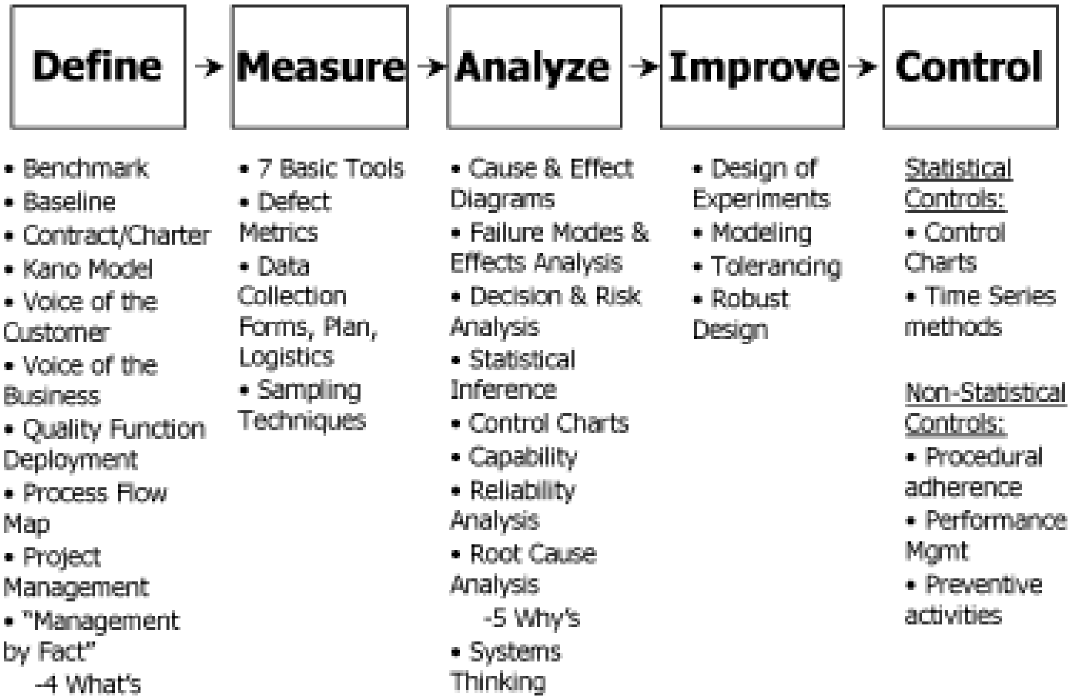
There is also a variation of the fundamental Six Sigma *DMAIC* sequence, called *DMADV*, applicable to the design of new processes. In the *DMADV* sequence, the **Define** stage is identical to the one in *DMAIC* (see above); the **Measure** stage focuses on the measurement of the customer and/or market/application needs, the **Analyze** stage deals with the analysis of the process options and, finally, the **Improve** and **Control** stages are replaced by the **Design** (design the process to meet the customer and/or market/application needs) and **Verify** (verify the design performance and ability to meet the criteria as set at the Design level) stages. Each of these steps involves using specific analytic (quantitative) methods from a wide selection of methods recommended by the Six Sigma approach (depending on the nature of the process).

The primary goal of Six Sigma is to improve customer satisfaction, and thereby profitability, by reducing and eliminating defects. Defects may be related to any aspect of customer satisfaction: high product quality, schedule adherence, cost minimization. Underlying this goal is the Taguchi Loss Function which shows that increasing defects leads to increased customer dissatisfaction and financial loss. Common Six Sigma metrics include defect rate (parts per million or ppm), sigma level, process capability indices, defects per unit, and yield. Many Six Sigma metrics can be mathematically related to the others.The Six Sigma drive for defect reduction, process improvement and customer satisfaction is based on the "statistical thinking" paradigm

- Everything is a process
- All processes have inherent variability
- Data is used to understand the variability and drive process improvement decisions

As the roadmap for actualizing the statistical thinking paradigm, the key steps in the Six Sigma improvement framework are Define - Measure - Analyze - Improve - Control (see Figure 1). Six Sigma distinguishes itself from other quality improvement programs immediately in the "Define" step. When a specific Six Sigma project is launched, the customer satisfaction goals have likely been established and decomposed into subgoals such as cycle time reduction, cost reduction, or defect reduction. (This may have been done using the Six Sigma methodology at a business/organizational level.) The Define stage for the specific project calls for baselining and benchmarking the process to be improved, decomposing the process into manageable sub-processes, further specifying goals/sub-goals and establishing infrastructure to accomplish the goals. It also includes an assessment of the cultural/organizational change that might be needed for success.

Once an effort or project is defined, the team methodically proceeds through Measurement, Analysis, Improvement, and Control steps. A Six Sigma improvement team is responsible for identifying relevant metrics based on engineering principles and models. With data/information in hand, the team then proceeds to evaluate the data/information for trends, patterns, causal relationships and "root cause," etc. If needed, special experiments and modeling may be done to confirm hypothesized relationships or to understand the extent of leverage of factors; but many improvement projects may be accomplished with the most basic statistical and non-statistical tools. It is often necessary to iterate through the Measure-Analyze-Improve steps. When the target level of performance is achieved, control measures are then established to sustain performance. A partial list of specific tools to support each of these steps is shown in Figure 1.



Note: Many tools can be effectively used in multiple steps of the framework. Tools that are not particularly relevant to software applications have not been included in this list.

Figure 1: Six Sigma Improvement Framework and Toolkit

An important consideration throughout all the Six Sigma steps is to distinguish which process substeps significantly contribute to the end result. The defect rate of the process, service or final product is likely more sensitive to some factors than others. The analysis phase of Six Sigma can help identify the extent of improvement needed in each substep in order to achieve the target in the final product. It is important to remain mindful that six sigma performance (in terms of the ppm metric) is not required for every aspect of every process, product and service. It is the goal only where it quantitatively drives (i.e, is a significant "control knob" for) the end result of customer satisfaction and profitability.

The current average industry runs at four sigma, which corresponds to 6210 defects per million opportunities. Depending on the exact definition of "defect" in payroll processing, for example, this sigma level could be interpreted as 6 out of every 1000 paychecks having an error. As "four sigma" is the average current performance, there are industry sectors running above and below this value. Chemists went testing for MTBE in water if operate at two sigma level then commit 308537 errors per million opportunities.

On the other extreme, in (U.S.) air quality fatality rates run at better than six sigma, which could be interpreted as fewer than 3.4 fatalities per million persons - that is, fewer than 0.00034 fatalities per 100 persons.

As just noted, flight fatality rates are "better than six sigma," where "six sigma" denotes the actual performance level rather than a reference to the overall combination of philosophy, metric, and improvement framework. Because customer demands will likely drive different performance expectations, it is useful to understand the mathematical origin of the measure and the term "six-sigma process." Conceptually, the sigma level of a process or product is where its customer-driven specifications intersect with its distribution. A centered six-sigma process has a normal distribution with mean=target and specifications placed 6 standard deviations to either side of the mean. At this point, the portions of the distribution that are beyond the specifications contain 0.002 ppm of the data (0.001 on each side). Practice has shown that air quality experienced by people shows a shift (due to drift over time) of 1.5 standard deviations so that the mean no longer equals target. When this happens in a six-sigma process, a larger portion of the distribution now extends beyond the specification limits: 3.4 ppm.

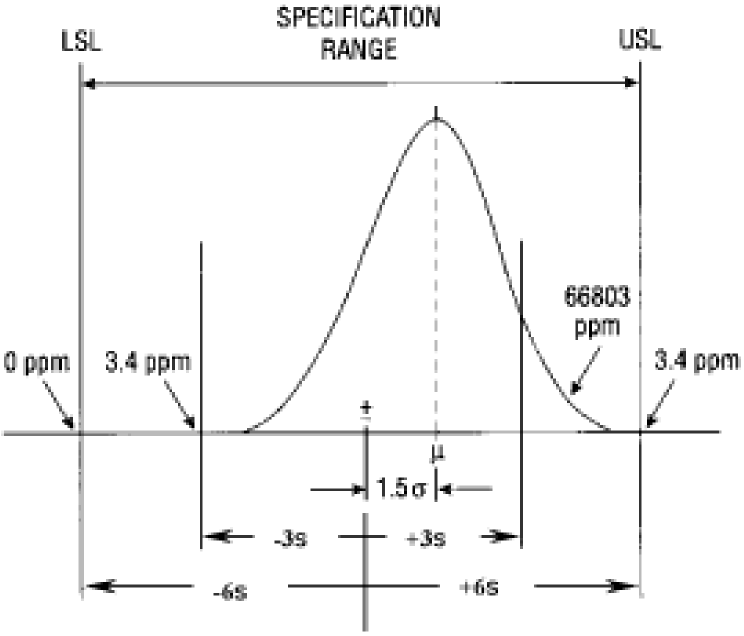


Figure 2: Six Sigma Process with Mean Shifted from Nominal by 1. 5

Assumptions:

- Normal Distribution
- Process Mean Shift of 1.5σ from Nominal is Likely
- Process Mean and Standard Deviation are known
- Defects are randomly distributed throughout units
- Parts and Process Steps are Independent

For this discussion, original nominal value = target

Key

σ =standard deviation
µ = center of the distribution
(shifted 1.5σ from its original , on-target location)
+/-3σ & +/-6σ show the specifications relative to

σ the original target

Figure 2 depicts a 1.5σ-shifted distribution with "6σ" annotations. In manufacturing, this shift results from things such as mechanical wear over time and causes the six-sigma defect rate to become 3.4 ppm. The magnitude of the shift may vary, but empirical evidence indicates that 1.5 is about average. Does this shift exist in the software process? While it will take time to build sufficient data repositories to verify this assumption within the software and systems sector, it is reasonable to presume that there are factors that would contribute to such a shift.

Process Map

A process map

- Graphically outlines the sequence of a process
- Shows how steps in a process relate to each other
- Identifies bottlenecks
- Pinpoints redundancies
- Locates waste in the process

THE COMMON METRIC: DEFECTS PER UNIT (DPU)

DPU is the best measure of the overall quality of the process.

- DPU is the independent variable.
- Process yields are dependent upon DPU.

We checked 500 samples of ambient air and these had 10 defects then,

d.p.u. = d/u = 10/500 = 0.02

In a sample of ambient air we check for the following:

a) organic compounds soluble in benzene
b) SO_2 solution
c) NO_2 concentration
d) CO concentration
e) Particulate matter

Then there are 5 opportunities for the defects to occur. Then, the total no. of opportunities = m u = 5x500 = 2500

Defects per opportunity, d.p.o. = d/(m u) = 10/2500 = 0.004
If expressed in terms of d.p.m.o. (defects per million opportunities) it becomes
d.p.m.o . = d.p.o. x 10^6 = 4000 PPM
From d.p.o., we go to the normal distribution tables and calculate Z_{LT} and corrected to Z_{ST} by adjusting for shift (1.5 s) then,

ÛZ_{LT} = 2.65; and
ÛZ_{ST} = 2.65 + 1.5 = 4.15

No. of opportunities = No. of points checked
If you don’t check some points then it becomes a passive opportunity. We should take only active opportunities into our calculation of d.p.o., and s level.

Customer Satisfaction and Defects Per Unit (DPU)

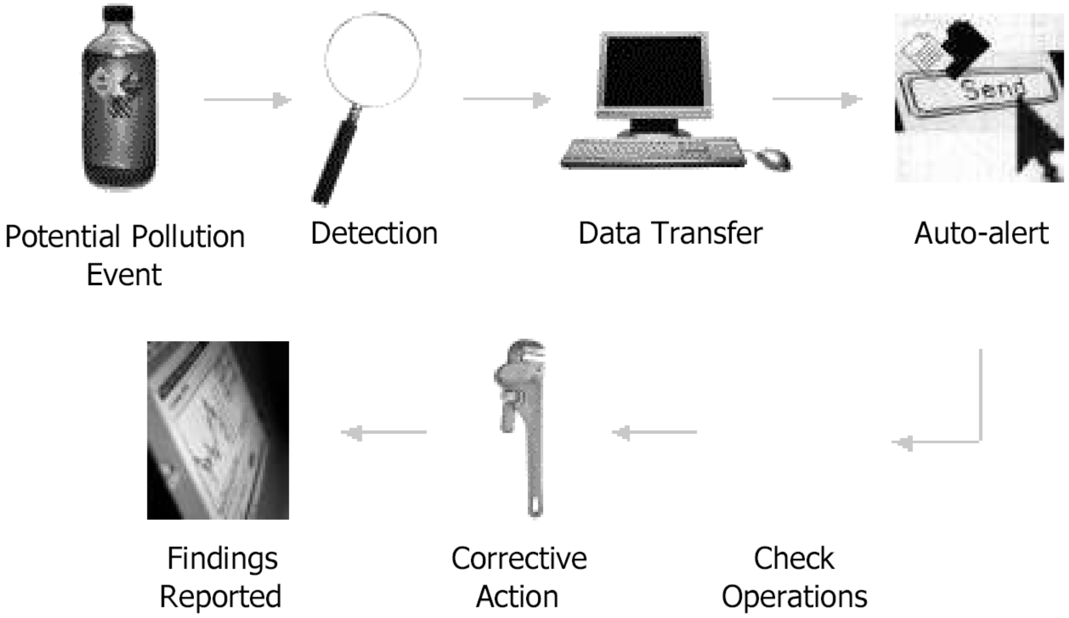
Reducing the Defects Per Unit (DPU) in the entire process:

- Reduces delivery delinquencies;
- Reduces delivered defects and early life failure rate

Process Cost and DPU

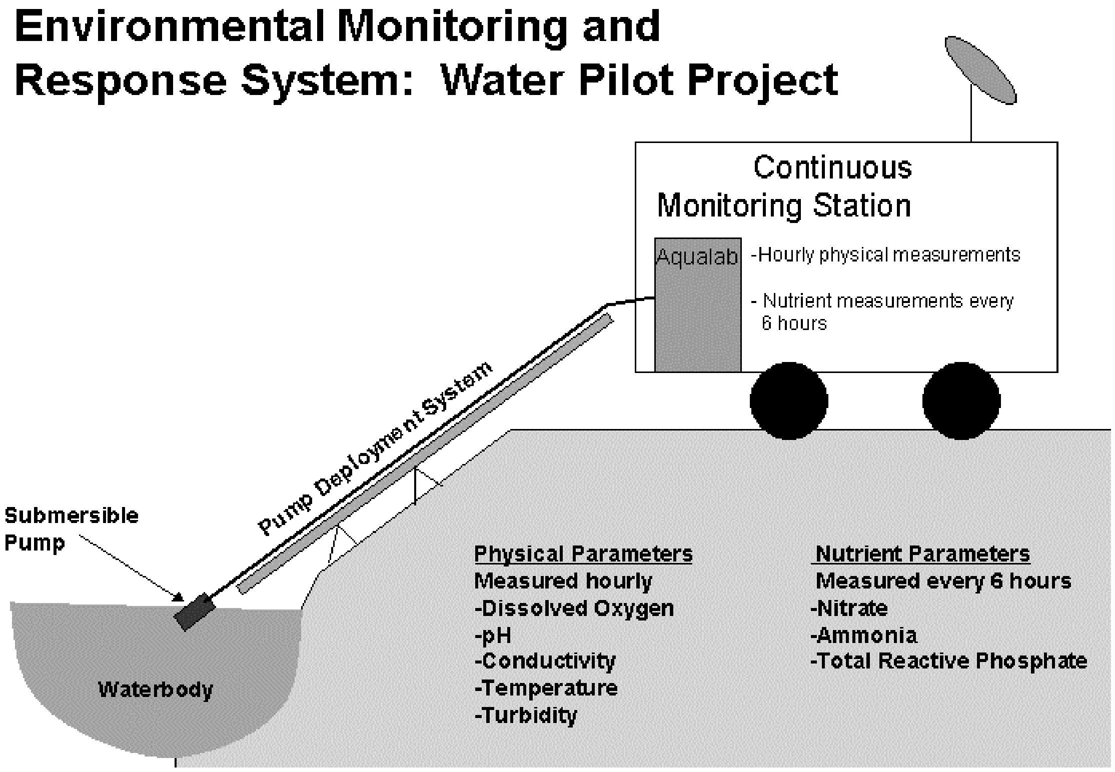
- Reducing the Defects Per Unit (DPU) in the entire process:
- Reduces the cycle time per unit….
- Reducing WIP (Work in Process);
- Reducing inventory carrying costs; and defect analysis and repair cost per unit; therefore…*Decreases Manufacturing Cost Per Unit*

Environmental Monitoring & Response System



A Vision for the Future

- Develop a warning system to prevent threats to human health and the environment and to act swiftly when such potential threats become a reality.
- Use monitoring data to develop better rules and to monitor their effectiveness.
- Be able to quickly respond to public health and environmental concerns raised by the public.
- Enhance the ability to provide accurate and timely information to the public concerning environmental quality.



Environmental Monitoring and Response System: Water Pilot Project

Continuous Monitoring Station

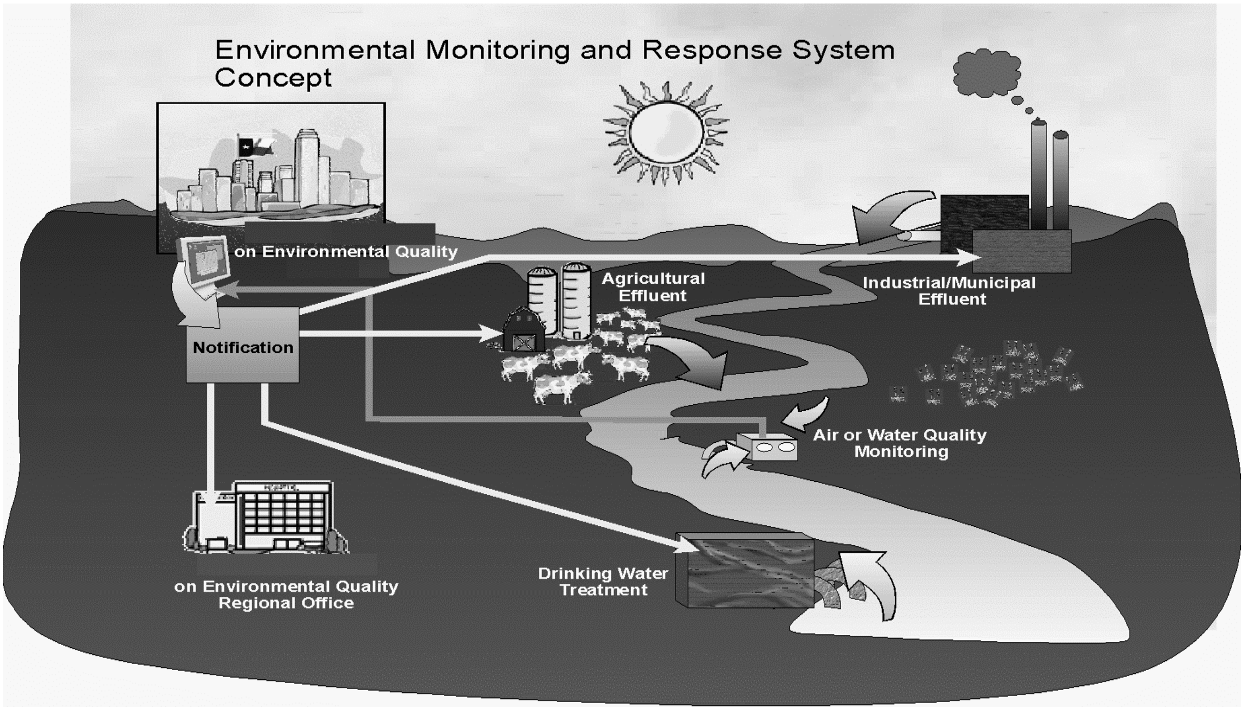
Aqualab

-Hourly physical measurements

- Nutrient measurements every 6 hours

Physical Parameters
Measured hourly
-Dissolved Oxygen
-pH
-Conductivity
-Temperature
-Turbidity

Nutrient Parameters
Measured every 6 hours
-Nitrate
-Ammonia
-Total Reactive Phosphate



Environmental Monitoring and Response System Concept

on Environmental Quality
Notification
Agricultural Effluent
Industrial/Municipal Effluent
Air or Water Quality Monitoring
on Environmental Quality Regional Office
Drinking Water Treatment

Managing for Environmental Results

- Environmental Results Management System
- Requires the agency to focus on achieving environmental results
- Connects results management with budget planning
- Based upon the Plan-Do-Check-Adapt cycle

- Agency's Strategic Plan
- Agency's Workplan
- Division Workplans
- Environmental Performance
- Quarterly Performance Report
- Quarterly Management Review
- Monthly Division Measures
- Department Results
- Legislative Reports

Six Sigma

- System for building and sustaining performance
- Uses specific tools for process improvement
- PCA uses Six Sigma because our resources are decreasing but our workload is growing
- Measurement of processes plays a significant role
- Requires calculation of cost/benefit and environmental benefit

Glossary of Definitions

DFSS – (Design for Six Sigma) is a systematic methodology utilizing tools, training and measurements to enable us to design products and processes that meet customer expectations and can be produced at Six Sigma quality levels.

DMAIC – (Define, Measure, Analyze, Improve and Control) is a process for continued improvement. It is systematic, scientific and fact based. This closed-loop process eliminates unproductive steps, often focuses on new measurements, and applies technology for improvement.

Six Sigma – A vision of quality which equates with only 3.4 defects per million opportunities for each product or service transaction. Strives for perfection.

Quality Tools

Associates are exposed to various tools and terms related to quality. Below are just a few of them.

Control Chart – Monitors variance in a process over time and alerts the business to unexpected variance which may cause defects.

Defect Measurement – Accounting for the number or frequency of defects that cause lapses in product or service quality.

Pareto Diagram – Focuses on efforts or the problems that have the greatest potential for improvement by showing relative frequency and/or size in a descending bar graph. Based on the proven Pareto principle: 20% of the sources cause 80% of any problems.

Process Mapping – Illustrated description of how things get done, which enables participants to visualize an entire process and identify areas of strength and weaknesses. It helps reduce cycle time and defects while recognizing the value of individual contributions.

Root Cause Analysis – Study of original reason for nonconformance with a process. When the root cause is removed or corrected, the nonconformance will be eliminated.

Statistical Process Control – The application of statistical methods to analyze data, study and monitor process capability and performance.

Tree Diagram – Graphically shows any broad goal broken into different levels of detailed actions. It encourages team members to expand their thinking when creating solutions.

Quality Terms

Black Belt – Leaders of team responsible for measuring, analyzing, improving and controlling key processes that influence customer satisfaction and/or productivity growth. Black Belts are full-time positions.

Control – The state of stability, normal variation and predictability. Process of regulating and guiding operations and processes using quantitative data.

CTQ: Critical to Quality (Critical "Y") – Element of a process or practice which has a direct impact on its perceived quality.

Customer Needs, Expectations – Needs, as defined by customers, which meet their basic requirements and standards.

Defects – Sources of customer irritation. Defects are costly to both customers and to manufacturers or service providers. Eliminating defects provides cost benefits.

Green Belt – Similar to Black Belt but not a full-time position.

Master Black Belt – First and foremost teachers. They also review and mentor Black Belts. Selection criteria for Master Black Belts are quantitative skills and the ability to teach and mentor. Master Black Belts are full-time postions.

Variance – A change in a process or business practice that may alter its expected outcome.

APPENDIX

STATISTICAL TABLES

TABLE A : AREAS UNDER THE STANDARD NORMAL CURVE

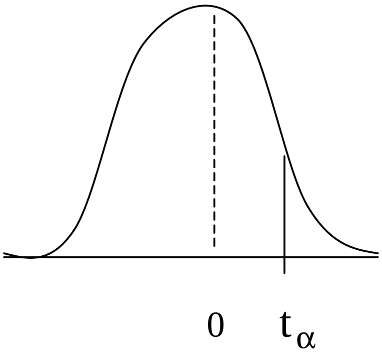
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4217
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table A: continued

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5010	0.5080	0.5120	0.5160	0.5199	0.5219	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6180	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9839	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9981	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

TABLE B t – DISTRIBUTION

Critical Values of the *t* Distribution α



	α				
v	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576

Table C : Percentage Points of the F distribution ($F_{.05, v_1, v_2}$)

V_1\	Degrees of freedom for the numerator (v$_1$)																		
v_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.0	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.9 0	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.33	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1 79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1 93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2 .24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1 58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Table D : Percentage of Points of the F Distribution ($F_{.10,v_1,v_2}$)

V₁\V₂	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2 .16
10	3.29	2.92	2.73	2.61	2.52	2.46	2. 41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2 .61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	1.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.67	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1 .96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	199	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1. 67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1 87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Table E : Statistical Constants for $\overline{X}$ and R Control Charts

n	A_2	D_3	D_4	d_2
2	1.880	0	3.268	1.128
3	1.023	0	2.574	1.693
4	0.729	0	2.282	2.059
5	0.577	0	2.114	2.326
6	0.483	0	2.004	2.534
7	0.419	0.076	1.924	2.704
8	0.373	0.136	1.864	2.847
9	0.337	0.184	1.816	2.970
10	0.308	0.223	1.777	3.078