



NATIONAL OPEN UNIVERSITY OF NIGERIA

**APPLIED QUANTITATIVE ANALYSIS
COURSE CODE: ECO 729**

**FACULTY OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS**

COURSE CONTENT DEVELOPER

**Dr. Ernest Simeon Odior
Department of Economics, Faculty of Social Sciences
University of Lagos, Lagos-Nigeria**

Course Content Editor

**Prof. Y. A. Zakari
Department of Economics
Faculty of Social Sciences
Usmanu Danfodiyo University, Sokoto
Nigeria**

© 2020 by NOUN Press
National Open University of Nigeria,
Headquarters,
University Village,
Plot 91, Cadastral Zone,
Nnamdi Azikiwe Expressway,
Jabi, Abuja.

Lagos Office
14/16 Ahmadu Bello Way,
Victoria Island, Lagos.

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher. Printed: 2020

CONTENTS

Introduction
Course Content
Course Aims
Course Objectives
Working through This Course
Course Materials
Study Units
Presentation Schedule
Assessment
Tutor-Marked Assignment (TMAs)
Final Examination and Grading
Facilitators/Tutors and Tutorials
Summary

INTRODUCTION

ECO 729 - Applied Quantitative Analysis is a two-credit and one-semester postgraduate Diploma course for Development Economics students. The course consists of up of thirteen units spread across twelve lectures weeks. This course guide gives you an insight to Quantitative Analysis in a broader way. It tells you about the course materials and how you can work your way through these materials. This course guide informs you of what you need to know about the aims and objectives of the course, composition of the course material, arrangement of the modules and study units, assignments, and examinations. It suggests some general guidelines for the amount of time required of you on each unit in order to achieve the course aims and objectives successfully.

COURSE CONTENT

This course is basically on Applied Quantitative Analysis because as you are aspiring to become an analytical economist, you must be able to apply the quantitative techniques to solve research problems. The topics covered include statistical theory, descriptive statistics and probability applications, quantitative techniques and linear programming quantitative techniques, forecasting, decision and inventory analysis, data analysis techniques and statistical software in applied quantitative techniques

COURSE AIMS

The course is aimed at acquainting students with applied practicable quantitative techniques tools in economic analysis and the understanding of applications of

quantitative techniques in economic and business decision making. To ensure the achievement of this aim, some important background information have been provided and discussed, including:

- Statistical Theory and Inference
- Overview of Descriptive Statistics
- Probability Applications
- Overview of Quantitative Techniques
- Linear Programming Graphical Method
- Linear Programming Simplex Method
- Linear Programming Transportation Model
- Forecasting and Decision Analysis
- Demonstrate Forecasting Methods
- Deterministic Inventory Control Models
- An Overview of Quantitative Research
- Quantitative Data Concepts
- Data Analysis Tools in Applied Quantitative Techniques

COURSE OBJECTIVES

To achieve the aims of this course, there are overall objectives which the course is out to achieve though, there are set out objectives for each unit. The unit objectives are included at the beginning of a unit; you should read them before you start working through the unit. On successful completion of the course, you should be able to:

- Know the importance of statistical theory
- Understand the meaning of statistical inference
- Make use of the basic statistical tools
- Understand the univariate methods for categorical variables
- Know bivariate methods for cases where both variables are categorical
- Define the meaning of descriptive statistics
- Explain the measure of central tendency
- Know the measure of central dispersion
- Define Probability concepts
- Distinguish between the different Probability Formulas
- Define and explain operational rules of Probability
- Explain the difference Probability Distribution
- Solve Questions Using Probability Formulas
- Describe the meaning of quantitative techniques
- Understand the various quantitative technique approaches
- Develop a quantitative analysis

- Understand the requirements of a Linear Programming Problem
- Formulate a typical Linear Programming Problem
- Solve a Linear Programming Problem using graphical method
- Understand the conditions that should be met before applying the simplex method
- Understand the steps to be followed in solving a linear program using the simplex method
- Solve a linear programming problem using the simplex method.
- Understand a typical transportation model
- Set up a transportation model
- Set up the mathematical model of a transportation problem
- Solve a transportation model using the Northwest Corner Rule
- Explain and understand the types of forecasts models
- Plot a scatter diagram with a time series.
- Measure forecast accuracy
- Understand time series forecasting models
- Monitor and control forecasts
- Understand the steps in Decision Making
- Know the types of decision-making environments.
- Demonstrate the various forecasting methods
- Forecast using data set and apply any of the forecast methods
- Discuss seasonality issues in forecasting
- Measure forecasting accuracy
- Define inventory and explain what inventory control is all
- Apply first order difference equations to estimate Inventory Control and EOQ model
- Apply modern inventory control models
- Define quantitative research.
- Distinguish between quantitative and qualitative research
- Determine the appropriate measurement scale for a research problem
- Describe how to identify the appropriate approach for a particular research problem
- Explain the difference between a research question and a research hypothesis and describe the appropriate use of each.
- Define validity, reliability, falsifiability, generalizability, and reproducibility as they relate to quantitative research
- Define and explain operational definitions and provide examples
- Define quantitative data and its characteristics.
- Describe common methods of quantitative data collection
- Distinguish between primary and secondary data in research methods
- Describe when quantitative research methods should be used to examine a research problem

- Define sampling and randomization.
- Demonstrate how to compute primary and secondary data
- Define the concept Statistical Software
- Describe the benefits and uses of software programs in statistical analysis of quantitative data.
- Compare and contrast the most commonly used software packages.
- Understand the use of Microsoft excel in statistical data analysis
- Explain role of computer software in quantitative analysis

WORKING THROUGH THE COURSE

To successfully complete this course a student will be required to read all the study units, read texts identified under further reading and other materials which may be provided. Each unit contains self-assessment exercises called Student Assessment Exercises (SAE). At some points in the course, a student will also be required to do self- assessment exercise to submit assignments for assessment purposes. At the end of the course there is a final examination. This course should take about 13 weeks to complete and some components of the course are outlined under the course material subsection.

COURSE MATERIAL

Major components of the course are:

1. Course guide
2. Study unit
3. Textbook
4. Assignment file
5. Presentation schedule

STUDY UNIT

There are four modules in this course broken into 13 units. The modules are made up of interconnected units as follows:

MODULE 1: STATISTICAL THEORY, DESCRIPTIVE STATISTICS AND PROBABILITY APPLICATIONS

Unit 1 Statistical Theory and Inference

Unit 2 Overview of Descriptive Statistics

Unit 3 Probability Applications

MODULE 2: QUANTITATIVE TECHNIQUES AND LINEAR PROGRAMMING

UNIT 1 Overview of Quantitative Techniques

UNIT 2 Linear Programming Graphical Method

- UNIT 3 Simplex Method
- UNIT 4 Transportation Model

MODULE 3: FORECASTING, DECISION AND INVENTORY ANALYSIS

- Unit 1: Forecasting and Decision Analysis
- Unit 2: Demonstrate Forecasting Methods
- Unit 3 Deterministic Inventory Control Models

MODULE 4: DATA ANALYSIS TECHNIQUES AND STATISTICAL SOFTWARE IN APPLIED QUANTITATIVE ANALYSIS

- Unit 1 An Overview of Quantitative Research
- Unit 2 Quantitative Data Concepts
- Unit 3 Data Analysis Tools in Applied Quantitative Techniques

Each study unit will take at least two hours, and it include the introduction, objective, main content, self-assessment exercise, conclusion, summary and reference. Other areas border on the Tutor-Marked Assessment (TMA) questions. Some of the self-assessment exercise will necessitate discussion, brainstorming and argument with some of your colleges. You are advised to do so in order to understand and get acquainted with historical economic event as well as notable periods.

There are also textbooks under the reference and other (on-line and off-line) resources for further reading. They are meant to give you additional information if only you can lay your hands on any of them. You are required to study the materials; practice the self-assessment exercise and tutor-marked assignment (TMA) questions for greater and in-depth understanding of the course. By doing so, the stated learning objectives of the course would have been achieved

ASSESSMENT

Each unit of the course has a Self-Assessment Exercise. You will be expected to attempt them as this exercise will enable you know if you understood the content of the unit.

TUTOR-MARKED ASSIGNMENT (TMAS)

The Tutor-Marked Assignment (TMA) at the end of each unit is designed to test your understanding and application of the concepts learnt. It is extremely important that you submit these assignments to your facilitators for assessment and comment. Tutor-Marked Assignment scores are often assigned through e-TMAs, organized by the University. The total score obtained is usually 30 percent.

EXAMINATION AND GRADING

At the end of the course, you will be expected to participate in the final examinations as scheduled. The final examination constitutes seventy (70) per cent of the total grading score for the course.

FACILITATORS/TUTORS AND TUTORIALS

There are tutorials provided in support of this course. The dates and venue will be worked out. The tutorials would provide the face to face physical contacts with your facilitator.

SUMMARY

This course, ECO 729: Applied Quantitative Analysis in Economics and Business is ideal for today's public and private-sector administrators, managers and the computerized business world in analyzing quantitative problems. It will enable students carry out research activities and effectively present arguments on the way forward for economic management of productive resources. Also, Students will gain some experience in applying these statistical tools to real world problems by collecting, processing, and analyzing their own data in two team writing assignments and a case study. Having successfully completed the activities as required by the course, you will be equipped with the global expectations in applying these statistical tools to real world problems. Enjoy the course.

ASSIGNMENT FILE

Assignment files and marking scheme will be made available to you. This file presents you with details of the work you must submit to your tutor for marking. The marks you obtain from these assignments shall form part of your final mark for this course. Additional information on assignments will be found in the assignment file and later in this Course Guide in the section on assessment.

There are four assignments in this course. The four course assignments will cover:

- Assignment 1 - All TMAs' question in Units 1 – 3 (Module 1)
- Assignment 2 - All TMAs' question in Units 4 – 7 (Module 2)
- Assignment 3 - All TMAs' question in Units 8 – 10 (Module 3)
- Assignment 4 - All TMAs' question in Units 11 – 13 (Module 4)

PRESENTATION SCHEDULE

The presentation schedule included in your course materials gives you the important dates for this year for the completion of tutor-marking assignments and attending tutorials. Remember, you are required to submit all your assignments by due date. You should guide against falling behind in your work.

ASSESSMENT

There are two types of the assessment of the course. First are the tutor-marked assignments; second, there is a written examination.

In attempting the assignments, you are expected to apply information, knowledge and techniques gathered during the course. The assignments must be submitted to your tutor for formal Assessment in accordance with the deadlines stated in the Presentation Schedule and the Assignments File. The work you submit to your tutor for assessment will count for 30% of your total course mark.

At the end of the course, you will need to sit for a final written examination of three hours' duration. This examination will also count for 70% of your total course mark.

TUTOR-MARKED ASSIGNMENTS (TMAS)

There are four tutor-marked assignments in this course. You will submit all the assignments. You are encouraged to work all the questions thoroughly. The TMAs constitute 30% of the total score.

Assignment questions for the units in this course are contained in the Assignment File. You will be able to complete your assignments from the information and materials contained in your set books, reading and study units. However, it is desirable that you demonstrate that you have read and researched more widely than the required minimum. You should use other references to have a broad viewpoint of the subject and also to give you a deeper understanding of the subject.

When you have completed each assignment, send it, together with a TMA form, to your tutor. Make sure that each assignment reaches your tutor on or before the deadline given in the Presentation File. If for any reason, you cannot complete your work on time, contact your tutor before the assignment is due to discuss the possibility of an extension. Extensions will not be granted after the due date unless there are exceptional circumstances.

FINAL EXAMINATION AND GRADING

The final examination will be of three hours' duration and have a value of 70% of the total course grade. The examination will consist of questions which reflect the types of self-assessment practice exercises and tutor-marked problems you have previously encountered. All areas of the course will be assessed

Revise the entire course material using the time between finishing the last unit in the module and that of sitting for the final examination to. You might find it useful to review your self-assessment exercises, tutor-marked assignments and comments on them before the examination. The final examination covers information from all parts of the course.

COURSE MARKING SCHEME

The Table presented below indicates the total marks (100%) allocation.

Assignment	Marks
Assignments (Best three assignments out of four that is marked)	30%
Final Examination	70%
Total	100%

COURSE OVERVIEW

The Table presented below indicates the units, number of weeks and assignments to be taken by you to successfully complete the course, **Applied Quantitative Analysis (ECO 729)**.

Units	Title of Work	Week's Activities	Assessment (end of unit)
	Course Guide		
Module 1: STATISTICAL THEORY, DESCRIPTIVE STATISTICS AND PROBABILITY APPLICATIONS			
1	Statistical Theory and Inference	Week 1	Assignment 1
2	Overview of Descriptive Statistics	Week 2	Assignment 2
3	Probability Applications	Week 3	Assignment 3
Module 2: QUANTITATIVE TECHNIQUES AND LINEAR PROGRAMMING			
1	Overview of Quantitative Techniques	Week 4	Assignment 1
2	Linear Programming Graphical Method	Week 5	Assignment 2
3	Simplex Method	Week 6	Assignment 3
4	Transportation Model	Week 7	Assignment 4
Module 3: FORECASTING, DECISION AND INVENTORY ANALYSIS			
1	Forecasting and Decision Analysis	Week 8	Assignment 1
2	Demonstrate Forecasting Methods	Week 9	Assignment 2
3	Deterministic Inventory Control Models	Week 10	Assignment 3
Module 4: DATA ANALYSIS TECHNIQUES AND STATISTICAL SOFTWARE IN APPLIED QUANTITATIVE TECHNIQUES			
1	An Overview of Quantitative Research	Week 11	Assignment 1
2	Quantitative Data Concepts	Week 12	Assignment 2

3	Data Analysis Tools in Applied Quantitative Analysis	Week 13	Assignment 3
	Examination	Week 14 & 15	

Contents

Module 1: Statistical Theory, Descriptive Statistics and Probability Applications

Unit 1 Statistical Theory and Inference

Unit 2 Overview of Descriptive Statistics

Unit 3 Probability Applications

Module 2: Quantitative Techniques and Linear Programming

Unit 1 Overview of Quantitative Techniques

Unit 2 Linear Programming Graphical Method

Unit 3 Simplex Method

Unit 4 Transportation Model

Module 3: Forecasting, Decision and Inventory Analysis

Unit 1 Forecasting and Decision Analysis

Unit 2 Demonstrate Forecasting Methods

Unit 3 Deterministic Inventory Control Models

Module 4: Data Analysis Techniques and Statistical Software in Applied Quantitative Analysis

Unit 1 An Overview of Quantitative Research

Unit 2 Quantitative Data Concepts

Unit 3 Data Analysis Tools in Applied Quantitative Techniques

MODULE 1: STATISTICAL THEORY, DESCRIPTIVE STATISTICS AND PROBABILITY APPLICATIONS

Unit 1	Statistical Theory and Inference
Unit 2	Overview of Descriptive Statistics
Unit 3	Probability Applications

UNIT 1: STATISTICAL THEORY AND INFERENCE CONTENTS

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Importance of Statistical Theory
3.2	Meaning of Statistical Inference
3.3	Basic Statistical Tools
3.3.1	F-test for precision
3.3.2	Linear Correlation and Regression
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

The theory of statistics provides a basis for the whole range of techniques, in both study design and data analysis, that are used within applications of statistics. The unit covers approaches importance of statistical theory and statistical inference, and the actions and deductions that satisfy the basic principles stated for these different approaches. Within a given approach, statistical theory gives ways of comparing statistical procedures; it can find a best possible procedure within a given context for given statistical problems, or can provide guidance on the choice between alternative procedures. Apart from philosophical considerations about how to make statistical inferences and decisions, much of statistical theory consists of mathematical statistics, and is closely linked to probability theory, to utility theory, and to optimization.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Know the importance of statistical theory
- Understand the meaning of statistical inference
- Make use of the basic statistical tools

3.0 MAIN CONTENT

3.1 Importance of Statistical Theory Scope and Modelling

Statistical theory provides an underlying rationale and provides a consistent basis for the choice of methodology used in applied statistics. Statistical models describe the sources of data and can have different types of formulation corresponding to these sources and to the problem being studied. Such problems can be of various kinds:

- Sampling from a finite population
- Measuring observational error and refining procedures
- Studying statistical relations

Statistical models, once specified, can be tested to see whether they provide useful inferences for new data sets. Testing a hypothesis using the data that was used to specify the model is a fallacy, according to the natural science of Bacon and the scientific method of Peirce.

Data collection

Statistical theory provides a guide to comparing methods of data collection, where the problem is to generate informative data using optimization and randomization while measuring and controlling for observational error. Optimization of data collection reduces the cost of data while satisfying statistical goals, while randomization allows reliable inferences. Statistical theory provides a basis for good data collection and the structuring of investigations in the topics of:

- Design of experiments to estimate treatment effects, to test hypotheses, and to optimize responses
- Survey sampling to describe populations

Summarising data

The task of summarising statistical data in conventional forms (also known as descriptive statistics) is considered in theoretical statistics as a problem of defining what aspects of statistical samples need to be described and how well they can be described from a typically limited sample of data. Thus, the problems theoretical statistics considers include:

- Choosing summary statistics to describe a sample
- Summarising probability distributions of sample data while making limited assumptions about the form of distribution that may be met
- Summarising the relationships between different quantities measured on the same items with a sample

Interpreting data

Besides the philosophy underlying statistical inference, statistical theory has the task of considering the types of questions that data analysts might want to ask about the problems they are studying and of providing data analytic techniques for answering them. Some of these tasks are:

- Summarising populations in the form of a fitted distribution or probability density function

- Summarising the relationship between variables using some type of regression analysis
- Providing ways of predicting the outcome of a random quantity given other related variables
- Examining the possibility of reducing the number of variables being considered within a problem (the task of Dimension reduction)

Statistical theory provides the basis for a number of data analytic methods that are common across scientific and social research. Some of these are: Interpreting data is an important objective of statistical research:

- Estimating parameters
- Testing statistical hypotheses
- Providing a range of values instead of a point estimate
- Regression analysis

Many of the standard methods for these tasks rely on certain statistical assumptions (made in the derivation of the methodology) actually holding in practice. Statistical theory studies the consequences of departures from these assumptions. In addition it provides a range of robust statistical techniques that are less dependent on assumptions, and it provides methods checking whether particular assumptions are reasonable for a give data-set.

3.2 Meaning of Statistical Inference

Statistical inference is the process of deducing properties of an underlying distribution by analysis of data. Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates. The population is assumed to be larger than the observed data set; in other words, the observed data is assumed to be sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and does not assume that the data came from a larger population.

Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (firstly) selecting a statistical model of the process that generates the data and (secondly) deducing propositions from the model.

The conclusion of a **statistical inference** is a statistical proposition.

Some common forms of statistical proposition are the following:

- a point estimate, i.e. a particular value that best approximates some parameter of interest;
- an interval estimate, e.g. a confidence interval (or set estimate), i.e. an interval constructed using a dataset drawn from a population so that, under repeated

sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated confidence level;

- a credible interval, i.e. a set of values containing, for example, 95% of posterior belief;
- rejection of a hypothesis;
- clustering or classification of data points into groups

Models and assumptions

Any statistical inference requires some assumptions. A **statistical model** is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to draw inference. Descriptive statistics are typically used as a preliminary step before more formal inferences are drawn.

Degree of models/assumptions

Statisticians distinguish between three levels of modelling assumptions;

- **Fully parametric:** The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters. For example, one may assume that the distribution of population values is truly Normal, with unknown mean and variance, and that datasets are generated by 'simple' random sampling. The family of generalized linear models is a widely used and flexible class of parametric models.
- **Non-parametric:** The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal. For example, every continuous probability distribution has a median, which may be estimated using the sample median or the Hodges–Lehmann–Sen estimator, which has good properties when the data arise from simple random sampling.
- **Semi-parametric:** This term typically implies assumptions 'in between' fully and non-parametric approaches. For example, one may assume that a population distribution has a finite mean. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but not make any parametric assumption describing the variance around that mean (i.e. about the presence or possible form of any heteroscedasticity). More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components.

Types of Statistical Data: Numerical and Categorical

When working with statistics, it's important to recognize the different types of data: numerical (discrete and continuous), categorical. Data are the actual pieces of

information that you collect through your study. Most data fall into one of two groups: numerical or categorical.

Numerical data. These data have meaning as a measurement, such as a person's height, weight, IQ, or blood pressure; or they're a count, such as the number of stock shares a person owns, how many teeth a dog has, or how many pages you can read of your favorite book before you fall asleep. (Statisticians also call numerical data quantitative data.). Numerical data can be further broken into two types: discrete and continuous.

- **Discrete data** represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity (making it countably infinite). For example, the number of heads in 100-coin flips takes on values from 0 through 100 (finite case), Its possible values are listed as 100, 101, 102, 103 . . . (representing the countably infinite case).
- **Continuous data** represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line. For example, the exact amount of gas purchased at the pump for cars with 20-gallon tanks would be continuous data from 0 gallons to 20 gallons, represented by the interval $[0, 20]$, inclusive. You might pump 8.40 gallons, or 8.41, or 8.414863 gallons, or any possible number from 0 to 20. In this way, continuous data can be thought of as being uncountably infinite. For ease of recordkeeping, statisticians usually pick some point in the number to round off.

Categorical data: Categorical data represent characteristics such as a person's gender, marital status, hometown, or the types of movies they like. Categorical data can take on numerical values (such as "1" indicating male and "2" indicating female), but those numbers don't have mathematical meaning. You couldn't add them together, for example. (Other names for categorical data are qualitative data, or Yes/No data.). Categorical variables can equally be referred to as discrete or qualitative variables. They (categorical variables) can take on exactly two values known as dichotomous variable and polychotomous variable.

- **Ordinal data** mixes numerical and categorical data. The data fall into categories, but the numbers placed on the categories have meaning. For example, rating a restaurant on a scale from 0 (lowest) to 4 (highest) stars gives ordinal data. Ordinal data are often treated as categorical, where the groups are ordered when graphs and charts are made. However, unlike categorical data, the numbers do have mathematical meaning. For example, if you survey 100 people and ask them to rate a restaurant on a scale from 0 to 4, taking the average of the 100 responses will have meaning. This would not be the case with categorical data.

- **Nominal variables** are variables that have two or more categories (that is dichotomous or polychromous) but do not have any basic order. For example, an estate agent in Lagos could classify the building properties under his/her control into categories such as mini flat, 2-bedrooms, 3-bedrooms or duplex. This shows that, the type of building property is nominal in nature because it has four (4) categories. It is pertinent to note that, the different categories of a nominal variable can also be called levels of the nominal variable.
- **Interval variables** are similar to ordinal variables, however, their distinctive features can be measured along a range and have numerical value assigned to it. Time is a good example of an interval variable or scale with known incremental values. These interval values are consistent and measurable. Another good instance is the temperature gauge calibrated in degrees Celsius or Fahrenheit. The difference between the degrees Celsius (10°C and 20°C , 80°C and 90°C) is of constant interval. In this instance, 10°C is the constant interval.
- **Ratio**, it is a version of data measurement scale and it is the highest level of data measurement in research. This is because; it possesses the attributes of nominal, ordinal and interval variables/data. Variables such as weight, area, speed, velocity, and many more are sets of variables which no other scale is appropriate except the Ratio. Ratio has an absolute or natural zero ("0") which has realistic implication.

SELF ASSESSMENT EXERCISE

Briefly explain the following types of Statistical Data: Numerical and Categorical data

3.3 Basic Statistical Tools

A multitude of different statistical tools is available, some of them simple, some complicated, and often very specific for certain purposes. In analytical work, the most important common operation is the comparison of data, or sets of data, to quantify accuracy (bias) and precision. Fortunately, with a few simple convenient statistical tools most of the information needed in regular laboratory work can be obtained: the "t-test", the "F-test", and regression analysis.

Therefore, some understanding of these statistics is essential and they will briefly be discussed here. The basic assumption to be made is that a set of data, obtained by repeated analysis of the same analyze in the same sample under the same conditions, has a normal or Gaussian distribution. (When the distribution is skewed statistical treatment is more complicated). The primary parameters used are the mean (or average) and the standard deviation and the main tools the F-test, the t-test, and regression and correlation analysis.

Mean

The average of a set of n data x_i

$$\bar{x} = \frac{\sum x_i}{n} \quad (1)$$

Standard deviation

This is the most commonly used measure of the spread or dispersion of data around the mean. The standard deviation is defined as the square root of the *variance* (V). The variance is defined as the sum of the squared deviations from the mean, divided by $n-1$. Operationally, there are several ways of calculation:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad \text{or} \quad \sqrt{\frac{\sum x_i^2 (\sum (x_i)^2 / n)}{n-1}} \quad \text{or} \quad \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} \quad (2)$$

The calculation of the mean and the standard deviation can easily be done on a calculator but most conveniently on a *PC* with computer programs such as *dBASE*, *Lotus 123*, *Quattro-Pro*, *Excel*, and others, which have simple ready-to-use functions.

Relative standard deviation. Coefficient of variation

Although the standard deviation of analytical data may not vary much over limited ranges of such data, it usually depends on the magnitude of such data: the larger the figures, the larger s . Therefore, for comparison of variations (e.g. precision) it is often more convenient to use the relative standard deviation (RSD) than the standard deviation itself. The RSD is expressed as a fraction, but more usually as a percentage and is then called coefficient of variation (CV). Often, however, these terms are confused.

$$\text{RSD} = \frac{S}{x} \quad \text{CV} = \frac{S}{x} \times 100\% \quad (3)$$

Note. When needed (e.g. for the F -test,) the variance can, of course, be calculated by squaring the standard deviation: $V = s^2$

Confidence limits of a measurement

The more an analysis or measurement is replicated, the closer the mean \bar{x} of the results will approach the "true" value μ , of the analyze content (assuming absence of bias).

A single analysis of a test sample can be regarded as literally sampling the imaginary set of a multitude of results obtained for that test sample. The uncertainty of such subsampling is expressed by

$$\mu = \bar{x} \pm t \frac{s}{\sqrt{n}} \quad (4)$$

where

μ = "true" value (mean of large set of replicates)

\bar{x} = mean of subsamples

t = a statistical value which depends on the number of data and the required confidence (usually 95%).

s = standard deviation of mean of subsamples

n = number of subsamples

(The term s/\sqrt{n} is also known as the standard error of the mean.)

The critical values for t are tabulated in Appendix 1 (they are, therefore, here referred to as t_{tab}). To find the applicable value, the number of *degrees of freedom* has to be established by: $df = n - 1$

Example: For the determination of the clay content in the particle-size analysis, a semi-automatic pipette installation is used with a 20 mL pipette. This volume is approximate and the operation involves the opening and closing of taps. Therefore, the pipette has to be calibrated, i.e. both the accuracy (trueness) and precision have to be established.

A tenfold measurement of the volume yielded the following set of data (in mL):

19.941	19.812	19.829	19.828	19.742
19.797	19.937	19.847	19.885	19.804

The mean is 19.842 mL and the standard deviation 0.0627 mL. $n = 10$ is $t_{tab} = 2.26$ ($df = 9$) and using Eq. (4) this calibration yields:

Pipette volume = $19.842 \pm 2.26 (0.0627/\sqrt{10}) = 19.84 \pm 0.04$ mL

(Note that the pipette has a systematic deviation from 20 mL as this is outside the found Equation (4) is then reduced to

$$\mu = x \pm t \cdot s \tag{5}$$

where

μ = "true" value

x = single measurement

t = applicable t_{tab}

s = standard deviation of set of previous measurements.

if the set of replicated measurements is large (say > 30), t is close to 2. Therefore, the (95%) confidence of the result x of a single test sample ($n = 1$ in Eq. 4) is approximated by the commonly used and well-known expression $\mu = x \pm 2s$

where S is the previously determined standard deviation of the large set of replicates. *Note:* This "method-s" or s of a control sample is not a constant and may vary for different test materials, analyze levels, and with analytical conditions. Running duplicates

will, according to Equation (4), increase the confidence of the (mean) result by a factor: $\sqrt{2}$

$$\mu = \bar{x} \pm 2 \frac{s}{\sqrt{2}} \quad (6)$$

where

\bar{x} = mean of duplicates

s = known standard deviation of large set

Similarly, triplicate analysis will increase the confidence by a factor $\sqrt{3}$, etc.

3.4 Statistical tests

Some of the most common and convenient statistical tools to quantify such comparisons are the F -test, the t -tests, and regression analysis.

Because the F -test and the t -tests are the most basic tests they will be discussed first. These tests examine if two sets of normally distributed data are similar or dissimilar (belong or not belong to the same "population") by comparing their *standard deviations* and *means* respectively. This is illustrated in Fig. 1.

Fig. 1 shows the three possible cases when comparing two sets of data ($n_1 = n_2$). A. Different mean (bias), same precision; B. Same mean (no bias), different precision; C. Both mean and precision are different. (The fourth case, identical sets, has not been drawn).

These tests for comparison, for instance between methods A and B , are based on the assumption that there is no significant difference (the "null hypothesis"). In other words, when the difference is so small that a tabulated *critical value* of F or t is not exceeded, we can be confident (usually at 95% level) that A and B are not different. Two fundamentally different questions can be asked concerning both the comparison of the standard deviations s_1 and s_2 with the F -test, and of the means \bar{x}_1 and \bar{x}_2 , with the t -test:

1. Are A and B different? (*two-sided* test)
2. Is A higher (or lower) than B ? (*One-sided* test).

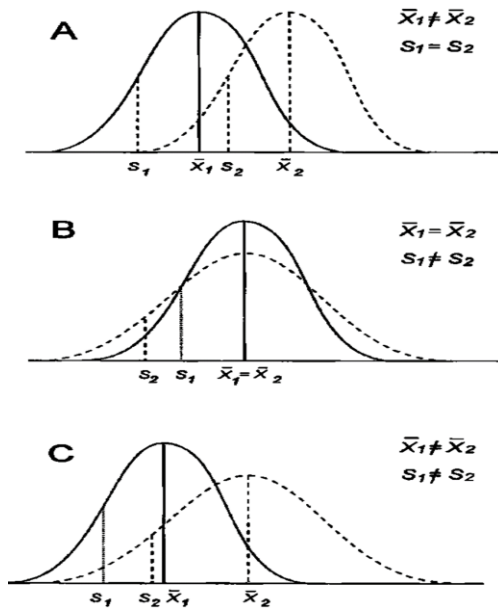


Figure 1: Two-sided vs. one-sided test

This distinction has an important practical implication as statistically the probabilities for the two situations are different: the chance that A and B are only different ("it can go two ways") is twice as large as the chance that A is higher (or lower) than B ("it can go only one way"). The most common case is the two-sided (also called *two-tailed*) test: there are no particular reasons to expect that the means or the standard deviations of two data sets are different. This difference in probability in the tests is expressed in the use of two tables of critical values for both F and t . In fact, the one-sided table at 95% confidence level is equivalent to the two-sided table at 90% confidence level.

3.3.1 F-test for precision

Because the result of the F -test may be needed to choose between the Student's t -test and the Cochran variant (see next section), the F -test is discussed first.

The F -test (or *Fisher's test*) is a comparison of the spread of two sets of data to test if the sets belong to the same population, in other words if the precisions are similar or dissimilar.

The test makes use of the ratio of the two variances:

$$F = \frac{s_1^2}{s_2^2} \quad (7)$$

where the larger s^2 must be the numerator by convention. If the performances are not very different, then the estimates s_1 , and s_2 , do not differ much and their ratio (and that of their squares) should not deviate much from unity. In practice, the calculated F is compared

with the applicable F value in the F-table. To read the table it is necessary to know the applicable number of degrees of freedom for s_1 , and s_2 . These are calculated by:

$$df_1 = n_1 - 1$$

$$df_2 = n_2 - 1$$

If F_{cal} and F_{tab} one can conclude with 95% confidence that there is no significant difference in precision (the "null hypothesis" that $s_1 = s_2$ is accepted). Thus, there is still a 5% chance that we draw the wrong conclusion. In certain cases more confidence may be needed, then a 99% confidence table can be used, which can be found in statistical textbooks.

Example I (two-sided test)

Table 1 gives the data sets obtained by two analysts for the cation exchange capacity (CEC) of a control sample. Using Equation (7) the calculated F value is 1.62. As we had no particular reason to expect that the analysts would perform differently, we use the F-table for the *two-sided* test and find $F_{tab} = 4.03$ (Appendix 2, $df_1 = df_2 = 9$). This exceeds the calculated value and the null hypothesis (no difference) is accepted. It can be concluded with 95% confidence that there is no significant difference in precision between the work of Analyst 1 and 2.

Table 1a. CEC values (in cmol_c/kg) of a control sample determined by two analysts.

1	2
10.2	9.7
10.7	9.0
10.5	10.2
9.9	10.3
9.0	10.8
11.2	11.1
11.5	9.4
10.9	9.2
8.9	9.8
10.6	10.2

Table 1b.

\bar{x} :	10.34	9.97
s :	0.819	0.644
n :	10	10
$F_{cal} = 1.62$	$t_{cal} = 1.12$	
$F_{tab} = 4.03$	$t_{tab} = 2.10$	

3.3.2 Linear Correlation and Regression

These also belong to the most common useful statistical tools to compare effects and performances X and Y . Although the technique is in principle the same for both, there is a fundamental difference in concept: *correlation analysis* is applied to independent factors: if X increases, what will Y do (increase, decrease, or perhaps not change at all)? In *regression analysis* a unilateral response is assumed: changes in X result in changes in Y , but changes in Y do not result in changes in X .

For example, in analytical work, correlation analysis can be used for comparing methods or laboratories, whereas regression analysis can be used to construct calibration graphs. In practice, however, comparison of laboratories or methods is usually also done by regression analysis.

The principle is to establish a statistical linear relationship between two sets of corresponding data by fitting the data to a straight line by means of the "least squares" technique. Such data are, for example, analytical results of two methods applied to the same samples (correlation), or the response of an instrument to a series of standard solutions (regression). *Note:* Naturally, non-linear higher-order relationships are also possible, but since these are less common in analytical work and more complex to handle mathematically, they will not be discussed here.

The resulting line takes the general form:

$$y = a + bx \tag{8}$$

Where a = intercept of the line with the y-axis

b = slope (tangent)

In laboratory work ideally, when there is perfect positive correlation without bias, the intercept $a = 0$ and the slope = 1. This is the so-called "1:1 line" passing through the origin (dashed line in Fig. 2).

If the intercept $a \neq 0$ then there is a systematic discrepancy (bias, error) between X and Y ; when $b \neq 1$ then there is a proportional response or difference between X and Y .

The correlation between X and Y is expressed by the correlation coefficient r which can be calculated with the following equation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{9}$$

where

x_i = data X

\bar{x} = mean of data X

y_i = data Y

\bar{y} = mean of data Y

It can be shown that r can vary from 1 to -1:

$r = 1$ perfect positive linear correlation

$r = 0$ no linear correlation (maybe other correlation)

$r = -1$ perfect negative linear correlation

Often, the correlation coefficient r is expressed as r^2 : the *coefficient of determination* or *coefficient of variance*. The advantage of r^2 is that, when multiplied by 100, it indicates the percentage of variation in Y associated with variation in X . Thus, for example, when $r = 0.71$ about 50% ($r^2 = 0.504$) of the variation in Y is due to the variation in X .

The line parameters b and a are calculated with the following equations:

$$b = \frac{\sum (x_1 - \bar{x})(y_1 - \bar{y})}{\sum (x_1 - \bar{x})^2} \quad (10)$$

$$a = \bar{y} - b\bar{x} \quad (11)$$

It is worth to note that r is independent of the choice which factor is the independent factory and which is the dependent Y . However, the regression parameters a and do depend on this choice as the regression lines will be different (except when there is ideal 1:1 correlation).

Construction of calibration graph

As an example, we take a standard series of P (0-1.0 mg/L) for the spectrophotometric determination of phosphate in a Bray-I extract ("available P"), reading in absorbance units. The data and calculated terms needed to determine the parameters of the calibration graph are given in Table 2. The line itself is plotted in Fig 2.

Table 2 is presented here to give an insight in the steps and terms involved. The calculation of the correlation coefficient r with Equation (9) yields a value of 0.997 ($r^2 = 0.995$). Such high values are common for calibration graphs. When the value is not close to 1 (say, below 0.98) this must be taken as a warning and it might then be advisable to repeat or review the procedure. Errors may have been made (e.g. in pipetting) or the used range of the graph may not be linear. On the other hand, a high r may be misleading as it does not necessarily indicate linearity. Therefore, to verify this, the calibration graph should always be plotted, either on paper or on computer monitor.

Using Equations (10 and 11) we obtain:

$$b = \frac{0.438}{0.70} = 0.626 \quad \& \quad a = 0.350 - 0.313 = 0.037$$

Thus, the equation of the calibration line is:

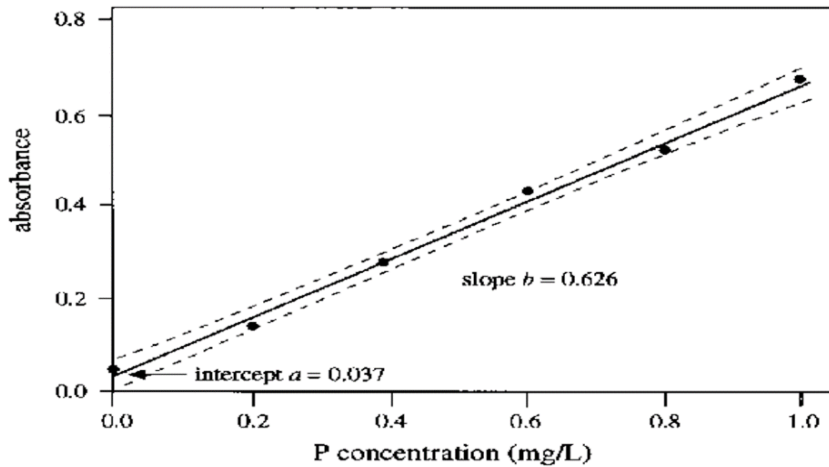
$$y = 0.626x + 0.037$$

(12)

Table 2. Parameters of calibration graph in Fig. 2

x_1	y_1	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$	$y_1 - \bar{y}$	$(y_1 - \bar{y})^2$	$(x_1 - \bar{x})(y_1 - \bar{y})$
0.0	0.05	-0.5	0.25	-0.30	0.090	0.15
0.2	0.14	-0.3	0.09	-0.21	0.044	0.063
0.4	0.29	-0.1	0.01	-0.06	0.004	0.006
0.6	0.43	0.1	0.01	0.08	0.006	0.008
0.8	0.52	0.3	0.09	0.17	0.029	0.051
1.0	0.67	0.5	0.25	0.32	0.102	0.160
3.0	2.10	0	0.70	0	0.2754	0.438 S
$\bar{x} = 0.5$	$\bar{y} = 0.35$					

Fig. 2. Calibration graph plotted from data of Table 2. The dashed lines delineate the 95% confidence area of the graph. Note that the confidence is highest at the centroid of the graph.



During calculation, the maximum number of decimals is used, rounding off to the last significant figure is done at the end.

Once the calibration graph is established, its use is simple: for each y value measured the corresponding concentration x can be determined either by direct reading or by calculation using Equation (12).

4.0 CONCLUSION

Statistical theory provides well-defined probability statements for the method when applied to all populations that could have arisen from the randomization used to generate the data. This provides an objective way of estimating parameters, estimating confidence intervals, testing hypotheses, and selecting the best. Even for observational data, statistical theory provides a way of calculating a value that can be used to interpret a

sample of data from a population, it can provide a means of indicating how well that value is determined by the sample, and thus a means of saying corresponding values derived for different populations are as different as they might seem; however, the reliability of inferences from post-hoc observational data is often worse than for planned randomized generation of data.

5.0 SUMMARY

In this unit, we have discussed the importance of statistical theory and the meaning of statistical inference. We also discussed some basic definitions of statistical terms like fully parametric, non-parametric, semi-parametric, numerical data, categorical data and ordinal data. Also, we explained some basic statistical tools such as f-test for precision linear correlation and regression. The value of statistics lies with organizing and simplifying data, to permit some objective estimate showing that an analysis is under control or that a change has occurred. Equally important is that the results of these statistical procedures are recorded and can be retrieved.

6.0 TUTORED MARKED ASSIGNMENTS

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams. Thus X = gestational age, Y = birth weight. The data summations are summarized below.

The gestational age data as $\sum X = 652.1$, $\sum(X - \bar{X}) = 0$, $\sum(X - \bar{X})^2 = 159.45$

The birth weight data as $\sum Y = 49,334$, $\sum(Y - \bar{Y}) = 0$, $\sum(Y - \bar{Y})^2 = 7,767.660$,

Where $n = 17$ and $\sum(X - \bar{X})(Y - \bar{Y}) = 28,768.4$

- i. Compute the variance of gestational age
- ii. Compute the variance of birth weight
- iii. Compute the covariance of gestational age and birth weight
- iv. Compute the sample correlation coefficient

7.0 REFERENCES/FURTHER READINGS

Abramson, J. H., & Abramson, Z. H. (2008). Scales of Measurement. Research Methods in Community Medicine: Surveys, Epidemiological Research, Programme Evaluation, Clinical Trials, Sixth Edition, 125-132.

Haessuler, E. F. and Paul, R. S. (1976). Introductory Mathematical Analysis for Students of Business and Economics, (2nd edition.) Reston Virginia: Reston Publishing Company.

Muaz, Jalil Mohammad (2013). Practical Guidelines for conducting research. Summarizing good research practice in line with the DCED Standard

UNIT 2: OVERVIEW OF DESCRIPTIVE STATISTICS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Univariate methods for categorical variables
 - 3.2 Bivariate methods for cases where both variables are categorical
 - 3.3 The meaning of descriptive statistics
 - 3.4 Measure of central tendency
 - 3.5 Measure of central dispersion
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

This unit introduces some common descriptive statistical methods. It is organized around two dichotomies: Methods that are used only for variables with small numbers of values, vs. methods that are used also or only for variables with many values. The former includes, in particular, descriptive methods for categorical variables, and the latter the methods for continuous variables. And Univariate descriptive methods which consider only one variable at a time, vs. bivariate methods which aim to describe the association between two variables.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Understand the univariate methods for categorical variables
- Know bivariate methods for cases where both variables are categorical
- Define the meaning of descriptive statistics
- Explain the measure of central tendency
- Know the measure of central dispersion

3.0 MAIN CONTENT

3.1 Single Categorical Variable

3.1.1 Describing the sample distribution

The term distribution is very important in statistics. In this section we consider the distribution of a single variable in the observed data, i.e. its sample distribution. The **sample distribution** of a variable consists of a list of the values of the variable which occur in a sample, together with the number of times each value occurs.

Sex	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	Total

3.1.2 Graphical methods: Bar charts

Graphical methods of describing data (statistical graphics) make use of our ability to process and interpret even very large amounts of visual information. The basic graph for summarizing the sample distribution of a discrete variable is a bar chart. It is the graphical equivalent of a one-way table of frequencies.

Figure 1 show the bar charts for region. Each bar corresponds to one category of the variable, and the height of the bar is proportional to the frequency of observations in that category. This visual cue allows us to make quick comparisons between the frequencies of different categories by comparing the heights of the bars.

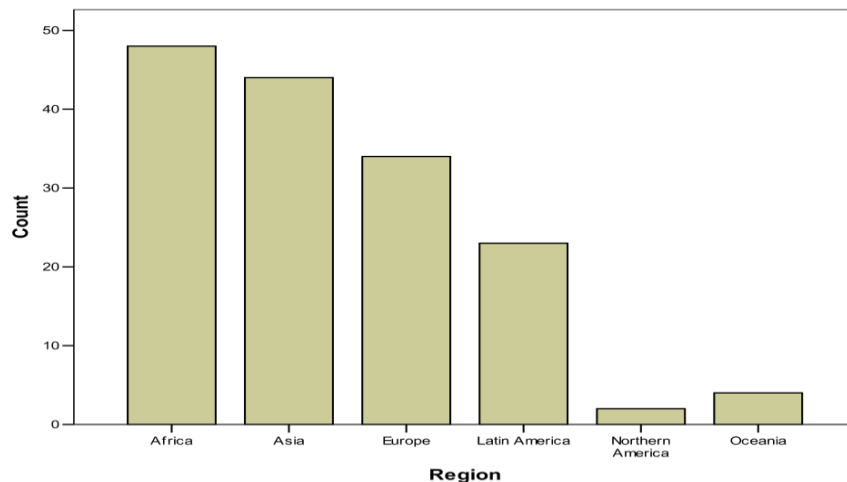


Figure 1: Bar chart of regions in the country data.

3.2 Two Categorical Variables

3.2.1 Two-Way Contingency Tables

The key tool is a table which shows the cross tabulation of the frequencies of the variables. This is also known as a contingency table. Table 3.1 shows such a table for the respondents' sex and attitude in our example below. Table 1 shows a two-way table of frequencies of respondents in the survey example, by sex and attitude towards income redistribution. We use it to introduce the basic structure and terminology of contingency tables:

Table 1: Data: Nigerian Social Survey, Round 5, 2015, Nigerian respondents only.

Male	160	439	187	200	41	1027
Female	206	651	239	187	34	1317
Total	366	1090	426	387	75	2344

Because a table like 3.1 summarizes the values of two variables, it is known as a two-way contingency table. It is also possible to construct tables involving more than two variables, i.e. three-way tables, four-way tables, and so on

3.3 The Meaning of Descriptive Statistics

Imagine that you are interested in measuring the level of anxiety of college students during finals week in one of your courses. You have 11 study participants rate their level of anxiety on a scale from 1 to 10, with 1 being 'no anxiety' and 10 being 'extremely anxious.' You collect the ratings and review them. The ratings are 8, 4, 9, 3, 5, 8, 6, 6, 7, 8, and 10. Your teacher asks you for a summary of your findings. How do you summarize this data? One way we could do this is by using descriptive statistics.

Descriptive statistics are used to describe or summarize data in ways that are meaningful and useful. For example, it would not be useful to know that all of the participants in our example wore blue shoes. However, it would be useful to know how spreads out their anxiety ratings were. Descriptive statistics is at the heart of all quantitative analysis. So how do we describe data? There are two ways: measures of central tendency and measures of variability, or dispersion.

This unit is concerned with two numerical ways of describing data, namely, measures of central tendency and measures of dispersion. Measures of location are often referred to as averages. The purpose of a measure of location is to pinpoint the center of a set of values.

3.4 Measures of Central Tendency

You are probably somewhat familiar with the mean, but did you know that it is a measure of central tendency? Measures of central tendency use a single value to describe the center of a data set. The mean, median, and mode are all the three measures of central tendency.

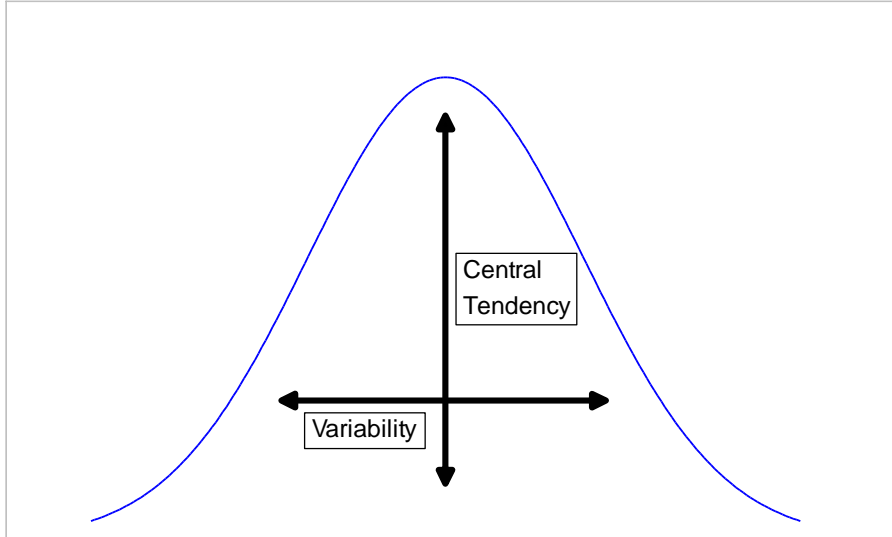


Fig 2: Central tendency describes the central point in a data set. Variability describes the spread of the data.

Mean

The mean, or average, is calculated by finding the sum of the study data and dividing it by the total number of data. The mode is the number that appears most frequently in the set of data.

Examples of a population mean would be: the mean closing price for Johnson and Johnson stock for the last 5 days is \$55.25; the mean annual rate of return for the last 10 years for Berger Funds is 5.21 percent; and the mean number of hours of overtime worked last week by the six welders in the welding department of Butts Welding, Inc., is 6.45 hours.

For raw data, that is, data that has not been grouped in a frequency distribution, the population mean is the sum of all the values in the population divided by the number of values in the population. To find the population mean, we use the following formula. This is given as:

$$\text{PopulationMean} = \frac{\text{Sum of all the values in the population}}{\text{Number of values in the Population}}$$

Instead of writing out in words the full directions for computing the population mean (or any other measure), it is more convenient to use the shorthand symbols of mathematics. The mean of a population using mathematical symbols is:

$$\text{PopulationMean} : \mu = \frac{\sum X}{N} \tag{1}$$

Where

μ = Population Mean, the Greek lowercase letter “mu.”

N = Number of values in the population

X = any particular value

\sum = the Greek capital letter "sigma" and indicates the operation of adding.

Example 1

There are 12 Yoghurt manufacturing companies in Nigeria. Listed below is the number of patents granted by the Nigerian government to each company in a recent year.

Table 2: Patent Distribution to Yoghurt Companies in Nigeria

Company	Number of Patents Granted
A	511
B	385
C	275
D	257
E	249
F	234
G	210
H	97
I	50
J	36
K	23
L	13

Question: Is this information a sample or a population? What is the arithmetic mean number of patents granted?

Solution

This is a population because we are considering all the Yoghurt companies obtaining patents. We add the number of patents for each of the 12 companies. The total number of patents for the 12 companies is 2,340. To find the arithmetic mean, we divide this total by 12. Therefore, the arithmetic mean is 195, found by $2340/12$. From formula;

$$\mu = \frac{511 + 385 + \dots + 13}{12} = \frac{2340}{12} = 195$$

How do we interpret the value of 195? The typical number of patents received by Yoghurt company is 195. Because we considered all the companies receiving patents, this value is a population parameter.

Properties of the Arithmetic Mean

The arithmetic mean is a widely used measure of location. It has several important properties:

1. Every set of interval- or ratio-level data has a mean.
2. All the values are included in computing the mean.
3. A set of data has only one mean. The mean is unique.

4. The *sum of the deviations of each value from the mean will always be zero.*
Expressed symbolically: $\sum(X - \bar{X}) = 0$

Weighted Mean

The weighted mean is a special case of the arithmetic mean. It occurs when there are several observations of the same value. To explain, suppose the nearby Wendy's Restaurant sold medium, large, and Biggie-sized soft drinks for \$.90, \$1.25, and \$1.50, respectively. Of the last 10 drinks sold, 3 were medium, 4 were large, and 3 were Biggie-sized. To find the mean price of the last 10 drinks sold, we could use formula:

$$\bar{X} = \frac{0.9+0.9+0.9+1.25+1.25+1.25+1.25+1.50+1.50+1.50}{10}$$

$$\bar{X} = \frac{12.20}{10} = 1.22$$

The mean selling price of the last 10 drinks is \$1.22. An easier way to find the mean selling price is to determine the weighted mean. That is, we multiply each observation by the number of times it happens. We will refer to the weighted mean as \bar{X}_w . This is read "X bar sub w."

$$\bar{X}_w = \frac{3(0.90) + 4(1.25) + 3(1.50)}{10} = \frac{12.20}{10} = 1.22$$

In this case the weights are frequency counts. However, any measure of importance could be used as a weight. In general the weighted mean of a set of numbers designated $X_1, X_2, X_3, \dots, X_n$ with the corresponding weights $W_1, W_2, W_3, \dots, W_n$ is computed by:

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \dots + w_nX_n}{w_1 + w_2 + w_3 + \dots + w_n}$$

$$\bar{X}_w = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \quad (2)$$

Geometric Mean

The geometric mean is useful in finding the average of percentages, ratios, indexes, or growth rates. It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the Gross Domestic Product, which compound or build on each other. The geometric mean of a set of n positive numbers is defined as the n th root of the product of n values. The formula for the geometric mean is written:

$$\text{Geometric Mean: } GM = \sqrt[n]{(X_1)(X_2)\dots(X_n)} \quad (3)$$

The geometric mean will always be less than or equal to (never more than) the arithmetic mean. Also, all the data values must be positive.

As an example of the geometric mean, suppose you receive a 5 percent increase in salary this year and a 15 percent increase next year. The average annual percent increase is 9.886, not 10.0. Why is this so? We begin by calculating the geometric mean. Recall, for example, that a 5 percent increase in salary is 105 percent. We will write it as 1.05.

$$\text{Geometric Mean: } GM = \sqrt{(1.05)(1.15)} = 1.09886$$

Median

The **median** is the middle value in a set of data. It is calculated by first listing the data in numerical order then locating the value in the middle of the list. When working with an **odd** set of data, the median is the middle number. For example, the median in a set of 9 data is the number in the fifth place. When working with an **even** set of data, you find the average of the two middle numbers. For example, in a data set of 10, you would find the average of the numbers in the fifth and sixth places.

Example 2

Suppose you are seeking to buy a condominium in Palm Aire. Your real estate agent says that the average price of the units currently available is \$110,000. Would you still want to look? If you had budgeted your maximum purchase price at \$75,000, you might think they are out of your price range. However, checking the individual prices of the units might change your mind. They are \$60,000, \$65,000, \$70,000, \$80,000, and a super deluxe penthouse costs \$275,000. The arithmetic mean price is \$110,000, as the real estate agent reported, but one price (\$275,000) is pulling the arithmetic mean upward, causing it to be an unrepresentative average. It does seem that a price around \$70,000 is a more typical or representative average, and it is. In cases such as this, the median provides a more valid measure of location.

The data must be at least ordinal level of measurement. The median price of the units available is \$70,000. To determine this, we ordered the prices from low (\$60,000) to high (\$275,000) and selected the middle value (\$70,000).

Table 3: Housing Prices

Prices Ordered from Low to High (\$)	Prices Ordered from High to Low (\$)
60,000	275,000
65,000	80,000
70,000 ←	70,000 →
80,000	65,000
275,000	60,000

Note that there is the same number of prices below the median of \$70,000 as above it. There are as many values below the median as above. The median is, therefore, unaffected by extremely low or high prices. Had the highest price been \$90,000, or \$300,000, or even \$1 million, the median price would still be \$70,000. Likewise, had the lowest price been \$20,000 or \$50,000, the median price would still be \$70,000.

Mode

The mean and median can only be used with numerical data. The **mode** can be used with both numerical and **nominal** data, and data in the form of names or labels. Eye color, gender, and hair color are all examples of nominal data. The mean is the preferred measure of central tendency since it considers all of the numbers in a data set; however, the mean is extremely sensitive to **outliers**, or extreme values that are much higher or lower than the rest of the values in a data set. The median is preferred in cases where there are outliers, since the median only considers the middle values. The mode is especially useful in describing nominal and ordinal levels of measurement.

As an example of its use for nominal-level data, a company has developed five bath oils. Figure 1 shows the results of a marketing survey designed to find which bath oil consumers prefer. The largest number of respondents favored Lamoure, as evidenced by the highest bar. Thus, Lamoure is the mode.

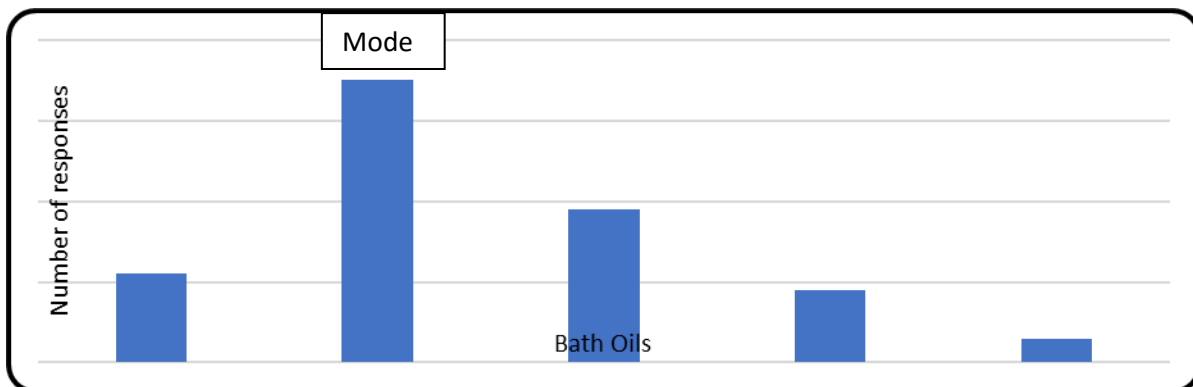


Figure 3: Number of Respondents Favoring Various Bath Oils

Example 3

Knowing what we know, let's calculate the mean, median, and mode using the example from before. Again, the anxiety ratings of your classmates are 8, 4, 9, 3, 5, 8, 6, 6, 7, 8, and 10.

Mean: $(8 + 4 + 9 + 3 + 5 + 8 + 6 + 6 + 7 + 8 + 10) / 11 = 74 / 11 =$ the mean is 6.73.

Median: In a data set of 11, the median is the number in the sixth place. 3, 4, 5, 6, 6, 7, 8, 8, 8, 9, 10. The median is 7.

Mode: The number 8 appears more than any other number. The mode is 8.

3.5 Measures of Dispersion

We've got some pretty solid numbers on our data now, but let's say that you wanted to look at how spread out the study data are from a central value, i.e. the mean. In this case, you would look at measures of dispersion, which include the range, variance, and standard deviation.

The simplest measure of dispersion is the **range**. This tells us how spread out our data is. In order to calculate the range, you subtract the smallest number from the largest (or highest) number. Just like the mean, the range is very sensitive to **outliers**.

Relevance of Studying Dispersion

A measure of location, such as the mean or the median, only describes the center of the data. It is valuable from that standpoint, but it does not tell us anything about the spread of the data. For example, if your nature guide told you that the river ahead averaged 3 feet in depth, would you want to wade across on foot without additional information? Probably not. You would want to know something about the variation in the depth. Is the maximum depth of the river 3.25 feet and the minimum 2.75 feet? If that is the case, you would probably agree to cross. What if you learned the river depth ranged from 0.50 feet to 5.5 feet? Your decision would probably be not to cross. Before making a decision about crossing the river, you want information on both the typical depth and the dispersion in the depth of the river.

A small value for a measure of dispersion indicates that the data are clustered closely, say, around the arithmetic mean. The mean is therefore considered representative of the data. Conversely, a large measure of dispersion indicates that the mean is not reliable.

Range

The simplest measure of dispersion is the range. It is the difference between the largest and the smallest values in a data set. In the form of an equation:

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

Mean Deviation

A defect of the range is that it is based on only two values, the highest and the lowest; it does not take into consideration all of the values. The mean deviation does. It measures the mean amount by which the values in a population, or sample, vary from their mean. In terms of a definition, mean deviation is the arithmetic mean of the absolute values of the deviations from the arithmetic mean.

$$\text{Mean Deviation: } MD = \frac{\sum |X - \bar{X}|}{n}$$

Where

X is the value of each observation

\bar{X} Is the arithmetic mean of the values

n is the number of observations in the sample
 $|$ indicates the absolute value

Why do we ignore the signs of the deviations from the mean? If we didn't, the positive and negative deviations from the mean would exactly offset each other, and the mean deviation would always be zero. Such a measure (zero) would be a useless statistic.

Example 4

The number of cappuccinos sold at the Starbucks location in the Orange County Airport between 4 P.M. and 7 P.M. for a sample of 5 days last year were: 103, 97, 101, 106, and 103. Determine the mean deviation and interpret.

The mean deviation is the mean of the amounts that individual observations differ from the arithmetic mean. To find the mean deviation of a set of data, we begin by finding the arithmetic mean. The mean number of cappuccinos sold is 102, found by $(103 + 97 + 101 + 106 + 103)/5$. Next we find the amount by which each observation differs from the mean. Then we sum these differences, ignoring the signs, and divide the sum by the number of observations. The result is the mean amount the observations differ from the mean. A small value for the mean deviation indicates that the data are clustered near the mean, whereas a large value for the mean deviation indicates greater dispersion in the data. Here are the details of the calculations using formula

Number of Cappuccinos Sold	$(X - \bar{X})$	Absolute Deviation
103	$(103 - 102) = 1$	1
97	$(97 - 102) = -5$	5
101	$(101 - 102) = -1$	1
106	$(106 - 102) = 4$	4
103	$(103 - 102) = 1$	1
		Total = 12

$$\text{Mean Deviation : MD} = \frac{\sum |X - \bar{X}|}{n} = \frac{12}{5} = 2.4$$

The mean deviation is 2.4 cappuccinos per day. The number of cappuccinos deviates, on average, by 2.4 cappuccinos from the mean of 102 cappuccinos per day.

Variance

The **variance** is a measure of the average distance that a set of data lies from its mean. The variance is not a stand-alone statistic. It is typically used in order to calculate other statistics, such as the standard deviation.

$$\text{Variance: } \sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

The higher the variance, the more spread out your data are.

There are four steps to calculate the variance:

- i. Calculate the mean
- ii. Subtract mean from each data value. This tells you how far each lies from the mean
- iii. Square each of the values so that you now have positive values, then find the sum of the squares
- iv. Divide the sum of the squares by the total number of data in the set

Example 5

The number of traffic citations issued during the last five months in Beaufort South Carolina, is: 38, 26, 13, 41, and 22. What is the population variance? We consider these data a population because the last five months' citations are all the values possible over the period. The details of the calculations follow:

Number (X)	$(X - \mu)$	$(X - \mu)^2$
38	10	100
26	-2	4
13	-15	225
41	13	169
22	-6	36
140		Total = 534

$$\mu = \frac{\sum X}{N} = \frac{140}{5} = 28$$

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N} = \frac{534}{5} = 106.8$$

Standard Deviation

The **standard deviation** is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range. The standard deviation is calculated by taking the square root of the variance.

$$\text{Standard Deviation: } \sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

Following example 5, the standard deviation is:

$$\sigma = \sqrt{\frac{534}{5}}$$
$$=10.3$$

SELF ASSESSMENT EXERCISE

Using the above data set, you are required to calculate the range, variance and standard deviation and interpret your result.

4.0 CONCLUSION

So far, this unit concludes that descriptive statistics are used to describe or summarize data in ways that are meaningful and useful.

5.0 SUMMARY

In this unit, we have discussed univariate methods for categorical variables, bivariate methods for cases where both variables are categorical, meaning of descriptive statistics, measure of central tendency and measure of central dispersion.

6.0 TUTORED MARKED ASSIGNMENTS

The number of ATM transactions per day were recorded at 16 locations in a large city. The data were: 13, 35, 49, 225, 50, 30, 65, 40, 55, 52, 75, 48, 324, 47, 32, and 60. Find (a) the median number of transactions and (b) the mean number of transactions. Hence obtain the variance and standard deviation of the data set.

7.0 REFERENCES/FURTHER READINGS

Lind, D.A., Marchal, W.G., & Wathen, S.A. (2006). *Basic Statistics for Business & Economics* (5th ed.). New York, USA: McGraw-Hill

Schaum's Outline of Statistics, Theory and Problems of Statistics, 4th Edition

UNIT 3: PROBABILITY APPLICATIONS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Probability Concepts
 - 3.2 Probability Formulas and Rules
 - 3.3 Solving Probability Problems
 - 3.4 Probability Distribution
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

The unit introduces the important topic of probability. The word probability has several meanings in ordinary conversation. Two of these are particularly important for the development and applications of the mathematical theory of probability. One is the interpretation of probabilities as relative frequencies, for which simple games involving coins, cards, dice, and roulette wheels provide examples. The distinctive feature of games of chance is that the outcome of a given trial cannot be predicted with certainty, although the collective results of a large number of trials display some regularity. For example, the statement that the probability of “heads” in tossing a coin equals one-half, according to the relative frequency interpretation, implies that in a large number of tosses the relative frequency with which “heads” actually occurs will be approximately one-half, although it contains no implication concerning the outcome of any given toss.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Define Probability concepts
- Distinguish between the different Probability Formulas
- Define and explain operational rules of Probability
- Explain the difference Probability Distribution
- Solve Questions Using Probability Formulas

3.0 MAIN CONTENT

3.1 Probability Concepts

Probability is a numerical description of how likely an event is to occur or how likely it is that a proposition is true. Probability is a number between 0 and 1, where, roughly

speaking, 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur.

These concepts have been given an axiomatic mathematical formalization in probability theory, which is used widely in such areas of study as mathematics, statistics, finance, gambling, science (in particular physics), artificial intelligence/machine learning, computer science, game theory, and philosophy to, for example, draw inferences about the expected frequency of events. Probability theory is also used to describe the underlying mechanics and regularities of complex systems.

A simple example is the tossing of a fair (unbiased) coin. Since the coin is fair, the two outcomes ("heads" and "tails") are both equally probable; the probability of "heads" equals the probability of "tails"; and since no other outcomes are possible, the probability of either "heads" or "tails" is $1/2$ (which could also be written as 0.5 or 50%).

Although it is not possible to perfectly predict random events, much can be said about their behaviour. Two major results in probability theory describing such behaviour are the law of large numbers and the central limit theorem. Central subjects in probability theory include discrete and continuous random variables, probability distributions, and stochastic processes, which provide mathematical abstractions of non-deterministic or uncertain processes or measured quantities that may either be single occurrences or evolve over time in a random fashion.

Sets and Probability

As we learned in the previous lesson, probability is all about statistical experiments. When a researcher conducts a statistical experiment, he or she cannot know the outcome in advance. The outcome is determined by chance. However, if the researcher can list all the possible outcomes of the experiment, it may be possible to compute the probability of a particular outcome. The list of all possible outcomes from a statistical experiment is called the **sample space**. And a particular outcome or collection of outcomes is called an **event**.

You can see that a sample space is a type of set. It is a well-defined listing of all possible outcomes from a statistical experiment. And an event in a statistical experiment is a subset of the sample space.

- A **set** is a well-defined collection of objects.
- Each object in a set is called an **element** of the set.
- Two sets are **equal** if they have exactly the same elements in them.
- A set that contains no elements is called a **null set** or an **empty** set.
- If every element in Set A is also in Set B , then Set A is a **subset** of Set B .

Statistical Experiment

The term "statistical experiment" is used to describe any process by which several chance observations are obtained. All possible outcomes of an experiment comprise a set that is called the **sample space**.

Statistical experiments have three things in common:

- The experiment can have more than one possible outcome.
- Each possible outcome can be specified in advance.
- The outcome of the experiment depends on chance.

A coin toss has all the attributes of a statistical experiment. There is more than one possible outcome. We can specify each possible outcome (i.e., heads or tails) in advance. And there is an element of chance, since the outcome is uncertain.

The Sample Space

- A **sample space** is a set of elements that represents all possible outcomes of a statistical experiment.
- A **sample point** is an element of a sample space.
- An **event** is a subset of a sample space - one or more sample points

Set Notation

- An element of a set is usually denoted by a small letter, such as x , y , or z .
- A set may be described by listing all of its elements enclosed in braces. For example, if Set A consists of the numbers 2, 4, 6, and 8, we may say: $A = \{2, 4, 6, 8\}$.
- The null set is denoted by $\{ \}$ or \emptyset .
- Sets may also be described by stating a rule. We could describe Set A from the previous example by stating: Set A consists of all the even single-digit positive integers.

Set Operations

Before discussing the rules of probability, we state the following definitions:

Suppose we have a sample space S defined as follows: $S = \{1, 2, 3, 4, 5, 6\}$. Within that sample space, suppose we define two subsets as follows: $A = \{1, 2\}$ and $B = \{2, 3, 4\}$.

- The **union** of two sets is the set of elements that belong to one or both of the two sets. Thus, if A is $\{1, 2\}$ and Y is $\{2, 3, 4\}$, the union of sets A and B is:
 $A \cup B = \{1, 2, 3, 4\}$. Symbolically, the union of A and B is denoted by $P(A \cup B)$.
- The **intersection** of two sets is the set of elements that are common to both sets. Thus, if A is $\{1, 2\}$ and B is $\{2, 3, 4\}$, the intersection of sets A and B is: $A \cap B = \{2\}$
Symbolically, the intersection of A and B is denoted by $P(A \cap B)$

- The **complement** of an event is the set of all elements in the sample space but not in the event. Thus, if the sample space is $\{1, 2, 3, 4, 5, 6\}$, and B is $\{2, 3, 4\}$, the complement of set B is: $B' = \{1, 5, 6\}$
- If the occurrence of Event A changes the probability of Event B , then Events A and B are **dependent**. On the other hand, if the occurrence of Event A does not change the probability of Event B , then Events A and B are **independent**.
- Two events are **mutually exclusive** or **disjoint** if they cannot occur at the same time. If Events A and B are mutually exclusive, $P(A \cap B) = 0$.
- The probability that Event A occurs, given that Event B has occurred, is called a **conditional probability**. The conditional probability of Event A , given Event B , is denoted by the symbol $P(A|B)$.

Example

What is the set of men with four arms?

Solution

Since all men have two arms at most, the set of men with four arms contains no elements. It is the null set (or empty set).

SELF-ASSESSMENT EXERCISE

Set $A = \{1, 2, 3\}$ and Set $B = \{1, 2, 4, 5, 6\}$. Is Set A a subset of Set B ?

3.2 Probability Formulas and Rules

The probability formula is used to compute the probability of an event to occur. To recall, the likelihood of an event happening is called probability. When a random experiment is entertained, one of the first questions that come in our mind is: What is the probability that a certain event occurs? A probability is a chance of prediction. When we assume that, let's say, x be the chances of happening an event then at the same time $(1-x)$ are the chances for "not happening" of an event.

Similarly, if the probability of an event occurring is "a" and an independent probability is "b", then the probability of both the event occurring is "ab". We can use the formula to find the chances of happening of an event.

The probability of an event tells that how likely the event will happen. Situations in which each outcome is equally likely, then we can find the probability using probability formula. Probability is a chance of prediction. If the probability that an event will occur is "x", then the probability that the event will not occur is "1 - x". If the probability that one event will occur is "a" and the independent probability that another event will occur is "b", then the probability that both events will occur is "ab".

Probability of an Event:

Probability of an event “A” can be written as:

$$P(A) = \frac{\text{Number of Favourable Outcome}}{\text{Total Number of Favourable Outcome}}$$

Probability is the measure of how likely an event is. And an event is one or more outcomes of an experiment. Probability formula is the ratio of number of favourable outcomes to the total number of possible outcomes.

$$\text{Or } P(A) = \frac{n(E)}{n(S)}$$

Where,

- $P(A)$ is the probability of an event “A”
- $n(E)$ is the number of favourable outcomes
- $n(S)$ is the total number of events in the sample space

Measures the likelihood of an event in the following way:

- If $P(A) > P(B)$ then event A is more likely to occur than event B.
- If $P(A) = P(B)$ then events A and B are equally likely to occur.

Note: Here, the favourable outcome means that the outcome of interest.

Think about the toss of a single die. The sample space consists of six possible outcomes (1, 2, 3, 4, 5, and 6). And each outcome is equally likely to occur. Suppose we defined Event A to be the die landing on an odd number. There are three odd numbers (1, 3, and 5). So, the probability of Event A would be $3/6$ or 0.5.

Types of Events

- Two events are **mutually exclusive** if they have no sample points in common.
- Two events are **independent** when the occurrence of one does not affect the probability of the occurrence of the other.

Basic Probability Rules (Definitions and Notation)

Let A and B are two events. The probability formulas are listed below:

Table 1: All Probability Formulas List in Maths.

Probability Range:	$0 \leq P(A) \leq 1$
Rule of Addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Rule of Complementary Events	$P(A') + P(A) = 1$
Disjoint Events: A and B are disjoint if	$P(A \cap B) = 0$
Independent Events: A and B are independent if	$P(A \cap B) = P(A) \cdot P(B)$

Conditional Probability	$P(A B) = P(A \cap B) / P(B)$
Bayes Formula	$P(A B) = P(B A) \cdot P(A) / P(B)$
Cumulative Distribution Function	$F_X(x) = P(X \leq x)$

- Probability Range: Probability of an event A is symbolized by P(A). Probability of an event A is lies between $0 \leq P(A) \leq 1$
- The **complement** of an event is the event not occurring. The probability that Event A will not occur is denoted by P(A').
- The probability that Events A and B *both* occur is the probability of the **intersection** of A and B. The probability of the intersection of Events A and B is denoted by P(A \cap B). If Events A and B are mutually exclusive, P(A \cap B) = 0.
- The probability that Events A or B occur is the probability of the **union** of A and B. The probability of the union of Events A and B is denoted by P(A \cup B) .
- If the occurrence of Event A changes the probability of Event B, then Events A and B are **dependent**. On the other hand, if the occurrence of Event A does not change the probability of Event B, then Events A and B are **independent**
- The probability that Event A occurs, given that Event B has occurred, is called a **conditional probability**. The conditional probability of Event A, given Event B, is denoted by the symbol P(A|B).

Rule of Subtraction

In a previous lesson, we learned two important properties of probability:

- The probability of an event ranges from 0 to 1.
- The sum of probabilities of all possible events equals 1.

The rule of subtraction follows directly from these properties.

Rule of Subtraction. The probability that event A will occur is equal to 1 minus the probability that event A will not occur. $P(A) = 1 - P(A')$

Suppose, for example, the probability that Bill will graduate from college is 0.80. What is the probability that Bill will not graduate from college? Based on the rule of subtraction, the probability that Bill will not graduate is 1.00 - 0.80 or 0.20.

Rule of Multiplication

The rule of multiplication applies to the situation when we want to know the probability of the intersection of two events; that is, we want to know the probability that two events (Event A and Event B) both occur. The probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred. $P(A \cap B) = P(A) P(B|A)$

Example

An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *without replacement* from the urn. What is the probability that both of the marbles are black?

Solution: Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

- In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, $P(A) = 4/10$.
- After the first selection, there are 9 marbles in the urn, 3 of which are black. Therefore, $P(B|A) = 3/9$.

Therefore, based on the rule of multiplication:

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

$$P(A \cap B) = (4/10) * (3/9) = 12/90 = 2/15 = 0.133$$

Rule of Addition

The rule of addition applies to the following situation. We have two events, and we want to know the probability that either event occurs. The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

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

Note: Invoking the fact that $P(A \cap B) = P(A)P(B|A)$, the Addition Rule can also be expressed as: $P(A \cup B) = P(A) + P(B) - P(A)P(B | A)$

3.3 Solving Probability Problems

A probability is determined from an experiment, which is any activity that has an observable outcome like tossing a coin and observing whether it lands heads up or tails up. The possible outcomes of an experiment are called sample space of the experiment.

Steps to find the probability:

Step 1: List the outcomes of the experiment.

Step 2: Count the number of possible outcomes of the experiment.

Step 3: Count the number of favourable outcomes.

Step 4: Use the probability formula

Questions Using Probability Formulas

Examples

Question 1: Suppose we conduct a simple statistical experiment. We flip a coin one time. The coin flip can have one of two equally-likely outcomes - heads or tails. Together, these outcomes represent the sample space of our experiment. Individually, each outcome represents a sample point in the sample space. What is the probability of each sample point?

Solution: The sum of probabilities of all the sample points must equal 1. And the probability of getting a head is equal to the probability of getting a tail. Therefore, the

probability of each sample point (heads or tails) must be equal to $1/2$. The **probability** of a sample point is a measure of the likelihood that the sample point will occur.

Question 2: Let's repeat the experiment of Example 1, with a die instead of a coin. If we toss a fair die, what is the probability of each sample point?

Solution: For this experiment, the sample space consists of six sample points: $\{1, 2, 3, 4, 5, 6\}$. Each sample point has equal probability. And the sum of probabilities of all the sample points must equal 1. Therefore, the probability of each sample point must be equal to $1/6$.

Question 3: What is the probability that a card taken from a standard deck, is an Ace?

Solution:

Total number of cards a standard pack contains = 52

A deck of cards contains Ace = 4 cards

So, the number of favourable outcome = 4

Now, by looking at the formula,

Probability of finding an ace from a deck is,

$P(\text{Ace}) = (\text{Number of favourable outcomes}) / (\text{Total number of favourable outcomes})$

$P(\text{Ace}) = 4/52 = 1/13$

So, we can say that the probability of getting an ace is $1/13$.

Question 4: Calculate the probability of getting an odd number if a dice is rolled?

Solution: Sample space (S) = $\{1, 2, 3, 4, 5, 6\}$

Let "E" be the event of getting an odd number, $E = \{1, 3, 5\}$

So, the Probability of getting an odd number $P(E) = (\text{Number of outcomes favourable}) / (\text{Total number of outcomes}) = n(E)/n(S) = 3/6 = 1/2$

Question 5: Two dice are rolled once. Calculate the probability that the sum of the numbers on the two dice is 5.

Solution:

Possible outcomes (Sample Space) = $\{(1, 1), (1, 2), \dots, (1, 6), (2, 1), (2, 2), \dots, (2, 6), (3, 1), (3, 2), \dots, (3, 6), \dots, (4, 1), (4, 2), \dots, (4, 6), (5, 1), (5, 2), \dots, (5, 6), (6, 1), (6, 2), \dots, (6, 6)\}$

Total possible outcomes = 36

Number of outcomes of the experiment that are favourable to the event that a sum of two events is 6

=> Favourable outcomes are: (1, 5), (2, 4), (3, 3), (4, 2) and (5, 1)

Number of favourable outcomes = 5

Use, probability formula = $P(A) = \frac{\text{Number of Favourable Outcome}}{\text{Total Number of Favourable Outcome}} = \frac{5}{36}$

The probability of a sum of 6 is $5/36$

Question 6: Suppose we draw a card from a deck of playing cards. What is the probability that we draw a spade?

Solution: The sample space of this experiment consists of 52 cards, and the probability of each sample point is $1/52$. Since there are 13 spades in the deck, the probability of drawing a spade is

$$P(\text{Spade}) = (13)(1/52) = 1/4$$

Question 7: Suppose a coin is flipped 3 times. What is the probability of getting two tails and one head?

Solution: For this experiment, the sample space consists of 8 sample points.

$$S = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$$

Each sample point is equally likely to occur, so the probability of getting any particular sample point is $1/8$. The event "getting two tails and one head" consists of the following subset of the sample space. $A = \{TTH, THT, HTT\}$

The probability of Event A is the sum of the probabilities of the sample points in A. Therefore, $P(A) = 1/8 + 1/8 + 1/8 = 3/8$

3.4 Probability Distribution

Definition

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on a number of factors. These factors include the distribution's mean (average), standard deviation, skewness, and kurtosis.

Types of Probability Distributions

There are many different classifications of probability distributions. Some of them include the normal distribution, chi square distribution, binomial distribution, and Poisson distribution. The different probability distributions serve different purposes and represent different data generation processes. The binomial distribution, for example, evaluates the probability of an event occurring several times over a given number of trials and given the event's probability in each trial. and may be generated by keeping track of

how many free throws a basketball player makes in a game, where 1 = a basket and 0 = a miss. Another typical example would be to use a fair coin and figuring the probability of that coin coming up heads in 10 straight flips. A binomial distribution is *discrete*, as opposed to continuous, since only 1 or 0 is a valid response.

The most commonly used distribution is the normal distribution, or "bell curve," although several distributions exist that are commonly used, which is used frequently in finance, investing, science, and engineering. The normal distribution is fully characterized by its mean and standard deviation, meaning the distribution is not skewed and does exhibit kurtosis. This makes the distribution symmetric and it is depicted as a bell-shaped curve when plotted. A normal distribution is defined by a mean (average) of zero and a standard deviation of 1.0, with a skew of zero and kurtosis = 3. Typically, the data generating process of some phenomenon will dictate its probability distribution. This process is called the probability density function.

Example of a Probability Distribution

As a simple example of a probability distribution, let us look at the number observed when rolling two standard six-sided dice. Each die has a 1/6 probability of rolling any single number, one through six, but the sum of two dice will form the probability distribution depicted in the image below. Seven is the most common outcome (1+6, 6+1, 5+2, 2+5, 3+4, 4+3). Two and twelve, on the other hand, are far less likely (1+1 and 6+6).

By Probability Distributions we are interested in some numerical description of the outcome. For example, when we toss a coin 3 times, and we are interested in the number of heads that fall, then a numerical value of 0,1,2,3 will be assigned to each sample point. The numbers 0, 1, 2 and 3 are random quantities determined by the outcome of an experiment.

They may be thought of as the values assumed by some **random variable** x , which in this case represents the number of heads when a coin is tossed 3 times. So we could write $x_1 = 0$, $x_2 = 1$, $x_3 = 2$ and $x_4 = 3$

1. A **random variable** is a variable whose value is determined by the outcome of a random experiment.
2. A **discrete random variable** is one whose set of assumed values is **countable** (arises from **counting**).
3. A **continuous random variable** is one whose set of assumed values is uncountable (arises from **measurement**).

We shall use: A capital (**upper case**) X for the random variable and Lower case $x_1, x_2, x_3 \dots$ for the **values** of the random variable in an experiment. These x_i then

represent an event that is a subset of the sample space. The **probabilities** of the events are given by: $P(x_1), P(x_2), P(x_3)$...

We also use the notation $P(X)$. For example, we may need to find some of the probabilities involved when we throw a die. We would write for the probability of obtaining a "5" when we roll a die as: $P(X = 5) = 1/6$

Discrete vs. Continuous Variables

All probability distributions can be classified as discrete probability distributions or as continuous probability distributions, depending on whether they define probabilities

If a variable can take on any value between two specified values, it is called a **continuous variable**; otherwise, it is called a **discrete variable**.

Some examples will clarify the difference between discrete and continuous variables.

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

Discrete Probability Distributions

If a random variable is a discrete variable, its probability distribution is called a **discrete probability distribution**.

An example will make this clear. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

The probability distribution for this statistical experiment appears below.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represents a *discrete* probability distribution because it relates each value of a discrete random variable with its probability of occurrence. **Note:** With a discrete

probability distribution, each possible value of the discrete random variable can be associated with a non-zero probability. Thus, a discrete probability distribution can always be presented in tabular form.

Continuous Probability Distributions

If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution. A continuous probability distribution differs from a discrete probability distribution in several ways.

- The probability that a continuous random variable will assume a particular value is zero.
- As a result, a continuous probability distribution cannot be expressed in tabular form.
- Instead, an equation or formula is used to describe a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a probability density function. Sometimes, it is referred to as a density function, a PDF, or a pdf. For a continuous probability distribution, the density function has the following properties:

- Since the continuous random variable is defined over a continuous range of values (called the domain of the variable), the graph of the density function will also be continuous over that range.
- The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
- The probability that a random variable assumes a value between a and b is equal to the area under the density function bounded by a and b .

For example, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable X was less than or equal to a . The probability that X is less than or equal to a is equal to the area under the curve bounded by a and minus infinity - as indicated by the shaded area.

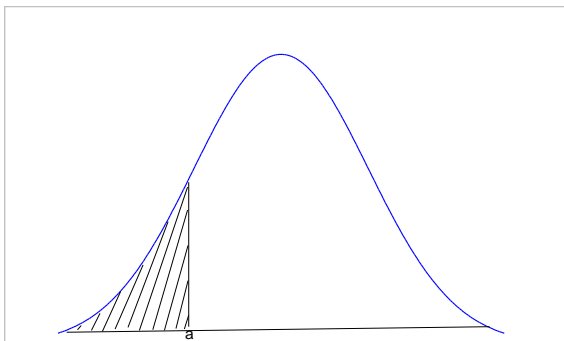


Figure 1: area under the shade

Note: The shaded area in the graph represents the probability that the random variable X is less than or equal to a . This is a cumulative probability. However, the probability that X is *exactly* equal to a would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as a) is always zero.

Example 1 - Discrete Random Variable

Two balls are drawn at random in succession without replacement from an urn containing 4 red balls and 6 black balls. Find the probabilities of all the possible outcomes. Find the probabilities of all the possible outcomes.

Solution

Let X denote the number of red balls in the outcome.

Possible Outcomes

	RR	RB	BR	BB
X	2	1	1	0

Here, $x_1 = 2, x_2 = 1, x_3 = 1, x_4 = 0$

Now, the probability of getting 2 red balls when we draw out the balls one at a time is:

Probability of first ball being red = $4/10$

Probability of second ball being red = $3/9$ (because there are 3 red balls left in the urn, out of a total of 9 balls left). So

$$P(x_1) = \frac{4}{10} \times \frac{3}{9} = \frac{2}{15}$$

Likewise, for the probability of red first is $4/10$ followed by black is $3/9$ (because there are 6 black balls still in the urn and 9 balls all together). So:

$$P(x_2) = \frac{4}{10} \times \frac{6}{9} = \frac{4}{15}$$

Similarly, for black then red:

$$P(x_3) = \frac{6}{10} \times \frac{4}{9} = \frac{4}{15}$$

Finally, for 2 black balls

$$P(x_4) = \frac{6}{10} \times \frac{5}{9} = \frac{1}{3}$$

As a check, if we have found all the probabilities, then they should add up to 1.

$$\frac{2}{15} + \frac{4}{15} + \frac{4}{15} + \frac{1}{3} = \frac{15}{15} = 1$$

So, we have found them all.

Example 2 - Continuous Random Variable

A jar of coffee is picked at random from a filling process in which an automatic machine is filling coffee jars each with 1 kg of coffee. Due to some faults in the automatic process, the weight of a jar could vary from jar to jar in the range 0.9 kg to 1.05 kg, excluding the latter

Let X denote the weight of a jar of coffee selected. What is the range of X ?

Solution: Possible outcomes: $0.9 \leq X < 1.05$ That's all there is to it!

Distribution Function

1. A **discrete probability distribution** is a table (or a formula) listing all possible values that a discrete variable can take on, together with the associated probabilities.
2. The function $f(x)$ is called a **probability density function** for the continuous random variable X where the total area under the curve bounded by the x -axis is equal to 1. i.e.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

The area under the curve between any two ordinates $x = a$ and $x = b$ is the probability that X lies between a and b .

$$\int_a^b f(x)dx = P(a \leq X \leq b)$$

Probabilities as Relative Frequency

If an experiment is performed a sufficient number of times, then in the long run, the **relative frequency** of an event is called the **probability** of that event occurring.

Example 3

Refer to the previous example. The weight of a jar of coffee selected is a continuous random variable. The following table gives the weight in kg of 100 jars recently filled by the machine. It lists the observed values of the continuous random variable and their corresponding frequencies.

Find the probabilities for each weight category.

Weight X	Number of Jars
0.900 - 0.925	1
0.925 - 0.950	7
0.950 - 0.975	25
0.975 - 1.000	32
1.000 - 1.025	30
1.025 - 1.050	5

Total	100
-------	-----

Solution

We simply divide the number of jars in each weight category 100 to give the probabilities

Weight X	Number of Jars	Probability $P(a \leq X < b)$
0.900 - 0.925	1	0.01
0.925 - 0.950	7	0.07
0.950 - 0.975	25	0.25
0.975 - 1.000	32	0.32
1.000 - 1.025	30	0.30
1.025 - 1.050	5	0.05
Total	100	1.00

Expected Value of a Random Variable

Let X represent a discrete random variable with the probability distribution function $P(X)$. Then the expected value of X denoted by $E(X)$, or μ , is defined as:

$$E(X) = \mu = \sum (x_i \times P(x_i))$$

To calculate this, we multiply each possible value of the variable by its probability, then add the results.

$$\sum (x_i \times P(x_i)) = \{ x_1 \times P(x_1) \} + \{ x_2 \times P(x_2) \} + \{ x_3 \times P(x_3) \} + \dots$$

$E(X)$ is also called the mean of the probability distribution

Example 4

In Example 1 above, we had an experiment where we drew 2 balls from an urn containing 4 red and 6 black balls. What is the expected number of red balls?

Solution

We already worked out the probabilities before:

Possible outcome	RR	RB	BR	BB
x_i	2	1	1	0
$P(x_i)$	2/15	4/15	4/15	1/3

$$E(X) = \left[(x_i \cdot P(x_i)) \right] = 2 \times \frac{2}{15} + 1 \times \frac{4}{15} + 0 \times \frac{1}{3} = \frac{4}{5} = 0.8$$

This means that if we performed this experiment 1000 times, we would expect to get 800 red balls.

Example 5

I throw a die and get \$1 if it is showing 1, and get \$2 if it is showing 2, and get \$3 if it is showing 3, etc. What is the amount of money I can expect if I throw it 100 times?

Solution

$$E(X) = \left[(x_i \cdot P(x_i)) \right] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{7}{2} = 3.5$$

So, for 100 throws, I can expect to get \$350.

Example 6

The number of persons X , in a Singapore family chosen at random has the following probability distribution:

X	1	2	3	4	5	6	7	8
P(X)	0.34	0.44	0.11	0.06	0.02	0.01	0.01	0.01

Find the average family size $E(X)$.

$$E(X) = \left[(x_i \cdot P(x_i)) \right] = 1 \times 0.34 + 2 \times 0.44 + 3 \times 0.11 + 4 \times 0.06 \\ + 5 \times 0.02 + 6 \times 0.01 + 7 \times 0.01 + 8 \times 0.01 = 2.1$$

So, the average family size is $E(X) = \mu = 2.1$ people.

Variance of a Random Variable

Let X represent a discrete random variable with probability distribution function

$P(X)$ The **variance** of X denoted by $V(X)$ or σ^2 is defined as:

$$V(X) = \sigma^2 \\ = \sum \left[\{X - E(X)\}^2 \times P(X) \right]$$

Since $\mu = E(X)$, (or the average value), we could also write this as:

$$V(X) = \sigma^2 \\ = \sum \left[\{X - E(\mu)\}^2 \times P(X) \right]$$

Another way of calculating the variance is:

$$V(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

Standard Deviation of the Probability Distribution

$\sigma = \sqrt{V(X)}$ is called the **standard deviation** of the probability distribution. The standard deviation is a number which describes the **spread** of the distribution. Small standard deviation means small spread, large standard deviation means large spread.

In the following 3 distributions, we have the same **mean** ($\mu = 4$), but the standard deviation becomes bigger, meaning the spread of scores is greater.

Example 7

Find $V(X)$ for the following probability distribution:

X	8	12	16	20	24
P(X)	1/8	1/6	3/8	1/4	1/12

Solution

We have to find $E(X)$ first:

$$E(X) = 8 \times \frac{1}{8} + 12 \times \frac{1}{6} + 16 \times \frac{3}{8} + 20 \times \frac{1}{4} + 24 \times \frac{1}{12} = 16$$

$$\text{Then } V(X) = E\left[\{X - E(X)\}^2 \cdot P(X)\right]$$

$$= (8-16)^2 \times \frac{1}{8} + (12-16)^2 \times \frac{1}{6} + (16-16)^2 \times \frac{3}{8} + (20-16)^2 \times \frac{1}{4} + (24-16)^2 \times \frac{1}{12} = 20$$

Checking this using the other formula:

$$V(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

For this, we need to work out the expected value of the **squares** of the random variable X .

X	8	12	16	20	24
X	64	144	256	400	576
P(X)	1/8	1/6	3/8	1/4	1/12

$$E(X^2) = \sum X^2 P(X)$$

$$= 64 \times \frac{1}{8} + 144 \times \frac{1}{6} + 256 \times \frac{3}{8} + 400 \times \frac{1}{4} + 576 \times \frac{1}{12} = 276$$

We found $E(X)$ before: $E(X)=16$

$$V(X) = E(X^2) - [E(X)]^2 = 276 - 16^2 = 20 \text{ as before}$$

Normal Distribution

The **normal distribution** refers to a family of continuous probability distributions described by the normal equation.

The normal distribution is defined by the following equation:

The Normal Equation. The value of the random variable Y is:

$$Y = \left\{1 / \left[\sigma * \text{sqrt}(2\pi)\right]\right\} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where X is a normal random variable, μ is the mean, σ is the standard deviation, π is approximately 3.14159, and e is approximately 2.71828.

The random variable X in the normal equation is called the **normal random variable**. The normal equation is the probability density function for the normal distribution.

The Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the centre of the graph, and the standard deviation determines the height and width of the graph. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

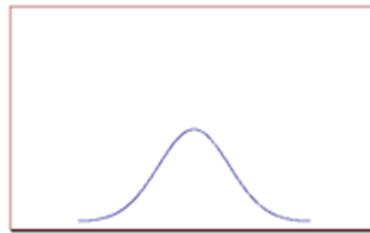
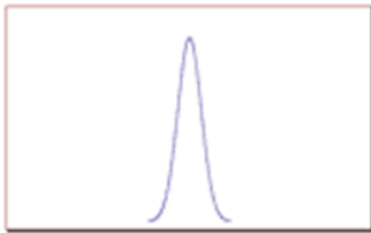


Fig 2: Smaller standard deviation

Bigger standard deviation

When the standard deviation is small, the curve is tall and narrow; and when the standard deviation is big, the curve is short and wide (see above)

Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable X equals any particular value is 0.
- The probability that X is greater than a equals the area under the normal curve bounded by a and plus infinity (as indicated by the *non-shaded* area in the figure below).
- The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (as indicated by the *shaded* area in the figure below).

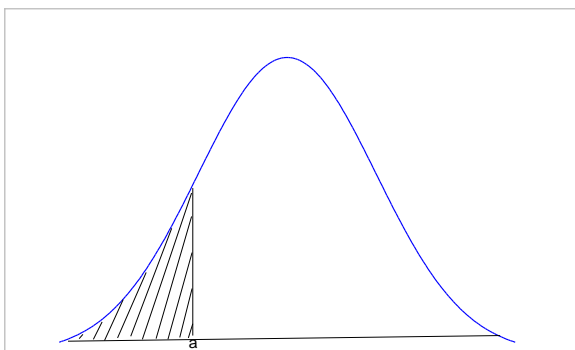


Figure 3: Area under the normal curve

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

Collectively, these points are known as the empirical rule or the 68-95-99.7 rule. Clearly, given a normal distribution, most outcomes will be within 3 standard deviations of the mean. To find the probability associated with a normal random variable, use a graphing calculator, an online normal distribution calculator, or a normal distribution table.

Example 8

An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

Solution: Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

- The value of the normal random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The Solution is: $P(X < 365) = 0.90$. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

Problem 9

Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

Solution: Here, we want to know the probability that the test score falls between 90 and 110. The "trick" to solving this problem is to realize the following:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$

We use the Normal Distribution Calculator to compute both probabilities on the right side of the above equation.

- To compute $P(X < 110)$, we enter the following inputs into the calculator: The value of the normal random variable is 110, the mean is 100, and the standard deviation is 10. We find that $P(X < 110)$ is 0.84.
- To compute $P(X < 90)$, we enter the following inputs into the calculator: The value of the normal random variable is 90, the mean is 100, and the standard deviation is 10. We find that $P(X < 90)$ is 0.16.

We use these findings to compute our final Solution as follows:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$

$$P(90 < X < 110) = 0.84 - 0.16$$

$$P(90 < X < 110) = 0.68$$

Thus, about 68% of the test scores will fall between 90 and 110.

Poisson Distribution

A Poisson distribution is the probability distribution that results from a Poisson experiment.

Attributes of a Poisson Experiment

A Poisson experiment is a statistical experiment that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (μ) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation

The following notation is helpful, when we talk about the Poisson distribution.

- e : A constant equal to approximately 2.71828. (Actually, e is the base of the natural logarithm system.)
- μ : The mean number of successes that occur in a specified region.
- x : The actual number of successes that occur in a specified region.
- $P(x; \mu)$: The Poisson probability that exactly x successes occur in a Poisson experiment, when the mean number of successes is μ .

Poisson Distribution

A Poisson random variable is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a Poisson distribution.

Given the mean number of successes (μ) that occur in a specified region, we can compute the Poisson probability based on the following formula:

Poisson Formula. Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ . Then, the Poisson probability is:

$$p(x, \mu) = (e^{-\mu})(\mu^x) / x!$$

where x is the actual number of successes that result from the experiment, and e is approximately equal to 2.71828.

The Poisson distribution has the following properties:

- The mean of the distribution is equal to μ
- The variance is also equal to μ

Poisson Distribution Example

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- $\mu = 2$; since 2 homes are sold per day, on average.
- $x = 3$; since we want to find the likelihood that 3 homes will be sold tomorrow.
- $e = 2.71828$; since e is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$$P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$$

$$P(3; 2) = (2.71828^{-2}) (2^3) / 3!$$

$$P(3; 2) = (0.13534) (8) / 6$$

$$P(3; 2) = 0.180$$

Thus, the probability of selling 3 homes tomorrow is 0.180

On this website “STAT TREK”, you cover the following continuous probability distributions. They can be found in the Stat Trek main menu under the Stat Tools tab

- Student's t distribution
- Chi-square distribution
- F distribution
- Hypergeometric probability distribution
- Multinomial probability distribution
- Negative binomial distribution
- Cumulative Probability and the F Distribution

How to Interpret Probability

Mathematically, the probability that an event will occur is expressed as a number between 0 and 1. Notationally, the probability of event A is represented by $P(A)$.

- If $P(A)$ equals zero, event A will almost definitely not occur.
- If $P(A)$ is close to zero, there is only a small chance that event A will occur.
- If $P(A)$ equals 0.5, there is a 50-50 chance that event A will occur.
- If $P(A)$ is close to one, there is a strong chance that event A will occur.
- If $P(A)$ equals one, event A will almost definitely occur.

In a statistical experiment, the sum of probabilities for all possible outcomes is equal to one. This means, for example, that if an experiment can have three possible outcomes (A, B, and C), then $P(A) + P(B) + P(C) = 1$.

SELF-ASSESSMENT EXERCISE

A student goes to the library. The probability that she checks out (a) a work of fiction is 0.40, (b) a work of non-fiction is 0.30, and (c) both fiction and non-fiction is 0.20. What is the probability that the student checks out a work of fiction, non-fiction, or both?

4.0 CONCLUSION

Probability is based on observations of certain events. Probability of an event is the ratio of the number of observations of the event to the total numbers of the observations. An experiment is a situation involving chance or probability that leads to results called outcomes. An outcome is the result of a single trial of an experiment. The probability of an event is a measure of the likelihood that the event will occur. Statisticians have agreed on the following rules and conventions. The probability of any event can range from 0 to 1 and the sum of probabilities of all sample points in a sample space is equal to 1. The probability of event A is the sum of the probabilities of all the sample points in event A. The probability of event A is denoted by $P(A)$. Thus, if event A were very unlikely to occur, then $P(A)$ would be close to 0. And if event A were very likely to occur, then $P(A)$ would be close to 1.

5.0 SUMMARY

This unit contains a description of the important mathematical concepts of probability theory, illustrated by some of the applications that have stimulated their development. The unit overviewed the probability concepts and gives the formulas and rules in probability theories. The unit shows that the probability of an event is the measure of the chance that the event will occur as a result of an experiment by solving some probability problems and explaining some probability distribution concepts. Since applications inevitably involve simplifying assumptions that focus on some features of a problem at the expense of others, it is advantageous to begin by thinking about simple experiments, such as tossing a coin or rolling dice, and later to see how these apparently frivolous investigations relate to important scientific questions.

6.0 TUTOR-MARKED ASSIGNMENT

i) Suppose we have a bowl with 10 marbles - 2 red marbles, 3 green marbles, and 5 blue marbles. We randomly select 4 marbles from the bowl, with replacement. What is the probability of selecting 2 green marbles and 2 blue marbles? (Apply the multinomial formula)

ii) Suppose you randomly select 7 women from a population of women, and 12 men from a population of men. The table below shows the standard deviation in each sample and in each population.

Population	Population Standard Deviation	Sample Standard Deviation
Women	30	35
Men	50	45

Compute the f statistic.

7.0 References/Further Readings

- Alan Stuart and Keith (2009). Kendall's Advanced Theory of Statistics, Distribution Theory, Ord, 6th Ed Volume 1
- Broemeling, Lyle D. (1 November 2011). An Account of Early Statistical Inference in Arab Cryptology". *The American Statistician*. 65 (4): 255–257.
- Freund, John. (1973). *Introduction to Probability*. Dickenson
- Grinstead, Charles Miller; James Laurie Snell. "Introduction". *Introduction to Probability*. pp. vii.
- Hacking, I. (2006). *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*, Cambridge University Press,
- Jeffrey, R.C. (1992). *Probability and the Art of Judgment*, Cambridge University Press. pp. 54–55
- William Feller (1968). *An Introduction to Probability Theory and Its Applications*, (Vol 1), 3rd Ed, Wiley,

MODULE 2: QUANTITATIVE TECHNIQUES & LINEAR PROGRAMMING

UNIT 1	Overview of Quantitative Techniques
UNIT 2	Linear Programming Graphical Method
UNIT 3	Simplex Method
UNIT 4	Transportation Model

UNIT 1: OVERVIEW OF QUANTITATIVE TECHNIQUES CONTENTS

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	The Meaning of Quantitative Techniques
3.2	The Quantitative Techniques Approach
3.3	How to Develop a Quantitative Analysis
4.0	Conclusion
5.0	Summary
1.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

In economic analysis as well as business management, people have been using mathematical tools to help solve problems for thousands of years; however, the formal study and application of quantitative techniques to practical decision making is largely a product of the twentieth century. The techniques we study in this module have been applied successfully to an increasingly wide variety of complex problems in business, government, health care, education, and many other areas. Many such successful uses are discussed throughout this book. It isn't enough, though, just to know the mathematics of how a particular quantitative technique works; you must also be familiar with the limitations, assumptions, and specific applicability of the technique.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Describe the meaning of quantitative techniques
- Understand the various quantitative technique approaches
- Develop a quantitative analysis

3.0 MAIN CONTENT

3.1 The Meaning of Quantitative Techniques

Quantitative Techniques is the scientific approach to managerial decision making. Whim, emotions, and guesswork are not part of the quantitative analysis approach. The approach starts with data. Like raw material for a factory, these data are manipulated or processed

into information that is valuable to people making decisions. This processing and manipulating of raw data into meaningful information is the heart of quantitative analysis.

In solving a problem, managers must consider both qualitative and quantitative factors. For example, we might consider several different investment alternatives, including certificates of deposit at a bank, investments in the stock market, and an investment in real estate. We can use quantitative analysis to determine how much our investment will be worth in the future when deposited at a bank at a given interest rate for a certain number of years. Quantitative analysis can also be used in computing financial ratios from the balance sheets for several companies whose stock we are considering. Some real estate companies have developed computer programs that use quantitative analysis to analyze cash flows and rates of return for investment property.

Because of the importance of qualitative factors, the role of quantitative analysis in the decision-making process can vary. When there is a lack of qualitative factors and when the problem, model, and input data remain the same, the results of quantitative analysis can *automate* the decision-making process. For example, some companies use quantitative inventory models to determine automatically when to order additional new materials. In most cases, however, quantitative analysis will be an *aid* to the decision-making process. The results of quantitative analysis will be combined with other (qualitative) information in making decisions.

SELF ASSESSMENT EXERCISE

Describe in your own words what you understand by quantitative analysis.

3.2 The Quantitative Techniques Approach

The quantitative techniques approach consists of defining a problem, developing a model, acquiring input data, developing a solution, testing the solution, analyzing the results, and implementing the results. One step does not have to be finished completely before the next is started; in most cases one or more of these steps will be modified to some extent before the final results are implemented. This would cause all of the subsequent steps to be changed. In some cases, testing the solution might reveal that the model or the input data are not correct. This would mean that all steps that follow defining the problem would need to be modified.

Step One- Defining the Problem

The first step in the quantitative approach is to develop a clear, concise statement of the **problem**. This statement will give direction and meaning to the following steps.

In many cases, defining the problem is the most important and the most difficult step. It is essential to go beyond the symptoms of the problem and identify the true causes. One problem may be related to other problems; solving one problem without regard to other

related problems can make the entire situation worse. Thus, it is important to analyze how the solution to one problem affects other problems or the situation in general.

It is likely that an organization will have several problems. However, a quantitative analysis group usually cannot deal with all of an organization's problems at one time. Thus, it is usually necessary to concentrate on only a few problems. For most companies, this means selecting those problems whose solutions will result in the greatest increase in profits or reduction in costs to the company. The importance of selecting the right problems to solve cannot be overemphasized. Experience has shown that bad problem definition is a major reason for failure of management science or operations research groups to serve their organizations well.

When the problem is difficult to quantify, it may be necessary to develop *specific, measurable* objectives. A problem might be inadequate health care delivery in a hospital. The objectives might be to increase the number of beds, reduce the average number of days a patient spends in the hospital, increase the physician-to-patient ratio, and so on. When objectives are used, however, the real problem should be kept in mind. It is important to avoid obtaining specific and measurable objectives that may not solve the real problem.

Step Two-Developing a Model

Once we select the problem to be analyzed, the next step is to develop a **model**. Simply stated, a model is a representation (usually mathematical) of a situation.

Even though you might not have been aware of it, you have been using models most of your life. You may have developed models about people's behavior. Your model might be that friendship is based on reciprocity, an exchange of favors. If you need a favor such as a small loan, your model would suggest that you ask a good friend.

Of course, there are many other types of models. Architects sometimes make a *physical model* of a building that they will construct. What sets quantitative analysis apart from other techniques is that the models that are used are mathematical. A **mathematical model** is a set of mathematical relationships. In most cases, these relationships are expressed in equations and inequalities, as they are in a spreadsheet model that computes sums, averages, or standard deviations.

Step Three- Acquiring Input Data

Once we have developed a model, we must obtain the data that are used in the model (*input data*). Obtaining accurate data for the model is essential; even if the model is a perfect representation of reality, improper data will result in misleading results. This situation is called *garbage in, garbage out*. For a larger problem, collecting accurate data can be one of the most difficult steps in performing quantitative analysis. There are a number of sources that can be used in collecting data (see module 4). Sampling and direct measurement provide other sources of data for the model. You may need to know how

many pounds of raw material are used in producing a new photochemical product. This information can be obtained by going to the plant and actually measuring with scales the amount of raw material that is being used. In other cases, statistical sampling procedures can be used to obtain data.

Step Four- Developing a Solution

Developing a solution involves manipulating the model to arrive at the best (optimal) solution to the problem. In some cases, this requires that an equation be solved for the best decision. In other cases, you can use a *trial and error* method, trying various approaches and picking the one that results in the best decision. For some problems, you may wish to try all possible values for the variables in the model to arrive at the best decision. This is called *complete enumeration*.

Step Five – Testing the Solution

Before a solution can be analyzed and implemented, it needs to be tested completely. Because the solution depends on the input data and the model, both require testing.

Testing the input data and the model includes determining the accuracy and completeness of the data used by the model. Inaccurate data will lead to an inaccurate solution. There are several ways to test input data. One method of testing the data is to collect additional data from a different source. If the original data were collected using interviews, perhaps some additional data can be collected by direct measurement or sampling. These additional data can then be compared with the original data, and statistical tests can be employed to determine whether there are differences between the original data and the additional data.

Step Six – Analyzing the Results and Sensitivity Analysis

Analyzing the results starts with determining the implications of the solution. In most cases, a solution to a problem will result in some kind of action or change in the way an organization is operating. The implications of these actions or changes must be determined and analyzed before the results are implemented.

Because a model is only an approximation of reality, the sensitivity of the solution to changes in the model and input data is a very important part of analyzing the results. This type of analysis is called sensitivity analysis or post optimality analysis. It determines how much the solution will change if there were changes in the model or the input data. When the solution is sensitive to changes in the input data and the model specification, additional testing should be performed to make sure that the model and input data are accurate and valid. If the model or data are wrong, the solution could be wrong, resulting in financial losses or reduced profits.

Step Seven - Implementing the Results

The final step is to *implement* the results. This is the process of incorporating the solution into the company. This can be much more difficult than you would imagine. Even if the solution is optimal and will result in millions of dollars in additional profits, if managers resist the new solution, all of the efforts of the analysis are of no value.

SELF ASSESSMENT EXERCISE

Using a typical problem in a Nigerian Banking Industry, show how the issue can be analyzed using a quantitative technique approach.

3.3 How to Develop a Quantitative Analysis

Developing a model is an important part of the quantitative analysis approach. Let's see how we can use the following mathematical model, which represents profit:

$$\text{Profit} = \text{Revenue} - \text{Expenses}$$

In many cases, we can express revenues as price per unit multiplied times the number of units sold. Expenses can often be determined by summing fixed costs and variable cost. Variable cost is often expressed as variable cost per unit multiplied times the number of units. Thus, we can also express profit in the following mathematical model:

$$\text{Profit} = \text{Revenue} - (\text{Fixed cost} + \text{Variable cost})$$

Profit = (Selling price per unit)(Number of units sold) – [Fixed cost + (Variable cost per unit)(Number of units sold)]

$$\text{Profit} = sX - [f + vX]$$

$$\text{Profit} = sX - f - vX$$

(2.1)

Where

s = Selling price per unit

f = fixed cost

v = variable cost per unit

X = number of units sold

The parameters in this model are f , and s , as these are inputs that are inherent in the model. The number of units sold (X) is the decision variable of interest.

In addition to the profit models shown here, decision makers are often interested in the **break-even point** (BEP). The BEP is the number of units sold that will result in ~~N~~0 profits. We set profits equal to \$0 and solve for X , the number of units at the break-even point:

$$0 = sX - f - vX$$

This can be written as

$$0 = (s - v)X - f$$

Solving for X , we have

$$f = (s - v)X$$

$$x = \frac{f}{s - v}$$

This quantity (X) that results in a profit of zero is the BEP, and we now have this model for the BEP:

$$BEP = \frac{\text{fixed cost}}{(\text{selling price per unit} - \text{variable cost per unit})}$$

SELF ASSESSMENT EXERCISE

Use a typical example to solve for the break-even point using the formula above

4.0 CONCLUSION

So far, this unit concludes that while mathematics is a relevant approach in solving managerial problems, quantitative techniques has been developed to enhance the solutions inherent in managerial decisions. Also, this unit explains the strategy that is employed in analyzing the quantitative analysis.

5.0 SUMMARY

In this unit, we have discussed the meaning of quantitative techniques, the various quantitative techniques approach which are defining a problem, developing a model, acquiring input data, developing a solution, testing the solution, analyzing the results, and implementing the results.

6.0 TUTORED MARKED ASSIGNMENTS

1. Briefly discuss quantitative techniques approach.
2. what are the steps to follow using quantitative analysis in analyzing a managerial problem?

7.0 REFERENCES/FURTHER READINGS

- Render, B., Stair, R.M., & Hanna, M.E. (2012). *Quantitative Analysis for Management* (11th ed.). Essex: England, Pearson.
- Murthy, P.R. (2007). *Operations Research* (2nd ed.). New Delhi, India: New Age International ltd.

UNIT 2 LINEAR PROGRAMMING GRAPHICAL METHOD

CONTENTS

- 1.0. Introduction
- 2.0. Objectives
- 3.0. Main Content
 - 3.1 Requirements of a Linear Programming Problem
 - 3.2 Formulating Linear Programming Problems
 - 3.3 Graphical Solution to a Linear Programming Problem
- 6.0 Conclusion
- 5.0 Summary
- 6.0. Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

Many management decisions involve trying to make the most effective use of an organization's resources. Resources typically include machinery, labor, money, time, warehouse space, and raw materials. These resources may be used to make products (such as machinery, furniture, food, or clothing) or services (such as schedules for airlines or production, advertising policies, or investment decisions). Despite its name, LP and the more general category of techniques called "**mathematical**" programming have very little to do with computer programming. In the world of management science, *programming* refers to modeling and solving a problem mathematically. Computer programming has, of course, played an important role in the advancement and use of LP. Real life LP problems are too cumbersome to solve by hand or with a calculator.

2.0 OBJECTIVE

At the end of this unit, you should be able to:

- Understand the requirements of a Linear Programming Problem
- Formulate a typical Linear Programming Problem
- Solve a Linear Programming Problem using graphical method

3.0 MAIN CONTENT

3.1. Requirements of a Linear Programming Problem

Linear programming (LP) is a widely used mathematical modeling technique designed to help managers in planning and decision making relative to resource allocation. All linear programming problems seek to *maximize* or *minimize* some quantity, usually profit or cost. We refer to this property as the **objective function** of an LP problem. The major objective of a typical manufacturer is to maximize naira profits. In the case of a trucking or railroad distribution system, the objective might be to minimize shipping costs. In any event, this objective must be stated clearly and defined mathematically. It does not

matter, by the way, whether profits and costs are measured in cents, dollars, or millions of naira.

The second property that LP problems have in common is the presence of restrictions, or **constraints**, that limit the degree to which we can pursue our objective. For example, deciding how many units of each product in a firm's product line to manufacture is restricted by available personnel and machinery. Selection of an advertising policy or a financial portfolio is limited by the amount of money available to be spent or invested. We want, therefore, to maximize or minimize a quantity (the objective function) subject to limited resources (the constraints).

There must be alternative courses of action to choose from. For example, if a company produces three different products, management may use LP to decide how to allocate among them its limited production resources (of personnel, machinery, and so on). Should it devote all manufacturing capacity to make only the first product, should it produce equal amounts of each product, or should it allocate the resources in some other ratio? If there were no alternatives to select from, we would not need LP.

The objective and constraints in LP problems must be expressed in terms of *linear* equations or inequalities. Linear mathematical relationships just mean that all terms used in the objective function and constraints are of the first degree (i.e., not squared, or to the third or higher power, or appearing more than once). Hence, the equation $2A+5B = 10$ is an acceptable linear function, while the equation $2A^2 + 5B^2 + 3AB = 10$ is not linear because the variable A is squared, the variable B is cubed, and the two variables appear again as a product of each other.

The term *linear* implies both proportionality and additivity. Proportionality means that if production of 1 unit of a product uses 3 hours, production of 10 units would use 30 hours. Additivity means that the total of all activities equals the sum of the individual activities. If the production of one product generated ₦3 profit and the production of another product generated ₦8 profit, the total profit would be the sum of these two, which would be ₦11.

We assume that conditions of *certainty* exist: that is, number in the objective and constraints are known with certainty and do not change during the period being studied.

We make the *divisibility* assumption that solutions need not be in whole numbers (integers). Instead, they are divisible and may take any fractional value. In production problems, we often define variables as the number of units produced per week or per month, and a fractional value (e.g., 0.3 chairs) would simply mean that there is work in process. Something that was started in one week can be finished in the next. However, in other types of problems, fractional values do not make sense. If a fraction of a product

cannot be purchased (for example, one-third of a submarine), an integer programming problem exists.

The properties of a linear program are:

- i. One objective function
- ii. One or more constraints
- iii. Alternative courses of action
- iv. Objective function and constraints are linear—proportionality and divisibility
- v. Certainty
- vi. Divisibility
- vii. Nonnegative variables

SELF ASSESSMENT EXERCISE

Briefly discuss the requirement needed in order to use a Linear Programming problem in solving a typical economic problem.

3.2 Formulating Linear Programming Problems

Formulating a linear program involves developing a mathematical model to represent the managerial problem. Thus, in order to formulate a linear program, it is necessary to completely understand the managerial problem being faced. Once this is understood, we can begin to develop the mathematical statement of the problem. The steps in formulating a linear program follow:

1. Completely understand the managerial problem being faced.
2. Identify the objective and the constraints.
3. Define the decision variables.
4. Use the decision variables to write mathematical expressions for the objective function and the constraints.

One of the most common LP applications is the **product mix problem**. Two or more products are usually produced using limited resources such as personnel, machines, raw materials, and so on. The profit that the firm seeks to maximize is based on the profit contribution per unit of each product. (Profit contribution, you may recall, is just the selling price per unit minus the variable cost per unit*) The company would like to determine how many units of each product it should produce so as to maximize overall profit given its limited resources.

SELF ASSESSMENT EXERCISE

Briefly discuss the steps required in formulating a linear programming problems.

3.3 Graphical Solution to a Linear Programming Problem

To be able to understand this method, we shall be taking an example. The Flair Furniture Company produces inexpensive tables and chairs. The production process for each is similar in that both require a certain number of hours of carpentry work and a certain number of labor hours in the painting and varnishing department. Each table takes 4

hours of carpentry and 2 hours in the painting and varnishing shop. Each chair requires 3 hours in carpentry and 1 hour in painting and varnishing. During the current production period, 240 hours of carpentry time are available and 100 hours in painting and varnishing time are available. Each table sold yields a profit of ₦70; each chair produced is sold for a ₦50 profit.

Flair Furniture's problem is to determine the best possible combination of tables and chairs to manufacture in order to reach the maximum profit. The firm would like this production mix situation formulated as an LP problem.

We begin by summarizing the information needed to formulate and solve this problem. This helps us understand the problem being faced. Next we identify the objective and the constraints. The objective is

Maximize profit

The constraints are

1. The hours of carpentry time used cannot exceed 240 hours per week.
2. The hours of painting and varnishing time used cannot exceed 100 hours per week.

The decision variables that represent the actual decisions we will make are defined as

T = number of tables to be produced per week

C = number of chairs to be produced per week

Now we can create the LP objective function in terms of *T* and *C*. The objective function is Maximize profit = ₦70*T* + ₦50*C*.

Our next step is to develop mathematical relationships to describe the two constraints in this problem. One general relationship is that the amount of a resource used is to be less than or equal to (\leq) the amount of resource *available*.

In the case of the carpentry department, the total time used is

(4 hours per table)(Number of tables produced) + (3 hours per chair)(Number of chairs produced)

So the first constraint may be stated as follows:

Carpentry time used \leq Carpentry time available

$4T + 3C \leq 240$ (hours of carpentry time)

Similarly, the second constraint is as follows:

Painting and varnishing time used \leq Painting and varnishing time available

→ $2T + 1C \leq 100$ (hours of painting and varnishing time)

└─ (This means that each table produced takes two hours of the painting and varnishing resource.)

Both of these constraints represent production capacity restrictions and, of course, affect the total profit. For example, Flair Furniture cannot produce 80 tables during the production period because if $T = 80$, both constraints will be violated. It also cannot make $T = 50$ tables and $C = 10$ chairs. Why? Because this would violate the second constraint that no more than 100 hours of painting and varnishing time be allocated.

To obtain meaningful solutions, the values for T and C must be nonnegative numbers. That is, all potential solutions must represent real tables and real chairs. Mathematically, this means that

$$T \geq 0 \text{ (number of tables produced is greater than or equal to 0)}$$

$$C \geq 0 \text{ (number of chairs produced is greater than or equal to 0)}$$

The complete problem may now be restated mathematically as

$$\text{Maximize profit} = \text{N}70T + \text{N}50C$$

subject to the constraints

$$4T + 3C \leq 240 \text{ (carpentry constraint)}$$

$$2T + 1C \leq 100 \text{ (painting and varnishing constraint)}$$

$$T \geq 0 \text{ (first nonnegativity constraint)}$$

$$C \geq 0 \text{ (second nonnegativity constraint)}$$

While the nonnegativity constraints are technically separate constraints, they are often written on a single line with the variables separated by commas. In this example, this would be written as:

$$T, C \geq 0$$

The easiest way to solve a small LP problem such as that of the Flair Furniture Company is with the graphical solution approach. The graphical procedure is useful only when there are two decision variables (such as number of tables to produce, T , and number of chairs to produce, C) in the problem. When there are more than two variables, it is not possible to plot the solution on a two-dimensional graph and we must turn to more complex approaches. But the graphical method is invaluable in providing us with insights into how other approaches work.

To find the optimal solution to an LP problem, we must first identify a set, or region, of feasible solutions. The first step in doing so is to plot each of the problem's constraints on a graph. The variable T (tables) is plotted as the horizontal axis of the graph and the variable C (chairs) is plotted as the vertical axis. The notation (T, C) is used to identify the points on the graph. The **nonnegativity constraints** mean that we are always working in the first (or northeast) quadrant of a graph.

To represent the first constraint graphically, $4T + 3C \leq 240$, we must first graph the equality portion of this, which is

$$4T + 3C = 240$$

As you may recall from elementary algebra, a linear equation in two variables is a straight line. The easiest way to plot the line is to find any two points that satisfy the equation, then draw a straight line through them. The two easiest points to find are generally the points at which the line intersects the T and C axes.

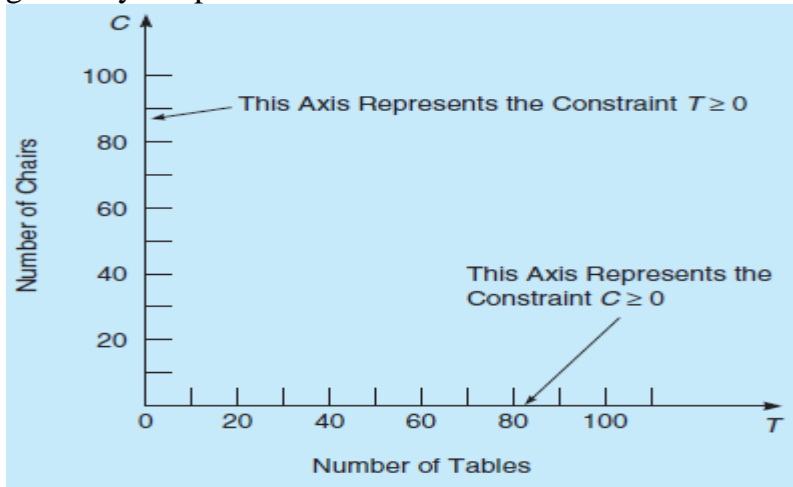


Figure 1: Quadrant Containing All Positive Values

When Flair Furniture produces no tables, namely $T = 0$, it implies that

$$4(0) + 3C = 240$$

$$3C = 240$$

$$C = 80$$

In other words, if *all* of the carpentry time available is used to produce chairs, 80 chairs *could* be made. Thus, this constraint equation crosses the vertical axis at 80. To find the point at which the line crosses the horizontal axis, we assume that the firm makes no chairs, that is, $C = 0$, Then

$$4T + 3(0) = 240$$

$$4T = 240$$

$$T = 60$$

Hence, when $C = 0$, we see that and that $4T = 240$ and that $T = 60$.

The carpentry constraint is illustrated in Figure 1. It is bounded by the line running from point $(T = 0, C = 80)$ to point $(T = 60, C = 0)$.

Recall, however, that the actual carpentry constraint was the **inequality** $4T + 3C \leq 240$. How can we identify all of the solution points that satisfy this constraint? It turns out that there are three possibilities. First, we know that any point that lies on the line $4T + 3C = 240$ satisfies the constraint. Any combination of tables and chairs on the line will use up all 240 hours of carpentry time. Now we must find the set of solution points that would use less than the 240 hours. The points that satisfy the portion of the \leq constraint (i.e., $4T + 3C < 240$) will be all the points on one side of the line, while all the points on the other side of the line will not satisfy this condition. To determine which side of the line this is, simply choose any point on either side of the constraint line shown in Figure 2 and check

to see if it satisfies this condition. For example, choose the point (30, 20), as illustrated in Figure 3:

$$4(30) + 3(20) = 180$$

Since $180 < 240$, this point satisfies the constraint, and all points on this side of the line will also satisfy the constraint. This set of points is indicated by the shaded region in Figure 3.

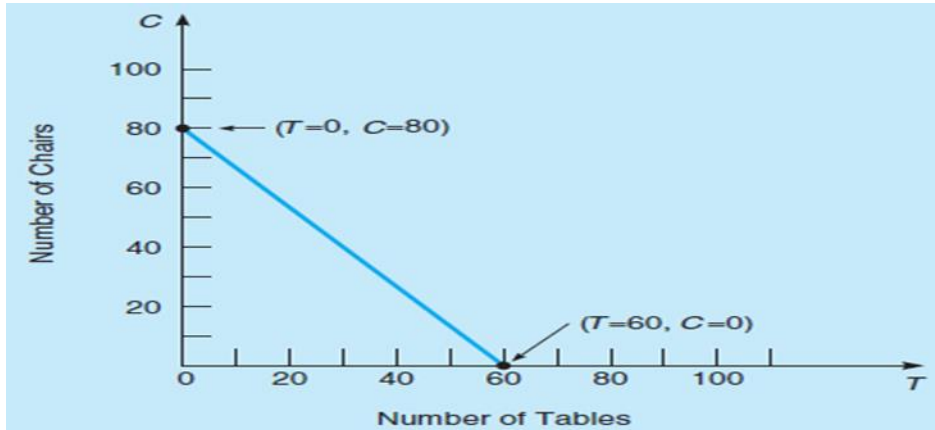


Figure 2: Graph of Carpentry Constraint Equation $4T + 3C = 240$

To see what would happen if the point did not satisfy the constraint, select a point on the other side of the line, such as (70, 40). This constraint would not be met at this point as

$$4(70) + 3(40) = 400$$

Since $400 > 240$, this point and every other point on that side of the line would not satisfy this constraint. Thus, the solution represented by the point would require more than the 240 hours that are available. There are not enough carpentry hours to produce 70 tables and 40 chairs.

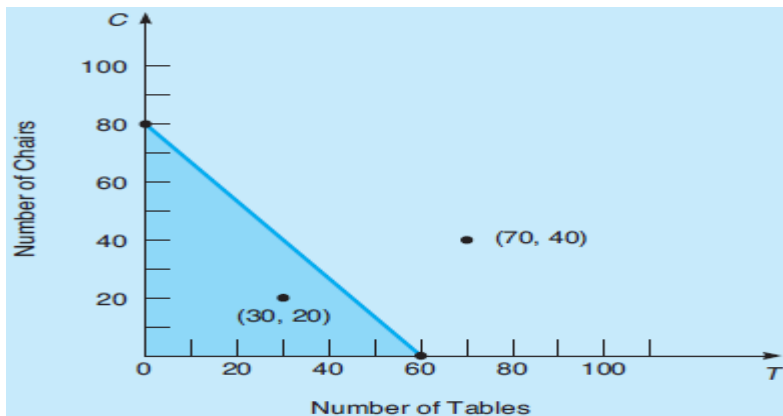


Figure 3: Region that satisfies the Carpentry Constraint

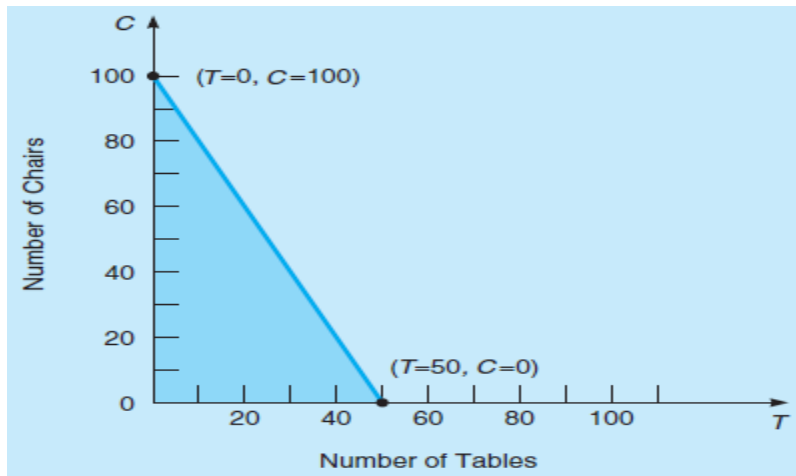


Figure 4: Region that Satisfies the Painting and Varnishing Constraint

Next, let us identify the solution corresponding to the second constraint, which limits the time available in the painting and varnishing department. That constraint was given as $2T + 1C \leq 100$. As before, we start by graphing the equality portion of this constraint, which is

$$2T + 1C = 100$$

To find two points on the line, select $T = 0$ and solve for C :

$$2(0) + 1C = 100$$

$$C = 100$$

So, one point on the line is $(0, 100)$. To find the second point, select $C = 0$ and solve for T :

$$2T + 1(0) = 100$$

$$T = 50$$

The second point used to graph the line is $(50, 0)$. Plotting this point, $(50, 0)$, and the other point, $(0, 100)$, results in the line representing all the solutions in which exactly 100 hours of painting and varnishing time are used, as shown in Figure 4.

To find the points that require less than 100 hours, select a point on either side of this line to see if the inequality portion of the constraint is satisfied. Selecting $(0, 0)$ give us

$$2(0) + 1(0) = 0 < 100$$

This indicates that this and all the points below the line satisfy the constraint, and this region is shaded in Figure 4.

Now that each individual constraint has been plotted on a graph, it is time to move on to the next step. We recognize that to produce a chair or a table, both the carpentry and painting and varnishing departments must be used. In an LP problem we need to find that set of solution points that satisfies all of the constraints *simultaneously*. Hence, the constraints should be redrawn on one graph (or superimposed one upon the other). This is shown in Figure 5.

The shaded region now represents the area of solutions that does not exceed either of the two Flair Furniture constraints. It is known by the term *area of feasible solutions* or, more simply, the **feasible region**. The feasible region in an LP problem must satisfy *all* conditions specified by the problem's constraints, and is thus the region where all constraints overlap. Any point in the region would be a **feasible solution** to the Flair Furniture problem; any point outside the shaded area would represent an **infeasible solution**. Hence, it would be feasible to

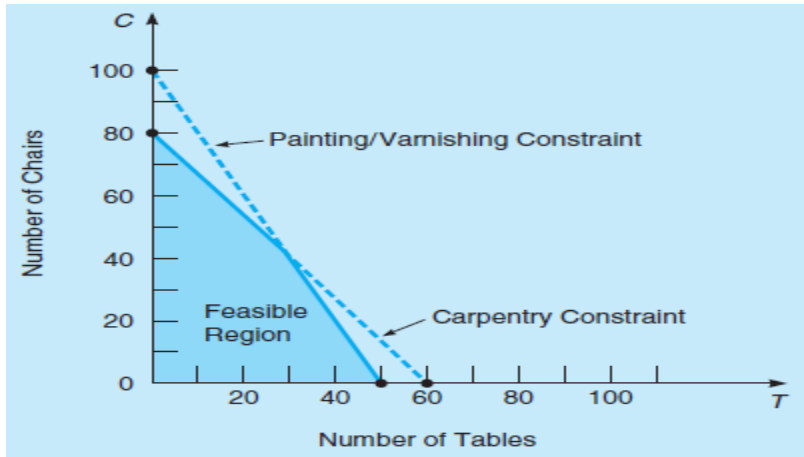


Figure 5: Feasible Solution Region for the Flair Furniture Company Problem

Manufacture 30 tables and 20 chairs ($T = 30, C = 20$) during a production period because both constraints are observed:

Carpentry Constraint	$4T + 3C \leq 240$ hours available
	$(4)(30) + (3)(20) = 180$ hours used
Painting Constraint	$2T + 1C \leq 100$ hours available
	$(2)(30) + (1)(20) = 80$ hours used

But it would violate both of the constraints to produce 70 tables and 40 chairs, as we see here mathematically:

Carpentry Constraint	$4T + 3C \leq 240$ hours available
	$(4)(70) + (3)(40) = 400$ hours used
Painting Constraint	$2T + 1C \leq 100$ hours available
	$(2)(70) + (1)(40) = 180$ hours used

Furthermore, it would also be infeasible to manufacture 50 tables and 5 chairs ($T = 50, C = 5$). Can you see why?

Carpentry Constraint	$4T + 3C \leq 240$ hours available
	$(4)(50) + (3)(5) = 205$ hours used
Painting Constraint	$2T + 1C \leq 100$ hours available
	$(2)(50) + (1)(5) = 105$ hours used

This possible solution falls within the time available in carpentry but exceeds the time available in painting and varnishing and thus falls outside the feasible region.

SELF-ASSESSMENT EXERCISE

1. Use the graphical method to find a feasible region for these constraints $X + 2Y \leq 3$ and $X + Y \leq 3$.

4.0 CONCLUSION

In this unit, we examined the requirements that is needed in constructing a linear program problem. Also, we went further to develop a simple linear program problem and then used the graphical method to solve the solution of a typical linear programming problem.

4.0 SUMMARY

In this unit, we have discussed the requirements that is needed in constructing a linear program problem. We further showed that the properties of a linear function problem are one objective function, one or more constraints, alternative courses of action, objective function and constraints are linear—proportionality and divisibility, certainty, divisibility and Nonnegative variables. Also, we went further to develop a simple linear program problem and then used the graphical method to solve the solution of a typical linear programming problem by identifying the feasible regions.

6.0 TUTOR-MARKED ASSIGNMENT

1. Mention and explain the properties of a linear program.
2. Kelechi warehouses is planning to expand its successful Cashew business into Tampa. In doing so, the company must determine how many storage rooms of each size to build. Its objective and constraints follow:

$$\text{Maximize monthly earnings} = 50X_1 + 20X_2$$

$$\text{subject to} \quad 2X_1 + 4X_2 \leq 400 \quad (\text{advertising budget available})$$

$$100X_1 + 50X_2 \leq 8,000 \quad (\text{square footage required})$$

$$X_1 \leq 60 \quad (\text{rental limit expected})$$

$$X_1, X_2 \geq 0$$

Where

X_1 = number of large spaces developed

X_2 = number of small spaces developed

Find the feasible combination of X_1 and X_2 to achieve the objective.

7.0 REFERENCES/FURTHER READINGS

Render, B., Stair, R.M., & Hanna, M.E. (2012). *Quantitative Analysis for Management* (11th ed.). Essex: England, Pearson.

Murthy, P.R. (2007). *Operations Research* (2nd ed.). New Delhi, India: New Age International ltd.

UNIT 3 SIMPLEX METHOD

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Conditions to be met before Applying the Simplex Method
 - 3.2 Steps involved in Simplex Method
 - 3.3 Classical Illustration of the Simplex Method
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0. INTRODUCTION

The Simplex Method or Simplex Algorithm is used for calculating the optimal solution to the linear programming problem. In other words, the simplex algorithm is an iterative procedure carried systematically to determine the optimal solution from the set of feasible solutions. Firstly, to apply the simplex method, appropriate variables are introduced in the linear programming problem, and the primary or the decision variables are equated to zero. The iterative process begins by assigning values to these defined variables. The value of decision variables is taken as zero since the evaluation in terms of the graphical approach begins with the origin. Therefore, x_1 and x_2 is equal to zero.

The decision maker will enter appropriate values of the variables in the problem and find out the variable value that contributes maximum to the objective function and removes those values which give undesirable results. Thus, the value of the objective function gets improved through this method. This procedure of substitution of variable value continues until any further improvement in the value of the objective function is possible.

2.0. OBJECTIVES

At the end of this unit, you should be able to:

- Understand the conditions that should be met before applying the simplex method
- Understand the steps to be followed in solving a linear program using the simplex method.
- Solve a linear programming problem using the simplex method.

3.0 MAIN CONTENT

3.1 Conditions to be met before Applying the Simplex Method

To be able to solve a simplex programming problem, the following conditions must be satisfied:

1. The right-hand side of each constraint inequality should be non-negative. In case, any linear programming problem has a negative resource value, then it should be converted into positive value by multiplying both the sides of constraint inequality by “-1”.
2. The decision variables in the linear programming problem should be non-negative. Thus, the simplex algorithm is efficient since it considers few feasible solutions, provided by the corner points, to determine the optimal solution to the linear programming problem.

SELF ASSESSMENT EXERCISE

Explain the conditions to be met before applying the simplex method in solving a linear programming problem.

3.2 Steps involved in Simplex Method

The first step of the simplex method requires that we convert each inequality constraint in an LP formulation into an equation. Less-than-or-equal-to constraints (\leq) can be converted to equations by adding *slack variables*, which represent the amount of an unused resource. The other steps are outlined below:

Step One: Present your problem in the initial tableau

Step Two: Find the column with the largest negative indicator on the bottom row.

Step Three: Divide all possible entries in this column into the corresponding entries of the final column

Step Four: Choose the entry which gives the smallest answer in step 3 as the pivot element.

Step Five: Use an elementary row transformation to change the value of the pivot element to 1 or divide all value in the row by an appropriate constant.

Step Six: Use elementary row transformation to change the other values in the pivot column to zero by adding or subtracting the pivot elements the appropriate number of times.

Step Seven: Examine the resultant values in the bottom row. If there are no negative indicators, the final solution has been reached, stop; Otherwise; return to step 2 and repeat the procedure.

SELF ASSESSMENT EXERCISE

Explain the steps involved in solving a linear programming problem using the simplex approach.

3.3 Classical Illustration of the Simplex Method

We shall examine the example below in order to understand how linear programming problem can be solved using the simplex method. There are three factors of production used to produce good p and q, the factors are labour, physical capital and technology. 1.5units of labour, 1 unit of physical capital and 1 unit of technology is needed to produce good p while 2 units of labour, 3 units of physical capital is needed to produce good q. if the total available resources in the factory are 2400 units of labour, 2400 units of physical capital and 1200 units of technology. What will be the maximum output of good p and q produced to maximize profit if the selling price of good p is ₦200 and the selling price of good q is ₦300. What will be the maximum profit earned?

In order to solve this problem, we first classify the problem into a presentable form:

Let labour = L

Physical Capital = PC

Technology =T

The objective function is categorized as:

$$\text{Maximize profit} = 200p + 300q$$

$$\text{subject to } \frac{3}{2}p + 2q \leq 2,400 \quad (\text{labour available})$$

$$p + 3q \leq 2,400 \quad (\text{Physical Capital required})$$

$$p \leq 1,200 \quad (\text{Technology Required})$$

$$p, q \geq 0 \quad (\text{Non - negativity Constraint})$$

We then follow the steps below

Step One:

Present your problem in the initial tableau. The information is below:

		C ₁	C ₂	C ₃	C ₄	C ₅	C ₅
	Inputs	P	Q	S ₁	S ₂	S ₃	Constraint
R ₁	L	3/2	2	1	0	0	2400
R ₂	PC	1	3	0	1	0	2400
R ₃	T	1	0	0	0	1	1200
R ₄	Profit	-200	-300	0	0	0	0

Step Two:

Find the column with the largest negative indicator on the bottom row. The column is C₂, with a negative indicator of -300.

Step Three:

Divide all possible entries in this column into the corresponding entries of the final column. The division is:

we have $\frac{C_5}{C_2}$,

$$\text{for } R_1, \frac{2400}{2} = 1200$$

$$\text{for } R_2, \frac{2400}{3} = 800$$

$$\text{for } R_3, \frac{1200}{0} = \infty$$

Step Four:

Choose the entry which gives the smallest answer in step 3 as the pivot element. The entry with the smallest number is row 2 and the pivot element is now 3.

Step Five:

Use an elementary row transformation to change the value of the pivot element to 1 or divide all value in the row by an appropriate constant.

This can be done by dividing the fourth row all through by 3.

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₅
Inputs	P	Q	S ₁	S ₂	S ₃	Constraint
PC	$\frac{1}{3}$	1	0	$\frac{1}{3}$	0	800

Step Six:

Use elementary row transformation to change the other values in the pivot column to zero by adding or subtracting the pivot elements the appropriate number of times.

This can be done by:

$$\text{For } R_1, \frac{3}{2} - 2\left(\frac{1}{3}\right) = \frac{5}{6}$$

$$2 - 2(1) = 0$$

$$1 - 2(0) = 1$$

$$0 - 2\left(\frac{1}{3}\right) = -\frac{2}{3}$$

$$0 - 2(0) = 0$$

$$2400 - 2(800) = 800$$

Row 3 is already zero and does not need any transformation.

$$\text{For } R_4, -200 + 300\left(\frac{1}{3}\right) = -100$$

$$-300 + 300(1) = 0$$

$$0 + 300(0) = 0$$

$$0 + 300\left(\frac{1}{3}\right) = 100$$

$$0 + 300(0) = 0$$

$$0 + 300(800) = 240,000$$

Step Seven:

Examine the resultant values in the bottom row. If there are no negative indicators, the final solution has been reached, Stop; Otherwise; return to step 2 and repeat the procedure.

		C ₁	C ₂	C ₃	C ₄	C ₅	C ₅
	Inputs	P	Q	S ₁	S ₂	S ₃	Constraint
R ₁	L	$\frac{5}{6}$	0	1	$-\frac{2}{3}$	0	800
R ₂	PC	$\frac{1}{3}$	1	0	$\frac{1}{3}$	0	800
R ₃	T	1	0	0	0	1	1,200
R ₄	Profit	-100	0	0	100	0	240,000

Since the last column is -100, we return to step 2.

Rerun of Step Two: Find the column with the largest negative indicator on the bottom row. The column is C₁, with a negative indicator of -100.

Rerun of Step Three: Divide all possible entries in this column into the corresponding entries of the final column. The division is:

$$\text{we have } \frac{C_5}{C_1},$$

$$\text{for } R_1, \frac{800}{\frac{5}{6}} = 960$$

$$\text{for } R_2, \frac{800}{\frac{1}{3}} = 2400$$

$$\text{for } R_3, \frac{1200}{1} = 1200$$

Rerun of Step Four:

Choose the entry which gives the smallest answer in step 3 as the pivot element. The entry with the smallest number is row 1 and the pivot element is now $\frac{5}{6}$.

Rerun of Step Five:

Use an elementary row transformation to change the value of the pivot element to 1 or divide all value in the row by an appropriate constant.

This can be done by dividing the fourth row all through by 3.

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₅
Inputs	P	Q	S ₁	S ₂	S ₃	Constraint
L	1	0	$\frac{6}{5}$	$-\frac{4}{5}$	0	960

Rerun of Step Six:

Use elementary row transformation to change the other values in the pivot column to zero by adding or subtracting the pivot elements the appropriate number of times.

This can be done by:

$$\text{For } R_2, \frac{1}{3} - 1\left(\frac{1}{3}\right) = 0$$

$$1 - 0\left(\frac{1}{3}\right) = 1$$

$$0 - \frac{1}{3}\left(\frac{6}{5}\right) = -\frac{2}{5}$$

$$\frac{1}{3} - \frac{1}{3}\left(-\frac{4}{5}\right) = \frac{3}{5}$$

$$0 - 0\left(\frac{1}{3}\right) = 0$$

$$800 - \frac{1}{3}(960) = 480$$

$$\text{For } R_3, 1 - 1(1) = 0$$

$$0 - 1(0) = 0$$

$$0 - 1\left(\frac{6}{5}\right) = -\frac{6}{5}$$

$$0 - 1\left(-\frac{4}{5}\right) = \frac{4}{5}$$

$$1 - 1(0) = 1$$

$$1200 - 1(960) = 240$$

$$\text{For } R_4, -100 + 100(1) = 0$$

$$0 + 100(0) = 0$$

$$0 + 100\left(\frac{6}{5}\right) = 120$$

$$100 + 100\left(-\frac{4}{5}\right) = 20$$

$$0 + 100(0) = 0$$

$$240,000 + 100(960) = 336,000$$

Rerun of Step Seven:

Examine the resultant values in the bottom row. If there are no negative indicators, the final solution has been reached, Stop; Otherwise; return to step 2 and repeat the procedure.

		C ₁	C ₂	C ₃	C ₄	C ₅	C ₅
	Inputs	P	Q	S ₁	S ₂	S ₃	Constraint
R ₁	L	1	0	$\frac{6}{5}$	$-\frac{4}{5}$	0	960
R ₂	PC	0	1	$-\frac{2}{5}$	$\frac{3}{5}$	0	480
R ₃	T	0	0	$-\frac{6}{5}$	$\frac{4}{5}$	1	240
R ₄	Profit	0	0	120	20	0	336,000

From the final tableau, since there is no negative value in the final column, the final solution has been reached. The maximum profit is ₦336,000 and 960 units of labour is employed, 480 units of physical capital and 240 units of technology is employed.

SELF ASSESMENT EXERCISE

State the steps to follow in using simplex approach to solve linear programming problem.

4.0. CONCLUSION

In this unit, we examined simplex approach to linear programming optimization and the various steps to employ in linear programming. The unit further demonstrated how the simplex method can be solved using an iterative method.

5.0 SUMMARY

In this unit, we have discussed simplex approach to linear programming optimization and the various steps to employ in linear programming. The steps to follow are Present your problem in the initial tableau, Find the column with the largest negative indicator on the bottom row, Divide all possible entries in this column into the corresponding entries of the final column, Choose the entry which gives the smallest answer in step 3 as the pivot element, Use an elementary row transformation to change the value of the pivot element to 1 or divide all value in the row by an appropriate constant, Use elementary row transformation to change the other values in the pivot column to zero by adding or subtracting the pivot elements the appropriate number of times, Examine the resultant values in the bottom row. If there are no negative indicators, the final solution has been reached, Stop; Otherwise; return to step 2 and repeat the procedure. The unit further demonstrated how the simplex method can be solved using an iterative method.

6.0 TUTOR MARKED ASSIGNMENT

Winkler Furniture manufactures two different types of china cabinets: a French Provincial model and a Danish Modern model. Each cabinet produced must go through three departments: carpentry, painting, and finishing. The table below contains all relevant information concerning production times per cabinet produced and production capacities for each operation per day, along with net revenue per unit produced. The firm has a contract with an Indiana distributor to produce a minimum of 300 of each cabinet per week (or 60 cabinets per day). Owner Bob Winkler would like to determine a product mix to maximize his daily revenue. Formulate as an LP problem.

7.0 REFERENCES/FURTHER READINGS

- Render, B., Stair, R.M., & Hanna, M.E. (2012). *Quantitative Analysis for Management* (11th ed.). Essex: England, Pearson.
- Murthy, P.R. (2007). *Operations Research* (2nd ed.). New Delhi, India: New Age International ltd.

UNIT 4: TRANSPORTATION MODEL

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 The Transportation Problem
 - 3.2 Setting up a Transportation problem
 - 3.3 Mathematical Model of a Transportation Problem
 - 3.4 Initial solution - Northwest Corner Rule
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0. INTRODUCTION

In this unit we extend the theory of linear programming to a special linear programming problem, the Transportation Problem. This problem can be solved by the simplex algorithm, but the process would result in very large simplex tableaux and numerous simplex iterations. Because of the special characteristics of each problem, however, alternative solution methods requiring significantly less mathematical manipulation have been developed.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- Understand a typical transportation model
- Set up a transportation model
- Set up the mathematical model of a transportation problem
- Solve a transportation model using the Northwest Corner Rule

3.0 MAIN CONTENT

3.1 The Transportation Problem

The general transportation problem is concerned with determining an optimal strategy for distributing a commodity from a group of supply centres, such as factories, called sources, to various receiving centers, such as warehouses, called destinations, in such a way as to minimise total distribution costs.

Each source is able to supply a fixed number of units of the product, usually called the capacity or availability, and each destination has a fixed demand, often called the requirement. Transportation models can also be used when a firm is trying to decide where to locate a new facility. Good financial decisions concerning facility location also attempt to minimize total transportation and production costs for the entire system.

The transportation problem deals with the distribution of goods from several points of supply (*origins* or sources) to a number of points of demand (destinations). Usually we are given a capacity (supply) of goods at each source, a requirement (demand) for goods at each destination, and the shipping cost per unit from each source to each destination. An example is shown in Figure 6. The objective of such a problem is to schedule shipments so that total transportation costs are minimized. At times, production costs are included also. Transportation models can also be used when a firm is trying to decide where to locate a new facility. Before opening a new warehouse, factory, or sales office, it is good practice to consider a number of alternative sites. Good financial decisions concerning the facility location also attempt to minimize total transportation and production costs for the entire system.

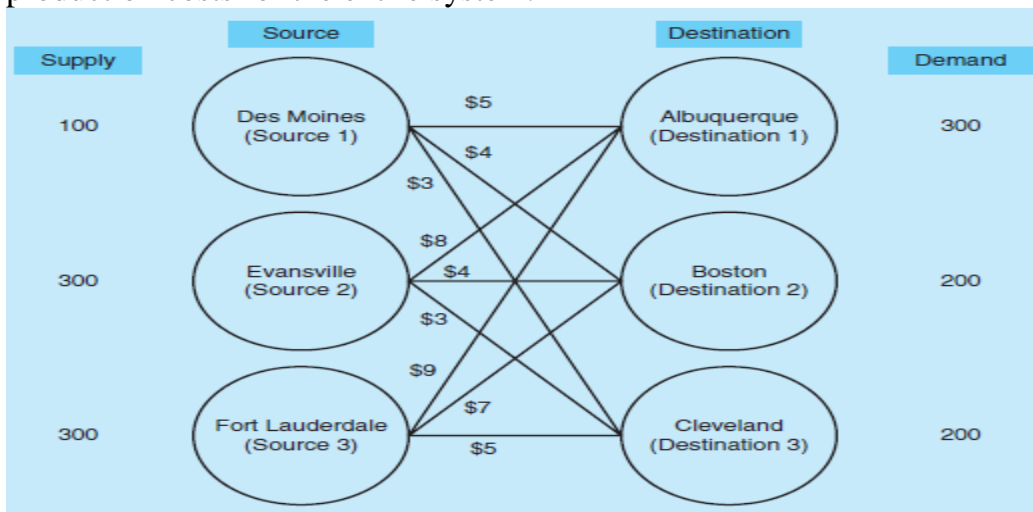


Figure 6: Network Representation of a Transportation problem, with Costs, Demands, and Supplies

SELF ASSESSMENT EXERCISE

Give 5 typical economic and managerial decision problems in which transportation model can be used to resolve the issue.

3.2. Setting up a Transportation Problem

To illustrate how to set up a transportation problem we consider the following example.

Example 1.

A concrete company transports concrete from three plants, 1, 2 and 3, to three construction sites, A, B and C. The plants are able to supply the following numbers of tons per week:

Table 1

Plant	Supply (Capacity)
1	300
2	300
3	100

The requirements of the sites, in number of tons per week, are:

Table 2

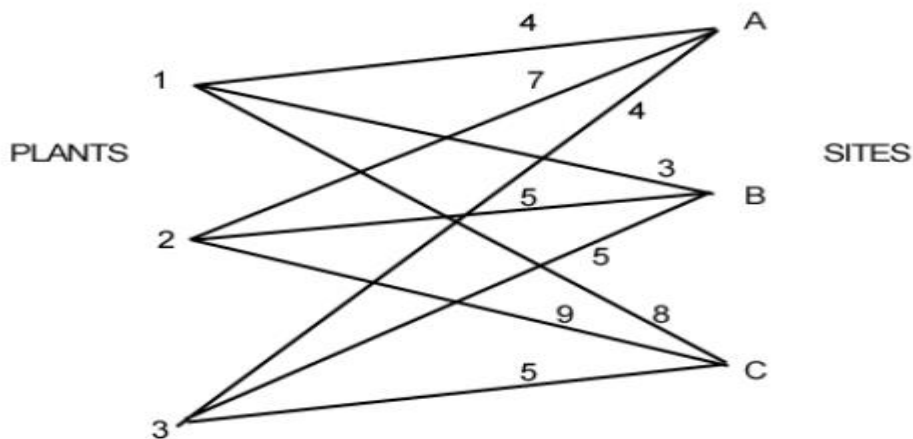
Construction site	Supply (Capacity)
A	200
B	200
C	300

The cost of transporting 1 ton of concrete from each plant to each site is shown in the figure 7 in Emalangen per ton. For computational purposes it is convenient to put all the above information into a table, as in the simplex method. In this table each row represents a source and each column represents a destination.

Table 3

		Sites				
		To	A	B	C	Supply (Availability)
From						
1		4	3	8	300	
2		7	5	9	300	
3		4	5	5	100	
	Demand (Requirement)	200	200	300		

Plants

**Figure 7: Constructing a Transportation Problem**

SELF ASSESSMENT EXERCISE

Set up this transportation model. A flower company transports flour from three plants, 1, 2 and 3, to three bakery shops, A, B and C. The plants are able to supply the following numbers of tons per week:

Table 4

Plant	Supply (Capacity)
1	800
2	1250
3	1100

The requirements of the sites, in number of tons per week, are:

Table 5

Bakery	Supply (Capacity)
A	1650
B	1050
C	450

It cost ₦80 to transport a 1 kg of flour from plant 1 to bakery A, ₦39 from plant 1 to bakery B and ₦73 from plant 1 to bakery C. it costs ₦72 to transport 1 kg of flour from plant 2 to bakery A, ₦32 to transport same 1 kg from plant 2 to bakery B and then ₦55 from plant 2 to bakery C. It however costs ₦123 to transport 1 kg of flour from plant 3 to bakery A, ₦28 from plant 3 to bakery B and ₦55 from plant 3 to bakery C.

3.3. Mathematical Model of a Transportation Problem

Before we discuss the solution of transportation problems we will introduce the notation used to describe the transportation problem and show that it can be formulated as a linear programming problem. We use the following notation;

x_{ij} = the number of units to be distributed from source i to destination j

($i = 1, 2, \dots, m; j = 1, 2, \dots, n$)

s_i = supply from source i ;

d_j = demand at destination j ;

c_{ij} = cost per unit distributed from source i to destination j

With respect to Example 4.1 the decision variables x_{ij} are the numbers of tons transported from plant i (where $i = 1, 2, 3$) to each site j (where $j = A, B, C$).

A basic assumption is that the distribution costs of units from source i to destination j is directly proportional to the number of units distributed. A typical cost and requirements table has the form shown on Table 1.

Let Z be total distribution costs from all the m sources to the n destinations. In example 1 each term in the objective function Z represents the total cost of tonnage transported on one route. For example, in the route $2 \rightarrow C$, the term is $9x_{2c}$, that is:

$$(\text{Cost per ton} = 9) \times (\text{number of tons transported} = x_{2c})$$

	1	2	...	n	Supply
1	c_{11}	c_{12}	...	c_{1n}	s_1
2	c_{21}	c_{22}	...	c_{2n}	s_2
Source
.
.
m	c_{m1}	c_{m2}	...	c_{mn}	s_m
Demand	d_1	d_2	...	d_n	

Table 6: Cost and requirements table

Hence the objective function is:

$$\begin{aligned} Z = & 4x_{1A} + 3x_{1B} + 8x_{1C} \\ & + 7x_{2A} + 5x_{2B} + 9x_{2C} \\ & + 4x_{3A} + 5x_{3B} + 5x_{3C} \end{aligned}$$

Notice that in this problem the total supply is $300 + 300 + 200 = 700$ and the total demand is $200 + 200 + 300 = 700$. Thus

$$\text{Total supply} = \text{total demand.}$$

In mathematical form this expressed as:

$$\sum_{i=1}^m s_i = \sum_{j=1}^n d_j$$

This is called a balanced problem. In this unit our discussion shall be restricted to the balanced problems.

In a balanced problem all the products that can be supplied are used to meet the demand. There are no slacks and so all constraints are equalities rather than inequalities as was the case in the previous unit.

The formulation of this problem as a linear programming problem is presented as

$$\text{Minimize } Z = \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} \quad (1)$$

Subject to

$$\sum_{j=1}^n x_{ij} = s_i \quad \text{for } i = 1, 2, \dots, m \quad (2)$$

$$\sum_{i=1}^m x_{ij} = d_j \quad \text{for } j = 1, 2, \dots, n \quad (3)$$

and

$$x_{ij} \geq 0, \text{ for all } i \text{ and } j$$

Any linear programming problem that fits this special formulation is of the transportation type, regardless of its physical context. For many applications, the supply and demand quantities in the model will have integer values and implementation will require that the distribution quantities also be integers. Fortunately, the unit coefficients of the unknown variables in the constraints guarantee an optimal solution with only integer values.

SELF ASSESSMENT EXERCISE

In converting the transportation problem into mathematical form, what is the relationship between the total demand and the total supply?

3.4 Initial solution - Northwest Corner Rule

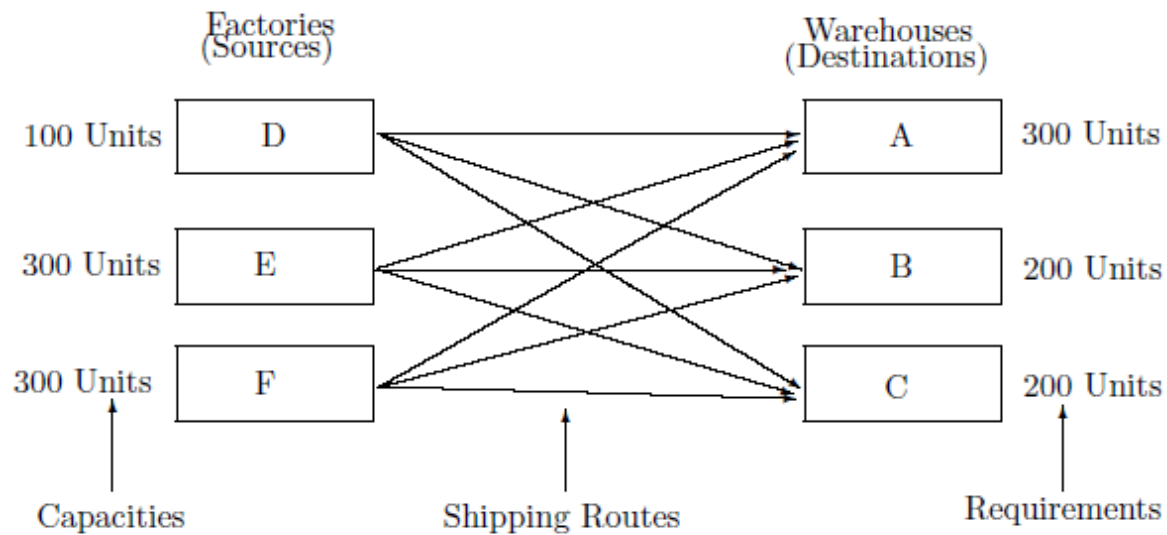
The initial basic feasible solution can be obtained by using one of several methods. We will consider only the North West corner rule of developing an initial solution. Other methods can be found in standard texts on linear programming.

The procedure for constructing an initial basic feasible solution selects the basic variables one at a time. The North West corner rule begins with an allocation at the top left-hand corner of the tableau and proceeds systematically along either a row or a column and make allocations to subsequent cells until the bottom right-hand corner is reached, by which time enough allocations will have been made to constitute an initial solution.

The procedure for constructing an initial solution using the North West Corner rule is as follows:

1. Start by selecting the cell in the most "North-West" corner of the table.
2. Assign the maximum amount to this cell that is allowable based on the requirements and the capacity constraints.
3. Exhaust the capacity from each row before moving down to another row.
4. Exhaust the requirement from each column before moving right to another column.
5. Check to make sure that the capacity and requirements are met.

Let us begin with an example dealing with Executive Furniture Corporation, which manufactures office desks at three locations: D, E and F. The firm distributes the desks through regional warehouses located in A, B and C (see the Network format diagram below)



It is assumed that the production costs per desk are identical at each factory. The only relevant costs are those of shipping from each source to each destination. The costs are shown in Table 7.

From \ To	A	B	C
D	₦5	₦4	₦3
E	₦8	₦4	₦3
F	₦9	₦7	₦5

Table 7: Transportation Costs per desk for Executive Furniture Corp.

We proceed to construct a transportation table and label its various components as show in Table 8. We can now use the Northwest corner rule to find an initial feasible solution to the problem. We start in the upper left hand cell and allocate units to shipping routes as follows:

From \ To	A	B	C	Capacity
D	5	4	3	100
E	8	4	3	300
F	9	7	5	300
Requirements	300	200	200	700

Table 8: Transportation Table for Executive Furniture Corporation

1. Exhaust the supply (factory capacity) of each row before moving down to the next row.
2. Exhaust the demand (warehouse) requirements of each column before moving to the next column to the right.
3. Check that all supply and demand requirements are met.
4. The initial shipping assignments are given in Table 9

From \ To	A	B	C	Factory (Capacity)
D	100			100
E	200	100		300
F		100	200	300
Warehouse (Requirement)	300	200	200	

Table 9: Initial Solution of the North West corner Rule

This initial solution can also be presented together with the costs per unit as shown in the Table 8. We can compute the cost of this shipping assignment as follows; therefore, the initial feasible solution for this problem is ₦4200.

Optimality Test - the Stepping Stone method

The next step is to determine whether the current allocation at any stage of the solution process is optimal. We will present one of the methods used to determine optimality of and improve a current solution. The method derives its name from the analogy of crossing a pond using stepping stones. The occupied cells are analogous to the stepping stones, which are used in making certain movements in this method.

The five steps of the Stepping-Stone Method are as follows:

1. Select an unused square to be evaluated.
2. Beginning at this square, trace a closed path back to the original square via squares that are currently being used (only horizontal or vertical moves allowed). You can only change directions at occupied cells.
3. Beginning with a plus (+) sign at the unused square, place alternative minus (-) signs and plus signs on each corner square of the closed path just traced.
4. Calculate an improvement index, I_{ij} by adding together the unit cost figures found in each square containing a plus sign and then subtracting the unit costs in each square containing a minus sign.
5. Repeat steps 1 to 4 until an improvement index has been calculated for all unused squares.
 - a. If all indices computed are greater than or equal to zero, an optimal solution has been reached.

- b. If not, it is possible to improve the current solution and decrease total shipping costs.

SELF ASSESSMENT

Briefly explain the five steps of the Stepping-Stone Method.

4.0. CONCLUSION

In this unit, we explained the transportation problem, how to set up a transportation problem, deriving the mathematical model of a transportation problem as well as solving the initial solution - northwest corner rule. We also solved a practical example of how to allocate using the northwest corner solution.

5.0. SUMMARY

This unit explained transportation problem, how to set up a transportation problem, deriving the mathematical model of a transportation problem as well as solving the initial solution - northwest corner rule. The procedure for constructing an initial solution using the North West Corner rule is start by selecting the cell in the most "North-West" corner of the table, then assign the maximum amount to this cell that is allowable based on the requirements and the capacity constraints, exhaust the capacity from each row before moving down to another row, exhaust the requirement from each column before moving right to another column and check to make sure that the capacity and requirements are met.

6.0. TUTOR-MARKED ASSIGNMENT

Use the northwest corner solution to allocate the following and then use the stepping stone method to check if optimality is reached.

To \ From	A	B	C	D	Supply
1	10	30	25	15	14
2	20	15	20	10	10
3	10	30	20	20	15
4	30	40	35	45	13
Demand	10	15	12	15	

7.0 REFERENCES/FURTHER READINGS

Render, B., Stair, R.M., & Hanna, M.E. (2012). *Quantitative Analysis for Management* (11th ed.). Essex: England, Pearson.

Murthy, P.R. (2007). *Operations Research* (2nd ed.). New Delhi, India: New Age International Ltd.

Online Source from

http://www.maths.unp.ac.za/coursework/MATH331/2012/transportation_assignment.pdf

MODULE 3: FORECASTING, DECISION AND INVENTORY ANALYSIS

Unit 1: Forecasting and Decision Analysis

Unit 2: Demonstrate Forecasting Methods

Unit 3 Deterministic Inventory Control Models

UNIT ONE: FORECASTING AND DECISION ANALYSIS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Forecasting analysis
 - 3.1.1 Types of Forecasts
 - 3.1.2 Scatter Diagrams and Time Series
 - 3.1.3 Measures of Forecast Accuracy
 - 3.1.4 Time-Series Forecasting Models
 - 3.1.5 Monitoring and Controlling Forecasts
 - 3.2 Decision Analysis
 - 3.2.1 The Steps in Decision Making
 - 3.2.2 Types of Decision-Making Environments
 - 3.2.3 Decision Making Under Uncertainty
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

Every day, economists make decisions without knowing what will happen in the future. Inventory is ordered though no one knows what sales will be, new equipment is purchased though no one knows the demand for products, and investments are made though no one knows what profits will be. Economists are always trying to reduce this uncertainty and to make better estimates of what will happen in the future. Accomplishing this is the main purpose of forecasting. Decision theory is an analytic and systematic approach to the study of decision making. In this unit, we present the mathematical models useful in helping managers make the best possible decisions. What makes the difference between good and bad decisions? A good decision is one that is based on logic, considers all available data and possible alternatives, and applies the quantitative approach we are about to describe. Occasionally, a good decision results in an unexpected or unfavourable outcome while, a bad decision is one that is not based on logic, does not use all available information, does not consider all alternatives, and does not employ appropriate quantitative techniques.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Explain and understand the types of forecasts models
- Plot a scatter diagram with a time series.
- Measure forecast accuracy
- Understand time series forecasting models
- Monitor and control forecasts
- Understand the steps in Decision Making
- Know the types of decision-making environments.

3.0 MAIN CONTENT

3.1 Forecasting Analysis

3.1.1 Types of Forecasts

In this unit, we consider forecasting models that can be classified into one of three categories: time-series models, causal models, and qualitative models

1. Time-Series Models

Time-series models attempt to predict the future by using historical data. These models make the assumption that what happens in the future is a function of what has happened in the past. In other words, time-series models look at what has happened over a period of time and use a series of past data to make a forecast. Thus, if we are forecasting weekly sales for lawn mowers, we use the past weekly sales for lawn mowers in making the forecast. The time-series models we examine in this unit are moving average, exponential smoothing, trend projections, and decomposition. Regression analysis can be used in trend projections and in one type of decomposition model.

2. Causal Models

Causal models incorporate the variables or factors that might influence the quantity being forecasted into the forecasting model. For example, daily sales of a cola drink might depend on the season, the average temperature, the average humidity, whether it is a weekend or a weekday, and so on. Thus, a causal model would attempt to include factors for temperature, humidity, season, day of the week, and so on. Causal models may also include past sales data as time series models do, but they include other factors as well.

3. Qualitative Models

Whereas time-series and causal models rely on quantitative data, qualitative models attempt to incorporate judgmental or subjective factors into the forecasting model. Opinions by experts, individual experiences and judgments, and other subjective factors may be considered. Qualitative models are especially useful when subjective factors are expected to be very important or when accurate quantitative data are difficult to obtain.

Here is a brief overview of four different qualitative forecasting techniques:

a. **Delphi method.** This iterative group process allows experts, who may be located in different places, to make forecasts. There are three different types of participants in the Delphi process: decision makers, staff personnel, and respondents. The decision-making group usually consists of 5 to 10 experts who will be making the actual forecast. The staff personnel assist the decision makers by preparing, distributing, collecting, and summarizing a series of questionnaires and survey results. The respondents are a group of people whose judgments are valued and are being sought. This group provides inputs to the decision makers before the forecast is made.

In the Delphi method, when the results of the first questionnaire are obtained, the results are summarized and the questionnaire is modified. Both the summary of the results and the new questionnaire are then sent to the same respondents for a new round of responses. The respondents, upon seeing the results from the first questionnaire, may view things differently and may modify their original responses. This process is repeated with the hope that a consensus is reached.

b. **Jury of executive opinion.** This method takes the opinions of a small group of high-level managers, often in combination with statistical models, and results in a group estimate of demand.

c. **Sales force composite.** In this approach, each salesperson estimates what sales will be in his or her region; these forecasts are reviewed to ensure that they are realistic and are then combined at the district and national levels to reach an overall forecast.

d. **Consumer market survey.** This method solicits input from customers or potential customers regarding their future purchasing plans. It can help not only in preparing a forecast but also in improving product design and planning for new products.

The types of forecast models is presented in figure 1 below

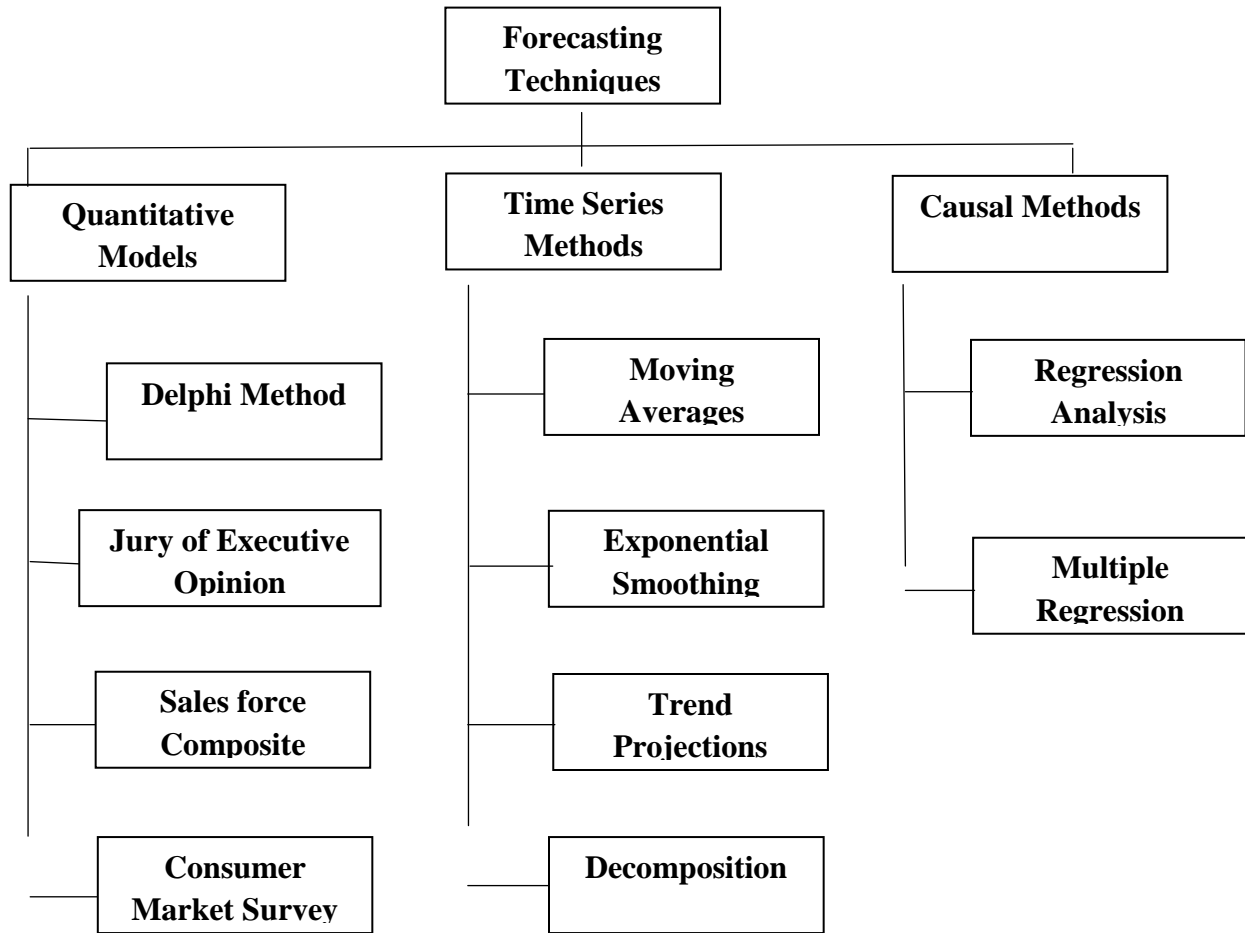


Figure 1: Forecasting Models

Steps in the Development of a Forecasting System

The following steps can help in the development of a forecasting system. While steps 5 and 6 may not be as relevant if a qualitative model is selected in step 4, data are certainly necessary for the quantitative forecasting models presented in this module. The steps are:

- i.** Determine the use of the forecast—what objective are we trying to obtain?
- ii.** Select the items or quantities that are to be forecasted.
- iii.** Determine the time horizon of the forecast—is it 1 to 30 days (short term), 1 month to 1 year (medium term), or more than 1 year (long term)?
- iv.** Select the forecasting model or models.
- v.** Gather the data or information needed to make the forecast.
- vi.** Validate the forecasting model.

- vii. Make the forecast.
- viii. Implement the results.

These steps present a systematic way of initiating, designing, and implementing a forecasting system. When the forecasting system is to be used to generate forecasts regularly over time, data must be collected routinely, and the actual computations or procedures used to make the forecast can be done automatically.

SELF ASSESSMENT EXERCISE

Describe in your own words the types of forecast you know.

3.1.2 Scatter Diagrams and Time Series

As with regression models, scatter diagrams are very helpful when forecasting time series. A scatter diagram for a time series may be plotted on a two-dimensional graph with the horizontal axis representing the time period. The variable to be forecast (such as sales) is placed on the vertical axis. Let us consider the example of a firm that needs to forecast sales for three different products.

Benson Distributors notes that annual sales for three of its products—television sets, radios, and compact disc players—over the past 10 years are as shown in Table 1. One simple way to examine these historical data, and perhaps to use them to establish a forecast, is to draw a scatter diagram for each product (Figure 2 - 4). This picture, showing the relationship between sales of a product and time, is useful in spotting trends or cycles. An exact mathematical model that describes the situation can then be developed if it appears reasonable to do so.

Table 1: Annual Sales of Three Products

Year	Television Sets	Radios	Compact Disc Players
2010	250	300	110
2011	300	310	100
2012	275	320	126
2013	367	330	138
2014	200	356	167
2015	259	254	144
2016	280	218	152
2017	312	301	173
2018	240	284	98
2019	223	276	109

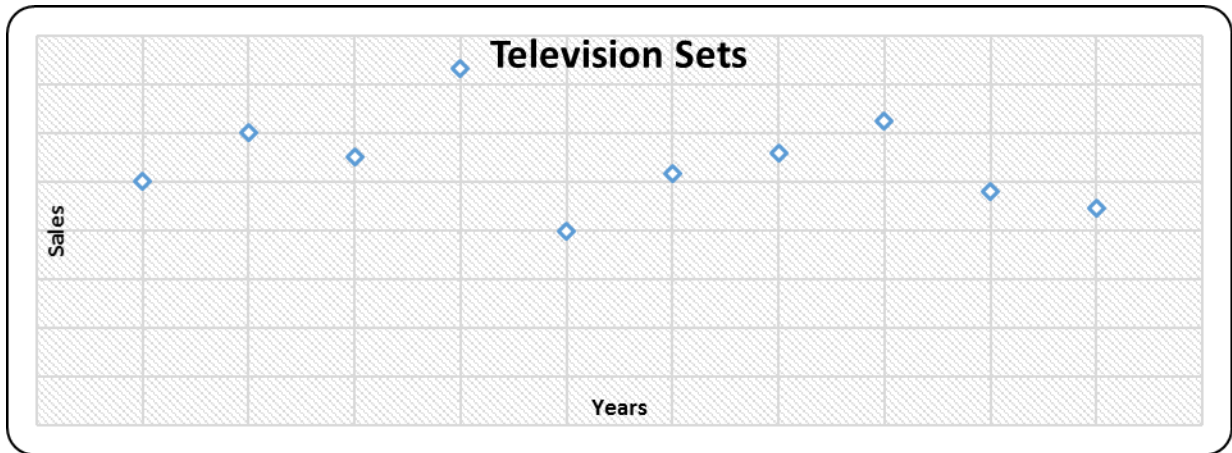


Figure 2: Scatter Diagram for Sales of Television Sets

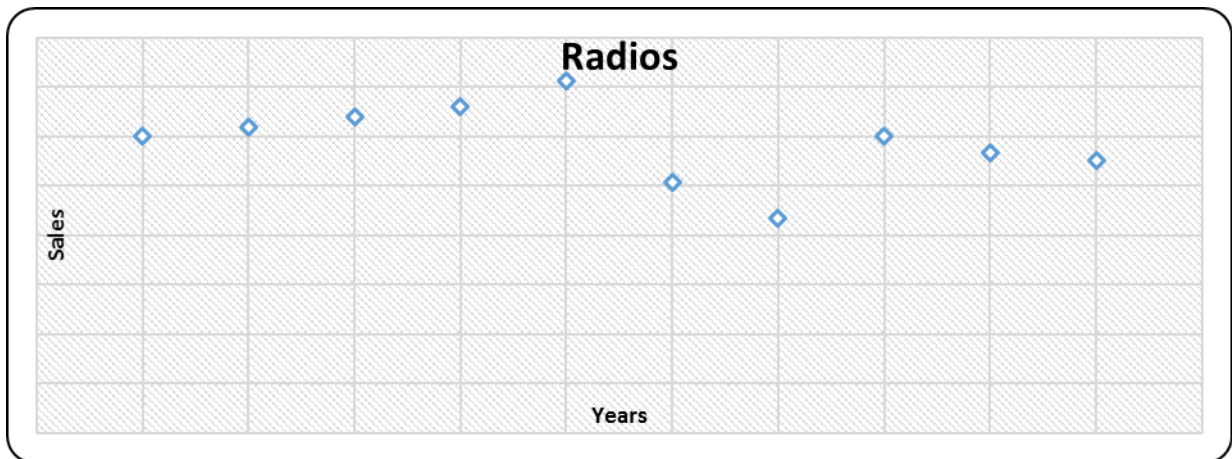


Figure 3: Scatter Diagram for Sales of Radios

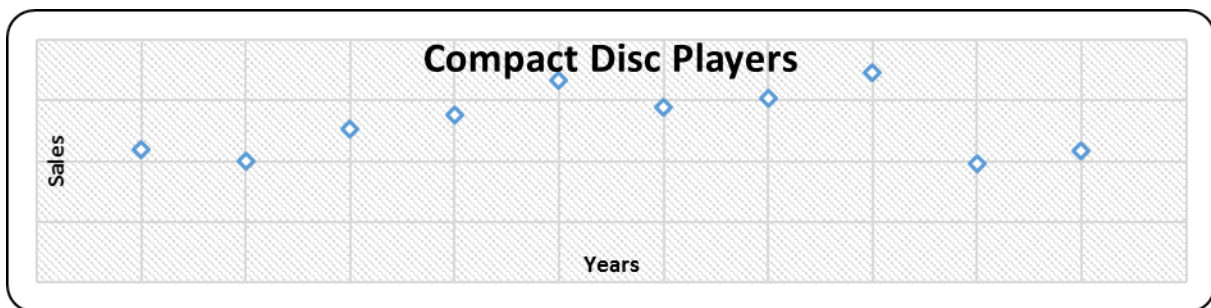


Figure 4: Scatter Diagram for Sales of Compact Disc Players

SELF ASSESSMENT EXERCISE

Plot the scatter graph of your daily consumption of bottle of water for the next one week and what can you say about the pattern?

3.1.3 Measures of Forecast Accuracy

We shall further discuss several different forecasting models in this unit. To see how well one model works, or to compare that model with other models, the forecasted values are compared with the actual or observed values. The forecast error (or deviation) is defined as follows:

$$\text{Forecast error} = \text{Actual value} - \text{forecast value}$$

One measure of accuracy is the mean absolute deviation (MAD). This is computed by taking the sum of the absolute values of the individual forecast errors and dividing by the numbers of errors (n):

$$MAD = \frac{\sum |forecast\ error|}{n} \quad (1)$$

Consider the Benson Distributors sales of CD players shown in Table 1. Suppose that in the past, Benson had forecast sales for each year to be the sales that were actually achieved in the previous year. This is sometimes called a naïve model. Table 2 gives these forecasts as well as the absolute value of the errors. In forecasting for the next time period (year 2020), the forecast would be 190. Notice that there is no error computed for year 2010 since there was no forecast for this year, and there is no error for year 2020 since the actual value of this is not yet known. Thus, the number of errors (n) is 9.

Table 2: Computing the Mean Absolute Deviation (MAD)

Year	Actual Sales of Compact Disc Players	Forecast sales	Absolute value of errors (Deviation) $ Actual - forecast $
2010	110	-	-
2011	100	90	$ 100 - 90 = 10$
2012	126	131	$ 126 - 131 = 5$
2013	138	127	$ 138 - 127 = 11$
2014	167	155	$ 167 - 155 = 12$
2015	144	148	$ 144 - 148 = 4$
2016	152	133	$ 152 - 133 = 19$
2017	173	160	$ 173 - 160 = 13$
2018	98	93	$ 98 - 93 = 5$
2019	109	110	$ 109 - 110 = 1$
2020		125	-
			Sum of $ errors = 80$ MAD = $\frac{80}{9} = 8.89$

From this, we see the following:

$$\begin{aligned}MAD &= \frac{\sum |forecast\ error|}{n} \\ &= \frac{80}{9} = 8.89\end{aligned}$$

This means that on the average, each forecast missed the actual value by 8.89 units. Other measures of the accuracy of historical errors in forecasting are sometimes used besides the MAD. One of the most common is the mean squared error (MSE), which is the average of the squared errors

$$MSE = \frac{\sum |error|^2}{n} \quad (2)$$

Besides the MAD and MSE, the mean absolute percent error (MAPE) is sometimes used. The MAPE is the average of the absolute values of the errors expressed as percentages of the actual values. This is computed as follows:

$$MAPE = \frac{\sum \left| \frac{error}{actual} \right|}{n} 100\% \quad (3)$$

There is another common term associated with error in forecasting. Bias is the average error and tells whether the forecast tends to be too high or too low and by how much. Thus, bias may be negative or positive. It is not a good measure of the actual size of the errors because the negative errors can cancel out the positive errors.

SELF ASSESSMENT EXERCISE

Use the Mean Square Error (MSE) method to compute the forecast error of the sales of compact disc players using the forecasted sales in table 2.

3.1.4 Time-Series Forecasting Models

A time series is based on a sequence of evenly spaced (weekly, monthly, quarterly, and so on) data points. Examples include weekly sales of HP personal computers, quarterly earnings reports of Microsoft Corporation, daily shipments of Eveready batteries, and annual Nigeria consumer price indices. Forecasting time-series data implies that future values are predicted *only* from past values of that variable (such as we saw in Table 1) and that other variables, no matter how potentially valuable, are ignored.

Components of a Time Series

Analyzing time series means breaking down past data into components and then projecting them forward. A time series typically has four components:

1. *Trend (T)* is the gradual upward or downward movement of the data over time.
2. *Seasonality (S)* is a pattern of the demand fluctuation above or below the trend line that repeats at regular intervals.

3. *Cycles (C)* are patterns in annual data that occur every several years. They are usually tied into the business cycle.
4. *Random variations (R)* are “blips” in the data caused by chance and unusual situations; they follow no discernible pattern.

Figure 5 shows a time series and its components.

There are two general forms of time-series models in statistics. The first is a multiplicative model, which assumes that demand is the product of the four components. It is stated as follows:

$$\text{Demand} = T \times S \times C \times R \quad (4)$$

An additive model adds the components together to provide an estimate. Multiple regression is often used to develop additive models. This additive relationship is stated as follows:

$$\text{Demand} = T + S + C + R \quad (5)$$

There are other models that may be a combination of these. For example, one of the components (such as trend) might be additive while another (such as seasonality) could be multiplicative. Understanding the components of a time series will help in selecting an appropriate forecasting technique to use. If all variations in a time series are due to random variations, with no trend, seasonal, or cyclical component, some type of averaging or smoothing model would be appropriate. The averaging techniques in this unit are moving average, weighted moving average, and exponential smoothing. These methods will smooth out the forecasts and not be too heavily influenced by random variations. However, if there is a trend or seasonal pattern present in the data, then a technique which incorporates that particular component into the forecast should be used. Two such techniques are exponential smoothing with trend and trend projections. If there is a seasonal pattern present in the data, then a seasonal index may be developed and used with any of the averaging methods. If both trend and seasonal components are present, then a method such as the decomposition method should be used.

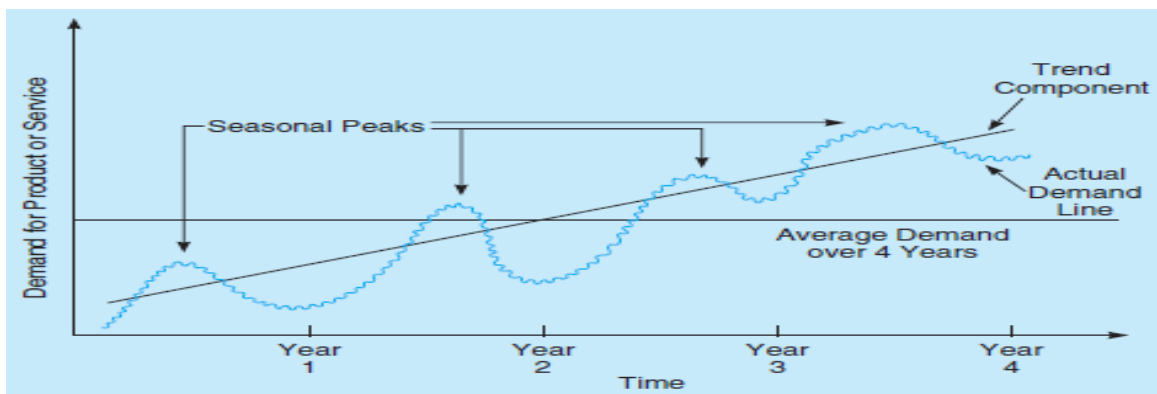


Figure 5: Product Demand Charted over 4 Yrs, with Trend and Seasonality Indicated

The time series forecasting models are discussed below:

3.1.4.1 Moving Averages

Moving averages are useful if we can assume that market demands will stay fairly steady over time. For example, a four-month moving average is found simply by summing the demand during the past four months and dividing by 4. With each passing month, the most recent month's data are added to the sum of the previous three months' data, and the earliest month is dropped. This tends to smooth out short-term irregularities in the data series. An n -period moving average forecast, which serves as an estimate of the next period's demand, is expressed as follows:

$$\text{Moving average forecast} = \frac{\text{Sum of demands in previous } n \text{ periods}}{n} \quad (6)$$

Mathematically, it can be written as:

$$F_{t+1} = \frac{Y_t + Y_{t-1} + \dots + Y_{t-n+1}}{n} \quad (7)$$

Where

F_{t+1} = Forecast for time period $t+1$

Y_t = Actual value in time period t

n = number of periods to average

A 4-month moving average has $n = 4$ a 5-month moving average has $n = 5$

Example

Storage shed sales at Wallace Garden Supply are shown in the middle column of Table 3. A 3-month moving average is indicated on the right. The forecast for the next January, using this technique, is 16. Were we simply asked to find a forecast for next January, we would only have to make this one calculation. The other forecasts are necessary only if we wish to compute the MAD or another measure of accuracy.

Table 3: Wallace Garden Supply Shed Sales

Month	Actual Shed Sales	3 Months Moving Average
January	10	
February	12	
March	13	
April	16	$(10+12+13)/3 = 11.67$
May	19	$(12+13+16)/3 = 13.67$
June	23	$(13+16+19)/3 = 16.00$
July	26	$(16+19+23)/3 = 19.33$
August	30	$(19+23+26)/3 = 22.67$
September	28	$(23+26+30)/3 = 26.33$
October	18	$(26+30+28)/3 = 28.00$
November	16	$(30+28+18)/3 = 25.33$
December	14	$(28+18+16)/3 = 20.67$
January	-	$(18+16+14)/3 = 16.00$

A modified version of the moving averages can be applied called the weighted moving average. A simple moving average gives the same weight to each of the past observations being used to develop the forecast. On the other hand, a weighted moving average allows different weights to be assigned to the previous observations. As the weighted moving average method typically assigns greater weight to more recent observations, this forecast is more responsive to changes in the pattern of the data that occur. However, this is also a potential drawback to this method because the heavier weight would also respond just as quickly to random fluctuations. A *weighted moving average* may be expressed as:

$$F_{t+1} = \frac{(\text{Weight in period } i)(\text{Actual value in period } i)}{\sum(\text{weights})}$$

Mathematically, it can be written as:

$$F_{t+1} = \frac{w_1 Y_t + w_2 Y_{t-1} + \dots + w_n Y_{t-n+1}}{w_1 + w_2 + \dots + w_n} \quad (8)$$

where

w_i = weight for i th observation

Wallace Garden Supply decides to use a 3-month weighted moving average forecast with weights of 3 for the most recent observation, 2 for the next observation, and 1 for the most distant observation.

3.1.4.2 Exponential Smoothing

Exponential smoothing is a forecasting method that is easy to use and is handled efficiently by computers. Although it is a type of moving average technique, it involves little record keeping of past data. The basic exponential smoothing formula can be shown as follows:

$$\text{New forecast} = \text{Last period's forecast} + \alpha (\text{Last period's actual demand} - \text{Last period's forecast}) \quad (9)$$

where α is a weight (or smoothing constant) that has a value between 0 and 1, inclusive.

Equation (9) can be re-written in mathematical form as:

$$F_{t+1} = F_t + \alpha(Y_t - F_t) \quad (10)$$

Where

F_{t+1} = new forecast (for time period $t + 1$)

F_t = previous forecast (for time period t)

α = smoothing constant ($0 \leq \alpha \leq 1$)

Y_t = previous period's actual demand

The concept here is not complex. The latest estimate of demand is equal to the old estimate adjusted by a fraction of the error (last period's actual demand minus the old estimate).

The smoothing constant, α , can be changed to give more weight to recent data when the value is high or more weight to past data when it is low. For example, $\alpha = 0.5$, when it can be shown mathematically that the new forecast is based almost entirely on demand in the past three periods. When $\alpha = 0.1$, the forecast places little weight on any single period, even the most recent, and it takes many periods (about 19) of historic values into account.

Example 1

For example, in January, a demand for 142 of a certain car model for February was predicted by a dealer. Actual February demand was 153 autos. Using a smoothing constant of $\alpha = 0.20$, we can forecast the March demand using the exponential smoothing model. Substituting into the formula, we obtain

$$\text{New forecast (for march demand)} = 142 + 0.2(153 - 142), = 144.2$$

Example 2

Let us apply this concept with a trial-and-error testing of one value of α in an example. The port of Baltimore has unloaded large quantities of grain from ships during the past eight quarters. The port’s operations manager wants to test the use of exponential smoothing to see how well the technique works in predicting tonnage unloaded. He assumes that the forecast of grain unloaded in the first quarter was 175 tons. Two values of α is examined. Table 4 shows the *detailed* calculations for $\alpha = 0.10$ and the already calculated own of $\alpha = 0.50$.

Table 4: Forecasting using the Exponential Smoothing Method

Quarter	Actual Tonnage Unloaded	Forecast using $\alpha = 0.10$	Forecast using $\alpha = 0.50$
1	180	175	175
2	168	$175.5 = 175.00 + 0.10(180 - 175)$	177.5
3	159	$174.75 = 175.50 + 0.10(168 - 175.50)$	172.75
4	175	$173.18 = 174.75 + 0.10(159 - 174.75)$	165.88
5	190	$173.36 = 173.18 + 0.10(175 - 173.18)$	170.44
6	205	$175.02 = 173.36 + 0.10(190 - 173.36)$	180.22
7	180	$178.02 = 175.02 + 0.10(205 - 175.02)$	192.61
8	182	$178.22 = 178.02 + 0.10(180 - 178.02)$	186.30
9	?	$178.60 = 178.22 + 0.10(182 - 178.22)$	184.15

3.1.4.3 Trend Projections

Another method for forecasting time series with trend is called trend projection. This technique fits a trend line to a series of historical data points and then projects the line into the future for medium- to long-range forecasts. There are several mathematical trend equations that can be developed (e.g., exponential and quadratic), but in this section we

look at linear (straight line) trends only. A trend line is simply a linear regression equation in which the independent variable (X) is the time period. The form of this is:

$$\hat{Y} = \beta_0 + \beta_1 X$$

Where

\hat{Y} = Predicted Value

β_0 = Intercept

β_1 = Slope of the line

X = time period (i.e. X = 1, 2, 3, ..., n)

The least squares regression method may be applied to find the coefficients that minimize the sum of the squared errors, thereby also minimizing the mean squared error (MSE). The formula for the coefficients are:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{11}$$

$$\hat{\beta}_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \tag{12}$$

The least squares regression method may be applied to find the coefficients that minimize the sum of the squared errors, thereby also minimizing the mean squared error (MSE).

Example

Let us consider the case of Midwestern Manufacturing Company. That firm’s demand for electrical generators over the period 2004–2010 was given. A trend line to predict demand (Y) based on the time period can be developed using a regression model. If we let 2004 be time period 1 then 2005 is time period 2 and so forth. The regression line can be developed using the formula. From this we get:

Table 5: Forecasting using trend

Year	X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
2004	1	74	-3	-24.8571	74.57143	9
2005	2	79	-2	-19.8571	39.71429	4
2006	3	80	-1	-18.8571	18.85714	1
2007	4	90	0	-8.85714	0	0
2008	5	105	1	6.142857	6.142857	1
2009	6	142	2	43.14286	86.28571	4
2010	7	122	3	23.14286	69.42857	9
	28	692			295	28

Following equation (11) and (12)

$$\hat{\beta}_0 = 56.71,$$

$$\hat{\beta}_1 = 10.54$$

Therefore, the trend equation is given as

$$\hat{Y} = 56.71 + 10.54X$$

To project demand in 2011, we first denote the year 2011 in our new coding system as $X = 8$;

$$(\text{sales in 2011}) = 56.71 + 10.54(8) = 141.03 \text{ or } 141 \text{ generators}$$

We can estimate demand for 2012 by inserting in the same equation:

$$(\text{sales in 2012}) = 56.71 + 10.54(9) = 151.57, \text{ or } 152 \text{ generators}$$

3.1.4.4 Seasonal Variations

Time-series forecasting such as that in the example of Midwestern Manufacturing involves looking at the *trend* of data over a series of time observations. Sometimes, however, recurring variations at certain seasons of the year make a *seasonal* adjustment in the trend line forecast necessary.

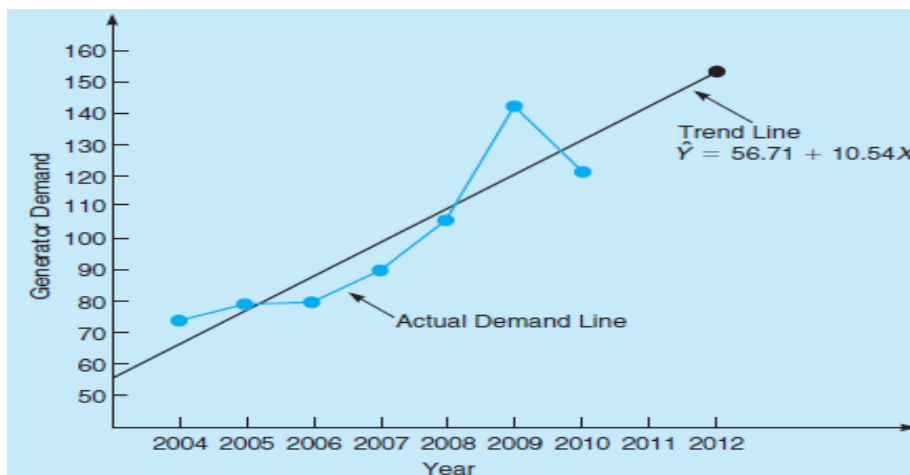


Figure 6: Electrical Generators and the Computed Trend Line

Steps Used to Compute Seasonal Indices Based on Centred Moving Averages

1. Compute a Centered Moving Averages for each observation (where possible).
2. Compute seasonal ratio = Observation/ Centered Moving Averages for that observation.
3. Average seasonal ratios to get seasonal indices.
4. If seasonal indices do not add to the number of seasons, multiply each index by (Number of seasons)/(Sum of the indices).

SELF ASSESSMENT

Show the workings of the last column in table 4.

3.1.5 Monitoring and Controlling Forecasts

After a forecast has been completed, it is important that it not be forgotten. No manager wants to be reminded when his or her forecast is horribly inaccurate, but a firm needs to determine why the actual demand (or whatever variable is being examined) differed significantly from that projected

One way to monitor forecasts to ensure that they are performing well is to employ a **tracking signal**. A tracking signal is a measurement of how well the forecast is predicting actual values. As forecasts are updated every week, month, or quarter, the newly available demand data are compared to the forecast values.

The tracking signal is computed as the **running sum of the forecast errors (RSFE)** divided by the mean absolute deviation:

$$\begin{aligned} \text{Tracking Signal} &= \frac{RSFE}{MAD} \\ &= \frac{\sum(\text{forecast error})}{MAD} \end{aligned} \tag{13}$$

Where:

$$MAD = \frac{\sum|\text{forecast error}|}{n}$$

as seen earlier in Equation 1.

Positive tracking signals indicate that demand is greater than the forecast. Negative signals mean that demand is less than forecast. A good tracking signal—that is, one with a low RSFE—has about as much positive error as it has negative error. In other words, small deviations are okay, but the positive and negative deviations should balance so that the tracking signal centers closely around zero. When tracking signals are calculated, they are compared with predetermined control limits. When a tracking signal exceeds an upper or lower limit, a signal is tripped. This means that there is a problem with the forecasting method, and management may want to reevaluate the way it forecasts demand. Figure 7 shows the graph of a tracking signal that is exceeding the range of acceptable variation. If the model being used is exponential smoothing, perhaps the smoothing constant needs to be readjusted.

How do firms decide what the upper and lower tracking limits should be? There is no single answer, but they try to find reasonable values—in other words, limits not so low as to be triggered with every small forecast error and not so high as to allow bad forecasts to be regularly overlooked. George Plossl and Oliver Wight, two inventory control experts, suggested using maximums of $\pm 4MADs$ for high-volume stock items and $\pm 8MADs$ for lower-volume items.

Other forecasters suggest slightly lower ranges. One MAD is equivalent to approximately 0.8 standard deviation, so $\pm 2MADs = 1.6$ standard deviations, $\pm 3MADs = 2.4$ standard deviations, and $\pm 4MADs = 3.2$ standard deviations. This suggests that for a forecast to be

“in control,” 89% of the errors are expected to fall within $\pm 2MADs$, 98% within $\pm 3MADs$ or 99.9% within $\pm 4MADs$ whenever the errors are approximately normally distributed.

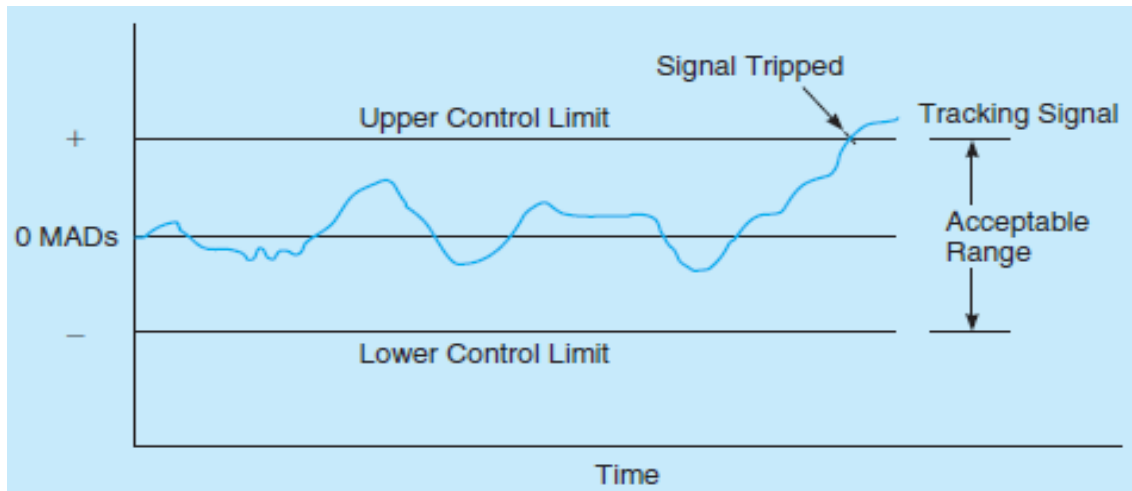


Figure 7: Plot of Tracking Signals

Adaptive Smoothing

A lot of research has been published on the subject of adaptive forecasting. This refers to computer monitoring of tracking signals and self-adjustment if a signal passes its preset limit. In exponential smoothing, the α and β coefficients are first selected based on values that minimize error forecasts and are then adjusted accordingly whenever the computer notes an errant tracking signal; this is called adaptive smoothing.

3.2 Forecasting Analysis

3.2.1 The Steps in Decision Making

Whether you are deciding about getting a haircut today, building a multimillion plant, or buying a new camera, the steps in making a good decision are basically the same:

Six Steps in Decision Making are:

1. Clearly define the problem at hand.
2. List the possible alternatives.
3. Identify the possible outcomes or states of nature.
4. List the payoff (typically profit) of each combination of alternatives and outcomes.
5. Select one of the mathematical decision theory models.
6. Apply the model and make your decision.

We use the Bola Company case as an example to illustrate these decision theory steps. Bola is the founder and president of Bola Company, a profitable firm located in Lagos.

Step 1. The problem that Bola identifies is whether to expand his product line by manufacturing and marketing a new product, backyard storage sheds. Bola’s second step is to generate the alternatives that are available to him. In decision theory, an **alternative** is defined as a course of action or a strategy that the decision maker can choose.

Step 2. Bola decides that his alternatives are to construct (1) a large new plant to manufacture the storage sheds, (2) a small plant, or (3) no plant at all (i.e., he has the option of not developing the new product line). One of the biggest mistakes that decision makers make is to leave out some important alternatives. Although a particular alternative may seem to be inappropriate or of little value, it might turn out to be the best choice. The next step involves identifying the possible outcomes of the various alternatives. A common mistake is to forget about some of the possible outcomes. Optimistic decision makers tend to ignore bad outcomes, whereas pessimistic managers may discount a favorable outcome. If you don’t consider all possibilities, you will not be making a logical decision, and the results may be undesirable. If you do not think the worst can happen, you may design another Edsel automobile. In decision theory, those outcomes over which the decision maker has little or no control are called **states of nature**.

Table 6: Decision Table with Conditional Values for Bola Company

Alternative	State of Nature	
	Favourable Market	Unfavourable Market
Construct a large plant	200,000	-180,000
Construct a small plant	100,000	-20,000
Do nothing	0	0

Step 3. Bola determines that there are only two possible outcomes: the market for the storage sheds could be favorable, meaning that there is a high demand for the product, or it could be unfavorable, meaning that there is a low demand for the sheds. Once the alternatives and states of nature have been identified, the next step is to express the payoff resulting from each possible combination of alternatives and outcomes. In decision theory, we call such payoffs or profits **conditional values**. Not every decision, of course, can be based on money alone—any appropriate means of measuring benefit is acceptable.

Step 4. Because Bola wants to maximize his profits, he can use *profit* to evaluate each consequence. Bola has already evaluated the potential profits associated with the various outcomes. With a favorable market, he thinks a large facility would result in a net profit of \$200,000 to his firm. This \$200,000 is a *conditional value* because Bola’s receiving the money is conditional upon both his building a large factory and having a good market. The conditional value if the market is unfavorable would be a \$180,000 net loss.

A small plant would result in a net profit of \$100,000 in a favorable market, but a net loss of \$20,000 would occur if the market was unfavorable. Finally, doing nothing would result in \$0 profit in either market. The easiest way to present these values is by constructing a **decision table**, sometimes called a **payoff table**. A decision table for Bola's conditional values is shown in Table 6. All of the alternatives are listed down the left side of the table, and all of the possible outcomes or states of nature are listed across the top. The body of the table contains the actual payoffs.

Steps 5 and 6. The last two steps are to select a decision theory model and apply it to the data to help make the decision. Selecting the model depends on the environment in which you're operating and the amount of risk and uncertainty involved.

SELF ASSESSMENT

Discuss the six Steps in decision making

3.2.2 Types of Decision-Making Environments

The types of decisions people make depend on how much knowledge or information they have about the situation. There are three decision-making environments:

1. Decision making under certainty
2. Decision making under uncertainty
3. Decision making under risk

1. Decision Making Under Certainty

In the environment of decision making under certainty, decision makers know with certainty the consequence of every alternative or decision choice. Naturally, they will choose the alternative that will maximize their well-being or will result in the best outcome.

For example, let's say that you have \$1,000 to invest for a 1-year period. One alternative is to open a savings account paying 6% interest and another is to invest in a government Treasury bond paying 10% interest. If both investments are secure and guaranteed, there is a certainty that the Treasury bond will pay a higher return. The return after one year will be \$100 in interest.

2. Decision Making Under Uncertainty

In decision making under uncertainty, there are several possible outcomes for each alternative, and the decision maker does not know the probabilities of the various outcomes. As an example, the probability that a Democrat will be president of the United States 25 years from now is not known. Sometimes it is impossible to assess the probability of success of a new undertaking or product.

3. Decision Making Under Risk

In decision making under risk, there are several possible outcomes for each alternative, and the decision maker knows the probability of occurrence of each outcome. We know, for example, that when playing cards using a standard deck, the probability of being dealt a club is 0.25. The probability of rolling a 5 on a die is 1/6. In decision making under risk, the decision maker usually attempts to maximize his or her expected wellbeing. Decision theory models for business problems in this environment typically employ two equivalent criteria: maximization of expected monetary value and minimization of expected opportunity loss.

Let's see how decision making under certainty (the type 1 environment) could affect Bola. Here we assume that John knows exactly what will happen in the future. If it turns out that he knows with certainty that the market for storage sheds will be favorable, what should he do? Look again at Bola's conditional values in Table 6. Because the market is favorable, he should build the large plant, which has the highest profit, \$200,000.

Few managers would be fortunate enough to have complete information and knowledge about the states of nature under consideration. Decision making under uncertainty, discussed next, is a more difficult situation. We may find that two different people with different perspectives may appropriately choose two different alternatives.

SELF ASSESSMENT

Discuss the types of decision-making environments

3.2.3 Decision Making Under Uncertainty

When several states of nature exist and a manager *cannot* assess the outcome probability with confidence or when virtually no probability data are available, the environment is called decision making under uncertainty. Several criteria exist for making decisions under these conditions. The ones that we cover in this section are as follows:

1. Optimistic (maximax)
2. Pessimistic (maximin)
3. Criterion of realism (Hurwicz)
4. Equally likely (Laplace)
5. Minimax regret

The first four criteria can be computed directly from the decision (payoff) table, whereas the minimax regret criterion requires use of the opportunity loss table.

The presentation of the criteria for decision making under uncertainty (and also for decision making under risk) is based on the assumption that the payoff is something in which larger values are better and high values are desirable. For payoffs such as profit, total sales, total return on investment, and interest earned, the best decision would be one that resulted in some type of maximum payoff. However, there are situations in which lower payoff values (e.g., cost) are better, and these payoffs would be minimized rather

than maximized. The statement of the decision criteria would be modified slightly for such minimization problems.

SELF ASSESSMENT

Discuss decision making under uncertainty.

4.0 CONCLUSION

So far, this unit concludes that there are different types of forecasting and these are time-series models, causal models, and qualitative models. The unit carefully explains time-series forecasting models which are moving averages, exponential smoothing and seasonal variations. Also, the unit concludes that decisions can be made under uncertainty, certainty and risk.

5.0 SUMMARY

In this unit, we have discussed the types of forecasts, scatter diagrams and time series, measures of forecast accuracy, Time-Series Forecasting Models, Monitoring and Controlling Forecasts, Steps in Decision Making, types of Decision-Making Environments and decision Making Under Uncertainty.

6.0 TUTORED MARKED ASSIGNMENTS

Data collected on the yearly demand for 50-pound bags of fertilizer at Wallace Garden Supply are shown in the following table. Develop a 3-year moving average to forecast sales. Then estimate demand again with a weighted moving average in which sales in the most recent year are given a weight of 2 and sales in the other 2 years are each given a weight of 1. Which method do you think is best?

Year	Demand for Fertilizer (1,000s of Bags)
1	4
2	6
3	4
4	5
5	10
6	8
7	7
8	9
9	12
10	14
11	15

7.0 REFERENCES/FURTHER READINGS

- Render, B., Stair, R.M., & Hanna, M.E. (2012). *Quantitative Analysis for Management* (11th ed.). Essex: England, Pearson.
- Murthy, P.R. (2007). *Operations Research* (2nd ed.). New Delhi, India: New Age International ltd.

UNIT 2: DEMONSTRATE FORECASTING METHODS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Data Set to Demonstrate Forecasting Methods
 - 3.2 Seasonality in Forecasting
 - 3.3 Measuring Forecasting Accuracy
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

One way of classifying forecasting problems is to consider the timescale involved in the forecast i.e. how far forward into the future we are trying to forecast. Short, medium and long-term are the usual categories but the actual meaning of each will vary according to the situation that is being studied, e.g. in forecasting energy demand in order to construct power stations 5-10 years would be short-term and 50 years would be long-term, whilst in forecasting consumer demand in many business situations up to 6 months would be short-term and over a couple of years long-term. The tables below show the timescale associated with business decisions. The basic reason for the above classification is that different forecasting methods apply in each situation, e.g. a forecasting method that is appropriate for forecasting sales next month (a short-term forecast) would probably be an inappropriate method for forecasting sales in five years' time (a long-term forecast).

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Demonstrate the various forecasting methods
- Forecast using data set and apply any of the forecast methods
- Discuss seasonality issues in forecasting
- Measure forecasting accuracy

3.0 MAIN CONTENT

3.1 Data Set to Demonstrate Forecasting Methods

Qualitative methods: These types of forecasting methods are based on judgments, opinions, intuition, emotions, or personal experiences and are subjective in nature. They do not rely on any rigorous mathematical computations.

Quantitative methods: These types of forecasting methods are based on mathematical (quantitative) models, and are objective in nature. They rely heavily on

mathematical computations. **Quantitative Forecasting Methods is divided in two. Time-Series Models and Associative Models.** Time series models look at past patterns of data and attempt to predict the future based upon the underlying patterns contained within those data. Associative models (often called causal models) assume that the variable being forecasted is related to other variables in the environment. They try to project based upon those associations.

Time Series Models

Model	Description
Naïve	Uses last period's actual value as a forecast
Simple Mean (Average)	Uses an average of all past data as a forecast
Simple Moving Average	Uses an average of a specified number of the most recent observations, with each observation receiving the same emphasis (weight) Uses an average of a specified number of the most
Weighted Moving Average	recent observations, with each observation receiving a different emphasis (weight)
Exponential Smoothing	A weighted average procedure with weights declining exponentially as data become older
Trend Projection	Technique that uses the least squares method to fit a straight line to the data
Seasonal Indexes	A mechanism for adjusting the forecast to accommodate any seasonal patterns inherent in the data

The following data set represents a set of hypothetical demands that have occurred over several consecutive years. The data have been collected on a quarterly basis, and these quarterly values have been amalgamated into yearly totals.

For various illustrations that follow, we may make slightly different assumptions about starting points to get the process started for different models. In most cases we will assume that each year a forecast has been made for the subsequent year. Then, after a year has transpired, we will have observed what the actual demand turned out to be (and we will surely see differences between what we had forecasted and what actually occurred, for, after all, the forecasts are merely educated guesses).

Finally, to keep the numbers at a manageable size, several zeros have been dropped off the numbers (i.e., these numbers represent demands in thousands of units).

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4	Total Annual Demand
1	62	94	113	41	310
2	73	110	130	52	365
3	79	118	140	58	395
4	83	124	146	62	415
5	89	135	161	65	450
6	94	139	162	70	465

Illustration of the Naïve Method

Naïve method: The forecast for next period (period $t+1$) will be equal to this period's actual demand (A_t).

In this illustration we assume that each year (beginning with year 2) we made a forecast, then waited to see what demand unfolded during the year. We then made a forecast for the subsequent year, and so on right through to the forecast for year 7.

	Actual Demand	Forecast	
Year	(A_t)	(F_t)	Notes
1	310	--	There was no prior demand data on which to base a forecast for period 1
2	365	310	From this point forward, these forecasts were made on a year-by-year basis.
3	395	365	
4	415	395	
5	450	415	
6	465	450	
7		465	

Mean (Simple Average) Method

Mean (simple average) method: The forecast for next period (period $t+1$) will be equal to the average of all past historical demands.

In this illustration we assume that a simple average method is being used. We will also assume that, in the absence of data at startup, we made a guess for the year 1 forecast (300). At the end of year 1 we could start using this forecasting method. In this illustration we assume that each year (beginning with year 2) we made a forecast, then waited to see what demand unfolded during the year. We then made a forecast for the subsequent year, and so on right through to the forecast for year 7.

	Actual Demand	Forecast	
Year	(At)	(Ft)	Notes
1	310	300	This forecast was a guess at the beginning.
2	365	310.000	From this point forward, these forecasts were made on a year-by-year basis using a simple average approach.
3	395	337.500	
4	415	356.667	
5	450	371.250	
6	465	387.000	
7		400.000	

Simple Moving Average Method

Simple moving average method: The forecast for next period (period $t+1$) will be equal to the average of a specified number of the most recent observations, with each observation receiving the same emphasis (weight).

In this illustration we assume that a 2-year simple moving average is being used. We will also assume that, in the absence of data at startup, we made a guess for the year 1 forecast (300). Then, after year 1 elapsed, we made a forecast for year 2 using a naïve method (310). Beyond that point we had sufficient data to let our 2-year simple moving average forecasts unfold throughout the years.

	Actual Demand	Forecast	
Year	(At)	(Ft)	Notes
1	310	300	This forecast was a guess at the beginning.
2	365	310	This forecast was made using a naïve approach. From this point forward, these forecasts
3	395	337.500	Were made on a year-by-year basis using a 2-yr moving average approach.
4	415	380.000	
5	450	405.000	
6	465	432.500	
7		457.500	

Another Simple Moving Average Illustration

In this illustration we assume that a 3-year simple moving average is being used. We will also assume that, in the absence of data at startup, we made a guess for the year 1 forecast

(300). Then, after year 1 elapsed, we used a naïve method to make a forecast for year 2 (310) and year 3 (365). Beyond that point we had sufficient data to let our 3-year simple moving average forecasts unfold throughout the years.

	Actual Demand	Forecast	
Year	(At)	(Ft)	Notes
1	310	300	This forecast was a guess at the beginning.
2	365	310	This forecast was made using a naïve approach
3	395	365	This forecast was made using a naïve approach
4	415	356.667	From this point forward, these forecasts were made on a year-by-year basis using a 3-yr moving average approach.
5	450	391.667	
6	465	420.000	
7		433.333	

Weighted Moving Average Method

Weighted moving average method: The forecast for next period (period $t + 1$) will be equal to a weighted average of a specified number of the most recent observations.

In this illustration we assume that a 3-year weighted moving average is being used. We will also assume that, in the absence of data at startup, we made a guess for the year 1 forecast (300). Then, after year 1 elapsed, we used a naïve method to make a forecast for year 2 (310) and year 3 (365). Beyond that point we had sufficient data to let our 3-year weighted moving average forecasts unfold throughout the years. The weights that were to be used are as follows: Most recent year, .5; year prior to that, .3; year prior to that, .2

	Actual Demand	Forecast	
Year	(At)	(Ft)	Notes
1	310	300	This forecast was a guess at the beginning.
2	365	310	This forecast was made using a naïve approach.
3	395	365	This forecast was made using a naïve approach.

4	415	369.000	From this point forward, these forecasts were made on a year-by-year basis Using a 3-yr weighted moving avg. approach.
5	450	399.000	
6	465	428.500	
7		450.500	

Exponential Smoothing Method

Exponential smoothing method: The new forecast for next period (period t) will be calculated as follows:

New forecast = Last period's forecast + α (Last period's actual demand – Last period's forecast)

(This box contains all you need to know to apply exponential smoothing)

$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1}) \text{ (equation 1)}$$

$$F_t = \alpha A_{t-1} + (1 - \alpha) F_{t-1} \text{ (alternate equation 1 – } \alpha \text{ bit more user friendly)}$$

Where α is a smoothing coefficient whose value is between 0 and 1.

The exponential smoothing method only requires that you dig up two pieces of data to apply it (the most recent actual demand and the most recent forecast).

An attractive feature of this method is that forecasts made with this model will include a portion of every piece of historical demand. Furthermore, there will be different weights placed on these historical demand values, with older data receiving lower weights. At first glance this may not be obvious, however, this property is illustrated on the following page.

Demonstration: Exponential Smoothing Includes All Past Data

Note: the mathematical manipulations in this box are not something you would ever have to do when applying exponential smoothing. All you need to use is equation 1 on the previous page. This demonstration is to convince the skeptics that when using equation 1, all historical data will be included in the forecast, and the older the data, the lower the weight applied to that data.

To make a forecast for next period, we would use the user-friendly alternate equation 1:

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

When we made the forecast for the current period (F_{t-1}), it was made in the following fashion:

$$F_{t-1} = \alpha A_{t-2} + (1 - \alpha) F_{t-2} \quad (2)$$

If we substitute equation 2 into equation 1 we get the following:

$$F_t = \alpha A_{t-1} + (1 - \alpha)[\alpha A_{t-2} + (1 - \alpha) F_{t-2}]$$

Which can be cleaned up to the following:

$$F_t = \alpha A_{t-1} + \alpha(1 - \alpha) A_{t-2} + (1 - \alpha)^2 F_{t-2} \quad (3)$$

$$\text{We could continue to play that game by recognizing that } F_{t-2} = \alpha A_{t-3} + (1 - \alpha) F_{t-3} \quad (4)$$

If we substitute equation 4 into equation 3 we get the following:

$$F_t = \alpha A_{t-1} + \alpha(1 - \alpha) A_{t-2} + (1 - \alpha)^2 [\alpha A_{t-3} + (1 - \alpha) F_{t-3}] \quad (5)$$

Which can be cleaned up to the following:

$$F_t = \alpha A_{t-1} + \alpha(1 - \alpha) A_{t-2} + \alpha(1 - \alpha)^2 A_{t-3} + (1 - \alpha)^3 F_{t-3}$$

If you keep playing that game, you should recognize that

$$F_t = \alpha A_{t-1} + \alpha(1 - \alpha) A_{t-2} + \alpha(1 - \alpha)^2 A_{t-3} + \alpha(1 - \alpha)^3 A_{t-4} + \alpha(1 - \alpha)^4 A_{t-5} + \alpha(1 - \alpha)^5 A_{t-6} \dots \quad (6)$$

As you raise those decimal weights to higher and higher powers, the values get smaller and smaller.

Exponential Smoothing Illustration

In this illustration we assume that, in the absence of data at startup, we made a guess for the year 1 forecast (300). Then, for each subsequent year (beginning with year 2) we made a forecast using the exponential smoothing model. After the forecast was made, we waited to see what demand unfolded during the year. We then made a forecast for the subsequent year, and so on right through to the forecast for year 7.

This set of forecasts was made using an α value of .1

Year	Actual Demand (A)	Forecast (F)	Notes
1	310	300	This was a guess, since there was no prior demand data.
2	365	301	From this point forward, these forecasts were made on a year-by-year basis using exponential smoothing with $b=.1$
3	395	307.4	
4	415	316.16	
5	450	326.044	
6	465	338.4396	
7		351.09564	

A Second Exponential Smoothing Illustration

In this illustration we assume that, in the absence of data at startup, we made a guess for the year 1 forecast (300). Then, for each subsequent year (beginning with year 2) we made a forecast using the exponential smoothing model. After the forecast was made, we waited to see what demand unfolded during the year. We then made a forecast for the subsequent year, and so on right through to the forecast for year 7.

This set of forecasts was made using an α value of .2

	Actual Demand	Forecast	
Year	(A)	(F)	Notes
1	310	300	This was a guess, since there was no prior demand data.
2	365	302	From this point forward, these forecasts were made on a year-by-year basis using exponential smoothing with $\alpha = .2$
4	415	330.68	
5	450	347.544	
6	465	368.0352	
7		387.42816	

A Third Exponential Smoothing Illustration

In this illustration we assume that, in the absence of data at startup, we made a guess for the year 1 forecast (300). Then, for each subsequent year (beginning with year 2) we made a forecast using the exponential smoothing model. After the forecast was made, we waited to see what demand unfolded during the year. We then made a forecast for the subsequent year, and so on right through to the forecast for year 7.

This set of forecasts was made using an α value of .4

	Actual Demand	Forecast	
Year	(A)	(F)	Notes
1	310	300	This was a guess, since there was no prior demand data.
2	365	304	From this point forward, these forecasts were made on a year-by-year basis using exponential smoothing with $\alpha = .4$
3	395	328.4	
4	415	355.04	
5	450	379.024	
6	465	407.4144	
7		430.44864	

Trend Projection

Trend projection method: This method is a version of the linear regression technique. It attempts to draw a straight line through the historical data points in a fashion that comes as close to the points as possible. (Technically, the approach attempts to reduce the vertical deviations of the points from the trend line, and does this by minimizing the squared values of the deviations of the points from the line). Ultimately, the statistical formulas compute a slope for the trend line (b) and the point where the line crosses the y-axis (a). This results in the straight-line equation $Y = a + bX$

Where X represents the values on the horizontal axis (time), and Y represents the values on the vertical axis (demand).

For the demonstration data, computations for b and a reveal the following (NOTE: I will not require you to make the statistical calculations for b and a; these would be given to you. However, you do need to know what to do with these values when given to you.)

$$b = 30 \text{ and } a = 295$$

$$Y = 295 + 30X$$

This equation can be used to forecast for any year into the future. For example:

$$\text{Year 7: Forecast} = 295 + 30(7) = 505$$

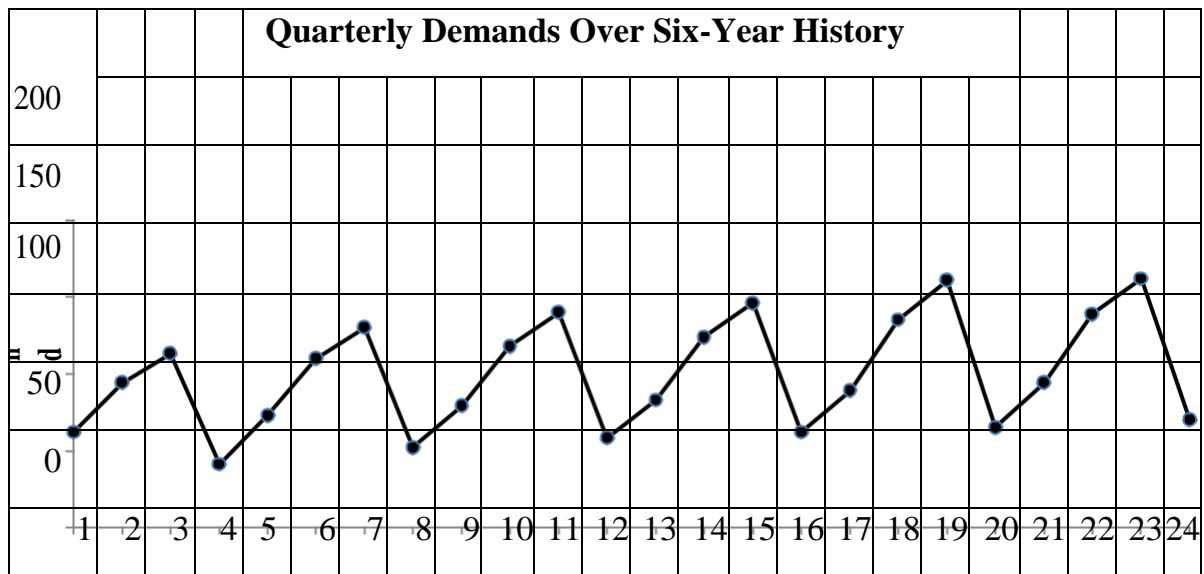
$$\text{Year 8: Forecast} = 295 + 30(8) = 535$$

$$\text{Year 9: Forecast} = 295 + 30(9) = 565$$

$$\text{Year 10: Forecast} = 295 + 30(10) = 595$$

3.2 Seasonality Issues in Forecasting

Up to this point we have seen several ways to make a forecast for an upcoming year. In many instances, managers may want more detail than just a yearly forecast. They may like to have a projection for individual time periods within that year (e.g., weeks, months, or quarters). Let's assume that our forecasted demand for an upcoming year is 480, but management would like a forecast for each of the quarters of the year. A simple approach might be to simply divide the total annual forecast of 480 by 4, yielding 120. We could then project that the demand for each quarter of the year will be 120. But of course, such forecasts could be expected to be quite inaccurate, for an examination of our original table of historical data reveals that demand is not uniform across each quarter of the year. There seem to be distinct peaks and valleys (i.e., quarters of higher demand and quarters of lower demand). The graph below of the historical quarterly demand clearly shows those peaks and valleys during the course of each year.



Sequential Quarters Over Six Years

Mechanisms for dealing with seasonality are illustrated over the next several pages.

Calculating Seasonal Index Values

This is the way you will find seasonal index values calculated in the textbook.

Begin by calculating the average demand in each of the four quarters of the year.

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Year	Q1	Q2	Q3	Q4	Annual Demand
1	62	94	113	41	310
2	73	110	130	52	365
3	79	118	140	58	395
4	83	124	146	62	415
5	89	135	161	65	450
6	94	139	162	70	465
Avg. Demand Per Qtr.	(62+73+79+83+89+94) ÷6=80	(94+110+118+124+135+139) ÷6=120	(113+130+140+146+161+162) ÷6=142	(41+52+58+62+65+70) ÷6=58	

Next, note that the total demand over these six years of history was 2400 (i.e., 310 + 365 + 395 + 415 + 450 + 465), and if this total demand of 2400 had been evenly spread over each of the 24 quarters in this six-year period, the average quarterly demand would have

been 100 units. Another way to look at this is the average of the quarterly averages is 100 units, i.e.

$$(80 + 120 + 142 + 58)/4 = 100 \text{ units.}$$

But, the numbers above indicate that the demand wasn't evenly distributed over each quarter. In Quarter 1 the average demand was considerably below 100 (it averaged 80 in Quarter 1). In Quarters 2 and 3 the average demand was considerably above 100 (with averages of 120 and 142, respectively). Finally, in Quarter 4 the average demand was below 100 (it averaged 58 in Quarter 4). We can calculate a seasonal index for each quarter by dividing the average quarterly demand by the 100 that would have occurred if all the demand had been evenly distributed across the quarters.

This would result in the following alternate seasonal index values:

Year	Q1	Q2	Q3	Q4
Seasonal	80/100 =	120/100 =	142/100 =	58/100 =
Index	.80	1.20	1.42	.58

A quick check of these alternate seasonal index values reveals that they average out to 1.0 (as they should). $(.80 + 1.20 + 1.42 + .58)/4 = 1.000$

Using seasonal index

The following forecasts were made for the next 4 years using the trend projection line approach (the trend projection formula developed was $Y = 295 + 30X$, where Y is the forecast and X is the year number).

Year	Forecast
7	505
8	535
9	565
10	595

If these annual forecasts were evenly distributed over each year, the quarterly forecasts would look like the following:

Year	Q1	Q2	Q3	Q4	Annual Forecast	Annual/4
7	126.25	126.25	126.25	126.25	505	126.25
8	133.75	133.75	133.75	133.75	535	133.75
9	141.25	141.25	141.25	141.25	565	141.25
10	148.75	148.75	148.75	148.75	595	148.75

However, seasonality in the past demand suggests that these forecasts should not be evenly distributed over each quarter. We must take these even splits and multiply them by the seasonal index (S.I.) values to get a more reasonable set of quarterly forecasts.

The results of these calculations are shown below.

S.I.	.80	1.20	1.42	.58	
Year	Q1	Q2	Q3	Q4	Annual
					Forecast
7	101.000	151.500	179.275	73.225	505
8	107.000	160.500	189.925	77.575	535
9	113.000	169.500	200.575	81.925	565
10	119.000	178.500	211.225	86.275	595

If you check these final splits, you will see that the sum of the quarterly forecasts for a particular year will equal the total annual forecast for that year (sometimes there might be a slight rounding discrepancy).

Other Methods for Making Seasonal Forecasts

Let's go back and reexamine the historical data we have for this problem. I have put a little separation between the columns of each quarter to let you better visualize the fact that we could look at any one of those vertical strips of data and treat it as a time series. For example, the Q1 column displays the progression of quarter 1 demands over the past six years. One could simply peel off that strip of data and use it along with any of the forecasting methods we have examined to forecast the Q1 demand in year 7. We could do the same thing for each of the other three quarterly data strips.

Year	Q1	Q2	Q3	Q4
1	62	94	113	41
2	73	110	130	52
3	79	118	140	58
4	83	124	146	62
5	89	135	161	65
6	94	139	162	70

To illustrate, I have used the linear trend line method on the quarter 1 strip of data, which would result in the following trend line:

$$Y = 58.8 + 6.0571X$$

For year 7, $X = 7$, so the resulting Q1 forecast for year 7 would be 101.200

We could do the same thing with the Q2, Q3, and Q4 strips of data. For each strip we would compute the trend line equation and use it to project that quarter's year 7 demand. Those results are summarized here:

Q2 trend line: $Y = 89.4 + 8.7429X$; Year 7 Q2 forecast would be 150.600

Q3 trend line: $Y = 107.6 + 9.8286X$; Year 7 Q3 forecast would be 176.400

Q4 trend line: $Y = 39.2 + 5.3714X$; Year 7 Q4 forecast would be 76.800

Total forecast for year 7 = $101.200 + 150.600 + 176.400 + 76.800 = 505.000$

These quarterly forecasts are in the same ballpark as those made with the seasonal index values earlier. They differ a bit, but we cannot say one is correct and one is incorrect. They are just slightly different predictions of what is going to happen in the future. They do provide a total annual forecast that is equal to the trend projection forecast made for year 7. (Don't expect this to occur on every occasion, but since it corroborates results obtained with a different method, it does give us confidence in the forecasts we have made.)

3.3 Measuring Forecast Accuracy

Mean Forecast Error (MFE): Forecast error is a measure of how accurate our forecast was in a given time period. It is calculated as the actual demand minus the forecast, or $E_t = A_t - F_t$

Forecast error in one time period does not convey much information, so we need to look at the accumulation of errors over time. We can calculate the average value of these forecast errors over time (i.e., a **Mean Forecast Error**, or **MFE**). Unfortunately the accumulation of the E_t values is not always very revealing, for some of them will be positive errors and some will be negative. These positive and negative errors cancel one another, and looking at them alone (or looking at the MFE over time) might give a false sense of security. To illustrate, consider our original data, and the accompanying pair of hypothetical forecasts made with two different forecasting methods.

		Hypothetical Forecasts	Forecast	Hypothetical Forecasts	Forecast
	Actual	Made With	Error With	Made With	Error With
	Demand	Method 1	Method 1	Method 2	Method 2
Year	A_t	F_t	$A_t - F_t$	F_t	$A_t - F_t$
1	310	315	-5	370	-60
2	365	375	-10	455	-90
3	395	390	5	305	90
4	415	405	10	535	-120
5	450	435	15	390	60

6	465	480	-15	345	120
Accumulated Forecast Errors			0		0
Mean Forecast Error, MFE			0/6=0		0/6=0

Based on the accumulated forecast errors over time, the two methods look equally good. However, most observers would judge that Method 1 is generating better forecasts than Method 2 (i.e., smaller misses).

Mean Absolute Deviation (MAD): To eliminate the problem of positive errors canceling negative errors, a simple measure is one that looks at the absolute value of the error (size of the deviation, regardless of sign). When we disregard the sign and only consider the size of the error, we refer to this deviation as the absolute deviation. If we accumulate these absolute deviations over time and find the average value of these absolute deviations, we refer to this measure as the mean absolute deviation (MAD). For our hypothetical two forecasting methods, the absolute deviations can be calculated for each year and an average can be obtained for these yearly absolute deviations, as follows:

		Hypothetical Forecasting Method 1			Hypothetical Forecasting Method 2		
	Actual		Forecast	Absolute		Forecast	Absolute
	Demand	Forecast	Error	Deviation	Forecast	Error	Deviation
Year	At	Ft	At - Ft	At - Ft	Ft	At - Ft	At - Ft
1	310	315	-5	5	370	-60	60
2	365	375	-10	10	455	-90	90
3	395	390	5	5	305	90	90
4	415	405	10	10	535	-120	120
5	450	435	15	15	390	60	60
6	465	480	-15	15	345	120	120
		Total Absolute Deviation		60			540
		Mean Absolute Deviation		60/6=10			540/6=90

The smaller misses of Method 1 have been formalized with the calculation of the MAD. Method 1 seems to have provided more accurate forecasts over this six year horizon, as evidenced by its considerably smaller MAD

Mean Squared Error (MSE): Another way to eliminate the problem of positive errors canceling negative errors is to square the forecast error. Regardless of whether the forecast error has a positive or negative sign, the squared error will always have a positive sign. If we accumulate these squared errors over time and find the average value of these squared errors, we refer to this measure as the mean squared error (MSE). For our hypothetical two forecasting methods, the squared errors can be calculated for each year and an average can be obtained for these yearly squared errors, as follows:

		Hypothetical Forecasting Method 1			Hypothetical Forecasting Method 2		
	Actual		Forecast	Squared		Forecast	Squared
	Demand	Forecast	Error	Error	Forecast	Error	Error
Year	A_t	F_t	$A_t - F_t$	$(A_t - F_t)^2$	F_t	$A_t - F_t$	$(A_t - F_t)^2$
1	310	315	-5	25	370	-60	3600
2	365	375	-10	100	455	-90	8100
3	395	390	5	25	305	90	8100
4	415	405	10	100	535	-120	14400
5	450	435	15	225	390	60	3600
6	465	480	-15	225	345	120	14400
		Total Squared Error		700			52200
		Mean Squared Error		$700/6 = 116.67$			$52200/6 = 8700$

Method 1 seems to have provided more accurate forecasts over this six-year horizon, as evidenced by its considerably smaller MSE.

The Question often arises as to why one would use the more cumbersome MSE when the MAD calculations are a bit simpler (you don't have to square the deviations). MAD does have the advantage of simpler calculations. However, there is a benefit to the MSE method. Since this method squares the error term, large errors tend to be magnified. Consequently, MSE places a higher penalty on large errors. This can be useful in situations where small forecast errors don't cause much of a problem, but large errors can be devastating.

Mean Absolute Percent Error (MAPE): A problem with both the MAD and MSE is that their values depend on the magnitude of the item being forecast. If the forecast item is measured in thousands or millions, the MAD and MSE values can be very large. To avoid this problem, we can use the MAPE. MAPE is computed as the average of the absolute difference between the forecasted and actual values, expressed as a percentage of the actual values. In essence, we look at how large the miss was relative to the size of the actual value. For our hypothetical two forecasting methods, the absolute percentage error can be calculated for each year and an average can be obtained for these yearly values, yielding the MAPE, as follows:

		Hypothetical Forecasting Method 1			Hypothetical Forecasting Method 2		
	Actual		Forecast			Forecast	
	Demand	Forecast	Error		Forecast	Error	
Year	A_t	F_t	$A_t - F_t$	Absolute % Error	F_t	$A_t - F_t$	Absolute % Error
				$100 A_t -$			$100 A_t -$

				F_t/A_t			F_t/A_t
1	310	315	-5	1.16%	370	-60	19.35%
2	365	375	-10	2.74%	455	-90	24.66%
3	395	390	5	1.27%	305	90	22.78%
4	415	405	10	2.41%	535	-120	28.92%
5	450	435	15	3.33%	390	60	13.33%
6	465	480	-15	3.23%	345	120	17.14%
		Total Absolute % Error		14.59%			134.85%
		Mean Absolute % Error		14.59/6=2.43%			134.85/6=22.48%

Method 1 seems to have provided more accurate forecasts over this six-year horizon, as evidenced by the fact that the percentages by which the forecasts miss the actual demand are smaller with Method 1 (i.e., smaller MAPE).

Illustration of the Four Forecast Accuracy Measures

Here is a further illustration of the four measures of forecast accuracy, this time using hypothetical forecasts that were generated using some different methods than the previous illustrations (called forecasting methods A and B; actually, these forecasts were made up for purposes of illustration). These calculations illustrate why we cannot rely on just one measure of forecast accuracy.

	Actual	Hypothetical Forecasting Method A					Hypothetical Forecasting Method B				
		Forecast	Absolute Error	Squared Deviation	Absolute % Error	Forecast	Absolute Error	Squared Deviation	Absolute % Error		
Year	A_t	F_t	$A_t - F_t$	$ A_t - F_t $	$(A_t - F_t)^2$	$ A_t - F_t /A_t$	F_t	$A_t - F_t$	$ A_t - F_t $	$(A_t - F_t)^2$	$ A_t - F_t /A_t$
1	310	330	-20	20	400	6.45%	310	0	0	0	0%
2	365	345	20	20	400	5.48%	365	0	0	0	0%
3	395	415	-20	20	400	5.06%	395	0	0	0	0%
4	415	395	20	20	400	4.82%	415	0	0	0	0%
5	450	430	20	20	400	4.44%	390	60	60	3600	13.33%
6	465	485	-20	20	400	4.30%	525	-60	60	3600	12.90%
		Totals	0	120	2400	30.55%	Totals	0	120	7200	26.23%
			MFE =	MAD =	MSE =	MAPE =		MFE =	MAD =	MSE =	MAPE =

			$0/6 =$	$120/6 =$	$2400/6$			$0/6 =$	$120/6 =$	$7200/6$	
			0	20	= 400	30.55/6		0	20	= 1200	26.23/6 4.37%

You can observe that for each of these forecasting methods, the same MFE resulted and the same MAD resulted. With these two measures, we would have no basis for claiming that one of these forecasting methods was more accurate than the other. With several measures of accuracy to consider, we can look at all the data in an attempt to determine the better forecasting method to use. Interpretation of these results will be impacted by the biases of the decision maker and the parameters of the decision situation. For example, one observer could look at the forecasts with method A and note that they were pretty consistent in that they were always missing by a modest amount (in this case, missing by 20 units each year).

However, forecasting method B was very good in some years, and extremely bad in some years (missing by 60 units in years 5 and 6). That observation might cause this individual to prefer the accuracy and consistency of forecasting method A. This causal observation is formalized in the calculation of the MSE. Forecasting method A has a considerably lower MSE than forecasting method B. The squaring magnified those big misses that were observed with forecasting method B. However, another individual might view these results and have a preference for method B, for the sizes of the misses relative to the sizes of the actual demand are smaller than for method A, as indicated by the MAPE calculations.

4.0 CONCLUSION

The key in forecasting nowadays is to understand the different forecasting methods and their relative merits and so be able to choose which method to apply in a particular situation (for example consider how many time series forecasting methods the package has available). All forecasting methods involve tedious repetitive calculations and so are ideally suited to be done by a computer. Forecasting packages, many of an interactive kind (for use on pc's) are available to the forecaster.

5.0 SUMMARY

Too many factors in the business environment cannot be predicted with certainty. Therefore, rather than search for the perfect forecast, it is far more important to establish the practice of continual review of forecasts and to learn to live with inaccurate forecasts. This is not to say that we should not try to improve the forecasting model or methodology, but that we should try to find and use the best forecasting method available, within reason. Because forecasts deal with past data, our forecasts will be less reliable the further into the future we predict. That means forecast accuracy decreases as time horizon increases. The accuracy of the forecast and its costs are interrelated. In general, the higher the need for accuracy translates to higher costs of developing forecasting models.

6.0 TUTOR MARKED ASSIGNMENT

A manager wanted to establish the seasonal pattern of the units of a particular product X demanded by his client. The following table contains the quarterly sales, that is, the average number of units sold during each quarter of the last 5 year

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2011	1861	2203	2415	1908
2012	1921	2343	2514	1986
2013	1834	2154	2098	1799
2014	1837	2025	2304	1965
2015	2073	2414	2339	1967

Calculate the seasonal indices and deseasonalize the time series

7.0 REFERENCES AND FURTHER READING

- Finch, Byron J. (2006). *Operations Now: Profitability, Processes, Performance*. 2 ed. Boston: McGraw-Hill Irwin.
- Green, William H. (2003). *Econometric Analysis*. 5 ed. Upper Saddle River, NJ: Prentice Hall.
- Joppe, Dr. Marion. "The Nominal Group Technique." *The Research Process*.
- Stevenson, William J. (2005). *Operations Management*. 8 ed. Boston: McGraw-Hill Irwin, 2005.

UNIT 3 DETERMINISTIC INVENTORY CONTROL MODELS

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Some concepts in the Inventory Control Theories

3.2 Model derivation Approaches:

3.2.1 Graphical method

3.2.2 Calculus approach

3.3 The Inventory Control Systems

3.3.1 Re-order Level System

3.3.2 The Periodic Review System

3.4 Inventory Control Models

3.4.1 Basic EOQ Models

3.4.2 The Adapted Model

4.0 Conclusion

5.0 Summary

6.0 Tutor-Marked Assignment

7.0 References/Further Readings

1.0 INTRODUCTION

An inventory must be carried by a firm for various reasons and some of the reason why are, for anticipating normal demand and taking advantage of bulk-purchase discounts. Other reason could be to meet emergency shortages and as a natural part of the production process and for absorbing wastages and unpredictable fluctuations. Decisions involving inventories have become an essential part of management decisions. It is important to note that a business firm must know how it can efficiently and effectively manage its inventories if one of its major objectives is business survival. The aim of an inventory control system is to minimise costs and establish, ordered the optimum amount of stock and the period between orders.

2.0 OBJECTIVES

- Define inventory and explain what inventory control is all
- Apply first order difference equations to estimate Inventory Control and EOQ model
- Apply modern inventory control models.

3.0 MAIN CONTENT

3.1 Some concepts in the Inventory Control Theories

A firm's inventory is defined in terms of the total stocks of various kinds including, the basic raw materials, partly finished goods and materials, subassemblies, office and workshop supplies and finished goods.

Like any other business decision variables these are some of concept associated with inventory control theories you need to know these to enable you make meanings out of inventory theories. These include:

- i. Ordering (Replacement) Costs:** These are such costs as transportation costs, clerical and administrative costs associated with the physical movement of the purchased external goods. Where the goods are manufactured internally, there are alternative initial costs to be borne with each production run referred to as set-up costs
- ii. Holding (Carrying) Costs** These are:
 - (a) storage costs in terms of staffing, equipment maintenance, and handling;
 - (b) storage overheads (heat, light, rent, and the like);
 - (c) cost of capital tied up in inventory;
 - (d) insurance, security and pilferage;
 - (e) deterioration or breakage
- iii. Stock out Costs**

These are costs associated with running out of stock. These include penalty payments, loss of goodwill, idle manpower and machine, and the like.
- iv. Lead Time:** This is the time between ordering of goods and their replenishment. Orders may be internal (requiring a production run) or external.
- v. Economic Ordering Quantity (EOQ):** This refers to the external order quantity that minimises total inventory costs.
- vi. Economic Batch Quantity (EBQ):** This refers to the size of the internal production run that minimises total inventory costs.
- vii. Safety Stock:** This is a term used to describe the stock held to cover possible deviations in demand or supply during the lead time. It is sometimes referred to as buffer or minimum stock.
- viii. Maximum Stock:** This is a level used as an indicator above which stocks are deemed to be too high.
- ix. Reorder Level:** This is the level of stock, which when reached, signals replenishment order.
- x. Reorder Quantity:** This is the level of replenishment order.

3.2 Model derivation Approaches:

There are two methods of Model derivation Approaches Graphical method and Calculus approach.

3.2.1 Graphical method

The purpose of inventory graph is to present the inventory control problems in graphical terms. It plots the relationship between quantity of stock held (Q) and time (t).

Figure 1 presents a general inventory graph with various features. It shows an initial inventory of 100 items, replenished by a further 100 items continuously over a given time period. Observe as indicated that for the next time period, there was no activity, but at time period 2, 100 items were demanded, followed, over the next two periods, by a continuous demand which used up the last 100 items. This stock out position led to the delivery of additional 150 items.

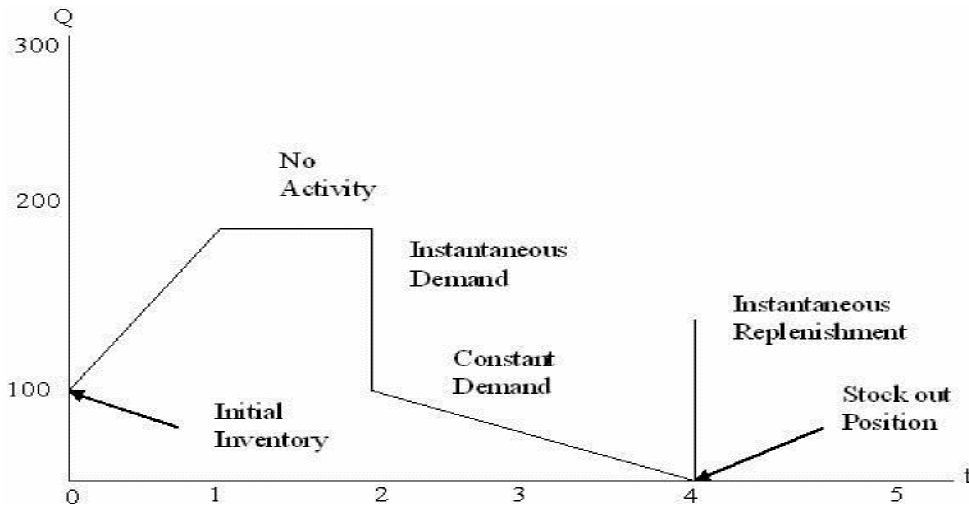


Figure 1: General Inventory Graph

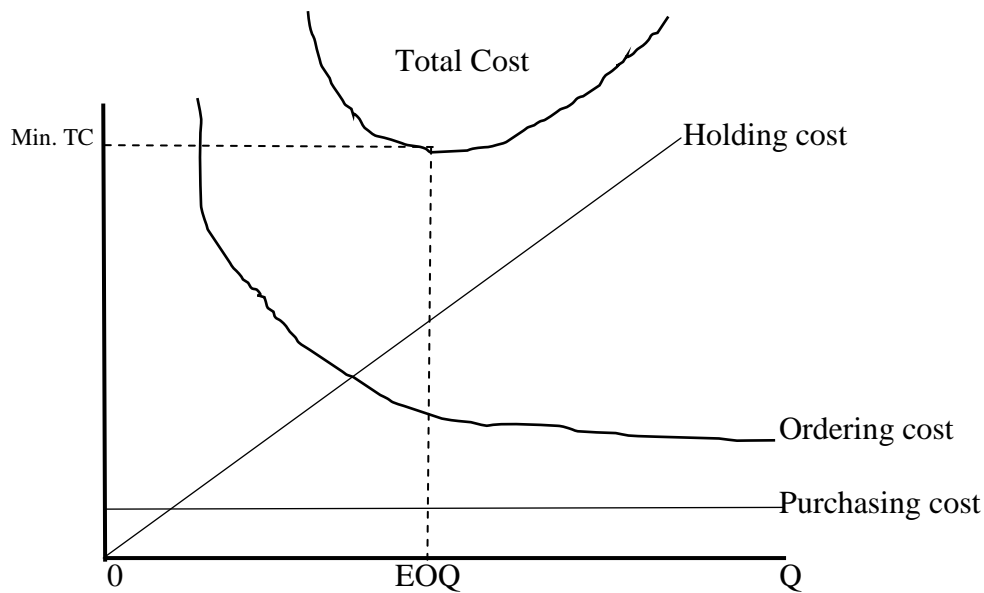


Figure 2: The General Inventory Graph

Observations about graph

- TC function shape is not influenced by the purchase cost i.e it is irrelevant for EOQ determination where there are no discounts.
- Total cost is minimum where holding cost = ordering cost.

3.3.2 Calculus approach

TC = Purchase Cost + Holding Cost + Ordering Cost (recall there are no shortage costs)

In symbolic form:

$$TC = DC_p + \frac{QCh}{2} + \frac{D}{Q}C_o$$

The objective now is to find Q which minimizes TC. We apply first order and second order conditions as follows:

$$\text{FOC: } \frac{dTC}{dQ} = \frac{C_h}{2} - DC_oQ^{-2} = 0$$

$$\frac{Ch}{2} - DC_o/Q^2 = 0$$

TC = Purchase Cost + Holding Cost + Ordering Cost (recall there are no shortage costs)

In symbolic form:

$$TC = DC_p + \frac{QCh}{2} + \frac{D}{Q}C_o$$

The objective now is to find Q which minimizes TC. We apply first order and second order conditions as follows:

$$\text{FOC: } \frac{dTC}{dQ} = \frac{C_h}{2} - DC_oQ^{-2} = 0$$

$$\frac{Ch}{2} - DC_o/Q^2 = 0$$

$$C_h/2 = DC_o/Q^2$$

Re-arranging:

$$Q^2 = 2DC_o/C_h$$

$$\text{Hence } Q = \sqrt{\frac{2DC_o}{C_h}}$$

To confirm the turning point is a minimum, we apply SOC as follows;

$$\text{SOC: } d^2TC/dQ^2 = 2DC_oQ^{-3} = 2DC_o/Q^3 > 0 \text{ i.e +ve,}$$

since D, Q, C_o are all positive values. Hence turning point is minimum.

$$\text{Note: } EOQ = \sqrt{\frac{2DC_o}{C_{pi}}} \text{ Or } EOQ = \sqrt{\frac{2DC_o}{C_h}}$$

Example

Star Logistics Ltd has established that annual quantity for a given item is 4000 units. The cost of placing an order is ₦5000 and the price per unit is ₦ 2000. Inventory holding cost percentage is 20% of purchase cost.

Required:

- a) Formulate the best (optimal) entry policy for this item i.e.
 - Quantity to order (EOQ)
 - Frequency for ordering and when to order
 - Re-order level/point; For ROP take lead-time to be 15 days while one year has 300 working days
 - Total cost associated with the policy.
- b) Suppose it actually turns out that
- c) Ordering cost per order = ₦6000 and
- d) Inventory hold cost percentage $I = 15\%$ and yet the policy formulated in (a) above is implemented for a year determine the cost of prediction error.

a) **From the illustration we can see that;** annual demand, $D = 4000$ units; cost of ordering, $C_o = ₦5000$; carrying cost percentage, $i = 20\%$ of unit cost; unit purchase cost, $C_p = ₦200$. Then;

$$EOQ = \sqrt{\frac{2DC_o}{C_{pi}}} = \sqrt{\frac{2 \times 4000 \times 5000}{200 \times 0.2}} = 1000 \text{ units}$$

Frequency of ordering

This is related to the annual number of orders, N which is given as;

$$N = \frac{D}{Q} = \frac{4000}{1000} = 4 \text{ orders}$$

Given that one year is 12 months, therefore make an order after every, $12/4 = 3$ months or quarterly.

Alternatively

If one year is 300 working days make an order every $300/4 = 75$ days or 2.5 months

Re-order level /re-order point (ROP)

ROP- represents the quantity remaining when an order is being made

= Usage during lead time period

= daily usage x lead time

= annual usage/No of days in the year * Lead time

$$= \frac{4000}{300} \times 15 = 200 \text{ units}$$

TC = purchase cost + Holding cost + Ordering cost

$$\begin{aligned}
&= DCp + \frac{Q}{2} \times Cpi + \frac{D}{Q} C_o \\
&= 4000 \times 200 + \left(\frac{1000}{2} \times 200 \times 0.2 \right) + \frac{4000}{1000} \times 5000 \\
&= 800,000 + 20,000 + 20,000 = \text{₦}840,000
\end{aligned}$$

Optimal policy

Given that some parameter estimates changed from predicted ones, there is need to calculate EOQ afresh using all correct parameters.

$$EOQ = \sqrt{\frac{DCo}{Cpi}} = \sqrt{\frac{2 \times 4000 \times 6000}{200 \times 0.15}} = 1265 \text{ units}$$

Relevant total cost for Optimal policy = $1265/2 * 200 * 0.15 + 4000/1265 * 6000 = \text{₦}37947$

Cost of prediction error = $39,000 - 37,947 = \text{₦} 1053$

3.3 The Inventory Control Systems

3.3.1 Re-order Level System

This is the most commonly used control system. It generally results in lower stocks. The system also enables items to be ordered in more economic quantities and is more responsive to fluctuations in demand than the second system discussed below. The system sets the value of three important levels of stock as warning t or action triggers for management:

Re-order Level: This is an action level of stock which leads to the replenishment order, normally the Economic Order Quantity (EOQ). For a particular time period, the re-order level is computed as follows:

(i) $L_{ro} = \text{maximum usage per period} \times \text{maximum lead time (in periods)}$

(ii) **Minimum Level:** This is a warning level set such that only in extreme cases (above average demand or late replenishment) should it be breached. It is computed as follows:

$$L_{min} = \text{Re-Order Level} - (\text{normal Usage} \times \text{Average lead time})$$

(iii) **Maximum Level:** This is another warning level set such that only in extreme cases (low levels of demand) should it be breached. It is computed by:

$$L_{max} = \text{Re-order Level} + \text{EOQ} - (\text{Minimum usage} \times \text{Minimum lead time})$$

Example:

Suppose for a particular inventory, there exists:

- (a) the weekly minimum, normal and maximum usage of 600, 1000, and 1400 respectively;

- (b) the lead time which vary between 4 and 8 weeks (average = 6 weeks); and,
- (c) the normal ordering quantity (EOQ) of 20,000.

It follows that:

The Re-order Level (L_{ro}) = $1400 \times 8 = 11,200$ units

Minimum Stock Level (L_{min}) = $11,200 - 1000 \times 6 = 5,200$ units Maximum

Stock Level (L_{max}) = $11,200 + 20,000 - 600 \times 4 = 28,800$

3.3.2 The Periodic Review System

- i. It enables stock positions to be reviewed periodically so that the chances of obsolete stock items are minimised.
- ii. Economies of scale are possible when many items are ordered at the same time or in the same sequence.

3.4 Inventory Control Models

Two basic inventory control models are currently in use. These include:

The basic EOQ Models and the Adapted Basic Model (with gradual replenishment)

3.4.1 Basic EOQ Models (Basic EOQ Model without and with Discounts)

Economic order quantity (EOQ) is the ideal order quantity a company should purchase to minimize inventory costs such as holding costs, shortage costs, and order costs.

The basic inventory control model is based on the following assumptions:

- The rate of demand (that is, the number of items demanded per year), D , is constant and continuous over a given period, and no excess demands.
- The ordering cost ($C_o = N/\text{circle}$) is constant and independent of the quantity ordered.
- Only one type of stock item is considered and its price ($P = N/\text{item}$) is constant.
- The holding cost ($C_h = N/\text{item}$) is the cost of carrying one article in stock for one year.
- The quantity ordered per circle (q) is supplied to store instantaneously whenever the inventory level becomes zero.

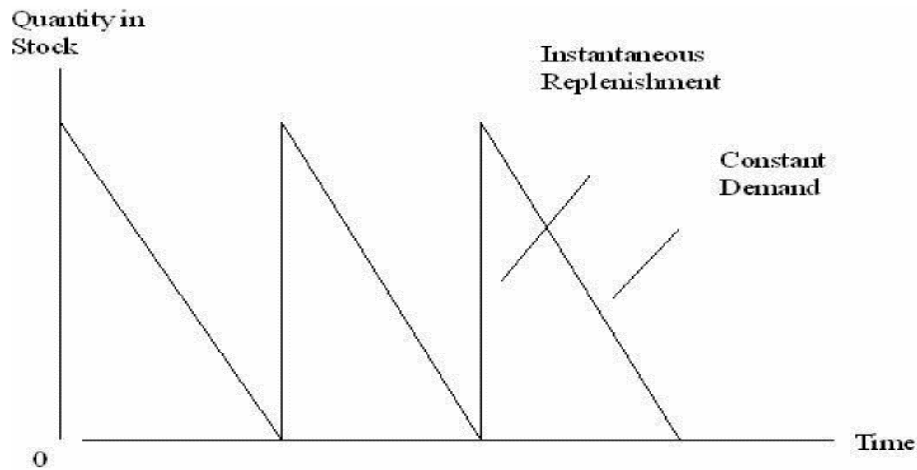


Figure 3: Inventory Graph for the Basic Inventory Control Model

For the **Basic EOQ Model**, the total annual inventory cost is minimised when the Economic Ordering Quantity (EOQ) takes the following value:

$$EOQ = \sqrt{\frac{2DC_o}{C_h}}$$

Where D = annual demand

C_o = Order Cost per circle

C_h = Holding Cost per item

The total annual inventory cost, C , is defined by:

C = Total Ordering Cost + Total Holding Cost = number of orders per year x order cost + average inventory level x holding cost per item

$$C = \frac{D}{q} \times C_o + \frac{q}{2} \times C_h$$

Notice from the above equation that as q gets larger:

- Annual ordering cost becomes smaller and
- Annual holding cost becomes larger.

The basic model also involves the calculation of the following statistics:

- i. Number of orders per year = $\frac{\text{Yearly Demand}}{\text{EOQ}}$
- ii. Length of cycle (days) = $\frac{\text{Number of Days per Year}}{\text{Number of Orders per Year}}$
- iii. Average Inventory Level = $\frac{\text{EOQ}}{2}$

Characteristics of Basic Economic Order Quantity (EOQ) Model

1. Durable (not perishable) product model.
2. Merchandizing (not manufacturing) firm.
3. Single product (not multiple product) model.

Illustration: Suppose, ordering cost per order is ₦30,000, the following schedule for ordering cost **per unit** for selected order quantities follow:

Quantity ordered	Ordering cost Per unit (Naira.)
1	30,000
10	3000
100	300
1000	30

Thus, ordering cost per unit is declining non-linear function of quantity ordered.

Basic EOQ Model with Discounts

Presence of discount has certain advantage and disadvantages as follows;

Types of discounts: there are generally two types;

- 1) **Single discount offer** e.g. unit selling prices is ₦10 but purchases of 100 units and above will get a discount of 3%.
- 2) **Multiple discount offers** – also called price-breaks, here the supplier provides a list of price /quantity ranges such as the following;

Quantity Range	Unit selling price (₦)
1 – 100	10
101 – 200	9.70
201 – 400	9.50
Above 400	9.00

Examples:

A commodity has a steady rate of demand of 2,000 units per year. Placing an order costs N200 and it costs N50 to hold a unit for a year:

- (a) Estimate the Economic Order Quantity (EOQ)
- (b) Find the number of orders placed per year
- (c) What is the length of the inventory circle?

Solution

Note that from the given information, $D = 2000$; $C_o = 200$; and $C_h = 50$

(a) The Economic Order Quantity is determined by

$$EOQ = \sqrt{\frac{2DC_o}{C_h}} = \sqrt{\frac{2(2000)(200)}{50}} = \sqrt{16000} = 126.491$$

Thus, the economic order quantity is about 126.5 units

$$(b) \text{ Number of orders per year} = \frac{\text{Yearly Demand}}{EOQ} = \frac{2000}{126.491} = 15.81$$

Thus, the number of orders per year is approximately 16 orders.

$$(c) \text{ Length of cycle (days)} = \frac{\text{Number of Days per Year}}{\text{Number of Orders per Year}} = \frac{365}{15.81} = 23.1$$

The cycle length is about 23 days

Example

A Ltd buys 400 units of an item at a purchase cost ₦5000 per unit and ordering cost of ₦2,000 per order placed. The carrying cost is estimated at 24% of cost of an item p.a. The Co. has received a 2% discount offer for purchases of 100 or more units.

Required:

- Determine the best inventory policy for this item.
- Determine the discount level at which the firm will be indifferent between taking and not taking the discount offer.

Steps

- Calculate the EOQ with no discount and find the resulting TC
- Find TC when discount is taken. The EOQ will be lowest quantity to just qualify for discount
- Compare TC in 1 and 2 and hence make the decision

Solution

a)1) **No discount**

$$EOQ = \sqrt{\frac{2DC_o}{C_h}} = \sqrt{\frac{2 \times 400 \times 2000}{5000 \times 0.24}} = 37 \text{ units}$$

$$TC = DC_p + \frac{Q}{2} C_h + \frac{D}{Q} \times C_o$$

$$= 400 \times 5000 + \frac{37}{2} \times 5000 \times 0.24 + \frac{400}{37} \times 2000 = \text{₦}2,043,822$$

2) **Taking discount offer**

EOQ with discount = 100 units, since this is the least quantity to just qualify for discount.

$$C_p \text{ with discount} = 5000 \times (1 - 0.02) = 49,000$$

$$TC = 400 \times 4900 + 100/2 \times 4900 \times 0.24 + \frac{400}{100} \times 2000 = \text{N} 2,026,800$$

3) Decision:

Take the discount offer since it is less costly than NOT taking it.

∴ EOQ = 100 units

Timing of orders

$$\text{Annual No. of orders, } N = \frac{D}{Q} = \frac{400}{100} = 4 \text{ orders}$$

Therefore, make an order every $12/4 = 3$ months or quarterly

b) Let x represent discount level for cost indifference point. At this point the following conditions will hold:

$$TC \text{ (with discount)} = TC \text{ (no discount)}$$

$$400 \times 5000(1 - x) + \left(\frac{100}{2} \times 5000(1 - x) \times 0.24 \right) + \frac{400}{100} \times 2000 = 2043822$$

$$2000000(1 - x) + 250000(1 - x) \times 0.24 + 8000 = 2043822$$

$$2000000 - 2000000x + 60000 - 60000x + 8000 = 2043822$$

$$2068000 - 2060000x = 2043822$$

$$2068000 - 2043822 = 2060000x$$

$$x = 0.0117$$

$$x = 1.17\% = 1.2\% \text{ (1 dp)}$$

Decision Rule

If discount is greater than 1.2% take it but less than 1.2% do not take it.

Multiple discount offers (price breaks)

This is an extension of the single discount offer in the sense that a price / quantity schedule is available or provided instead of a single offer.

The solution approach can be broken down into the following steps:

1. Calculate EOQ for each price/quantity range.
2. The EOQ calculated in 1 will fall in one of 3 categories which will be differently treated as follows:
 - a) **Below range** – Ignore the calculated EOQ but calculate TC for the least quantity in the range.
 - b) **Within range** – Evaluate total cost for the EOQ calculated.
 - c) **Above range** – Ignore the range since there would be another range which will yield lower total cost.

3.4.2 The Adapted Model

The adapted model assumes gradual replenishment unlike the basic model. This assumption is on the premises that when stocks are received from the production line, it is very likely that finished items are received continuously over a period of time. Stock is therefore, subject to gradual replenishment. Like the basic model, this model also assumes constant demand rate, D , and the two cost factors: the order cost per cycle, C_s , and the holding cost per item, C_h .

The adapted model is often referred to as the *production run* model whereby:

- (i) a production run is started every time the inventory level decreases to zero, and stops when q items have been produced or supplied. The run lasts for time t and is known as the *run time*
- (ii) the quantity ordered per cycle is referred to as the *run size*, and items are supplied at the rate of P per annum
- (iii) the effective replenishment rate is defined as $P - D$ items

The adapted model uses the following optimal statistics:

- i. Number of run per year = $\frac{\text{Yearly Demand}}{\text{EBQ}}$
- ii. Length of cycle (days) = $\frac{\text{Number of Days per Year}}{\text{Number of Runs per Year}}$
- iii. Run time (days) = $\frac{\text{EBQ} - \text{Numbers of days per year}}{\text{Annual production rate}}$
- iv. Peak inventory level = Effective Replenishment rate – run time
- v. Average inventory level = $\frac{1}{2}$ X peak inventory level

The value of the run size that minimises inventory cost for the adapted model is defined by Economic Batch Quantity (EBQ),

Economic Batch Quantity (EBQ), also known as the optimum production quantity (EPQ), is the order size of a production batch that minimizes the total cost. **Economic Batch Quantity (EBQ)** is a formula for calculating the **quantity** of inventory that a company should order in cases where the resupply is gradual e.g. when the company produces its own inventory and takes a while to complete production. Think of a company assembling vehicles.

Batch production is a technique which is commonly used today for distributing the total production in a series of small batches rather than mass producing in one go.

Sometimes the production of goods in batches is necessary because, for example, certain equipment used in manufacturing (e.g. dyes) may wear out and need replacement before the production can run again.

Batch production may be desirable in other cases as well. For example, where the objects being produced are perishable, the entire production requirement for say a year can't be manufactured in a week as it might cause the goods to expire after some time. Batch production also reduces the risk of obsolescence as any minor changes required in the specification of goods (e.g. size, colour, etc.) can be made in future batches according to the feedback received from customers or retailers instead of producing everything in one go and hoping for the best.

Whereas EOQ is suitable for determining the **order size** when the parts, materials or finished goods are ready to be delivered by **external** suppliers when the order is placed, EBQ is used to determine the size of a production run (i.e. **batch size**) when the manufacturing takes place **internally** and any raw materials or parts required for production are either acquired internally or are supplied incrementally by other companies according to the production requirement.

Formula for EBQ is given as;

$$\text{Economic Batch Quantity (EBQ)} = \sqrt{\frac{2DC_s}{C_h(1-D/P)}}$$

Where:

- C_s is the setup cost of a batch
- D = Annual Demand
- P is the annual production capacity
- C_h is the annual cost of holding one unit of finished inventory

The formula for calculating EBQ is very similar to EOQ with one notable difference in the denominator. The cost of holding in EBQ formula is decreased by the amount of inventory that will be produced and sold on the same day therefore not contributing to the annual cost of holding the inventory.

Example

Abu owns and operates a small factory that manufactures plastic bottles which he sells to bottling companies.

Additional information:

- Annual demand is 1 million bottles spread evenly over the year
- Setup cost is \$5000 per batch
- Holding cost is \$3 per annum for each bottle
- Maximum production capacity is 2 million bottles per annum
- Currently, bottles are manufactured in 10 batches

- A. Find the optimum production quantity that Sarah should produce to minimize her costs
- B. Calculate the current annual holding cost and setup cost
- C. Calculate the savings to Sarah if she adopts the EBQ

Solution A: Optimum Production Quantity

Economic Batch Quantity

$$(EBQ) = \sqrt{\frac{2DC_s}{C_h(1-D/P)}} = \sqrt{\frac{2 \times 5000 \times 1,000,000}{3 \times (1 - (1,000,000/2,000,000))}}$$

$$= \sqrt{\frac{10,000,000,000}{1.5}} = \sqrt{6,666,666,666} = 81,650$$

Abu should manufacture bottles in batches of 81,650 units.

Solution B: Current Costs

$$\begin{aligned} \text{Batch Quantity} &= \text{Annual Demand} \div \text{Number of batches} \\ &= 1,000,000 \div 10 \\ &= 100,000 \text{ units} \end{aligned}$$

$$\begin{aligned} \text{Annual Holding Cost} &= (\text{Batch Quantity}/2) \times C_h \times (1 - D/P) \\ &= (100,000/2) \times 3 \times (1 - (1,000,000/2,000,000)) \\ &= \$75,000 \end{aligned}$$

$$\begin{aligned} \text{Setup Cost} &= \text{Number of setups} \times \text{setup cost} \\ &= 10 \times 5000 \\ &= \$50,000 \end{aligned}$$

$$\text{Total Current Cost} = (\$75,000 + \$50,000) = \$125,000$$

Solution C: Savings from EBQ

$$\begin{aligned} \text{Annual Holding Cost:} \\ &= (\text{Batch Quantity}/2) \times C_h \times (1 - D/P) \\ &= (81,650/2) \times 3 \times (1 - (1,000,000/2,000,000)) \\ &= \$61,238 \text{ (A)} \end{aligned}$$

Setup Cost

$$\text{Number of batches} = 1,000,000 \div 81,650 = 12.2475$$

$$\begin{aligned} \text{Setup Cost} &= \text{Number of batches} \times \text{Cost of setup} \\ &= 12.2475 \times \$5000 = \$61,23 \text{ (B)} \end{aligned}$$

$$\text{Total Cost (EBQ)} = (\text{A}) + (\text{B}) = \$122,476 \text{ (C)}$$

$$\text{Total Current Cost} = 125,000 \text{ (D)}$$

$$\text{Savings} = (\text{D}) - (\text{C}) = 2,524$$

SELF ASSESSMENT EXERISE

A manufacturing activity requires a continuous supply of 3000 items per year from store, replenished by production runs, each of which operates at the constant rate of 5000 items per year. Each production run has a set-up cost of N300, and the holding cost per item per annum is N25. Compute the Economic Batch Quantity (EBQ), and find the number of runs per year, the length of cycle, the run time, the peak inventory level, and the average inventory level.

4 CONCLUSION

For many small businesses, the primary source of revenue is inventory turnover. If a company maintains too large an inventory, inventory storage, spoilage and obsolescence costs can increase operating costs and decrease a company's income. The economic order quantity model minimizes these inventory costs by determining the optimal inventory quantity the company should keep on hand and ensuring inventory arrives in time to meet customer needs. The EOQ model accomplishes these goals by monitoring inventory and, when inventory levels fall below a certain point, ordering the inventory needed to avoid shortages. To meet these requirements, the EOQ model determines the optimal-reorder quantity for each order as well as the appropriate-reorder point. EBQ is basically a refinement of the **economic order quantity (EOQ)** model to take into account circumstances in which the goods are produced in **batches**. The goal of calculating EBQ is that the product is produced in the required **quantity** and required quality at the lowest cost.

5 SUMMARY

The aim of an inventory control system is to minimise costs and establish the optimum amount of stock to be ordered and the period between orders. In the EOQ model, it is assumed that the orders are received all at once. However, in the EBQ model, this assumption is relaxed. The unit shows that there are two types of costs: those which increase with the batch size such as working capital investment in materials and labour, cost of handling and storing materials, insurance and tax charges, interest on capital investment, etc., and those which decrease with the batch size such as cost (per unit) of setting up machines, cost of preparing paper work that enters and controls the production of the order.

6.0 TUTOR-MARKED ASSIGNMENT

a) A company buys 30,000 units for an item per year at an ordering cost of ₦2500 per order and holding cost chards are 20% of the cost of average inventory per annum.

The following price quality schedule is available from the supplier.

Quantity (Unit)	Unit price (₦)
1 3000	21:00
300 5000	19:00
5000 7000	17:00

7001 9000	15:50
9001 and above	13:50

Required: Recommend the best inventory policy for this item.

b) An inventory system follows the adapted basic inventory control model. The demand rate is constant at 1,600 items per year, the unit price of the items is N1.80, the holding cost is 10 percent of the unit price per year, and the set-up cost is N12 per run. If the production rate is 5,000 items per year, calculate:

- (a) the optimum run size (EBQ)
- (b) the run time
- (c) the cycle length
- (d) the average inventory level

7.0 REFERENCES/FURTHER READINGS

Abdullahi S. A. & Onyebuenyi, F. E. Course Code: BFN 728, Quantitative Techniques for Financial Decisions, National Open University of Nigeria, Faculty of Management Sciences

Laurence D. Hoffmann & Gerald L. Bradley (2007) Applied Calculus for Business, Economics, and the Social and Life Sciences, Expanded Ninth Edition

Michael K. Chirchir and Joash N. Mageto Dps 302 Inventory Management

Onwe, O. J. (2007) Statistical Methods for Business and Economic Decisions: A Practical Approach (Lagos: Samalice Publishers)

MODULE 4: DATA ANALYSIS TECHNIQUES AND STATISTICAL SOFTWARE IN APPLIED QUANTITATIVE TECHNIQUES

Unit 1 An Overview of Quantitative Research

Unit 2 Quantitative Data

Unit 3 Data Analysis Tools in Applied Quantitative Techniques

UNIT 1 AN OVERVIEW OF QUANTITATIVE RESEARCH

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Conceptual Issues in Quantitative Research

3.1.1 Background, Definition & Analyzing Quantitative Research

3.1.2 Differences Between Quantitative & Qualitative Research

3.1.3 Quantitative Scales of Measurement

3.1.4 Approaches to Quantitative Research

3.1.5 Research Questions and Hypotheses

3.2 Key Issues in Quantitative Research

3.3 Variables and Operational Definitions

4.0 Conclusion

5.0 Summary

6.0 Tutor-Marked Assignment

7.0 References/Further Readings

1.0 INTRODUCTION

This unit is aimed at acquainting learners with an overview of conceptual issues in quantitative research analysis, including a discussion of the necessary steps and types of statistical analyses. Quantitative Research studies result in data that provides quantifiable, objective, and easy to interpret results. The data can typically be summarized in a way that allows for generalizations that can be applied to the greater population and the results can be reproduced. The design of most quantitative research studies also helps to ensure that personal bias does not impact the research methods. This unit describes some of the most commonly used quantitative analysis procedures. The unit also discusses the basics of measurement and scales of measurement commonly used in quantitative research approaches and key issues related to quantitative research that must be addressed to ensure a quality research study that is valid, reliable, generalizable and reproducible. And the unit explains the different types of variables and discusses operational definitions of variables in quantitative research.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Define quantitative research.
- Distinguish between quantitative and qualitative research
- Determine the appropriate measurement scale for a research problem
- Describe how to identify the appropriate approach for a particular research problem
- Explain the difference between a research question and a research hypothesis and describe the appropriate use of each.
- Define validity, reliability, falsifiability, generalizability, and reproducibility as they relate to quantitative research
- Define and explain operational definitions and provide examples

3.0 MAIN CONTENT

3.1 Conceptual Issues in Quantitative Research Analysis

Here research refers to activities aimed at obtaining new knowledge about the world, in the case of the social sciences the social world of people and their institutions and interactions. Here we are concerned solely with empirical research, where such knowledge is based on information obtained by observing what goes on in that world. This unit briefly presented below, some basic or fundamental quantitative research concepts used in applied quantitative Analysis. This understanding is required for analysing.

3.1.1 Background, Definition and Analyzing Quantitative Research

a) Background and Definition

In natural and social sciences, and sometimes in other fields, quantitative research is the systematic empirical investigation of observable phenomena via statistical, mathematical, or computational techniques. The objective of quantitative research is to develop and employ mathematical models, theories, and hypotheses pertaining to phenomena. The process of measurement is central to quantitative research because it provides the fundamental connection between empirical observation and mathematical expression of quantitative relationships. Qualitative research, on the other hand, inquires deeply into specific experiences, with the intention of describing and exploring meaning through text, narrative, or visual-based data, by developing themes exclusive to that set of participants.

Quantitative research is widely used in psychology, economics, demography, sociology, marketing, community health, health & human development, gender studies, and political science; and less frequently in anthropology and history. Research in mathematical sciences, such as physics, is also "quantitative" by definition, though this use of the term differs in context. In the social sciences, the term relates to empirical methods originating

in both philosophical positivism and the history of statistics, in contrast with qualitative research methods.

In another definition, quantitative research refers to research conducted using numerical data. This research can be descriptive in nature or analyzed using inferential statistical methods for the purpose of hypothesis testing. This differs from qualitative research which tends to be exploratory in nature, using information that is not believed to be naturally quantifiable.

Applied research refers to a study conducted not for the purpose of mere curiosity, but for immediate application to a real-world problem. With that being said, applied quantitative research refers to research that provides statistical conclusions with the intentions of answering a specific question or solving a specific problem. The design of most quantitative studies also helps to ensure that personal bias does not impact the data. Quantitative data can be analyzed in several ways.

Once a researcher has written the research question, the next step is to determine the appropriate research methodology necessary to study the question. The three main types of research design methods are qualitative, quantitative and mixed methods. The focus of this set of units is quantitative research.

Quantitative methods are used to examine the relationship between variables with the primary goal being to analyze and represent that relationship mathematically through statistical analysis. This is the type of research approach most commonly used in scientific research problems. Following is a list of characteristics and advantages of using quantitative methods:

- i. The data collected is numeric, allowing for collection of data from a large sample size.
- ii. Statistical analysis allows for greater objectivity when reviewing results and therefore, results are independent of the researcher.
- iii. Numerical results can be displayed in graphs, charts, tables and other formats that allow for better interpretation.
- iv. Data analysis is less time-consuming and can often be done using statistical software.
- v. Results can be generalized if the data are based on random samples and the sample size was sufficient.
- vi. Data collection methods can be relatively quick, depending on the type of data being collected.
- vii. Numerical quantitative data may be viewed as more credible and reliable, especially to policy makers, decision makers, and administrators.

There are a variety of quantitative methods and sampling techniques that will be discussed in detail in the other units in this unit. However, following are examples of research questions where quantitative methods may be appropriately applied:

- What is the difference in the number of calories consumed between male and female high school students?
- What percentage of married couples seek couples counseling?
- What are the top 5 factors that influence a student's choice of college or university?

b) Analyzing Quantitative Research

The first step in quantitative data analysis is to identify the levels or scales of measurement as nominal, ordinal, interval or ratio. See the Research Ready: Scales of Measurement unit for more information on the scales of measurement. This is an important first step because it will help you determine how best to organize the data. The data can typically be entered into a spreadsheet and organized or “coded” in some way that begins to give meaning to the data.

The next step would be to use descriptive statistics to summarize or “describe” the data. It can be difficult to identify patterns or visualize what the data is showing if you are just looking at raw data. Following is a list of commonly used descriptive statistics:

- Frequencies – a count of the number of times a particular score or value is found in the data set
- Percentages – used to express a set of scores or values as a percentage of the whole
- Mean – numerical average of the scores or values for a particular variable
- Median – the numerical midpoint of the scores or values that is at the center of the distribution of the scores
- Mode – the most common score or value for a particular variable
- Minimum and maximum values (range) – the highest and lowest values or scores for any variable

It is now apparent why determining the scale of measurement is important before beginning to utilize descriptive statistics. For example, nominal scales where data is coded, as in the case of gender, would not have a mean score. Therefore, you must first use the scale of measurement to determine what type of descriptive statistic may be appropriate. The results are then expressed as exact numbers and allow you to begin to give meaning to the data. For some studies, descriptive statistics may be sufficient if you do not need to generalize the results to a larger population. For example, if you are comparing the percentage of teenagers that smoke in private versus public high schools, descriptive statistics may be sufficient.

However, if you want to utilize the data to make inferences or predictions about the population, you will need to go another step farther and use inferential statistics. Inferential statistics examine the differences and relationships between two or more samples of the population. These are more complex analyses and are looking for significant differences between variables and the sample groups of the population. Inferential statistics allow you to test hypotheses and generalize results to the population as a whole. Following is a list of basic inferential statistical tests:

- Correlation – seeks to describe the nature of a relationship between two variables, such as strong, negative, positive, weak, or statistically significant. If a correlation is found, it indicates a relationship or pattern, but keep in mind that it does not indicate or imply causation
- Analysis of Variance (ANOVA) – tries to determine whether or not the means of two sampled groups is statistically significant or due to random chance. For example, the test scores of two groups of students are examined and proven to be significantly different. The ANOVA will tell you if the difference is significant, but it does not speculate regarding “why”.
- Regression – used to determine whether one variable is a predictor of another variable. For example, a regression analysis may indicate to you whether or not participating in a test preparation program results in higher ECONOMICS scores for high school students.

SELF-ASSESSMENT EXERCISE

What are the steps involved in analysing quantitative data?

3.1.2 Differences between Quantitative & Qualitative Research

Research is a systematic investigation that aims to generate knowledge about a particular phenomenon. However, the nature of this knowledge varies and reflects your study objectives. Some study objectives seek to make standardised and systematic comparisons, others seek to study a phenomenon or situation in detail. These different intentions require different approaches and methods, which are typically categorised as either quantitative or qualitative. You have probably already made decisions about using qualitative or quantitative data for monitoring and evaluation. Perhaps you have had to choose between using a questionnaire or conducting a focus group discussion in order to gather data for a particular indicator.

1. Quantitative research

Quantitative research typically explores specific and clearly defined questions that examine the relationship between two events, or occurrences, where the second event is a consequence of the first event. Such a question might be: ‘what impact did the programme have on children’s school performance?’ To test the causality or link between the programme and children’s school performance, quantitative

researchers will seek to maintain a level of control of the different variables that may influence the relationship between events and recruit respondents randomly. Quantitative data is often gathered through surveys and questionnaires that are carefully developed and structured to provide you with numerical data that can be explored statistically and yield a result that can be generalised to some larger population.

2. Qualitative research

Research following a qualitative approach is exploratory and seeks to explain ‘how’ and ‘why’ a particular phenomenon, or programme, operates as it does in a particular context. As such, qualitative research often investigates i) local knowledge and understanding of a given issue or programme; ii) people’s experiences, meanings and relationships and iii) social processes and contextual factors (e.g., social norms and cultural practices) that marginalise a group of people or impact a programme. Qualitative data is non-numerical, covering images, videos, text and people’s written or spoken words. Qualitative data is often gathered through individual interviews and focus group discussions using semi- structured or unstructured topic guides.

Table 3. 1: Key differences between qualitative and quantitative research

	Qualitative research	Quantitative research
Type of knowledge	Subjective	Objective
Aim	Exploratory and observational	Generalisable and testing
Characteristics	Flexible	Fixed and controlled
	Contextual portrayal	Independent and dependent
	Dynamic, continuous view of	Pre- and post-measurement of
Sampling	Purposeful	Random
Data collection	Semi-structured or unstructured	Structured
Nature of data	Narratives, quotations, descriptions	Numbers, statistics
	Value uniqueness, particularity	Replication
Analysis	Thematic	Statistical

Qualitative research produces information only on the particular cases studied, and any more general conclusions are only hypotheses. Quantitative methods can be used to verify which of such hypotheses are true.

3.1.3 Quantitative Scales of Measurement

Quantitative research requires that measurements be both accurate and reliable. Researchers commonly assign numbers or values to the attributes of people, objects, events, perceptions, or concepts. This process is referred to as measurement. The variables that are measured are commonly classified as being measured on a nominal, ordinal, interval or ratio scale. The following discussion defines and provides examples of each of the four levels of measurement.

Nominal Scale: The nominal scales is essentially a type of coding that simply puts people, events, perceptions, objects or attributes into categories based on a common trait or characteristic. The coding can be accomplished by using numbers, letters, colours, labels or any symbol that can distinguish between the groups. The nominal scale is the lowest form of a measurement because it is used simply to categorize and not to capture additional information. Other features of a nominal scale are that each participant or object measured is placed exclusively into one category and there is no relative ordering of the categories. Some examples include distinguishing between smokers and nonsmokers, males and females, types of religious affiliations, and so on. In a study related to smoking, smokers may be assigned a value of 1 and nonsmokers may be assigned a value of 2. The assignment of the number is purely arbitrary and at the researcher's discretion.

Ordinal Scale: The ordinal scale differs from the nominal scale in that it ranks the data from lowest to highest and provides information regarding where the data points lie in relation to one another. An ordinal scale typically uses non-numerical categories such as low, medium and high to demonstrate the relationships between the data points. The disadvantage of the ordinal scale is that it does not provide information regarding the magnitude of the difference between the data points or rankings. An example of the use of an ordinal scale would be a study that examines the smoking rates of teenagers. The data collected may indicate that the teenage smokers in the study smoked anywhere from 15 to 40 cigarettes per day. The data could be arranged in order and examined in terms of the number of smokers at each level.

Interval Scale: An interval scale is one in which the actual distances, or intervals between the categories or points on the scale can be compared. The distance between the numbers or units on the scale are equal across the scale. An example would be a temperature scale, such as the Fahrenheit scale. The distance between 20 degrees and 40 degrees is the same as between 60 degrees and 80 degrees. A distinguishing feature of interval scales is that there is no absolute zero point because the key is simply the consistent distance or interval between categories or data points.

Ratio Scale: The ratio scale contains the most information about the values in a study. It contains all of the information of the other three categories because it categorizes the data, places the data along a continuum so that researchers can examine categories or data points in relation to each other, and the data points or categories are equal distances or intervals apart. However, the difference is the ratio scale also contains a non-arbitrary absolute zero point. The lowest data point collected serves as a meaningful absolute zero point which allows for interpretation of ratio comparisons. Time is one example of the use of a ration measurement scale in a study because it is divided into equal intervals and a ratio comparison can be made. For example, 20 minutes is twice as long as 10 minutes.

SELF-ASSESSMENT EXERCISE

List and describe the four types of scales of measurement used in quantitative research.

3.1.4 Approaches To Quantitative Research

There are four main types of quantitative research designs: descriptive, correlational, quasi-experimental and experimental. The differences between the four types primarily relates to the degree the researcher designs for control of the variables in the experiment. Following is a brief description of each type of quantitative research design, as well as chart comparing and contrasting the approaches.

A **Descriptive Design** seeks to describe the current status of a variable or phenomenon. The researcher does not begin with a hypothesis, but typically develops one after the data is collected. Data collection is mostly observational in nature.

A **Correlational Design** explores the relationship between variables using statistical analyses. However, it does not look for cause and effect and therefore, is also mostly observational in terms of data collection.

A **Quasi-Experimental Design** (often referred to as Causal-Comparative) seeks to establish a cause-effect relationship between two or more variables. The researcher does not assign groups and does not manipulate the independent variable. Control groups are identified and exposed to the variable. Results are compared with results from groups not exposed to the variable.

Experimental Designs, often called true experimentation, use the scientific method to establish cause-effect relationship among a group of variables in a research study. Researchers make an effort to control for all variables except the one being manipulated (the independent variable). The effects of the independent variable on the dependent variable are collected and analysed for a relationship.

4.1 KEY ISSUES IN QUANTITATIVE RESEARCH

If the results of quantitative research are to be considered useful and trustworthy, there are several key issues that must be considered and addressed as part of the experimental design and analysis. Following is a description of these issues:

1) Validity

The term validity refers to the strength of the conclusions that are drawn from the results. In other words, how accurate are the results? Do the results actually measure what was intended to be measured? There are several types of validity that are commonly examined and they are as follows:

- i. Conclusion validity looks at whether or not there is a relationship between the variable and the observed outcome.
- ii. Internal validity considers whether or not that relationship may be causal in nature.
- iii. Construct validity refers to whether or not the operational definition of a variable actually reflects the meaning of the concept. In other words, it is an attempt to generalize the treatment and outcomes to a broader concept.
- iv. External validity is the ability to generalize the results to another setting.

2) Reliability

Reliability is defined as the consistency of the measurements. To what level will the instrument produce the same results under the same conditions every time it is used? Reliability adds to the trustworthiness of the results because it is a testament to the methodology if the results are reproducible. The reliability is often examined by using a test and retest method where the measurement are taken twice at two different times. The reliability is critical for being able to reproduce the results, however, the validity must be confirmed first to ensure that the measurements are accurate. Consistent measurements will only be useful if they are accurate and valid.

3) Falsifiability

The term falsifiability mean that any for any hypothesis to have credence, it must be possible to test whether that hypothesis may be incorrect. A researcher should test his/her own hypothesis to prove or disprove it before releasing results to prevent another researcher from proving it wrong. If a theory or hypothesis cannot be tested in such a way that may disprove it, it will likely not be considered scientific or valuable to those in the field.

4) Generalizability

Generalizability refers to whether or not the research findings and conclusions that result from the study are generalizable to the larger population or other similar situations. The ability to generalize results allows researchers to interpret and apply findings in a broader context, making the finding relevant and meaningful.

5) Replication

Replication is the reproducibility of the study. Will the methodology produce the same results when used by different researchers studying similar subjects?

Replication is important because it ensures the validity and reliability of the results and allows the results to be generalized.

3.1.5 Research Questions and Hypotheses

Once you have chosen your research topic or subject, you will need to decide how you will approach the research process – by formulating a hypothesis or developing a research question. This can be determined by starting with the following questions. Is there a significant body of knowledge already available about your subject that allows you to make a prediction about the results of your study before you begin? If so, you will be using a hypothesis. Or is your research more exploratory and investigative in nature and will require that you collect data and analyse results before drawing any conclusions? If this describes your research topic, you will be developing a research question. Understanding this difference and choosing the correct approach will drive the rest of your research project. The following sections further describe research questions and hypotheses and provide examples of each.

1) Research Questions:

- Used to analyse and investigate a topic. It is written as a question and is inquisitive in nature.
- A properly written question will be clear and concise. It should contain the topic being studied (purpose), the variable(s), and the population.
- **Three main types of questions:**
 - Causal Questions – Compares two or more phenomena and determines if a relationship exists. Often called relationship research questions. Example: Does the amount of calcium in the diet of elementary school children effect the number of cavities they have per year?
 - Descriptive Questions – Seek to describe a phenomenon and often study “how much”, “how often”, or “what is the change”. Example: How often do college-aged students use Twitter?
 - Comparative Questions – Aim to examine the difference between two or more groups in relation to one or more variables. The questions often begin with “What is the difference in...” Example: What is the difference in caloric intake of high school girls and boys?
- The type of research question will influence the research design.
- Once data has been collected, it will be analyzed and conclusions can be made.

2) Hypothesis:

- It is predictive in nature and typically used when significant knowledge already exists on the subject which allows the prediction to be made.
- Data is then collected, analyzed, and used to support or negate the hypothesis, arriving at a definite conclusion at the end of the research.

- It is always written as a statement and should be developed before any data is collected.
- A complete hypothesis should include: the variables, the population, and the predicted relationship between the variables.
- Commonly used in quantitative research, but not qualitative research which often seeks answers to open-ended questions.
- Examples: A company wellness program will decrease the number sick days claimed by employees. Consuming vitamin C supplements will reduce the incidence of the common cold in teenagers.

SELF ASSESSMENT EXERISE

What is the relationship between hypothesis and research question?

4.2 Variables and Operational Definitions

The goal of quantitative research is to examine the relationships between variables. A variable is a characteristic or attribute of interest in the research study that can take on different values and is not constant. Variables may be straightforward and easy to measure including characteristics such as gender, weight, height, age, size, and time. Other variables may be more complex and more difficult to measure. Examples of these types of variables may include socioeconomic status, attitudes, achievement, education level, and performance.

This unit will focus on five types of variables: independent, dependent, extraneous, moderator and mediator variables. The two primary types of variables are **dependent and independent variables**. An independent variable is the variable manipulated or changed by the researcher. The independent variable affects or determines the values of dependent variable. The dependent variable is sometimes referred to as the outcome variable because the resulting outcome of manipulating the independent variable is typically the focus of the research study. The dependent variable is the one that the researcher is attempting to predict or explain. The distinction between independent and dependent variables is especially important when studying cause-effect relationships. Following are two examples:

- A researcher wants to study the effectiveness of different dosages of a particular antibiotic in clearing an infection.
 - The independent variable - varying dosages of antibiotic.
 - The dependent variable - the presence or absence of infection following a specific time period.
- The researcher plans to study the relationship between the amount of time spent in a study group and test scores.
 - The independent variable – number of hours spent in a study group.
 - The dependent variable – test scores.

Extraneous variables, sometimes referred to as nuisance or confounding variables, are not the variables of primary interest. However, they are believed to be related to the independent or dependent variable and therefore, may impact the results. Researchers should attempt to control extraneous variables in order to attain meaningful results. If they cannot be controlled, extraneous variables should at least be considered when interpreting results.

A **moderator variable** is a variable that interacts with the independent variable and may influence the strength of the relationship between the independent and dependent variables. This variable is measured and taken into consideration, making it different than an extraneous variable. For example, if studying the relationship between exercise and weight loss, the number of calories consumed maybe a moderating variable.

Mediating variables, commonly referred to as intervening variables, are processes that may not be observable but link the independent and dependent variables. An instructor may have a new teaching approach for a mathematical concept and plans to study the use of this approach and its relationship to test scores. The differing levels at which students in the class are able to process abstract mathematical concepts is a mediating variable.

While it is important to identify, understand, and consider the variables within a study, the researcher must also consider the measurement of those variables and the types of values that may be collected. When measuring the values of variables, there are two main classifications: categorical and quantitative variables. Categorical variables are those that express a qualitative attribute and do not express a numerical ordering. These variables refer to different types or categories of phenomenon or characteristic. Some examples would include gender, eye colour, race, religion, payment method, or social status. Quantitative variables vary in degree or amount and are expressed using numerical ordering. Height, weight, shoe size, income, and test scores are quantitative variables.

The specific way in which a variable is measured in a particular study is called the **operational definition**. It is critical to operationally define a variable in order to lend credibility to the methodology and to ensure the reproducibility of the results. Another study may measure the same variable differently. The operational definition also helps to control the variable by making the measurement constant. Therefore, when it comes to operational definitions of a variable, the more detailed the definition is, the better. For example, if the researcher was planning to weigh research subjects, there would several constructs that should be spelled out including what the subjects were to wear, whether or not they would wear shoes, what type of scale was being used, and time of day. It may also be important to define the measurement of the outcome. For example, if a study was examining the relationship of swimming on overall fitness, the researcher would need to define how the outcome of overall fitness would be measured.

SELF-ASSESSMENT EXERCISE

What are the key variables in saying whether the information is important or not?

4.0 CONCLUSION

Quantitative research is often contrasted with qualitative research, which purports to be focused more on discovering underlying meanings and patterns of relationships, including classifications of types of phenomena and entities, in a manner that does not involve mathematical models. Qualitative research, on the other hand, inquires deeply into specific experiences, with the intention of describing and exploring meaning through text, narrative, or visual-based data, by developing themes exclusive to that set of participants. Quantitative research is generally closely affiliated with ideas from 'the scientific method', which can include: the generation of models, theories and hypotheses; the development of instruments and methods for measurement; experimental control and manipulation of variables; collection of empirical data; modeling and analysis of data. Measurement is often regarded as being only a means by which observations are expressed numerically in order to investigate causal relations or associations. However, it has been argued that measurement often plays a more important role in quantitative research. For example, within quantitative research, the results that are shown can prove to be strange. This is because accepting a theory based on results of quantitative data could prove to be a natural phenomenon

5.0 SUMMARY

This unit briefly presented some basic or fundamental quantitative research concepts. This unit discusses the basics of measurement and scales of measurement commonly used in quantitative research. You now hopefully have a better understanding of the difference between quantitative and qualitative. The unit shows that measurement and scales, provides an excellent overview of measurement terminology. The four scales of measurements, a concept map of when to use each type of scale, specific examples and information concerning the development of a scale. In this unit, the four approaches to quantitative research are described and examples are provided. The type of data analysis will also depend on the number of variables in the study. Studies may be univariate, bivariate or multivariate in nature. Validity is seen by many as being the primary issue that should be examined. The unit explains the different types of variables in quantitative research and discusses operational definitions of variables. Identifying and defining variables is a critical first step in a research study and will impact the validity and reliability of the study.

6.0 TUTOR-MARKED ASSIGNMENT

- i. Understanding the Differences between Constructs, Variables, and Operational Definitions
- ii. Discuss the importance of validity, reliability, falsifiability, generalizability, and reproducibility in a quantitative study.

- iii. Discuss and provide examples of inferential statistical analyses in applied quantitative analysis

7.0 REFERENCES/FURTHER READINGS

- Abramson, J. H., & Abramson, Z. H. (2008). Scales of Measurement. *Research Methods in Community Medicine: Surveys, Epidemiological Research, Programme Evaluation, Clinical Trials*, Sixth Edition, 125-132.
- Adcock, R. (2001). Measurement validity: A shared standard for qualitative and quantitative research. In *American Political Science Association* (Vol. 95, No. 03, pp. 529-546). Cambridge University Press.
- Bennett, J. A. (2000). Mediator and moderator variables in nursing research: Conceptual and statistical differences. *Research in nursing & health*, 23(5), 415-420.
- Bernard, H. R., & Bernard, H. R. (2012). *Social Research Methods: Qualitative and Quantitative Approaches*. Sage.
- Blaikie, N. (2003). *Analysing quantitative data: From description to explanation*. Sage.
- Bryman, A., & Cramer, D. (1994). *Quantitative data analysis for social scientists* (rev. Taylor & Frances/Routledge).
- Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications, Incorporated.
- Monitoring, Evaluation, Accountability and Learning (MEAL), *Methods of data collection and analysis*, Save the child, Open University
- Neuman, W. L., & Robson, K. (2004). *Basics of social research*. Pearson.
- Onwuegbuzie, A. J. (2000). *Expanding the Framework of Internal and External Validity in Quantitative Research*.
- Punch, K. F. (2013). *Introduction to social research: Quantitative and qualitative approaches*. Sage.

UNIT 2 QUANTITATIVE DATA CONCEPTS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Quantitative Data Definition, Characteristics & Hypotheses
 - 3.1.1 **Error! Hyperlink reference not valid.**
 - 3.1.2 Quantitative Data Characteristics
 - 3.2 When To Use Quantitative Methods & Sampling Methods
 - 3.3 Examples of Primary & Secondary Data Problem
 - 3.3.1 Primary Data Analysis Using a Chi-Square Method
 - 3.3.2 Secondary Data Analysis Using a Regression Method
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

UNIT 2 QUANTITATIVE DATA CONCEPTS

CONTENTS

1.0 INTRODUCTION

There is a growing competition and need for companies and institutions to collect and analyze quantitative data (primary and secondary data collection). Many potential researchers are unsure where they can obtain data to begin their research and analysis. This is done by various means depending on the nature and type of data. Once the need of data collection and its format is decided, process of data collection starts. Secondary data is critical when you are defining the problems and objectives that are the focus of your Marketing Intelligence initiatives. However, in most cases, you will need to collect primary data of some kind in order to have the information you need to make real decisions. Quantitative data can be analyzed in several ways. This unit discusses some of the most commonly concepts surrounding quantitative data. Data is thought to be the lowest unit of information from which other measurements and analysis can be done. This unit explains the common methods of quantitative data collection distinguish between primary and secondary data in research methods.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Define quantitative data and its characteristics.

- Describe common methods of quantitative data collection
- Distinguish between primary and secondary data in research methods
- Describe when quantitative research methods should be used to examine a research problem
- Define sampling and randomization.
- Demonstrate how to compute primary and secondary data

3.0 MAIN CONTENT

3.1 Quantitative Data Definition, Characteristics and Hypotheses

Data collection is basically the process of collecting data from different sources under any specified work environment to get answers to some predefined questions trying to gather information about some certain issue. Data is collected for various purposes like educational & research purposes, business research, marketing purpose and almost everything you can think of.

3.1.1 [Error! Hyperlink reference not valid.](#)

Definition of Quantitative Data

Data can be defined as the quantitative or qualitative values of a variable or primary or secondary. Data is plural of datum which literally means to give or something given. Data can be numbers, images, words, figures, facts or ideas. Data in itself cannot be understood and to get information from the data one must interpret it into meaningful information. Quantitative data is any data that is in numerical form such as statistics, percentages, etc. The researcher analyses the data with the help of statistics and hopes the numbers will yield an unbiased result that can be generalized to some larger population.

The word **quantitative** in the title of the course indicates that the methods you will learn here are used to analyze quantitative data. This means that the data will enter the analysis in the form of *numbers* of some kind. In social sciences, for example, data obtained from administrative records or from surveys using structured interviews are typically quantitative.

Quantitative data is defined as the value of data in the form of counts or numbers where each data-set has a unique numerical value associated with it. This data is any quantifiable information that can be used for mathematical calculations and statistical analysis, such that real-life decisions can be made based on these mathematical derivations. Quantitative data is used to answer questions such as “How many?”, “How often?”, “How much?” This data can be verified and can also be conveniently evaluated using mathematical techniques.

For example, there are quantities corresponding to various parameters, for instance, “How much did that laptop cost?” is a question which will collect quantitative data. There are values associated with most measuring parameters such as pounds or kilograms for weight, dollars for cost etc. Quantitative data makes measuring various parameters controllable due to the ease of mathematical derivations they come with. Quantitative data is usually collected for statistical analysis using surveys, polls or questionnaires sent across to a specific section of a population. The retrieved results can be established across a population.

SELF-ASSESSMENT EXERCISE

- i. What does quantitative data mean?
- ii. Time, in hours, spent working per year is what kind of variable?

Types of Quantitative Data Collection Techniques

There are two main quantitative data collection techniques. Primary and Secondary data collection techniques

Primary Data

When someone refers to "primary data" they are referring to data collected by the researcher himself/herself. This is data that has never been gathered before, whether in a particular way, or at a certain period of time. Primary data means original data that has been collected specially for the purpose in mind. It means someone collected the data from the original source first hand. Researchers tend to gather this type of data when what they want cannot be found from outside sources. You can tailor your data questions and collection to fit the need of your research questions.

Primary data collection uses surveys, experiments or direct observations. Data collected this way is called primary data. Primary data has not been published yet and is more reliable, authentic and objective. Primary data has not been changed or altered by human beings; therefore, its validity is greater than secondary data.

This can be an extremely costly task and, if associated with a college or institute, requires permission and authorization to collect such data. Issues of consent and confidentiality are of extreme importance. Primary data actually follows behind secondary data because you should use current information and data before collecting more so you can be informed about what has already been discovered on a particular research topic.

Survey is most commonly used method in social sciences, management, marketing and psychology to some extent. Surveys can be conducted in different methods. Questionnaire is the most commonly used method in survey. Questionnaires are a list of questions either an open-ended or close-ended for which the respondent give answers. Questionnaire can be conducted via telephone, mail, live in a public area, or in an institute, through electronic mail or through fax and other methods. Interview is a face-to-

face conversation with the respondent. It is slow, expensive, and they take people away from their regular jobs, but they allow in-depth questioning and follow-up questions. Observations can be done while letting the observing person know that he is being observed or without letting him know. Observation can also be made in natural settings as well as in artificially created environment.

Methods of primary data collection:

- Personal investigation: The surveyor collects the data himself/herself. The data so collected is reliable but is suited for small projects.
- Collection via Investigators: Trained investigators are employed to contact the respondents to collect data.
- Questionnaires: Questionnaires may be used to ask specific questions that suit the study and get responses from the respondents. These questionnaires may be mailed as well.
- Telephonic Investigation: The collection of data is done through asking questions over the telephone to give quick and accurate information.
- Registration: registers and licenses are particularly valuable for complete enumeration, but are limited to variables that change slowly, such as numbers of fishing vessels and their characteristics.
- Direct observations: making direct measurements is the most accurate method for many variables, such as catch, but is often expensive. Many methods, such as observer programmes, are limited to industrial fisheries.

Advantages of Primary Data

- Data interpretation is better
- Targeted Issues are addressed
- Efficient Spending for Information
- Decency of Data.
- Addresses Specific Research Issues.
- Greater Control
- Proprietary Issues.

Disadvantages of Primary Research

- High Cost
- Time Consuming
- Inaccurate Feed-backs
- More number of resources is required

Secondary Data

Secondary data is the data that has been already collected by and readily available from other sources. When we use Statistical Method with Primary Data from another purpose

for our purpose, we refer to it as Secondary Data. It means that one purposes Primary Data is another purposes Secondary Data. So that secondary data is data that is being reused. This type of data typically comes from other studies done by other institutions or organizations. Such data are more quickly obtainable than the primary data. These secondary data may be obtained from many sources, including literature, industry surveys, compilations from computerized databases and information systems, and computerized or mathematical models of environmental processes.

Secondary data collection may be conducted by collecting information from a diverse source of documents or electronically stored information, census and market studies are examples of a common sources of secondary data. This is also referred to as "data mining." Secondary data may be more appropriate for your research.

Published Printed Sources. There are varieties of published printed sources. Their credibility depends on many factors. For example, on the writer, publishing company and time and date when published. New sources are preferred and old sources should be avoided as new technology and researches bring new facts into light. Books are available today on any topic that you want to research. The uses of books start before even you have selected the topic. After selection of topics books provide insight on how much work has already been done on the same topic and you can prepare your literature review. Books are secondary source but most authentic one in secondary sources. Journals/periodicals Journals and periodicals are becoming more important as far as data collection is concerned. The reason is that journals provide up-to-date information which at times books cannot and secondly, journals can give information on the very specific topic on which you are researching rather talking about more general topics. Magazines are also effective but not very reliable. Newspaper on the other hand is more reliable and in some cases the information can only be obtained from newspapers as in the case of some political studies.

Published Electronic Sources. As internet is becoming more advance, fast and reachable to the masses; it has been seen that much information that is not available in printed form is available on internet. In the past the credibility of internet was questionable but today it is not. The reason is that in the past journals and books were seldom published on internet but today almost every journal and book is available online. Some are free and for others you have to pay the price. E-journals: e-journals are more commonly available than printed journals. Latest journals are difficult to retrieve without subscription but if your university has an e-library you can view any journal, print it and those that are not available you can make an order for them. General Websites; Generally, websites do not contain very reliable information so their content should be checked for the reliability before quoting from them. Weblogs: Weblogs are also becoming common. They are actually diaries written by different people. These diaries are as reliable to use as personal written diaries.

Methods of Secondary Data Collection:

- Official publications such as the Ministry of Finance, Statistical Departments of the government, Federal Bureaus, Agricultural Statistical boards. E.g. World Bank Data, IMF Data etc.
- Semi-official sources include State Bank, Boards of Economic Enquiry, etc.
- Data published by Chambers of Commerce and trade associations and boards.
- Articles in the newspaper, from journals and technical publications.
- information collected through censuses or government departments like housing, social security, electoral statistics, tax records
- internet searches or libraries
- progress reports

Advantages of Secondary Data

- Inexpensive
- Easily accessible
- Immediately available
- Will provide essential background and help to clarify or refine research problem – essential for literature review
- Secondary data sources will provide research method alternatives
- Will also alert the researcher to any potential difficulties.

Disadvantages of Secondary Data

- Not immediately available – takes time to define problem, sampling frame, method and analysis.
- Not as readily accessible
- Incomplete Information

Tips for Sourcing Your Data

Good data is nothing without good sourcing. Next time you're scoping data for a project, follow these five tips to make sure your sources are solid.

1. **Use Recent Data.** The world changes quickly. Oftentimes, data that was produced or collected 10 or even 5 years ago is obsolete. Always use the most recently published version of the data available. In all cases, be upfront about how old the data set is.
2. **Use Only Reliable Data.** Verify that the source you choose is relevant, legitimate, and as non-biased as possible. Strong sources include data collected/produced by government agencies, such as the statistics compiled by the World Bank, IMF financial data, Nigeria National Bureau of Statistics, CBN, Nigeria National Population commission, OPEC, Other top-tier data sources can include industry white papers or academic publications. Remember that surveys conducted by polling agencies or think tanks, while usable, will often have a political agenda. Therefore, as with the case of aged data, use discretion.

3. **Use Only Primary Sources.** If you come across an interesting piece of information on Wikipedia or in a magazine article, never take the publication's word for it. Go to the primary data source. Without reviewing the primary document yourself, you'll never know if the methods were flawed, the sample size too small, or the questionnaire biased.
4. **Limit Your Sources.** Finding multiple data sets from multiple sources on one subject is exciting, but you can't create a consistent narrative with 15 different sources. Try to find one or two cohesive sets to work with.
5. **Use Complementary Sources.** Even if you only use two data sources, they can still create a lot of variance. Using two data sets that clash, such as data collected by think tanks on the opposite sides of the political spectrum, is just an easy way to make crafting a narrative difficult. Make sure that the sources you use complement each other and don't introduce inconsistent or biased information. Complementary sources are the same type of data, collected in the same time frame, using similar questionnaire designs.

Error! Hyperlink reference not valid.

Listed below are some examples of quantitative data that can help understand exactly what this pertains:

- 73 people downloaded the latest mobile application.
- My aunt lost 18 pounds last year.
- 140 respondents were of the opinion that the new product feature will not be successful.
- There will be 30% increase in revenue with the inclusion of a new product.
- 400 people attended the seminar.
- 58% people prefer shopping online instead of going to the mall.
- She has 6 holidays in this year.
- Product X costs ₦100,000
- I updated my phone 10 times in a quarter.
- My teenager grew by 4 inches last year.

As you can see in the above 10 examples, there is a numerical value assigned to each parameter and this is known as, quantitative data.

Error! Hyperlink reference not valid.

Some of advantages of quantitative data, are:

- **Conduct in-depth research:** Since quantitative data can be statistically analyzed, it is highly likely that the research will be detailed.
- **Minimum bias:** There are instances in research, where personal bias is involved which leads to incorrect results. Due to the numerical nature of quantitative data, the personal bias is reduced to a great extent.
- **Accurate results:** As the results obtained are objective in nature, they are extremely accurate.

Error! Hyperlink reference not valid.

Some of disadvantages of quantitative data, are:

- **Restricted information:** Because quantitative data is not descriptive, it becomes difficult for researchers to make decisions based solely on the collected information.
- **Depends on question types:** Bias in results is dependent on the question types included to collect quantitative data. The researcher's knowledge of questions and the objective of research are exceedingly important while collecting quantitative data.

SELF-ASSESSMENT EXERCISE

What is primary data? Discuss the various method of collecting primary data and what are the relative advantages

3.1.2 Quantitative Data Characteristics

Quantitative data is data that can be counted or expressed numerically. It is commonly used to ask “how much” or “how many” and can be used to study events or levels of occurrence. Because it is numerical in nature, quantitative data is both definitive and objective. It also lends itself to statistical analysis and mathematical computations and therefore, is typically illustrated in charts or graphs.

There are two main types of quantitative data: discrete and continuous. **Discrete data** is described as having a finite number of possible values. For example, if a teacher gives an exam that has 100 questions, the exam scores reflect the number of answers that were correct out of the 100 possible questions. Discrete data may also be defined as data where there is space between values on a number line, thus values must be a whole number. For example, if a study examined the number of vehicles owned by households in America, the data collected would be whole numbers. **Continuous data** is defined as data where the values fall on a continuum and it is possible to have fractions or decimals. Continuous data is usually a physical measurement. Examples may include measurements of height, age, or distance.

Quantitative data collection may include any method that will result in numerical values. Common examples of quantitative data collection strategies may include:

- Experiments and clinical trials
- Surveys, interviews and questionnaires that collect numerical information or count data by using closed-ended questions
- Observing or recording well-defined events such as the number of visits patients make to a doctor's office each year
- Obtaining information from a management information system.

The advantage of collecting quantitative data is that the numerical outcomes result in data that can be statistically analyzed that may be viewed as credible and useful in decision-

making. However, the disadvantage of quantitative data is that it may be superficial and fail to fully capture explanatory information.

SELF ASSESSMENT EXERISE

- i. State and discuss whether the actual data are discrete or continuous. The number of cars crossing the Third Mainland Bridge in Lagos each hour.
- ii. Name four variables from the healthcare field of study that are considered continuous and four that are discrete.

3.2 When to Use Quantitative Methods & Sampling Methods

1) When to Use Quantitative Methods

This unit describes when to choose quantitative methodology in research. The previous unit provided an overview and general definitions of quantitative research, as well as several examples.

Researchers should begin by asking themselves the following questions:

- What type of question am I asking?
- What type of data will I need to collect to answer the question?
- What type of results will I report?

For example, a researcher may want to determine the link between income and whether or not families have health insurance. This is a question that asks “how many” and seeks to confirm a hypothesis. The methods will be highly structured and consistent during data collection, most likely using a questionnaire with closed-ended questions. The results will provide numerical data that can be analyzed statistically as the researcher looks for a correlation between income and health insurance. Quantitative methodology would best apply to this research problem. A quantitative approach allows the researcher to examine the relationship between the two variables of income and health insurance. The data can be used to look for cause and effect relationships and therefore, can be used to make predictions.

Another researcher is interested in exploring the reasons that people choose not to have health insurance. This researcher wants to know the various reasons why people make that choice and what the possible barriers may be when people choose not to get insurance. This is an open-ended question that will not provide results that will lend themselves to statistical analysis.

2) Sampling Methods for Quantitative Research

The sample will be representative of the population if the researcher uses a **random selection procedure** to choose participants. The group of units or individuals who have a legitimate chance of being selected are sometimes referred to as the **sampling frame**. If a researcher studied developmental milestones of preschool children and target licensed preschools to collect the data, the sampling frame would be all preschool aged children in those preschools. Students in those preschools could then be selected at random through a systematic method to participate in the study. This does, however, lead to a discussion of **biases** in research. For example, low-income children may be less likely to be enrolled in preschool and therefore, may be excluded from the study. Extra care has to be taken to control biases when determining sampling techniques.

There are two main types of sampling: **probability and non-probability sampling**. The difference between the two types is whether or not the sampling selection involves randomization. Randomization occurs when all members of the sampling frame have an equal opportunity of being selected for the study. Following is a discussion of probability and non-probability sampling and the different types of each.

Probability Sampling – Uses randomization and takes steps to ensure all members of a population have a chance of being selected. There are several variations on this type of sampling and following is a list of ways probability sampling may occur:

- Random sampling – every member has an equal chance
- Stratified sampling – population divided into subgroups (strata) and members are randomly selected from each group
- Systematic sampling – uses a specific system to select members such as every 10th person on an alphabetized list
- Cluster random sampling – divides the population into clusters, clusters are randomly selected and all members of the cluster selected are sampled
- Multi-stage random sampling – a combination of one or more of the above methods

Non-probability Sampling – Does not rely on the use of randomization techniques to select members. This is typically done in studies where randomization is not possible in order to obtain a representative sample. Bias is more of a concern with this type of sampling. The different types of non-probability sampling are as follows:

- Convenience or accidental sampling – members or units are selected based on availability
- Purposive sampling – members of a particular group are purposefully sought after
- Modal instance sampling – members or units are the most common within a defined group and therefore are sought after

- Expert sampling – members considered to be of high quality are chosen for participation
- Proportional and non-proportional quota sampling – members are sampled until exact proportions of certain types of data are obtained or until sufficient data in different categories is collected
- Diversity sampling – members are selected intentionally across the possible types of responses to capture all possibilities
- Snowball sampling – members are sampled and then asked to help identify other members to sample and this process continues until enough samples are collected

3.3 Examples of Primary And Secondary Data Problems

3.3.1 Primary Data Analysis Using a Chi-Square Method

Chi-Square Test

Generally speaking, the chi-square test is a statistical test used to examine differences with categorical variables or determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. It is a nonparametric test that is performed on categorical (nominal or ordinal) data. That is, the Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis.

Do the number of individuals or objects that fall in each category differ significantly from the number you would expect? Is this difference between the expected and observed due to sampling error, or is it a *real* difference? There are a number of features of the social world we characterize through categorical variables - religion, political preference, etc. To examine hypotheses using such variables, use the chi-square test.

The chi-square test is used in two similar but distinct circumstances:

- a. for estimating how closely an observed distribution matches an expected distribution - we'll refer to this as the goodness-of-fit test
- b. for estimating whether two random variables are independent

Chi-Square Test Requirements

1. Quantitative data.
2. One or more categories.
3. Independent observations.
4. Adequate sample size (at least 10).
5. Simple random sample.
6. Data in frequency form.
7. All observations must be used.

Chi-Square Formula

Chi-square is the sum of the squared difference between observed (O) and the expected (E) data (or the deviation, d), divided by the expected data in all possible categories

$$\chi^2 = \frac{\sum (\text{Observed Value} - \text{Expected Value})^2}{\sum (\text{Expected Value})} \text{ Or } \frac{(O - E)^2}{E} \quad (1)$$

where O is the Observed value (Frequency) in each category

E is the Expected value (Frequency) in the corresponding category

Expected Frequencies: When you find the value for chi square, you determine whether the observed frequencies differ significantly from the expected frequencies. You find the expected frequencies for chi square in three ways.

You hypothesize that all the frequencies are equal in each category. For example, to study the contribution of sales department employees to the overall performance of A and Z insurance company. We hypothesize that the performance of sales department had no contribution to the overall performance

You determine the expected frequencies on the basis of some prior knowledge. Now let's take a situation, find the expected frequencies, and use the chi-square test to solve the problem.

df is the "degree of freedom" ($n-1$). Degrees of freedom refers to the number of values that are free to vary after restriction has been placed on the data.

χ^2 is the Chi Square

The steps in using the chi-square test may be summarized as follows:

- 1) Define Null and Alternative Hypotheses
- 2) Write the observed frequencies in column O
- 3) Figure the expected frequencies and write them in column E.
- 4) Use the formula to find the chi-square value:
- 5) Calculate Degrees of Freedom (df) ($N-1$)
- 6) State Alpha
- 7) Find the table value (consult the Chi Square Table.)
- 8) If your chi-square value is equal to or greater than the table value, reject the null hypothesis: differences in your data are not due to chance alone

Now let's take a situation of assessing a performance appraisal problem. We find the expected frequencies, and use the Chi-Square, Weighted Average and Simple Correlation. Percentage Analysis is another method but not included.

Set the Performance Appraisal Objective (s) from the onset

1. To study the contribution of sales department in the overall performance of A and Z insurance company.
2. To investigate into the representativeness of 90 or 180 or 270 or 360-degree feedback. This depends on the given situation
3. To forward recommendations based on research findings for improvement in HRD/HRM practices.

360 -Degree Feedback Appraisal Method: 360-degree feedback, also known as ‘multi-rater feedback’, is the most comprehensive appraisal where the feedback about the employees’ performance comes from all the sources that come in contact with the employee on his job.

These sources include superiors, subordinates, peers, team members, customers, and suppliers apart from the employee himself, who can provide feedback on the employee’s job performance.



Figure 1a: 360 -Degree Feedback Feedback

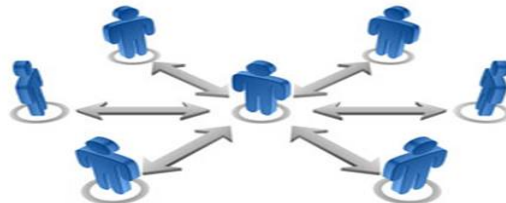


Figure 1b: 360 -Degree

Set the Performance Appraisal Hypothesis from the onset

1. Null hypothesis - H_0 : The performance of sales department employees of A and Z had no contribution to the overall performance of A and Z insurance company.
2. Alternate hypothesis - H_1 : Sales department employees of A and Z contribute considerably to the overall performance of A and Z insurance company.

Research methodology/ Design/Approach - Questionnaire Method

This study gathered the data through a series of interviews and questionnaires. It sought the perceptions and experiences of management and participants in relation to the implementation process and the outcomes of the research. Research methodology for this study is given as under.

Methods of collecting data: There are three main methods used to collect performance appraisal (PA) data:

1. Objective production,
2. Personnel, and
3. Judgmental evaluation. Judgmental evaluations are the most commonly used with a large variety of evaluation methods.

Primary Data: the different methods used in collection of primary data Questionnaire and Observation

Take a Sample Size: the sample for this study is taken as 120

Sampling technique: The Sampling technique applied is convenience sampling

Analytical Framework: Using Chi-Square

The chi-square test was probably the most popular goodness-of-fit used in the social sciences today. This judgment is based solely on the frequency with which this test was applied in the literature. There are statistical measures that evaluate the significance of difference between what is observed and what is expected according to a chance (i.e., expected according to some ideal model or standard).

Table 1: Chi square test the relation between the performance of sales department employees and the overall performance of A and Z Insurance Company.

7: 3 x 3 Contingency

Chi square test the relation between the performance of sales department employees and the overall performance of A and Z Insurance Company						
Observed			The Impacts of sales department on the overall performance of A and Z insurance company			
			Category 1	Category 2	Category 3	
			Great Extent	Some Extent	improves performance	Total
Satisfaction with the Sales department	Sample A	Highly satisfied	15	8	7	30
	Sample B	Satisfied	47	7	13	67

	Sample C	Neutral	10	5	8	23
	TOTAL	TOTAL	72	20	28	120

Expected		The Impacts of sales department on the overall performance of A and Z insurance company				
		Category 1	Category 2	Category 3		
		Great Extent	Some Extent	Improves performance	Total	
Satisfaction with the Sales department	Sample A	Highly satisfied	$72 \times \frac{30}{120} = 18$	$20 \times \frac{30}{120} = 5$	$28 \times \frac{30}{120} = 7$	30
	Sample B	Satisfied	$72 \times \frac{67}{120} = 40.2$	$20 \times \frac{67}{120} = 11.167$	$28 \times \frac{67}{120} = 15.6333$	67
	Sample C	Neutral	$72 \times \frac{23}{120} = 13.8$	$20 \times \frac{23}{120} = 3.833$	$28 \times \frac{23}{120} = 5.3667$	23
		TOTAL	72	20	28	120

Calculation				
Observed Frequencies (O)	Expected Frequencies (E)	O - E	(O - E)²	$\frac{(O - E)^2}{E}$
15	18	-3	9	0.5
8	5	3	9	1.8
7	7	0	0	0
47	40.2	6.8	46.24	1.15025
7	11.167	-4.167	17.36	1.5546
13	15.633	-2.633	6.94	0.4439
10	13.8	-3.8	14.44	1.0464
5	3.8333	1.1667	1.36	0.3548
8	5.3667	26.33	6.934	1.2920
TOTAL				8.142

$$c^2 = \frac{(15 - 18)^2}{18} + \frac{(8 - 5)^2}{5} + \frac{(7 - 7)^2}{7} + \frac{(47 - 40.2)^2}{40.2} + \frac{(7 - 11.167)^2}{11.167} + \frac{(13 - 15.633)^2}{15.633} + \frac{(10 - 13.8)^2}{13.8} + \frac{(5 - 3.833)^2}{3.833} + \frac{(8 - 5.3667)^2}{5.3667} = 8.142 \quad (2)$$

$$\text{Degree of freedom (df)} \quad df = (r - 1)(c - 1) = (3 - 1)(3 - 1) = 2 \times 2 = 4 \quad (3)$$

Find the table value for Chi Square. Begin by finding the df found in step 7 along the left-hand side of the table. Run your fingers across the proper row until you reach the predetermined level of significance (0.05) at 4 degree of freedom with Tabulated value of 9.488 value (see the chi-square distribution table below)

Level of significance is 5%; Using alpha (α) = 0.05
Tabulated value is 9.49 and Calculated Value is 8.142

Interpretation

Calculated $c^2 <$ Tabulated c^2

Since the Tabulated value is less than the Calculated Value Null hypothesis (H_0): which states: The performance of sales department employees of A and Z had no contribution to the overall performance of A and Z insurance company is rejected

The alternate hypothesis (H_1), which states that; the sales department employees of A and Z contribute considerably to the overall performance of A and Z insurance company is accepted.

Degrees of Freedom (<i>df</i>)	Probability (<i>p</i>)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46

7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32	
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12	
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88	
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59	
	Non-significant								Significant			

General notation for a 2 x 2 contingency table.

There are several types of chi square tests depending on the way the data was collected and the hypothesis being tested. Let also look at this simplest case: a 2 x 2 contingency table. If we set the 2 x 2 table to the general notation shown below in Table, using the letters a, b, c, and d to denote the contents of the cells, then we would have the following table:

Table 3: 2 x 2 contingency table

Observed	Data type 1	Data type 2	Totals
Category 1	a	b	a + b
Category 2	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

For a 2 x 2 contingency table the Chi Square statistic is calculated by the formula:

Note: notice that the four components of the denominator are the four totals from the table columns and rows.

$$\text{Chi square } (c^2) = \frac{N[(ad) - (bc)]^2}{(a + b)(c + d)(b + d)(a + c)} \quad (4)$$

Table 4: Calculate Expected Value on a 2 x 2

Expected	Data type 1	Data type 2	Totals
Category 1	(a + c)(a + b)/N	(b + d)(a + b)/N	a + b
Category 2	(a + c)(c + d)/N	(b + d)(c + d)/N	c + d
Total	a + c	b + d	a + b + c + d = N

Other Methods to Measure Performance Appraisal

Weighted Average (WA) Method

Table 5: Weighted average analysis of the satisfactoral level of performance of management system

S/No.	Attributes	Weightage	No. of Respondents	Weighted Average
1	Highly satisfied	5	14	70
2	Satisfied	4	67	268
3	Neutral	3	32	96
4	Dissatisfied	2	5	10
5	Highly Dissatisfied	1	2	2
Total			120	446

$$WA_{(\bar{x})} = \frac{(5 \cdot 14) + (4 \cdot 67) + (3 \cdot 32) + (2 \cdot 5) + (1 \cdot 2)}{120} \quad (5)$$

$$WA_{(\bar{x})} = \frac{446}{120} = 3.71$$

Inference: It is inferred that the most of the people are satisfied with existing Performance Management System (PMS)

Simple Correlation (r) Method

Table 6: Correlation for Showing the Correlation between the Existing PMS and Organisational Climate

X	Y	X ²	Y ²	XY
3	14	9	196	42
21	67	441	4489	1407
18	32	324	10304	576
64	5	4096	25	320
14	2	196	4	28
$\sum X = 120$	$\sum Y = 134$	$\sum X^2 = 5066$	$\sum Y^2 = 15018$	$\sum XY = 2373$

Analysis:

$$r = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}} \quad (6)$$

$$r = \frac{5(2373) - (120)(134)}{\sqrt{(5(5066) - (120)^2)(5(15018) - (134)^2)}} = - 0.1705396 \quad (7)$$

Inference:

The value lies between -1 to +1. So, there is strong correlation between the Existing PMS and Organisational Climate.

Suggestions:

The employees can be given feedback about their strong area and improvement required are every quarterly, which improves their performance more. Training can be given on soft skills and personality development which makes the employees to perform well in their job. It has been observed that the company rates their employee's performance through critical incident diary. Key result areas should be modified annually. PMS should be planned well in advance and executed in a much more systematic way. It is suggested to implement 360 degrees appraisal also.

Limitations: this survey is restricted to your organisation only. The options of the respondents are accepted are true and valid. Time spent on the research was limited. The conclusions drawn from the study is indicative and not exhaustive in nature.

3.3.2 Secondary Data Analysis Using A Regression Method**Example 1**

Madam Kemi has her Parlour Restaurants near Universities campuses in Nigeria and is interested in relationship between the size of the student population (in thousands) and the quarterly sales. The sample data she collected from ten restaurants is shown below

Restaurant <i>i</i>	1	2	3	4	5	6	7	8	9	10
Student Population (000s)	1	3	4	4	6	8	10	10	11	13
Quarterly Sales (000 naira)	29	52.5	44	59	58.5	68.5	78.5	84.5	74.5	101

- i. Develop a simple regression equation for this model with the identification of the regressand (y) and regressor (x) variables
- ii. From your solved regression equation, get an estimate of the quarterly sales for a campus with 8000 students?
- iii. Use the above estimated regression equation to proof that the sum of residuals $y_i - \hat{y}_i$ is = 0
- iv. Interpret your results

Solution to Question 1

(a) Let the quarterly sales y be the dependent variable (N1000), the independent variable x = campus student population (1000), then we have

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	29	-6	-36	216	36
3	52.5	-4	-12.5	50	16
4	44	-3	-21	63	9
4	59	-3	-6	18	9
6	58.5	-1	-6.5	6.5	1
8	68.5	1	3.5	3.5	1
10	78.5	3	13.5	40.5	9
10	84.5	3	19.5	58.5	9
11	74.5	4	9.5	38	16
13	101	6	36	216	36
70	650	0	0	710	142

So according to the above formula, we should have

$$n = 10$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{70}{10} = 7$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{650}{10} = 65,$$

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{710}{142} = 5,$$

$$a = \bar{y} - b\bar{x} = 65 - 5 * 7 = 65 - 35 = 30.$$

Thus, the regression equation for the model is: $\hat{y} = 30 + 5x$

(b) If we want to get an estimate of the quarterly sales for a campus with 8000 students, from the regression equation, $\hat{y} = 30 + 5x$, we have

$$\bar{y} = 30 + 5 * 8 = 30 + 40 = 70 \text{ (Thousand Naira)}$$

(d) Using the above estimated regression equation, to prove that the sum of residuals $y_i - \hat{y}_i = 0$

Residual: The difference between the observed value and the estimated value of the dependent variable y is called the residual, namely, $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \dots, y_n - \hat{y}_n$. The criterion for goodness of the regression equation is: the smaller residual, the better fitting of observation data.

Using the above estimated regression equation of Mr. Biggs Parlour Restaurant; we obtain the residuals as follows

Student population x_i	Quarterly sales y_i	Estimated sales $\hat{y}_i = 30 + 5x_i$	Residuals $y_i - \hat{y}_i$
1	29	35	-6
3	52.5	45	7.5
4	44	50	-6
4	59	50	9
6	58.5	60	-1.5
8	68.5	70	-1.5
10	78.5	80	-1.5
10	84.5	80	4.5
11	74.5	85	-10.5
13	101	95	6
			Total = 0

(d) Interpretation of the results

The regression equation is: $\hat{y} = 30 + 5x$: Which could be interpreted as: whenever the campus size is increased by thousand students, the quarterly sales at the Mr. Biggs Parlour restaurants will be expected to increase five thousand naira.

Example 2

If an insurance premium (x) is depending on driving experience (y) in a regression model. And the following summation results were computed from a random sample of eight drivers insured with a company and having similar auto insurance policies.

$$\sum x = 90, \sum y = 474, \sum xy = 4739, \sum x^2 = 1396 \text{ and } \sum y^2 = 29,642$$

- i. Compute SS_{xx} , SS_{yy} , and SS_{xy}
- ii. Find the least squares regression line
- iii. Interpret the meaning of the values of the parameters (the intercept and the slope)
- iv. Do you expect a positive or a negative relationship between these two variables?
- v. Calculate r and r^2 and explain what they mean.
- vi. Predict the monthly auto insurance premium for a driver with 10 years of driving experience.
- vii. Compute the standard deviation of errors.

Solution 2

- (i) Compute SS_{xx} , SS_{yy} , and SS_{xy}

The values of x and y are

$$\bar{x} = \Sigma x / n = 90 / 8 = 11.25$$

$$\bar{y} = \Sigma y / n = 474 / 8 = 59.25$$

$$SS_{xy} = \Sigma xy = \frac{(\Sigma x)(\Sigma y)}{n} = 4739 - \frac{(90)(474)}{8} = -593.5000$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1396 - \frac{(90)^2}{8} = 383.5000$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 29,642 - \frac{(474)^2}{8} = 1557.5000$$

(ii) To find the regression line, we calculate a and b as follows

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{-593.5000}{383.5000} = -1.5476$$

$$a = \bar{y} - b\bar{x} = 59.25 - (-1.5476)(11.25) = 76.6605$$

Thus, our estimated regression line $\hat{y} = a + bx$ is $\hat{y} = 76.6605 - 1.5476x$

(iii) The value of $a = 76.6605$ gives the value of \hat{y} for $x = 0$; that is, it gives the monthly auto insurance premium for a driver with no driving experience. However, as mentioned earlier in this unit, we should not attach much importance to this statement because the sample contains drivers with only two or more years of experience. The value of b gives the change in \hat{y} due to a change of one unit in x . Thus, $b = -1.5476$ indicates that, on average, for every extra year of driving experience, the monthly auto insurance premium decreases by \$1.55. Note that when b is negative, y decreases as x increases.

(iv) Based on theory and intuition, we expect the insurance premium to depend on driving experience. Consequently, the insurance premium is a dependent variable and driving experience is an independent variable in the regression model. A new driver is considered a high risk by the insurance companies, and he or she has to pay a higher premium for auto insurance. On average, the insurance premium is expected to decrease with an increase in the years of driving experience. Therefore, we expect a negative relationship between these two variables. In other words, both the population correlation coefficient ρ and the population regression slope B are expected to be negative.

(v) The values of r and r^2 are computed as follows

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{-593.5000}{\sqrt{(383.5000)(1557.5000)}} = -.77$$

$$r^2 = \frac{b SS_{xy}}{SS_{yy}} = \frac{(-1.5476)(-593.5000)}{1557.5000} = .59$$

The value of $r = -.77$ indicates that the driving experience and the monthly auto insurance premium are negatively related. The (linear) relationship is strong but not very strong. The value of $r^2 = .59$ states that 59% of the total variation in insurance premiums is explained by years of driving experience and 41% is not. The low value of r^2 indicates that there may be many other important variables that contribute to the determination of auto insurance premiums. For example, the premium is expected to depend on the driving record of a driver and the type and age of the car.

(vi) Using the estimated regression line, we find the predicted value of y for $x = 10$ is

$$\hat{y} = 76.6605 - 1.5476x = 76.6605 - 1.5476(10) = \$61.18$$

(vii) The standard deviation of errors is

$$s_e = \sqrt{\frac{SS_{yy} - b SS_{xy}}{n - 2}} = \sqrt{\frac{1557.5000 - (-1.5476)(-593.5000)}{8 - 2}} = 10.3199$$

SELF-ASSESSMENT EXERCISE

Due to a down-turn in the economy, your company has been experiencing financial losses in revenue. You have been asked to put together a team that will find 3-5 low-cost or no-cost ways to resource costs by 25% for your company's main product line. You are being given less 6 weeks to identify and test these cost-cutting measures. If they work in a pilot, you and your team guide the organization through the changes, if not your team will be "redeployed". Identify the evaluations and key decisions that need to be made in each of the next six weeks. (Reminder: this isn't about designing the solutions but the plan to get to solutions.

4.0 CONCLUSION

Data is one of the most important and vital aspect of any research studies. Researchers conducted in different fields of study can be different in methodology but every research is based on data which is analyzed and interpreted to get information. As quantitative data is in the form of numbers, mathematical and statistical analysis of these numbers can lead to establishing some conclusive results. Data is the basic unit in statistical studies. Statistical information like census, population variables, health statistics, and road accidents records are all developed from data. Quantitative studies result in data that provides quantifiable, objective, and easy to interpret results. The data can typically be summarized in a way that allows for generalizations that can be applied to the greater population and the results can be reproduced. However, if you want to utilize the data to make inferences or predictions about the population, you will need to go another step farther and use inferential statistics.

5.0 SUMMARY

This unit has taken you through the process of data collection methods, Sources and selecting appropriate methodologies. You now have a better understanding of the difference between primary data and secondary data collection methods and associated benefits and limitations. This unit further discusses the difference between discrete and continuous data and explains probability and non-probability sampling and describes the different types of each. Also, the unit describe when quantitative research methods should be used to examine a research problem and provide examples of the appropriate use of quantitative research methodology. The unit introduced you to some common methods and techniques of data analysis for both primary and secondary research and list the steps involved in analyzing quantitative data.

6.0 TUTOR-MARKED ASSIGNMENT

A political analyst is studying the effect of number of hours of campaigns (CH) and percentage of votes (VP) received by six candidates in a municipal election in Nigeria

Candidate	CH	VP
Abubakar	40	50
Pius	10	20
Adekule	30	30
Bola	60	80
Amadi	70	60
Michael	50	60

To study the effect of number of hours of campaigns on percentage of votes received, a researcher specified the following regression model:

$$VP_i = \beta_1 + \beta_2 CH_i + u_i$$

- Use the above data to obtain least squares estimates of β_1 and β_2
- Interpret the estimates of β_1 and β_2 in the model
- Draw scatter diagram and regression line
- Predict the percentage of votes received by a candidate who campaigns for 100 hours.

7.0 REFERENCES/FURTHER READINGS

- Bryman, A. (2012). Social research methods. Oxford university press.
- Bryman, A., & Cramer, D. (1994). Quantitative data analysis for social scientists (rev. Taylor & Frances/Routledge.
- Creswell, J. W. (2013). Research design: Qualitative, quantitative, and mixed methods approaches. Sage Publications, Incorporated.
- Gall, M. D., Borg, W. R., Gall, J. P. (2003). Educational research: An introduction. (7th Edition). White Plains, New York: Longman.
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). Data analysis: A model comparison approach. Routledge/Taylor & Francis Group.

- Neuman, W. L., & Neuman, W. L. (2006). Social research methods: Qualitative and quantitative approaches.
- Robson, C. (2002). Real world research (Vol. 2). Oxford: Blackwell publishers.
- Trochim, W. M., & Donnelly, J. P. (2001). Research methods knowledge base.

UNIT 3 DATA ANALYSIS TOOLS IN APPLIED QUANTITATIVE TECHNIQUES

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 What are Statistical Software?
 - 3.1.1 Top Statistical software for quantitative analysis
 - 3.1.2 Role of Computer Software in quantitative analysis
 - 3.2 Excel for Statistical Data Analysis
 - 3.2.1 Uses of Microsoft Excel
 - 3.2.2 The Excel Screen
 - 3.2.3 Entering Data
- 4.0 Conclusion
- 5.0 Summary
- 4.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

This unit explores the use of statistical software programs in analyzing quantitative data. Quantitative studies often result in large numerical sets that would be cumbersome to analyze without the assistance of computer software packages. Programs such as, EXCEL are relatively straight-forward and available to most researcher and are particularly useful descriptive statistics and less complicated analyses. Sometimes, the data analysis calls for a more sophisticated software package. Fortunately, there are several excellent statistical software packages available. This unit presentation and describes how the EXCEL Programs is used and includes examples.

2.0 OBJECTIVES

At the end of this unit, student should be able to:

- Define the concept Statistical Software
- Describe the benefits and uses of software programs in statistical analysis of quantitative data.
- Compare and contrast the most commonly used software packages.
- Understand the use of Microsoft excel in statistical data analysis
- Explain role of computer software in quantitative analysis.
- explain the role of computer in quantitative data analysis

3.0 MAIN CONTENT

3.1 What are Statistical Software?

Statistical software are programs which are used for the statistical analysis of the collection, organization, analysis, interpretation and presentation of data. Statistical Analysis is the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends and these are applied every day in research, industry and government to become more scientific about decisions that need to be made. Statistical Software helps in analysis of data. The software can either read data directly from an excel spreadsheet, the user can enter the data directly to the software, or the user can use a specialized data entry software to capture data.

The statistical software then manipulates the information they possess to discover patterns which can help the user uncover business opportunities and increase their business revenues and profits. It helps users with predictive analytics, data management, business intelligence and multivariate analysis.

- **Analysis of variance:** Features for Balanced and unbalanced designs, Multivariate analysis of variance and repeated measurements and linear models.
- **Bayesian analysis:** Built-in Bayesian modeling and inference for generalized linear models, accelerated failure time models, Cox regression models and finite mixture models.
- **Regression Analysis:** The statistical software helps the user understand which among the independent variables are related to the dependent variable and find out more about the forms of such relationships.
- **Reporting/Analytics:** Statistical software should have the ability to explore data and reports to extract useful insights which can, in turn, be used to understand and improve business performance.
- **Statistical Process control:** Statistics software helps in quality control which is performed by statistical methods to monitor and control the process.
- **Survey sampling and analysis:** Sample selection, Descriptive statistics, Linear and logistic regression, Proportional hazards regression and Missing value imputation.
- **Visualization:** Statistical data helps analyse data into visual imagery where it creates diagrams, graphs, images or animations to communicate the message from the data.

Statistical Software plays a vital role in business. It helps the user analyse vast volumes of data to derive meaning from it and then make the right decisions. You may also like to review the proprietary statistical software solutions list which is given below:

What is Quantitative Analysis?

Quantitative analysis is a scientific approach to managerial decision making in which raw data are processed and manipulated to produce meaningful information.

- Quantitative factors are data that can be accurately calculated. Examples include: – Different investment alternatives – Interest rates – Inventory levels – Demand – Labour cost
- Qualitative factors are more difficult to quantify but affect the decision process. Examples include: – The weather – State and federal legislation – Technological breakthroughs.

Examples of Quantitative Analyses

- In the mid-2000s, Taco Bell saved over ₦150 million using forecasting and scheduling quantitative analysis models.
- NBC television increased revenues by over ₦2 billion between 2014 and 2018 by using quantitative analysis to develop better sales plans.
- Continental Airlines saved over ₦4 billion in 2019 using quantitative analysis models to quickly recover from weather delays and other disruptions.

SELF-ASSESSMENT EXERCISE

What is quantitative data analysis?

3.1.1 Top Statistical software for quantitative analysis

In the last two decades more and more software packages have been designed to help with data analysis. The software is designed for questionnaire-based research, called quantitative research. Following is a list of seven of the most popular software for quantitative data analysis these include:

SPSS - One of the most well-known packages for quantitative research is called SPSS, which stands for Statistics Package for the Social Sciences. SPSS is perhaps the most widely used statistics software package in social science research. SPSS offers the ability to easily compile descriptive statistics, parametric and non-parametric analyses, as well as graphical depictions of results through the graphical user interface (GUI). It also includes the option to create scripts to automate analysis, or to carry out more advanced statistical processing. One advantage is that is comprehensive and compatible with nearly any type of data file. SPSS is very user-friendly and can be used to run both descriptive statistics and other more complicated analyses. Data can be entered directly into the program will also generate reports, graphs, plots, pie-charts and trend lines based on the data analyses.

Some software has been developed for commercial use, such as spreadsheet and database packages. Commercial packages are not (yet) capable of such analyses, and this is why programs such as SPSS are recommended above spreadsheets and databases.

Microsoft Excel: While not a cutting-edge solution for statistical analysis, MS Excel does offer a wide variety of tools for data visualization and simple statistics. It's simple to generate summary metrics and customizable graphics and figures, making it a usable tool for many who want to see the basics of their data. As many individuals and companies both own and know how to use Excel, it also makes it an accessible option for those looking to get started with statistics.

STATA – This is an interactive program that can also be used for both simple and complex analyses. It will also generate charts, graphs and plots of your data and results. This program may seem a bit more complicated to some researchers. It uses four different windows including the command window, the review window, the result window and the variable window. While it is a very useful program, the organization of this software may seem daunting.

SAS – The Statistical Analysis System (SAS) is another great statistical software package that can work with very large data sets. It has additional capabilities that make it commonly used in the business world because it can address issues such as business forecasting, quality improvement, planning, and so forth. It is a great program for data sets that need to incorporate strata, weighting, or groups. However, some knowledge of programming language is required to operate the software, making it a less appealing option for some. It is a premium solution that is widely used in business, healthcare, and human behavior research alike. It's possible to carry out advanced analyses and produce publication-worthy graphs and charts, although the coding can also be a difficult adjustment for those not used to this approach.

R (R Foundation for Statistical Computing): R is a free statistical software package that is widely used across both human behavior research and in other fields. Toolboxes (essentially plugins) are available for a great range of applications, which can simplify various aspects of data processing. While R is a very powerful software, it also has a steep learning curve, requiring a certain degree of coding. It does however come with an active community engaged in building and improving R and the associated plugins, which ensures that help is never too far away.

MATLAB - (The Mathworks): MatLab is an analytical platform and programming language that is widely used by engineers and scientists. As with R, the learning path is steep, and you will be required to create your own code at some point. A plentiful amount of toolboxes are also available to help answer your research questions.

GraphPad Prism - GraphPad Prism is premium software primarily used within statistics related to biology, but offers a range of capabilities that can be used across various fields. Similar to SPSS, scripting options are available to automate analyses, or carry out more complex statistical calculations, but the majority of the work can be completed through the GUI. But the majority of the work can be completed through the GUI.

Minitab - The Minitab software offers a range of both basic and fairly advanced statistical tools for data analysis. Similar to GraphPad Prism, commands can be executed through both the GUI and scripted commands, making it accessible to novices as well as users looking to carry out more complex analyses.

SELF-ASSESSMENT EXERCISE

What is the best software for statistical analysis?

3.1.2 Role of Computer Software in Quantitative Data Analysis

Looking at the various statistical software, it seems appropriate to note in advance that most of these Quantitative Analysis techniques can easily be implemented using computer software packages. Sometimes estimation of certain statistical models entails application of complex formulas which cannot be conveniently evaluated using a calculator or by manual computation. In such cases, knowledge of computing comes in handy. There are currently many Statistical software packages on the market which can be used to estimate even very complex Statistical models. For this purpose, many statistical software packages are currently commercially available in the market, both for the mainframe and the microcomputer.

Developing a solution, testing the solution, and analyzing the results are important steps in the quantitative analysis approach. Because we will be using mathematical models, these steps require mathematical calculations. Fortunately, we can use the computer to make these steps easier. A program that allow you to solve many of the problems is:

POM-QM for Windows is an easy-to-use decision support system that was developed for use with production/operations management (POM) and quantitative methods or quantitative management (QM) courses. POM for Windows and QM for Windows were originally separate software packages for each type of course. These are now combined into one program called POM-QM for Windows.

In fact, computers have changed the ways in which applied quantitative research is compiled and analyzed in that,

- **Complex Data Analysis:** Computers used in applied quantitative research have the ability to analyses data in ways and at speeds not possible with the human eye.
- **Solving Mathematical Equations:** Applied quantitative research research often requires that complex mathematical equations be solved

- Prediction Modelling: Researchers are able to use computer programs to model how data might manifest itself in the future.

3.2 EXCEL FOR STATISTICAL DATA ANALYSIS

Excel is the widely used statistical package, which serves as a tool to understand statistical concepts and computation to check your hand-worked calculation in solving your homework problems. The site provides an introduction to understand the basics of and working with the Excel. Redoing the illustrated numerical examples in this site will help improving your familiarity and as a result increase the effectiveness and efficiency of your process in statistics.

This site provides illustrative experience in the use of Excel for data summary, presentation, and for other basic statistical analysis. I believe the popular use of Excel is on the areas where Excel really can excel. This includes organizing data, i.e. basic data management, tabulation and graphics.

Excel QM, which can also be used to solve many of the problems discussed in this book, works automatically within Excel spreadsheets. Excel QM makes using a spreadsheet even easier by providing custom menus and solution procedures that guide you through every step.

3.2.1 Uses of Microsoft excel

Ms Excel is used very widely nowadays by everyone because it is very helpful and it helps in saving a lot of time. It is being used for so many years and it gets upgraded every year with new features. The most impressive thing about MS Excel is that it can be used anywhere for any kind of work. For example, it is used for billing, data management, analysis, inventory, finance, business tasks, complex calculations, etc. One can even do mathematical calculations using this and can also store important data in it in the form of charts or spreadsheets.

MS Excel provides security to your files so that no one else can see your files or ruin them. With the help of MS Excel, you can keep your files password protected. MS Excel can be accessed from anywhere and everywhere. You can even work on MS Excel using mobile if you don't have laptops. There are so many benefits of using MS Excel that it has become an inevitable part of lives of millions of people. MS Excel has numerous tools and features that make one's work easy and saves one's time also.

To use MS Excel to the best of its ability one must know its benefits and advantages. Following are the ten best uses of MS Excel:

Uses of Microsoft Excel: Analysing and storing data

One of the best uses of MS Excel is that you can analyse larger amounts of data to discover trends. With the help of graphs and charts, you can summarize the data and store it in an organized way so that whenever you want to see that data then you can easily see it. It becomes easier for you to store data and it will definitely save a lot of time for you. Once the data is stored in a systematic way, it can be used easily for multiple purposes. MS Excel makes it easier to implement various operations on the data through various tools that it possesses.

Uses of Microsoft Excel: Excel tools make your work easier

There are so many tools of MS Excel that make your work extremely easy and save your time as well. There are wonderful tools for sorting, filtering and searching which all the more make you work easy. If you will combine these tools with tables, pivot tables etc. then you will be able to finish your work in much less time. Multiple elements can be searched easily from large amounts of data to help solve a lot of problems and questions.

Uses of Microsoft Excel: Data recovery and spreadsheets

Another best use of MS Excel is that if your data gets lost then you can recover it without much inconvenience. Suppose, there is a businessman who has stored his important data in MS Excel and somehow it gets lost or the file gets damaged then he must not worry as with the new MS Excel XML format one can restore the lost or damaged file data. The next important use is that there are spreadsheets in MS Excel which also makes your work easy and with the help of new Microsoft MS Excel XML format you can reduce the size of the spreadsheet and make things compact easily.

Uses of Microsoft Excel: Mathematical formulas of MS Excel make things easier

Next best use of MS Excel is that it makes easy for you to solve complex mathematical problems in a much simpler way without much manual effort. There are so many formulas in MS Excel and by using these formulas you can implement lots of operations like finding sum, average, etc. on a large amount of data all at once. Therefore, people use MS Excel whenever they have to solve complex mathematical problems or they need to apply simple mathematical functions on tables containing larger data.

Uses of Microsoft Excel: Security

The chief use of MS Excel is that it provides security for excel files so people can keep their files safe. All the files of MS Excel can be kept password-protected through visual basic programming or directly within the excel file. People store their important data in the MS Excel so that they can keep their data in an organized way and save their time as

well. Almost every person wants his files to be password protected so that no one is able to see them or ruin them so here MS Excel solves this problem very efficiently.

Uses of Microsoft Excel: Add sophistication to data presentations

Next use of MS Excel is that it helps you in adding more sophistication to your data presentations which means that you can improve the data bars, you can highlight any specific items that you want to highlight and make your data much more presentable easily.

Suppose you have stored data in MS Excel and you want to highlight something that is important so then you can do that through the various features of data presentations available in MS Excel. You can even make the spreadsheets more attractive on which you have stored data.

Uses of Microsoft Excel: Online access

Another use of MS Excel is that it can be accessed online from anywhere and everywhere which means that you can access it from any device and from any location whenever you want. It provides the facility of working conveniently which means that if you don't have laptops then you can use mobile and do your work easily without any problem. Therefore, due to the large amount of flexibility that MS Excel provides, people like to work on MS Excel so that they can comfortably work without worrying about their device or location.

Uses of Microsoft Excel: Keeps data combined at one location

Another interesting use of MS Excel is that you can keep all your data at one location. This will help you in saving your data from getting lost. It will keep all your data in one place and then you will not have to waste your time in searching for the files. So it will save your time and whenever need be, you can look up the categorized and sorted data easily.

Uses of Microsoft Excel: Helps businessmen in developing future strategy

You can represent data in the form of charts and graphs so it can help in identifying different trends. With the help of MS Excel, trend lines can be extended beyond graph and therefore, it helps one in analysing the trends and patterns much easier. In business, it is very important to analyse the popularity of goods or the selling pattern that they follow to maximize sales. MS Excel simplifies this task and helps businessmen grow and maximize profits through the same.

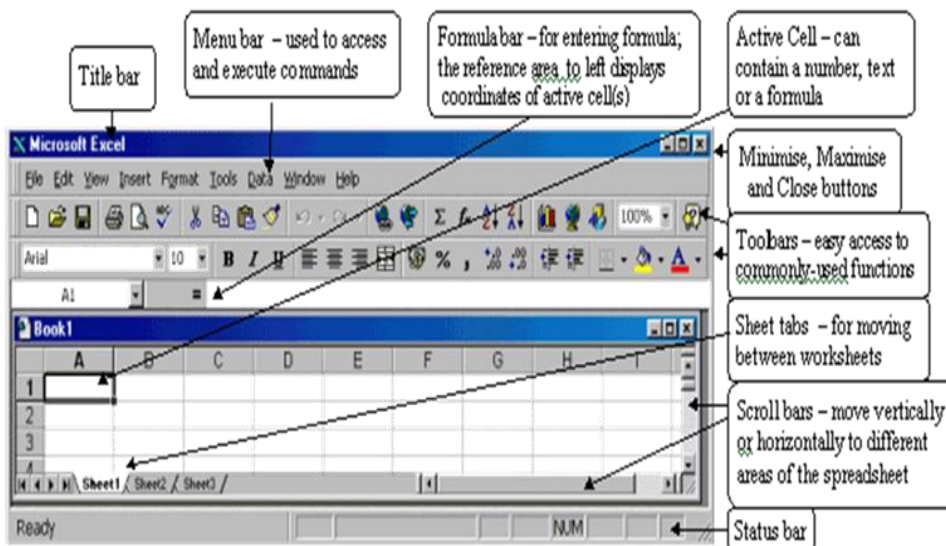
Uses of Microsoft Excel: Manage expenses

MS Excel helps in managing expenses. Suppose if a doctor is earning around 50,000 per month then he will make some expenses as well and if he wants to know how much he is exactly spending per month then he can do it with the help of MS Excel easily. He can

write his monthly income as well as expenses in the excel tables and then he can get to know that how much he is spending and he can thus, control his expenses accordingly.

There are a lot of benefits of using MS Excel, which is why it is used worldwide by people for performing so many tasks. It not only saves time but also it makes the work easier. It can almost perform every type of task. For example, you can do mathematical calculations and you can also make graphs as well as charts for storing the data. It becomes easy for the businessman to calculate things and store data in it.

3.2.2 The Excel screen



Workbooks and worksheets:

When you start Excel, a blank worksheet is displayed which consists of a multiple grid of cells with numbered rows down the page and alphabetically-titled columns across the page. Each cell is referenced by its coordinates (e.g., A3 is used to refer to the cell in column A and row 3; B10:B20 is used to refer to the range of cells in column B and rows 10 through 20).

Your work is stored in an Excel file called a workbook. Each workbook may contain several worksheets and/or charts - the current worksheet is called the active sheet. To view a different worksheet in a workbook click the appropriate Sheet Tab. You can access and execute commands directly from the main menu or you can point to one of the toolbar buttons (the display box that appears below the button, when you place the cursor over it, indicates the name/action of the button) and click once.

Moving Around the Worksheet:

It is important to be able to move around the worksheet effectively because you can only enter or change data at the position of the cursor. You can move the cursor by using the arrow keys or by moving the mouse to the required cell and clicking. Once selected the cell becomes the active cell and is identified by a thick border; only one cell can be active at a time. To move from one worksheet to another click the sheet tabs. (If your workbook contains many sheets, right-click the tab scrolling buttons then click the sheet you want.) The name of the active sheet is shown in bold.

Moving Between Cells:

Here are keyboard shortcuts to move the active cell:

- Home - moves to the first column in the current row
- Ctrl + Home - moves to the top left corner of the document
- End then Home - moves to the last cell in the document

To move between cells on a worksheet, click any cell or use the arrow keys. To see a different area of the sheet, use the scroll bars and click on the arrows or the area above/below the scroll box in either the vertical or horizontal scroll bars.

Note that the size of a scroll box indicates the proportional amount of the used area of the sheet that is visible in the window. The position of a scroll box indicates the relative location of the visible area within the worksheet.

3.2.3 Entering Data

Entering Data

A new worksheet is a grid of **rows** and **columns**. The rows are labelled with numbers, and the columns are labelled with letters. Each intersection of a row and a column is a **cell**. Each cell has an **address**, which is the column letter and the row number. The arrow on the worksheet to the right points to cell A1, which is currently **highlighted**, indicating that it is an **active cell**. A cell must be active to enter information into it. To highlight (select) a cell, click on it.

To select more than one cell:

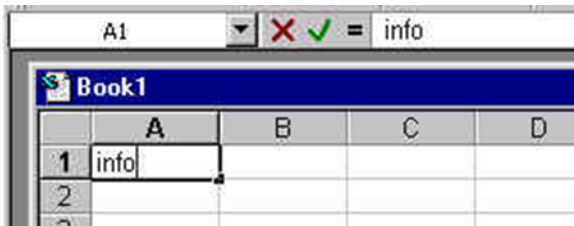
- Click on a cell (e.g. A1), then hold the shift key while you click on another (e.g. D4) to select all cells between and including A1 and D4.
- Click on a cell (e.g. A1) and drag the mouse across the desired range, unclicking on another cell (e.g. D4) to select all cells between and including A1 and D4.
- To select several cells which are not adjacent, press "control" and click on the cells you want to select. Click a number or letter labelling a row or column to select that entire row or column.

One worksheet can have up to 256 columns and 65,536 rows, so it'll be a while before you run out of space.

Each cell can contain a **label**, **value**, **logical value**, or **formula**.

- Labels can contain any combination of letters, numbers, or symbols.

- Values are numbers. Only values (numbers) can be used in calculations. A value can also be a date or a time
- Logical values are "true" or "false."
- Formulas automatically do calculations on the values in other specified cells and display the result in the cell in which the formula is entered (for example, you can specify that cell D3 is to contain the sum of the numbers in B3 and C3; the number displayed in D3 will then be a function of the numbers entered into B3 and C3).



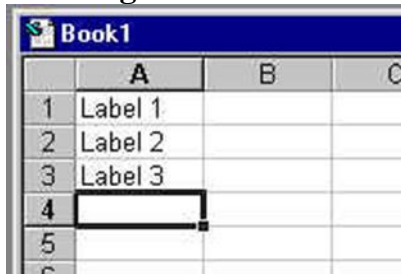
To enter information into a cell, select the cell and begin typing.

Note that as you type information into the cell, the information you enter also displays in the formula bar. You can also enter information into the formula bar, and the information will appear in the selected cell.

When you have finished entering the label or value:

- Press "Enter" to move to the next cell below (in this case, A2)
- Press "Tab" to move to the next cell to the right (in this case, B1)
- Click in any cell to select it

Entering Labels



Unless the information you enter is formatted as a value or a formula, Excel will interpret it as a label, and defaults to align the text on the left side of the cell.

If you are creating a long worksheet and you will be repeating the same label information in many different cells, you can use the **Auto Complete** function. This function will look at other entries in the same column and attempt to match a previous entry with your current entry. For example, if you have already typed "Wesleyan" in another cell and you type "W" in a new cell, Excel will automatically enter "Wesleyan." If you intended to type "Wesleyan" into the cell, your task is done, and you can move on to the next cell. If you intended to type something else, e.g. "Williams," into the cell, just continue typing to enter the term.

To turn on the AutoComplete function, click on "Tools" in the menu bar, then select "Options," then select "Edit," and click to put a check in the box beside "Enable AutoComplete for cell values."

Another way to quickly enter repeated labels is to use the **Pick List** feature. Right click on a cell, then select "Pick from List." This will give you a menu of all other entries in cells in that column. Click on an item in the menu to enter it into the currently selected cell.

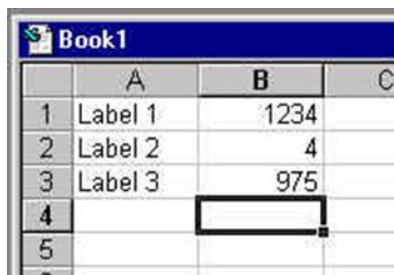
Entering Values

A value is a number, date, or time, plus a few symbols if necessary to further define the numbers [such as: . , + - () % \$ /].

Numbers are assumed to be positive; to enter a negative number, use a minus sign "-" or enclose the number in parentheses "()".

Dates are stored as MM/DD/YYYY, but you do not have to enter it precisely in that format. If you enter "Jan 9" or "Jan-9", Excel will recognize it at January 9 of the current year, and store it as 1/9/2002. Enter the four-digit year for a year other than the current year (e.g. "jan 9, 1999"). To enter the current day's date, press "control" and ";" at the same time.

Times default to a 24 hour clock. Use "a" or "p" to indicate "am" or "pm" if you use a 12 hour clock (e.g. "8:30 p" is interpreted as 8:30 PM). To enter the current time, press "control" and ":" (shift-semicolon) at the same time.



The screenshot shows a portion of an Excel spreadsheet titled "Book1". The spreadsheet has columns labeled A, B, and C, and rows numbered 1 through 5. The data is as follows:

	A	B	C
1	Label 1	1234	
2	Label 2	4	
3	Label 3	975	
4			
5			

An entry interpreted as a value (number, date, or time) is aligned to the right side of the cell, to reformat a value.

Rounding Numbers that Meet Specified Criteria: To apply colours to maximum and/or minimum values:

1. Select a cell in the region, and press Ctrl+Shift+* (in Excel 2003, press this or Ctrl+A) to select the Current Region.
2. From the Format menu, select Conditional Formatting.
3. In Condition 1, select Formula Is, and type =MAX(\$F:\$F)=\$F1.
4. Click Format, select the Font tab, select a colour, and then click OK.
5. In Condition 2, select Formula Is, and type =MIN(\$F:\$F)=\$F1.
6. Repeat step 4, select a different colour than you selected for Condition 1, and then click OK.

Note: Be sure to distinguish between absolute reference and relative reference when entering the formulas.

Rounding Numbers that Meet Specified Criteria

Problem: Rounding all the numbers in column A to zero decimal places, except for those that have "5" in the first decimal place.

Solution: Use the IF, MOD, and ROUND functions in the following formula:
`=IF(MOD(A2,1)=0.5,A2,ROUND(A2,0))`

To Copy and Paste All Cells in a Sheet

1. Select the cells in the sheet by pressing Ctrl+A (in Excel 2003, select a cell in a blank area before pressing Ctrl+A, or from a selected cell in a Current Region/List range, press Ctrl+A+A).
OR Click Select all at the top-left intersection of rows and columns.
2. Press Ctrl+C.
3. Press Ctrl+Page Down to select another sheet, then select cell A1.
4. Press Enter.

To Copy the Entire Sheet

Copying the entire sheet means copying the cells, the page setup parameters, and the defined range Names.

Option 1:

1. Move the mouse pointer to a sheet tab.
2. Press Ctrl, and hold the mouse to drag the sheet to a different location.
3. Release the mouse button and the Ctrl key.

Option 2:

1. Right-click the appropriate sheet tab.
2. From the shortcut menu, select Move or Copy. The Move or Copy dialog box enables one to copy the sheet either to a different location in the current workbook or to a different workbook. Be sure to mark the Create a copy checkbox.

Option 3:

1. From the Window menu, select Arrange.
2. Select Tiled to tile all open workbooks in the window.
3. Use Option 1 (dragging the sheet while pressing Ctrl) to copy or move a sheet.

Sorting by Columns

The default setting for sorting in Ascending or Descending order is by row. To sort by columns:

1. From the Data menu, select Sort, and then Options.
2. Select the Sort left to right option button and click OK.
3. In the Sort by option of the Sort dialog box, select the row number by which the columns will be sorted and click OK.

SELF-ASSESSMENT EXERCISE

- i. Is Excel a data analysis tool?
- ii. What is the benefit of using formula in Excel sheet?

4.0 CONCLUSION

There are a range of different software tools available, and each offers something slightly different to the user – what you choose will depend on a range of factors, including your research question, knowledge of statistics, and experience of coding. These factors could mean that you are at the cutting-edge of data analysis, but as with any research, the quality of the data obtained is reliant upon the quality of the study execution. It's therefore important to keep in mind that while you might have advanced statistical software (and the knowledge to use it) available to you, the results won't mean much if they weren't collected in a valid way. We've put together a guide to experimental design, helping you carry out quality research so that the results you collect can be relied on. For real statistical analysis one must learn using the professional commercial statistical packages such as SAS, and SPSS.

5.0 SUMMARY

A few words of caution regarding the use of statistical software packages are in order. First, it is good practice for the statistician to personally know how to solve a particular problem before resorting to statistical software packages. In other words, statistical software packages should be used as a tool for facilitating quantitative analysis rather than as an alternative to understanding the procedure involved. Second, the statistician must be aware of the advantages and shortcomings of the methods employed when analysing the results obtained using these software packages. A proper understanding of the advantages and drawbacks of these methods helps the econometrician to make informed judgement about the reliability of the results. The unit also analyses the software that can be used for quantitative analyses which are POM-QM and Excel-QM.

6.0 TUTOR-MARKED ASSIGNMENT

- 1) Discuss how to analyse data in Excel?
- 2) Discuss the role of computer in quantitative analysis.

7.0 REFERENCES/FURTHER READINGS

- Blaikie, N. (2003). *Analysing quantitative data: From description to explanation*. Sage.
- Bohrstedt, G. W., & Knoke, D. (1994). *Statistics for social data analysis*.
- Bryman, A., & Cramer, D. (1994). *Quantitative data analysis for social scientists* (rev. Taylor & Frances/Routledge).
- Cramer, D. (2003). *Advanced quantitative data analysis*. McGraw-Hill International.
- Glass, G. V., & Hopkins, K. D. (1970). *Statistical methods in education and psychology* (p. 534). Englewood Cliffs, NJ: Prentice-Hall.
- Hamilton, L. C. (2006), "Statistics with Stata", Brooks/Cole, Belmont, CA
- Shapiro, Fred (2000). "Origin of the Term Software: Evidence from the JSTOR Electronic Journal Archive" (PDF). *IEEE Annals of the History of Computing*. 22 (2): 69–71