NATIONAL OPEN UNIVERSITY OF NIGERIA

STATISTICS FOR ECONOMIST 1
ECO 253

FACULTY OF SOCIAL SCIENCES

COURSE GUIDE

Course Developer:
Okojie, Daniel Esene
Economics Department
University of Lagos

Edited By:
Dr. Ibrahim Bakare
Department of Economics,
Lagos State University

Course Reviewer
Dr. Mutiu Rasaki
Department of Economics,
Augustine University, Ilara, Epe, Lagos State

COURSE GUIDE

ECO 253
STATISTICS FOR ECONOMIST 1

Course Team Okojie, Daniel Esene (Course Developer) - UNILAG Dr Ibrahim Bakare(Course Editor)- Lagos State University

NATIONAL OPEN UNIVERSTIY OF NIGERIA

National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadastral Zone,
Nnamdi Azikwe Expressway
Jabi, Abuja

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

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

ISBN:

Introduction
Course Competences

Course Objectives
Working through this Course
Study Units
References and Further Readings
Presentation Schedule
Assessment
How to Get the Most from This Course
Online Facilitations

Course Information

Course Code: ECO 347

Course Title: Development Economics II

Course Unit: 2

Course Status: Compulsory

Course Blub:

Semester: Second Semester

Course Duration: Fifteen Lecture Weeks

Required Hours for Study**:** Two hours for each unit

Course Team

Course Developer: NOUN

Course Writer: **Okojie, Daniel Esene**

Content Editor: **Dr. Ibrahim Bakare**

Instructional Designer:

Learning Technologists:

Copy Editor

Introduction

Welcome to STATISTICS FOR ECONOMIST 1 (ECO 253)

Statistics for Economist 1 is a three-credit unit, first semester undergraduate course for Economics Students in the National Open University of Nigeria. This course focuses progressively on elementary understanding of distribution functions and other inferential statistical techniques. The course focuses on practical issues involved in the substantive interpretation of economic data using sampling, estimation, hypothesis testing, correlation, and regression. For this reason, empirical case studies that apply the techniques to real-life data are stressed and discussed throughout the course, and students are required to perform several statistical analyses on their own.

The course is a very useful material to you in your academic pursuit and helps to further broaden your understanding of the role of statistics in the study of economics. This course is developed to guide you on what statistics for economists' entails, what course materials in line with a course learning structure you will be using. The learning structure suggests some general guidelines for a time frame required of you on each unit in order to achieve the course aims and objectives effectively.

Course Competences

This course is basically an introduction to statistics and application of Statistics in Economics. The topics covered in this course include: The Normal, Binomial and Poisson Distributions, Estimate Theory, Test of Statistical hypothesis including t, f and chi- square tests analysis of least square method, correlation and Regression analyses. Others are elementary sampling theory and design of experiments, non-parametric methods, introduction to the central limit theory (CLT) and the law of large numbers

Course Objectives

To achieve the aims set above in addition with the overall slated course objectives, each unit would also have its specific objectives. The unit objectives are included at the beginning of a unit; you should read them before you start working through the unit. You may want to refer to them during your study of the unit to check on your progress. You should always look at the unit objectives after completing a unit. In this way, you can be certain you have done what was necessary of you by the unit. The course objectives are set below for you to achieve the aims of the course. On successful conclusion of the course, you should be able to:

- Identify and gather economic data
- Do basic data manipulation and hypothesis testing
- State statistical estimation of economic relationships
- Apply correlation and regression analyses models to data
- Understand non-parametric methods, elementary sampling theory and design of experiments
- Discuss in an introductory manner central limit theory and the law of large numbers
- Solve problems that could lead you to using some standard statistical software.

Working through the Course

To successfully complete this course, you are required to read the study units, referenced books and other materials on the course.

Each unit contains self-assessment exercises called Student Assessment Exercises (SAE). At some points in the course, you will be required to submit assignments for assessment purposes. At the end of the course there is a final examination. This course should take about twelve weeks to complete and some components of the course are outlined under the course material subsection.

Study Units

There are 20 units in this course which should be studied carefully and diligently.

Module 1: Probability and Statistic Distribution Functions

Unit 1: Bernoulli Distribution
Unit 2: Binomial Distribution
Unit 3: Normal Distribution
Unit 4 Poisson Distribution

Module 2: Statistical Hypothesis Test

Unit 5: T- test
Unit 6: F- test
Unit 7: Chi square test
Unit 8: ANOVA
Unit 9: Parametric and Non-Parametric test Methods

Module 3: Correlation and Regression Coefficient Analyses

Unit 10: Pearson's Correlation Coefficient
Unit 11: Spearman's Rank Correlation Coefficient
Unit 12: Methods of Curve and Eye Fitting of Scattered Plot and the Least Square Regression Line
Unit 13: Forecasting in Regression

Module 4: Introduction to the Central Limit Theory (CLT) Unit 14: Central Limit Theorems for Independent Sequences Unit 15: Central Limit Theorems for dependent Processes Unit 16: Relation to the law of large numbers
Unit 17: Extensions to the theorem and Beyond the Classical Framework

Module 5: Index Numbers and Introduction to Research Methods in Social Sciences

Unit 18: Index Number
Unit 19: Statistical Data
Unit 20: Sample and Sampling Techniques

Here module 1 (units 1-4) presents you with the common probability distribution functions as a general background on the course, statistics for economists; the discreteness of Bernoulli, Binomial and Poisson distributions and the continuous natures of Normal distribution are shown. Module 2 (units 5-9) explains some statistical hypothesis tests; the t-test, f-test, chi square test, analysis of variance (ANOVA), parametric and non-parametric test methods are all introduced. Their usage, significance, samples comparison and application for economists are also explained.

Correlation and Regression Coefficient Analyses are contained in module 3 (unit 10-13). This module explores Pearson's Correlation Coefficient, Spearman's Rank Correlation Coefficient, the Least Square Regression Line and Forecasting in Regression. The module 4 (unit 14-17) covers detail description of an introduction to central limit theory (CLT). CL theorems for independent sequences, dependent processes and the relation to law of large numbers brought to the students' knowledge here. Also, extensions to the theorem and beyond the classical framework are presented in units 17 of module 4. While basic concepts and notation of elementary Index Numbers and Introduction to Research Methods in Social Sciences are in units 18-20 of module 5. This module 5 (units 18-20) has present in it: Index Number, Statistical Data and Sample & Sampling Techniques.

.

Each study unit will take at least two hours, and it include the introduction, objective, main content, examples, In-Text Questions (ITQ) and their solutions, self-assessment exercise, conclusion, summary and reference. Other areas border on the Tutor-Marked Assessment (TMA) questions. Some of the ITQ and self-assessment exercise will require you brainstorming and solving with some of your colleagues. You are advised to do so in order to comprehend and get acquainted with how important statistics is in making the most of economics.

There are also statistical materials, textbooks under the reference and other (on-line and off-line) resources for further reading. They are meant to give you additional information whenever you avail yourself of such opportunity. You are required to study the materials; practise the ITQ and self-assessment exercise and 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.

References and Further Readings

For additional reading and more detailed information about the course, the following reference texts and materials are recommended:

Adebayo O. A., (2006). **Understanding Statistics:** Lagos, 5th Edition Lagos Nigeria JAS Publisher.

Spiegel, Murray R. and Walpole, Ronald E., (1992). **Theory and Problems of Statistics op. cit: Introduction to Statistic.** 2nd Ed. Collier Macmillan International Editions.

Walpole R. E., Richard Lerson and Morris Marx (1995), **An Introduction to Mathematical Statistics and Its Applications,** 5th Edition New York: John Wiley & Sons. Inc.

Dowling Edward T., (2001). **Mathematical Economics** 2nd Edition; Schaum Outline Series.

Esan E. O. and Okafor R. O., **Basic Statistical Methods,** Lagos**,** Nigeria**.** JAS Publishers. ISBN – 978 – 33180 – 0 – 4

Presentation Schedule
The presentation plan 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 assessments this course. First is the tutor-marked assignment and second, would be 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/lecturer 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.

An advantage of the distance learning is that the study units replace the university lecturer. You can read and work through specially designed study materials at your own tempo and at a time and place that goes well with you.
Consider doing it yourself in solving and providing solutions to statistical problems in the lecture instead of listening and copying solution being provided by a lecturer. In the same way that a lecturer might set you some practice exercises and ITQ to do, the study units tell you when to solve problems and read your books or other material, and when to embark on discussion with your colleagues. Just as a lecturer might give you an in-class exercise, your study units provides exercises for you to do at appropriate points.

Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole. Next is a set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit. You should use these objectives to guide your study. When you have finished the unit you must go back and check whether you have achieved the objectives. If you make a habit of doing this you will significantly improve your chances of passing the course and getting the best grade.

The main body of the unit guides you through the required understanding from other sources. This will usually be either from your set books or from a readings section.

Some units require you to undertake practical overview of real life statistical events. You will be directed when you need to embark on discussion and guided through the tasks you must do.
The purpose of the practical overview of real life statistical events is in twofold. First, it will enhance your understanding of the material in the unit. Second, it will give you practical experience and skills to evaluate economic arguments, and understand the roles of statistics in guiding current economic problems, calculations, analysis, solutions and debates outside your studies. In any event, most of the critical thinking skills you will develop during studying are applicable in normal working practice, so it is important that you encounter them during your studies.

Self-assessments are interspersed throughout the units, and answers are given at the ends of the units. Working through these tests will help you to achieve the objectives of the unit and prepare you for the assignments and the examination. You should do each self-assessment exercises as you come to it in the study unit. Also, ensure to master some major statistical theorems and models during the course of

studying the material.

The following is a practical strategy for working through the course. If you run into any trouble, consult your tutor. Remember that your tutor's job is to help you. When you need help, don't hesitate to call and ask your tutor to provide it.

1. Read this Course Guide thoroughly.
2. Organize a study schedule. Refer to the `Course overview' for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of the semester is available from study centre. You need to gather together all this information in one place, such as your dairy or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working breach unit.
3. Once you have created your own study schedule, do everything you can to stick to it. The major reason that students fail is that they get behind with their course work. If you get into difficulties with your schedule, please let your tutor know before it is too late for help.
4. Turn to Unit 1 and read the introduction and the objectives for the unit.
5. Assemble the study materials. Information about what you need for a unit is given in the `Overview' at the beginning of each unit. You will also need both the study unit you are working on and one of your set books on your desk at the same time.
6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading.
7. Up-to-date course information will be continuously delivered to you at the study centre.
8. Work before the relevant due date (about 4 weeks before due dates), get the Assignment File for the next required assignment. Keep in mind that you will learn a lot by doing the assignments carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the exam. Submit all assignments no later than the due date.
9. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.
10. When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.
11. When you have submitted an assignment to your tutor for marking, do not wait for its return `before starting on the next units. Keep to your schedule. When the assignment is returned, pay particular attention to your tutor's comments, botth on the tutor-marked assignment form and also written on the assignment. Consult your tutor as soon as possible if you have any questions or problems.
12. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at the beginning of each unit) and the course objectives (listed in this Course Guide).

Online Facilitations and Tutorials

There are some hours of tutorials (1-hour sessions) provided in support of this course.

You will be notified of the dates, times and location of these tutorials. Together with the name and phone number of your tutor, as soon as you are allocated a tutorial group.

Your tutor will mark and comment on your assignments, keep a close watch on your progress and on any difficulties you might encounter, and provide assistance to you during the course. You must mail your tutor-marked assignments to your tutor well before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible.

Do not hesitate to contact your tutor by telephone, e-mail, or discussion board if you need help. The following might be circumstances in which you would find help necessary. Contact your tutor if.

- You do not understand any part of the study units or the assigned readings
- You have difficulty with the self-assessment exercises
- You have a question or problem with an assignment, with your tutor's comments on an assignment or with the grading of an assignment.

You should try your best to attend the tutorials. This is the only chance to have face to face contact with your tutor and to ask questions which are answered instantly. You can raise any problem encountered in the course of your study. To gain the maximum benefit from course tutorials, prepare a question list before attending them. You will learn a lot from participating in discussions actively.

The general aim of this module is to provide learners‘ with a thorough understanding of Probability and Statistical Distribution Functions. The focus here is to provide learners‘with the common probability distribution functions as a general background to the course. The discreteness of Bernoulli, Binomial and Poisson distributions and the continuous natures of Normal distribution are presented in this module.

Module 1: Probability and Statistic Distribution Functions

The four units that constitute this module are statistically linked. At the end of this module, learners would have been able to list, differentiate and link these common probability distribution functions as well as identify and use them to solve related statistical problems. The units to be studied are;

Unit 1: Bernoulli Distribution
Unit 2: Binomial Distribution
Unit 3: Normal Distribution
Unit 4: Poisson Distribution

UNIT 1: BERNOULLI DISTRIBUTION UNIT STRUCTURE

1.1 Introduction
1.2 Learning Outcomes
1.3 Bernoulli Distribution
1.4 Bernoulli Trials
1.5 Bernoulli Process
1.5.1 Interpretation
1.5.2 Further Explanation
1.5.3 Solved Examples
1.6 Summary
1.7 References/Further Reading/ Web Resources
1.8 Possible Answers to Self-Assessment Exercises (SAEs)

1.1 Introduction

Bernoulli distribution is a discrete probability distribution, meaning it's concerned with discrete random variables. A discrete random variable is one that has a finite or countable number of possible values—the number of heads you get when tossing three coins at once. A Bernoulli distribution is a discrete distribution with only two possible values for the random variable. The distribution has only two possible outcomes and a single trial which is called a Bernoulli trial. The two possible outcomes in Bernoulli distribution are labeled by n=0 and n=1 in which n=1 (success) occurs with probability **p** and n=0 (failure) occurs with probability **1-p.**

1.2 Learning Outcomes

By the end of this unit, you will be able to:

- Define probability distribution
- Distinguish between discrete and continuous probability distributions.
- Solve problems on Bernoulli distribution

1.3 BERNOULLI DISTRIBUTION

In probability theory and statistics, the Bernoulli distribution, named after Swiss scientist Jacob Bernoulli, is a discrete probability distribution, which takes value 1 with success probability *P* and value 0 with failure probability q=1-*P*

If X is a random variable with this distribution, we have:

[ ] [ ]

A classical example of a Bernoulli experiment is a single toss of a coin. The coin might come up

heads with probability *P* and tails with probability *1-P*. The experiment is called fair if, *P*=0.5 indicating the origin of the terminology in betting (the bet is fair if both possible outcomes have the same probability).

The probability mass function of this distribution is given as;

$f(k;p) = \{$

This can also be expressed as

$$f(k;p) = pk(1\text{-}p)^{1-k} \quad \text{for k } \epsilon \text{ (0,1)}$$

i.e. given that k is an element of a set consisting of 0 and 1. This implies that k will take on value

zero or
1.

The expected value of a Bernoulli random variable X is $E(X) = P$, and its variance is

$$Var\ X = p(1\text{-}q)$$

It should be NOTED that Bernoulli distribution is a special case of the Binomial distribution with n=1.

The kurtosis goes to infinity for high and low values of *P*, but for $P = 0.5$ the Bernoulli distribution has a lower kurtosis than any other probability distribution, namely −2.

The Bernoulli distributions for form an exponential family

The maximum likelihood estimator based on a random sample is the sample mean.

1.4 Bernoulli trial

A Bernoulli trial is an instantiation of a Bernoulli event. It is one of the simplest experiments that can be conducted in probability and statistics. It's an experiment where there are two possible outcomes (Success and Failure).

Examples of Bernoulli trials:

(i) **Coin tosses:** Record how many tosses of coins resulted in heads and how many coin tosses resulted in tails. We can consider the result of getting heads as success and not getting head i.e., getting tails to be a failure.

(ii) **Football:** How many shots on a goal post resulted in the goal score, and how many shots were missed. We can call a goal scored as a "success" and a missed target to be a failure.

(iii) **Rolling Dice:** The probability of a roll of two dice resulting in a double six. A double six dice roll could be considered to be a success and everything else can be considered a failure.

Bernoulli process: A sequence of Bernoulli trials is called a Bernoulli process. Among other conclusions that could be reached, for n trials, the probability of n successes is p^n.

1.5 Bernoulli Process

A Bernoulli process is a finite or infinite sequence of binary random variable, so it is a discrete- time stochastic (involving or showing random behaviour) process that takes only two values specifically 0 and 1. The component Bernoulli variables X_i are identical and independent. In the ordinary sense, a Bernoulli process is a repeated coin flipping, possibly with an unfair coin (but with consistent unfairness). Every variable X_i in the sequence is associated with a Bernoulli trial or experiment. They all have the same Bernoulli distribution. Much of what can be said about the Bernoulli process can also be generalized to more than two outcomes (such as the process for a six-sided die); this generalization is known as the Bernoulli scheme.

The problem of determining the process, given only a limited sample of the Bernoulli trials, may be

called the problem of checking if a coin is fair.

Furthermore, a Bernoulli process is a finite or infinite sequence of independent random variables

X_1, X_2, X_3, ..., such
that

- For each i, the value of X_i is either 0 or 1;

- For all values of i, the probability that $X_i = 1$ is the same number p.

In other words, a Bernoulli process is a sequence of independent identically distributed Bernoulli trials. Independence of the trials implies that the process has no memory. Given that the probability p is known, past outcomes provide no information about future outcomes. (If p is unknown, however, the past informs about the future indirectly, through inferences about p). If the process is infinite, then from any point the future trials constitute a Bernoulli process identical to the whole process, the fresh-start property.

1.5.1 Interpretation

The two possible values of each X_i are often called "success" and "failure". Thus, when expressed as a number 0 or 1, the outcome may be called the number of successes on the ith "trial". Two other common interpretations of the values are true or false and yes or no. Under any interpretation of the two values, the individual variables X_i may be called Bernoulli trials with parameter p. In many applications time passes between trials, as the index i increases. In effect, the trials X_1, X_2, ... X_i, ... happen at "points in time" 1, 2, ..., i, ... However, passage of time and the associated notions of "past" and "future" are not necessary. Most generally, any X_i and X_j in the process are simply two from a set of random variables indexed by {1, 2... n} or by {1, 2, 3...}, the finite and infinite cases.

Several random variables and probability distributions beside the Bernoulli itself may be derived from the Bernoulli process:

- The number of successes in the first n trials, which has a Binomial distribution B(n, p)

- The number of trials needed to get r successes, which has a negative Binomial distribution NB(r, p)
- The number of trials needed to get one success, which has a geometric distribution

  NB(1, p), a special case of the negative binomial distribution

The negative Binomial variables may be interpreted as random waiting times.

The Bernoulli process can be formalized in the language of probability spaces as a random sequence of independent realisations of a random variable that can take values of heads or tails. The state space for an individual value is denoted by $2 = \{H, T\}$. Specifically, one considers the countable infinite direct product of copies of $2 = \{H, T\}$. It is common to examine either the one- sided set $\Omega = 2^{\mathbb{N}} = \{H, T\}^{\mathbb{N}}$ or the two-sided set $\Omega = 2^{\mathbb{Z}}$. There is a natural topology on this space,
called the product topology. The sets in this topology are finite sequences of coin flips, that is, finite-length strings of H and T, with the rest of (infinitely long) sequence taken as "don't care". These sets of finite sequences are referred to as cylinder sets in the product topology. The set of all such strings

form a sigma algebra, specifically, a Borel algebra. This algebra is then commonly written as $(\Omega,\mathcal{F})$ where the elements of $\mathcal{F}$ are the finite-length sequences of coin flips (the cylinder

sets).If the chances of flipping heads or tails are given by the probabilities $\{p, 1\text{-}p\}$, then one can define a natural measure on the product space, given by $P = \{p, 1\text{-}p\}^{N}$(or by $P = \{p, 1\text{-}p\}^{Z}$ for the two-sided process). Given a cylinder set, that is, a specific sequence of coin flip results $[w_1, w_2, w_3,.....w_n]$ at times 1, 2, 3.....n , the probability of observing this particular sequence is given by; $P([w_1, w_2, w_3,.....w_n]) = p^{k}(1\text{-}p)^{n-k}$

where*k* is the number of times that *H* appears in the sequence, and *n-k* is the number of times that *T* appears in the sequence. There are several different kinds of notations for the above; a common one is to write

$$P(X_1 = w_1, X_2 = w_2 ....., X_n = w_n) = p^{k}(1\text{-}p)^{n-k}$$

where each X_i is a binary-valued random variable. It is common to write x_i for w_i. This probability *P* is commonly called the Bernoulli measure.

Note that the probability of any specific, infinitely long sequence of coin flips is exactly zero; this is because for any. One says that any given infinite sequence has measure zero. Nevertheless, one can still say that some classes of infinite sequences of coin flips are far more likely than others; this is given by the asymptotic equipartition property.

To conclude the formal definition, a Bernoulli process is then given by the probability triple, $(\Omega, \mathcal{F}, P)$ (as defined above).

1.5.2 Further Explanation

A Bernoulli random variable is one which has only 0 and 1 as possible values. Let $p = P(X = 1)$
Thus a Bernoulli distribution X has the following ―table‖

Table 1.1 Bernoulli distribution table

Possible values of X	0	1
Probabilities	*1-p*	*p*

Definition: Say that

A Bernoulli random variable is the simplest random variable. It models an experiment in which there are only two outcomes.Generically, we say that X=1 is a success and X=0 is a failure. We say that p is the ―success‖ probability.

Mean and Variance: For a Bernoulli random variable with success probability p:

$$\mu_X = 0(1-p)+1p = p$$

$$\sigma^2_X = 0^2 (1-p)+1^2 \quad p - p^2$$

$$= p - p^2 = p(1-p)$$

1.5.3 Solved Examples

Example 1: A fair die is tossed. Let X = 1 only if the first toss shows a ―4‖ or ―5‖.

Solution:

Then

Example 2: Find the probability of getting a head in a single toss of a coin.
Solution: Since a fair coin is tossed. Let the variable *x* take values 1 and 0 according to as the toss results in ‗Head' or ‗Tail'. Then *X* is a Bernoulli variable with parameter .
Here, *X* denotes the number of heads obtained in the toss.

=> Probability of success and the probability of failure . Example 3:
Find the probability of getting 5 in a single throw of a dice.

Solution: In a single throw of a die, the outcome ―5‖ is called a success and any other outcome is called a failure, then the successive throws of a dice will contain Bernoulli trials. Therefore, the probability of success and the probability of failure

Self-Assessment Exercise 1 (SAE 1)

1. What is a Bernoulli trial?
2. State 3 examples of Bernoulli trail.
3. What is a Bernoulli process?

1.6 SUMMARY

In this unit, you are expected to have learnt the essentials and applications of Bernoulli distribution. Also, learners' by now would have been able toidentify Bernoulli distribution function problems and solve the problems accordingly.

1.7 REFERENCES /FURTHER READING/WEB RESOURCES

Spiegel, M. R.and Stephens L.J. (2008).*Statistics.*(4th ed.).New York: McGraw Hill press. Swift L., (1997).*Mathematics and Statistics for Business, Management and Finance.*(2nd ed.). London: Macmillan Publishers

McCullagh P. and Nelder J., (1989). *Generalized Linear Models,* (2nd ed.). Boca Raton: Chapman and Hall/CRC. ISBN 0-412-31760-5 Web site:http://en.wikipedia.org/wiki/Bernoulli_distribution

Johnson, N. L., Kotz, S. and Kemp A., (1993).*Univariate Discrete Distributions* (2nd ed.). Wiley. ISBN 0-471-54897-9 Web site:http://en.wikipedia.org/wiki/Bernoulli_distribution

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.
McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

 **1.8 Possible answers to the SAEs within the content**.

Answers to SAE 1

1. A Bernoulli trial is an instantiation of a Bernoulli event. It's an experiment where there are two possible outcomes -Success and Failure.
2. Examples of Bernoulli trials are Coin tosses, Football, and Rolling Dice.
3. A Bernoulli process is a sequence of Bernoulli trials. It is a finite or infinite sequence of binary random variable, involving or showing random behavior, and that takes only two values specifically 0 and 1.

UNIT 2: BINOMIAL DISTRIBUTION

Unit Structure

2.1 Introduction

2.2 Learning Outcomes

2.3 Binomial Distribution

2.3.1 Probability Density Function of Binomial Distribution

2.4 The Mean of a Binomial Distribution

2.5. Variance of Binomial Distribution
2.5.1 Solved Examples

2.6 Summary

2.7. References / Further Reading/Web Resources
2.8 Possible Answers to Self-Assessment Exercises (SAEs) within the content

2.1 INTRODUCTION

The binomial probability distribution is a discrete probability distribution that provides many applications. It is associated with a multiple-step experiment that we call the binomial experiment.

The Binomial distribution can be used under the following conditions:

(i) The random experiment is performed repeatedly a finite and fixed number of times. In other words n, number of trials, is finite and fixed.

(ii) The outcome of the random experiment (trial) results in the dichotomous classification of events. In other words, the outcome of each trial may be classified into two mutually disjoint categories, called success (the occurrence of the event) and failure (the non- occurrence of the event). i.e. no middle event.

(iii) All trials are independent i.e. the result of any trial is not affected in any way, by the result of any trial, is not affected in any way, by the preceding trials and does not affect the result of succeeding trials.

(iv) The probability of success (happening of an event) in any trial is p and is constant for each trial. $q = 1-p$, is then termed as the probability of failure (non-occurrence of the event) and is constant for each trial.

For example, if we toss a fair coin n times (which is fixed and finite) then the outcome of any trial is one of the mutually exclusive events, viz., head (success) and tail (failure). Furthermore, all the trials are independent, since the result of any throw of a coin does not affect and is not affected by the result of other throws. Moreover, the probability of success (head) in any trial is ½, which is constant for each trial. Hence, the coin tossing

problems will give rise to Binomial distribution.

Similarly, dice throwing problems will also conform to Binomial distribution. More precisely, we expect a Binomial distribution under the following conditions:

(i) *n*, the number of trials is finite.

(ii) Each trial results in mutually exclusive and exhaustive outcomes termed as success and failure.
(iii) Trials are independent.

(iv) *p*, the probability of success is constant for each trial. Then $q = 1-p$, is the probability of failure in any trial.

Note: The trials satisfying the above four conditions are also known as Bernoulli trials.

2.2 Learning Outcomes

By the end of this unit, you will be able to:

- compute probabilities using the binomial probability distribution.
- compute the mean of binomial random variables.
- compute the standard deviation of binomial random variables

2.3 BINOMIAL DISTRIBUTION

2.3.1 Probability Function of Binomial Distribution

If *X* denotes the number of success in n trials satisfying the above conditions, then X is a random variable which can take the values 0,1,2,3 ...n; since in *n* trials we may get no success (i.e. all failures), one success, two successes, ...*n* successes. The general expression for the probability of r successes is given by:

$$P(r) = P(X = r) = {}^{n}C_{r}.P^{r}.q^{n-r};\ r = 0,1,2,.....n \qquad \text{...equation (1)}$$

Proof: Let S_i denote the success and F_i denote the failures at the ith trial; $i=1,2,.....n$. Then we have: $P(S_i) = p$ and $P(F_i) = q$; i= 1,2,3,.....n ...equation (2)

The probability of r successes and consequently (n-r) failuresin a sequence of n-trials in any fixed specified order, say, $S_1\ F_2\ S_3\ S_4\ F_5\ F_6.....S_{n-1}F_n$where *S* occurs *r* times and *F* occurs *(n-r)* times is given by:

[

$$= P(S_1).\ P(F_2).\ P(S_3).\ P(S_4).\ P(F_5).\ P(F_6)\ .....P(S_{n-1}).P(F_n)$$

By compound probability theorem, since the trials are independent

= p. q. p. p. q. q.....p.q (from equation 2)

= [p x p x p x .....r times] X [q x q x q x..... (n-r) times

$= p^{r}\ .\ q^{n-r}$

…equation (3) But in n trials, the total number of possible ways of obtaining r successes and (n-r) failure is

———— $^{n}C_{r}$,

all of which are mutually disjoint. The probability for each of these $^{n}C_{r}$ mutually exclusive ways is

the same as given in equation (2), viz., $p^{r} q^{n-r}$.

Hence by the addition theorem of probability, the required probability of getting *r* successes and consequently (n-r) failures in *n* trials, in *any order what-so-ever* is given by:

$$P(X=r) = P^{r}q^{n-r} + P^{r}q^{n-r} + \text{.....} + p^{r}q^{n-r} \ (^{n}C_{r} \text{ terms})$$

$$= {}^{n}C_{r}P^{r}q^{n-r};\ r = 0, 1, 2\text{.....}n$$

Table 1.2 Binomial Probabilities

R	***P*(r) = *P*(X = r)**
0	$^{n}C_{0}P^{0}q^{n} = q^{n}$
1	$^{n}C_{1}P^{0}q^{n-1}$
2	$^{n}C_{2}P^{2}q^{n-2}$
.	.
.	.
.	.
.	.
N	$^{n}C_{0}P^{0}q^{n} = P^{n}$

Note:

1. Putting *r = 0,1,2,.....n* in equation 1, we get the probabilities of 0,1, 2, ...*n* success respectively in *n* trials and these are tabulated in the table above. Since these probabilities are the successive terms in the Binomial expansion $(q + p)^{n}$, it is called the **BINOMIAL DISTRIBUTION**
2. Total probability is unity, i.e. 1;

$$\sum$$

$$= q^{n} + {}^{n}C_{1}q^{n-1}p + {}^{n}C_{2}q^{n-1}P^{2} + \text{.....}P^{n}$$

$$= (q+p)^{n} = 1 \quad \textit{Therefore, } p+q = 1$$

3. The expression for P(X = r) in equation 1is known as the probability mass function of the Binomial distribution with *parameters n and p*. The random variable X following the probability law expressed in equation 1 is called the *Binomial Variate* with parameters *n* and *p*. The Binomial distribution is completely determined, i.e. all the probabilities can be obtained, if *n* and *p* are known. Obviously, *q* is known when *p* is given because *q = 1-p.*
4. Since the random variable X takes only integral values, Binomial distribution is a discrete

probability distribution.

5. For *n* trials, the binomial probability distribution consists of (n+1) terms, the successive binomial coefficients being,

$$^nC_0, \ ^nC_1, \ ^nC_2, \ ^nC_3, \ ....., \ ^nC_{n-1}, \ ^nC_n$$

Since $^nC_0 = \ ^nC_n = 1$, the first and last coefficient will always be 1.

Further, since nCr = $^nC_{n-r}$, the binomial coefficients will be symmetric. Moreover, we have for all values of x:

$$(1+x)n = \ ^nC_0 + \ ^nC_1 + \ ^nC_2 + \ldots\ldots\ldots + \ ^nC_n.$$

This implies that $(1+x)n = \ ^nC_0 + \ ^nC_1 + \ ^nC_2 + \ldots\ldots\ldots + \ ^nC_n = 2^n$

i.e. the sum of binomial coefficients is 2^n

2.4 The Mean of a Binomial Distribution

$$\text{Mean} = \Sigma rp(r) = \ ^nC_1q^{n-1}p + 2\ ^nC_2q^{n-2}.p^2 + 3\ ^nC_3q^{n-3}.p^3 + ... + np^n$$

$$= nq^{n-1}p + 2.\ \frac{n-2}{q}p^2 + \frac{n-3}{q}p^3 + ... + np^n$$

$$= np[q^{n-1} + (n-1)q^{n-2}p + \frac{n-3}{q}p^2 + ... + p^{n-1}]$$

$$= np[q^{n-1} + \ ^{n-1}C_1q^{n-2}p + \ ^{n-1}C_2q^{n-3}p^2 + ... P^{n-1}]$$

$= np\ (q+p)^{n-1}$ *(By Binomial expansion for positive integer index),*
Therefore, p+q = 1
Therefore, ***Mean= np.***

2.5 Variance of a Binomial

$$\text{Variance} = \Sigma r^2p(r) - [\Sigma rp(r)]^2 = \Sigma r^2p(r) - (mean)^2 \qquad .....(*)$$

$$\Sigma r^2 p(r) = 1^2 X\, {}^nC_1 q^{n-1} p + 2^2 {}^nC_2 q^{n-2}.p^2 + 3^2 {}^nC_3 q^{n-3}.p^3 + \ldots\ldots\ldots + n^2 p^n$$

$$= nq^{n-1}p + \frac{n-2}{q} p^2 + \frac{n-3}{q} p^3 + \ldots\ldots + n^2 p^n$$

$$np[q^{n-1} + 2(n-1)q^{n-2}p + {}^3/_2(n-1)(n-2)q^{n-3}p^2 + \ldots\ldots\ldots + np^{n-1}]$$

$$= np[\{q^{n-1} + (n-1)q^{n-2}p + \frac{n-3}{q} p^2 + \ldots\ldots + 1p^{n-1}\}$$

$$+\{(n-1)q^{n-2}p + (n-1)(n-2)q^{n-3}p^2 + \ldots\ldots\ldots + (n-1)P^{n-1}\}]$$

$$= np[\{(q+p)^{n-1} + (n-1)p\{q^{n-2} + (n-2)q^{n-3}p + \ldots\ldots\ldots p^{n-2}\}]$$

$$= np[(q+p)^{n-1} + (n-1)p(q+p)^{n-2}]$$

$$= np[1 + (n-1)p]$$

Substituting in (*) above we get

$$Variance = np[1 + np - p] - (np)^2 = np[1 + np - p - np] = np[1-p] = npq$$

Hence for the Binomial Distribution; Mean=np; and Variance = npq

2.5.1 SOLVED EXAMPLES

Example 1: Ten unbiased coins are tossed simultaneously. Find the probability of obtaining:

(i) Exactly six heads

(ii) At least eight heads

(iii) No head

(iv) At least one head

(v) Not more than three heads

(vi) At least four heads

Solution: *p*denotes the probability of a head, q denotes the probability of tail
In this case, p =q = ½ and n= 10

Recall the Binomial probability law that the probability of *r* heads is given by

$$p(r) = P(X=r) = {}^{n}C_r P^r q^{n-r}$$

(i) Probability of exactly *six* heads Here, n=10, r= 6, p= ½ , q = ½ *p*(*6 heads*) = ${}^{10}C_6 P^6 q^{10-6}$

But, recall that ${}^{n}C_r$ ———— ,

Therefore, ${}^{10}C_6$ ————

———— = 210

p(exactly *6 heads*) = 210 . $(½)^6$. $(½)^4$

= 210 X $^1/_{64}$ X $^1/_{16}$

= ——

= $^{210}/_{1024}$

= 0.205

p (exactly *6 heads*) = ——

(ii) Probability of at least eight heads = *P*(*X*≥*8*) = *p*(*8*)+*P*(*9*)+*p*(*10*)
i.e*P*(exactly 8 heads) + *P*(exactly 9 heads) + *P*(exactly 10 heads)

Here, we find the probability of each of the three separately using the formula ${}^{n}C_r P^r q^{n-r}$ and we add them together.

Therefore, *P*(exactly 8 heads) = ${}^{10}C_8 P^8 q^{10-8}$

———— $(½)^8$. $(½)^2$

= 45 X $^1/_{256}$ X ¼

= ——

= $^{45}/_{1024}$

$= 0.044$

$P(\text{exactly 9 heads}) = {}^{10}C_9 P^9 q^{10-9}$

$= 10 \times {}^{45}/_{1024}$

$= {}^{450}/_{1024}$

$= 0.439$

$P(\text{exactly 10 heads}) = {}^{10}C_{10} P^{10} q^{10-10}$

$= 1 \times {}^{1}/_{1024}$

$= {}^{1}/_{1024}$

$= 0.001$

Therefore, Probability of at least 8 heads = $+ + $

$P(\text{at least 8 heads}) =$

$= 0.044 + 0.439 + 0.001$

$= 0.484$

(iii) Probability of no head = P(X= r = 0)

$P(X = r) = {}^{n}C_r P^r q^{n-r}$

$P(\text{ 0 head}) = {}^{10}C_0 P^0 q^{10-0}$

$= 1 \times 1 \times$

$P(0 \text{ head}) =$

$= 1 \times 1 \times {}^{1}/_{1024}$

$= {}^{1}/_{1024}$

$= 0.001$

(iv) Probability of at least one head

= 1 – *P*[No head]

= 1 – *P*(0)

Recall that *P*(0) =

= 1 -

= 1 -

=

$= 1 - {}^{1}/_{1024}$

$= 1 - 0.001$

$= 0.999$

(v) Probability of not more than three heads

= *P*(X≤3) = *P*(0) + *P*(1) + *P*(2) + *P*(3)

= $[{}^{10}C_0 + {}^{10}C_1 + {}^{10}C_2 + {}^{10}C_3]$ =

=

$= C_0^{10}\, p^{10} q^{10-0} + C_1^{10}\, p^{10} q^{10-1} + C_2^{10}\, p^{10} q^{10-2} + C_3^{10}\, p^{10} q^{10-3}$

$= C_0^{10} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{10} + C_1^{10} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^9 + C_2^{10} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^8 + C_3^{10} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7$

$= 1 \times \left(\frac{1}{2}\right)^{10} + 10 \times \left(\frac{1}{2}\right)^{10} + 45 \times \left(\frac{1}{2}\right)^{10} + 120 \times \left(\frac{1}{2}\right)^{10}$

$= \frac{1}{1024} + \frac{10}{1024} + \frac{45}{1024} + \frac{120}{1024}$

$= \frac{176}{1024}$

$= 0.172$

(vi) Probability (at least 4 heads) = (X ≥4) = 1 – P (X≤3)

= 1 – [p(0) + p(1) + p(2) + p(3)] = 1 -

=

$= 1 - 0.172 = 0.828$

Example 2: There are five flights daily from Pittsburgh via US Airways into the Bradford, Pennsylvania Regional Airport. Suppose the probability that any flight arrives late is .20.

(a) What is the probability that none of the flights are late today?

(b) What is the probability that exactly one of the flights is late today?

The probability that a particular flight is late is .20.

$$p = 0.2; q = 1 - p = 1 - 0.2 = 0.8$$

There are five flights, so $n = 5$, and x, the random variable, refers to the number of successes. In this case a "success" is a plane that arrives late.

(a) What is the probability that none of the flights are late today?

$$p(\mathrm{r}) = P(X=r) = {}^{n}C_r P^r q^{n-r}$$

$$P(0) = C_0^5\,(0.2)^0(0.8)^{5-0}$$

$$P(0) = C_0^5\,(0.2)^0(0.8)^5$$

$$= (1)(1)(0.3277)$$

$$= 0.3277$$

(b) What is the probability that exactly one of the flights is late today?

$$P(1) = C_1^5\,(0.2)^1(0.8)^{5-1}$$

$$P(0) = C_1^5\,(0.2)^1(0.8)^4$$

$$= (5)(0.2)(0.496)$$

$$= 0.496$$

The entire probability distribution is shown in Table below:

Number of late flights	**Probability**
0	0.3277
1	0.4096
2	0.2048
3	0.0512
4	0.0064
5	0.0003
Total	1.000

Self-Assessment Exercises 1

1. The NCC survey shows that 70% of Nigerian households have mobile phones. If 15 households are chosen at random, what is the probability that

   (a) exactly 10 have mobile phones?

   (b) At least 13 have mobile phones?

2. According to the NCC, 70% of Nigerian households have mobile phones. If 300 households are chosen at random, determine the

   (a) mean (b) Variance (c.) standard deviation

2.6 SUMMARY

In this unit, learners have been made to understand that a Binomial Distribution is the sum of Independent Bernoulli Random Variables and that the Binomial distribution describes thedistribution of binary data from a finite sample. Thus it gives the probability of getting *r* events out of *n* trials. In summary, the binomial distribution describes the behaviour of a count variable *X* if the following conditions apply:

1. The number of observations *n* is fixed.

2. Each observation is independent.

3. Each observation represents one of two outcomes ("success" or "failure").

4. The probability of "success" *p* is the same for each outcome.

If in your application of Binomial these conditions are met, then *X* has a Binomial distribution with parameters *n* and *p,* abbreviated *B(n, p).*

2.7 REFERENCES/FURTHER READINGS/WEB RESOURCES

Spiegel, M. R. and Stephens L.J. (2008).*Statistics***.(**4th ed.).New York, McGraw Hill. Gupta S.C. (2011). *FundamentalsofStatistics***.(**6th Rev. and Enlarged ed.). Mumbai India: Himalayan Publishing House

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance***.**London,Macmillan

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

2.8 Possible Answers to SAEs

Possible answers to the SAEs within the content**.**
Answers to SAEs 1

1. (a) 0.2061
   (b) 0.1268

2. (a) 210
   (b) 63 (c) 7.9

UNIT 3: NORMAL DISTRIBUTION

Unit Structure

3.1 Introduction

3.2 Learning Outcomes

3.3 Normal Distribution

3.3.1 Properties of Normal Distribution
3.3.2 Definitions
3.4 The Standard Normal Distribution
3.4.1 Solved Examples
3.5 Relationship between Binomial and normal distribution

3.6 Summary

3.7 References/Further Reading/ Web Resources
3.8 Possible Answers to Self-Assessment Exercises (SAEs) within the content

3.1 INTRODUCTION

The Normal probability distribution commonly called the normal distribution is one of the most important continuous the theoretical distributions in Statistics. Most of the data relating to economic and business statistics or even in the social and physical sciences conform to this distribution. The normal distribution was first discovered by English Mathematician De-voire (1667-1754) in 1733 who obtained the mathematical equation for this distribution while dealing with problems arising in the game of chance. Normal distribution is also known as Gaussian distribution (Gaussian Law of Errors) after Karl Friedrich Gauss (1777-1855) who used the distribution to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies.
Today, normal probability model is one of the most important probability models in statistical analysis.
Its graph, called the normal curve is shown below:

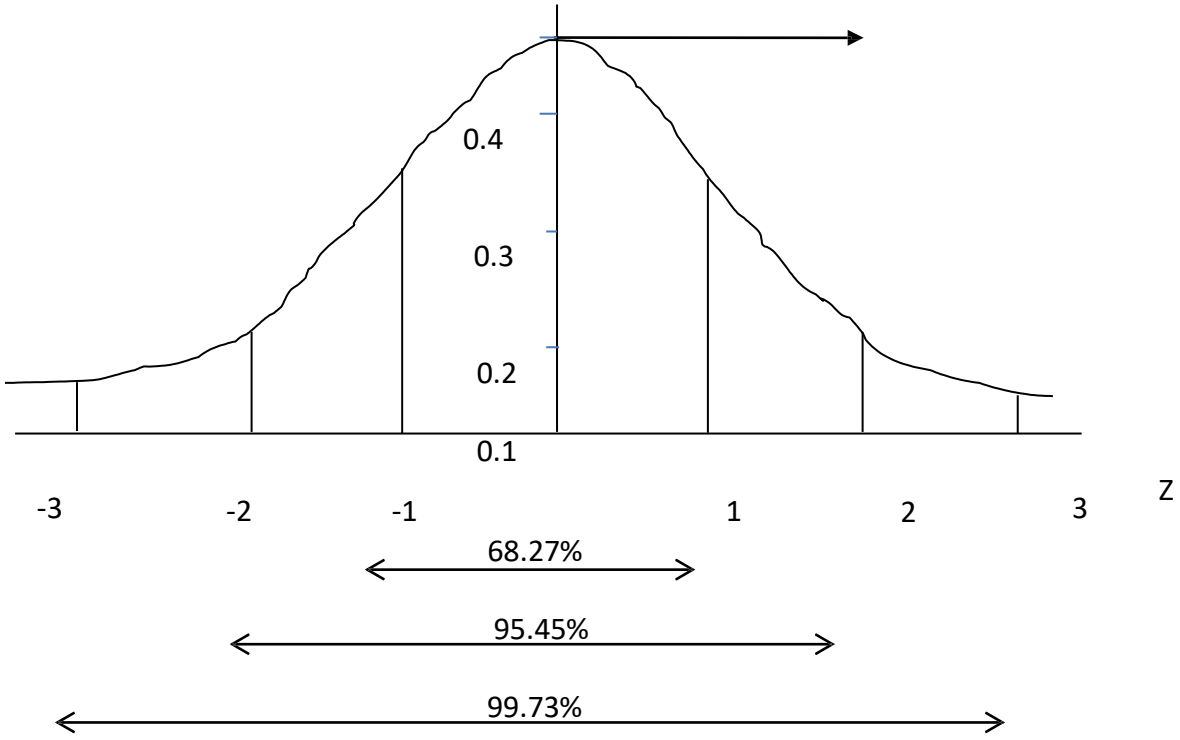

Fig. 1: Normal Curve

3.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- State the characteristics of normal probability distribution
- Compute value for normal probability distribution
- Compute probabilities from normal distribution

3.3 NORMAL DSITRIBUTION

3.3.1 Properties of the normal distribution curve

1. The mode which is the point on the horizontal axis where the curve is a maximum occurs at

   $X = \mu$ (i.e. at the mean)

2. The curve is symmetric about a vertical axis through the mean μ

3. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.
4. The total area under the curve and above the horizontal axis is equal to 1

When does normal distribution arise?

Because the normal probability density function (pdf) peaks at the mean and ―tails off‖ towards the extremes, the normal distribution provides a good approximation for many naturally occurring random variables. However, the normal distribution occurs even more widely due to the following:

1. The total (and also the average) of a large number of random variables which have the same probability distribution approximately has a normal distribution. For instance, if the amount taken by a shop in a day has particular (maybe unknown) distribution, the total of 100 days' takings is the sum of 100 identically distributed random variables and so it will (approximately) have a normal distribution. Many random variables are normal because of this. For example, the amount of rainfall which falls during a month is the total of the amounts of rainfall which have fallen each day or each hour of the month and so is likely to have a normal distribution. In the same way the average or total of a large sample will usually have a normal distribution. This can be explored further by further readings on populations and samples
2. The normal distribution provides approximate probabilities for the binomial distribution when n, the number of trials is large.

3.3.2 Definitions

1. A random variable X has normal Distribution, and it is referred to as a normal random variable, if and only if its probability density is given by:

$$\frac{}{\sqrt{}} \quad \overline{\quad}\ , -\infty < x < \infty \quad or$$

$$\frac{}{\sqrt{}} \quad \overline{\quad}\ , -\infty < x < \infty \text{ and } \zeta > 0$$

Where and are constants given by $= {}^{22}/_{7}$,$\sqrt{\quad} = 2.5066$

And = 2.71828 (which is the base system of Natural Logarithm)

2. The normal distribution with $\mu = 0$ and $\zeta = 1$ is referred to as the standard normal distribution.
3. If X has a normal distribution with the mean of μ and the deviation ζ, then is the standard normal distribution

Note

(i) Definition No.3 is used to determine probabilities relating to random variables having normal distribution other than the standard normal distribution.

(ii) Because a normal curve is symmetrical about its mean, $P(z < -a) = P(z > a)$

(iii) $P(z < a) + P(z > a) = 1.0000$

(iv) Only values of $P(z < a)$ are shown in most statistical tables. For $P(z > a)$, $1 - P(z < a)$ is used.

Students are implored to make copies of normal tables from any standard statistic textbook.

3.4 The Standard Normal Distribution

The number of normal distributions is unlimited, each having a different mean (IL), standard deviation (a), or both. While it is possible to provide probability tables for discrete distributions such as the binomial and the Poisson, providing tables for the infinite number of normal distributions is impossible. Fortunately, one member of the family can be used to determine the probabilities for all normal distributions. It is called the standard normal distribution, and it is unique because it has a mean of 0 and a standard deviation of 1.

Any normal distribution can be converted into a standard normal distribution by subtracting the mean from each observation and dividing this difference by the standard deviation. The results are called **z values**. They are also referred to as **z scores**, the **z statistics**, the **standard normal deviates, the standard normal values, or just the normal deviate.**

The formula for $z\ score\ or\ value$ is given as:

$$z = \frac{X - \mu}{\sigma}$$

where:
X is the value of any particular observation or measurement.
μ is the mean of the distribution.
σ is the standard deviation of the distribution

3.4.1 Solved Example

1. Using normal tables, find the values of the following probabilities: (a) P(z < 0.50)
   (b) P(z < -2.50)

   (c) P(1.62 < z < 2.20)
   (d) P(-1.50 < z < 2.50)
   (e) P(z > 0.50)

Solution

(a) P(z < 0.50) = 0.6915

i.e read directly from statistical table

(b) P(z < -2.50) = 0.0062

(c) P(1.62 < z < 2.20)
= P(z < 2.20) – P(z < 1.62)
0.9861 – 0.9474
0.0387

(d) P(-1.50 < z < 2.50)
= P(z < 2.50) – P(z < -1.50)
= 0.9938 – 0.0668
= 0.9270

(e) P(z > 0.50)
Because most tables only provide for P(z < 0.50), we shall therefore apply: P(z > 0.50) = 1- P(z < 0.50)
= 1 – 0.6915
= 0.3085

2. Given a normal distribution with mean of 230 and standard deviation of 20, what is the probability that an observation from this population is:
   (a) Greater than 280
   (b) Less than = 220
   (c) Lies between 220 and 280

Solution

(a) 280) = P(z > 2.50)

= 1 – P(z < 2.50)

X = 280,
μ= 230, ζ = 20

= 1 – 0.9938

= 0.0062

Therefore,
Z =

=
2
.
5
0

Therefore,
P(X >

(b) P(X < 220)

Z = ———

= -0.50

Therefore, P(X < 220) = P(z < -0.50)

= 0.3085

(c) P(220 < X < 280)

= P(-0.50 < z < 2.50)

= P(z < 2.50) - P(z < -0.50)

= 0.9938 – 0.3085

= 0.6853

3.. The weekly incomes of shift foremen in the glass industry are normally distributed with a mean of $1,000 and a standard deviation of $100. What is the z value for the income X of a foreman who earns:
(a) $1,100 per week?
(b) $900 per week?

Solution
Using formula

$$z = \frac{X - \mu}{\sigma}$$

(a) X = $1,100

$$z = \frac{X - \mu}{\sigma} = \frac{1100 - 1000}{100} = 1.00$$

(b. X = $900

$$z = \frac{X - \mu}{\sigma} = \frac{900 - 1000}{100} = -1.00$$

4 A random variable X is approximately normally distributed with $\mu = 20$ and $\sigma = 4$, compute Z_1 when $X_1 = 21$.

Solution

$$z = \frac{X - \mu}{\sigma} = \frac{21 - 20}{4} = 0.25$$

3.5 Relation between Binomial and Normal Distribution

Normal Distribution is a limiting case of the binomial probability distribution under the following conditions:

(i) *n*, the number of trials is indefinitely large, *i.e.n* ∞.

(ii) Neither *p* nor *q* is very small.

⟶

We know that for a binomial variate *X* with parameter *n* and *p*.

E(*X*) = *np* and Var(*X*) = *npq*

De-Moivre proved that under the above two conditions, the distribution of standard Binomial variate

——— √—

tends to the distribution of standard Normal variate.

If *p* and *q* are nearly equal (i.e., *p* is nearly ½), the normal approximation is surprisingly good even

for small values of *n*. However, then *p* and *q* are not equal, i.e. when *p* or *q* is small, even then the Binomial distribution tends to normal distribution but in this case the convergence is slow. By this we mean that if *p* and *q* are not equal then for Binomial distribution to tend to Normal distribution we need relatively larger value of n as compared to the value of *n* required in the case when *p* and *q* are nearly equal. Thus, the normal approximation to the Binomial distribution is better for increasing values of *n* and is exact in the limiting case as *n* ∞.

In the light of the above, binomial related problems can be solve through Poisson approximation using a combination of both.

Self-Assessment Exercises (SAEs 1)

1. State 2 characteristics of normal distribution
2. A random variable X is approximately normally distributed with $\mu = 50$ and $\sigma = 8$.

3.6 SUMMARY

In summary, learners would have understood that the normal distribution can be described completely by the two parameters. The mean is the center of the distribution and the standard deviation is the measure of the variation around the mean.

3.7 REFERENCES/FURTHER READINGS/WEB RESOURCES

Spiegel, M. R. and Stephens L.J. (2008).*Statistics***.(**4th ed.). New York:McGraw Hill Press. Gupta S.C (2011). *Fundamentals of Statistics.*(6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance.*London:Macmillan.

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

3.8 Possible Answers to SAEs

These are the possible answers to the SAEs within the content**.**

Answers to SAEs 1

1. (i) The curve is symmetric about a vertical axis through the mean μ
2. (ii) The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.

2.. (a) -0.25 (b) 1.25

UNIT 4: POISSON DISTRIBUTION CONTENTS

Unit Structure

4.1 Introduction
4.2 Learning Outcomes
4.3 Poisson Distribution
4.3.1 Characteristics of Poisson Distribution
4.4 Condition for using Poisson Distribution
4.5 Formula for Poisson Distribution
4.5.1 Solved Examples
4.6 Summary
4.7 References/Further Reading/ Web Resources
4.8 Possible Answers to Self-Assessment Exercise(s) within the content

4.1 INTRODUCTION

Poisson distribution was derived in 1837 by a French mathematician Simeon D. Poisson (1781 – 1840). Poisson distribution may be obtained as a limiting case of Binomial probability distribution under the following conditions:

(i) *n,* the number of trials is indefinitely large *i.e.n tends towards infinity*

(ii) *p,* the constant probability of success for each trial is indefinitely small *i.e.p tends towards zero.*

(iii) $np = \mu$, is finite

Under the above three conditions the Binomial probability function tends to the probability function of the Poisson distribution given as:

Where X or r is the number of success (occurrences of the event) $\mu = np$ and e =2.71828 (the base of the system of natural logarithm)

4.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- State the characteristics of Poisson distribution
- Compute probabilities from Poisson distribution

4.3 POISSON DSITRIBUTION

4.3.1 Characteristics of Poisson Distribution

1. The random variable is the number of times some event occurs during a defined interval.
2. The probability of the event is proportional to the size of the interval.
3. The intervals which do not overlap are independent.

4.4 Condition for Using Poisson Distribution

The condition under which Poisson distribution is obtained is in a limiting case of Binomial Distribution. It is applicable in fields such as Queuing Theory (waiting line problems), insurance, biology, business, Economics and Industry. Some of the practical situation in which the distribution can be applied include but not limited to:

(i) The number of vehicles arriving at a filling station
(ii) Number of patients arriving at a hospital
(iii) The number of accidents taking place per day on a busy road
(iv) The number of misprint per page of a typed material etc.

4.5 Formula for Poisson Distribution

The Poisson distribution can be described mathematically by the formula:

$$P(X) = \frac{\mu^X e^{-\mu}}{X!}$$

μ is the mean number of occurrences (successes) in a particular interval.
e is the constant 2.71828 (base of the Naperian logarithmic system).
X is the number of occurrences (successes).
P(x) is the probability for a specified value of *x.*

4.5.1 Solved Example

Example 1: The mean number of misprints per page in a book is 1.2. What is the probability of finding on a particular page?

(a) No misprints
(b) Three or more misprints

Solution

μ = 1.2

(a) Pr (No misprints)

=Pr(X=0)

$= e^{-1.2}$

= 0.301 (b) Pr(or more misprint)

= Pr(X ≥ 3)

= 1 – [Pr(0) + Pr(1) + Pr(2)]

Pr(0) = 0.301 as in (a) above

Pr(1) $\frac{}{}$

= 0.3612

Pr(2)= $\frac{}{}$

= 0.21672

Pr (0) + Pr(1) + Pr(2) = 0.87892

Therefore, Pr(X ≥ 3) = 1 - 0.87892

= 0.12108
= 0.121

Example 2: Suppose a random sample of 1,000 flights shows a total of 300 bags were lost. Thus, the arithmetic mean number of lost bags per flight is 0.3, found by *300/1,000.* Find the probability of not losing any bags.

Solution

$$P(X) = \frac{\mu^X e^{-\mu}}{X!}$$

$$X = 0;\ \mu = 0.3$$

$$P(X) = \frac{\mu^X e^{-\mu}}{X!}$$

$$P(0) = \frac{0.3^0 e^{-0.3}}{0!}$$
$$= 0.7408$$

Self-Assessment Exercises (SAEs 1)

1. State the features of Poisson distribution
2. Suppose a random sample of 1,000 flights shows a total of 300 bags were lost. Thus, the arithmetic mean number of lost bags per flight is 0.3, found by *300/1,000.* Find the probability of exactly 1 lost bag.

4.6 SUMMARY

In this unit, student must have learnt the rudiments and applications of Poisson distribution. Students are must have learnt how to solve problems using Poisson distribution.

4.7 REFERENCES/FURTHER READING/WEB RESOURCES

Spiegel, M. R. and Stephens L.J. (2008).*Statistics***.(**4th ed.). New York:McGraw Hill Press. Gupta S.C. (2011). *Fundamentals of Statistics.***(**6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House.

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance.*London:Macmillan.

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

4.8 Possible Answers to Self-Assessment Exercise(s) within the content

1. (i) The random variable is the number of times some event occurs during a defined interval.
   (ii) . The probability of the event is proportional to the size of the interval.
   (iii. The intervals which do not overlap are independent.

2.. 0.2222

MODULE 2: STATISTICAL HYPOTHESIS TEST

This module explains the method of making decisions using data from a scientific study. In statistics, a result is interpreted as being statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "test of significance" was coined by statistician Ronald Fisher. These tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance; this can help to decide whether results contain enough information to cast doubt on conventional wisdom, given that conventional wisdom has been used to establish the null hypothesis. The *critical region* of a hypothesis test is the set of all outcomes which cause the null hypothesis to be rejected in favour of the alternative hypothesis. Statistical hypothesis testing is sometimes called **confirmatory data analysis**, in contrast to exploratory data analysis, which may not have pre-specified hypotheses. Statistical hypothesis testing is a key technique of frequentist inference.

Statistics are helpful in analyzing most collections of data. This is equally true of hypothesis testing which can justify conclusions even when no scientific theory exists.
Common test Statistics are; t-test, z-test, chi-square test and f-test which is sometimes referred to as analysis of variance (ANOVA) test.
In this module, five statistical tests will be discussed and analyzed in order to make learners appreciate and understand of the different statistical hypothesis tests. These statistical tests are:

Unit 1: **T- test**
Unit 2: **F- test**
Unit 3: **Chi square test**
Unit 4: **ANOVA**
Unit 5: **Parametric and Non-Parametric test Methods**

UNIT 1: T–TEST

UNIT STRUCTURE

1.1 Introduction
1.2 Learning Outcomes
1.3 T-Test
1.3.1 Characteristics of T-test
1.3.2 Application of *t*-distribution
1.4 Test for single mean
1.4.1 Assumptions for Student's test
1.4.2 One-tailed test
1.4.3 Two-tailed test
1.5. Solved Examples
1.6 Summary
1.7 References/Further Reading/Web Resources
1.8 Possible Answers to Self-Assessment Exercises (SAEs) within the content

1.1 INTRODUCTION

A *t*-test (also known as Student's *t*-test) is a tool for evaluating the means of one or two populations using hypothesis testing. A t-test may be used to evaluate whether a single group differs from a known value (a one-sample t-test), whether two groups differ from each other (an independent two-sample t-test), or whether there is a significant difference in paired measurements (a paired, or dependent samples t-test).

If the population variance is unknown then for the large samples, its estimates provided by sample variance S^2 is used and normal test is applied. For small samples an unbiased estimate of population variance ζ^2 is given by:

It is quite conventional to replace ζ^2 by S^2 (for small samples) and then apply the normal test even

for small samples. W.S.Goset, who wrote under the pen name of Student, obtained the sampling

distribution of the statistic $\frac{}{\sqrt{}}$ for small samples and showed that it is far from normality. This

discovery started a new field, viz *Exact Sample Test* in the history of statistical inference.
Note: If $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ is a random sample of size *n* from a normal population with mean μ and variance ζ^2 then the Student's *t* statistic is defined as:

Where = is the sample mean and $\sum$ is an unbiased estimate of the population variance ζ^2

1.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Conduct a test of hypothesis about a population mean.
- State conclusion to hypothesis testing

1. 3 T-test

1.3.1 Characteristics of T-test

1. It is a continuous distribution.
2. It is bell-shaped and symmetrical.
3. There is a family of *t* distributions. Each time the degrees of freedom change, a new distribution is created.
4. As the number of degrees of freedom increases, the shape of the *t* distribution approaches that of the standard normal distribution.
5. The *t* distribution is flatter, or more spread out, than the standard normal distribution.

1.3.2 Applications of *t*-distribution

(i) *t*-test for the significance of single mean, population variance being unknown

(ii) *t*-test for the significance of the difference between two sample means, the population variances being equal but unknown

(iii) *t*-test for the significance of an observed sample correlation coefficient

1.4 Test for Single Mean

Sometimes, we may be interested in testing if:

(i) the given normal population has a specified value of the population mean, say μ_o.

(ii) the sample mean differ significantly from specified value of population mean.

(iii) A given random sample $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ of size n has been drawn from a normal population with specified mean μ_o.

Basically, all the three problems are the same. We set up the corresponding null hypothesis thus:

(a) H_O: $\mu = \mu_O$ i.e the population mean is μ_O

(b) H_O: There is no significant difference between the sample mean and the population mean.

In order words, the difference between and μ is due to fluctuations of sampling.

(c) H_O: The given random sample has been drawn from the normal population with mean μ_O.

The formula for the T-test is mathematically written as:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

with n - 1 degrees of freedom, where:
$\bar{X}$ is the mean of the sample.
μ is the hypothesized population mean.
s is the standard deviation of the sample.
n is the number of observations in the sample.

We compute the test-statistic using the formula above under H_O and compare it with the tabulated value of t for (n-1) $d.f$ at the given level of significance. If the absolute value of the calculated t is greater than tabulated t, we say it is significant and the null hypothesis is rejected. But if the calculated t is less than tabulated t, H_O may be accepted at the level of significance adopted.

1.4.1 Assumptions for Student's test

(i) The parent population from which the sample is drawn is normal

(ii) The sample observations are independent i.e the given sample is random. (iii) The population standard deviation ζ is unknown

1.4.2 One-Tailed Test

Let us consider an example of a one-tailed test about a population mean for the σ unknown case. A business travel magazine wants to classify transatlantic gateway airports according to the mean rating for the population of business travelers. A rating scale with a low score of 0 and a high score of 10 will be used, and airports with a population mean rating greater than 7 will be designated as superior service airports. The magazine staff surveyed a sample of 60 business travelers at each airport to obtain the ratings data. The sample for London's Heathrow Airport provided a sample mean rating of $\bar{x} = 7.25$ and a sample standard deviation of $s = 1.052$. Do the data indicate that Heathrow should be designated as a superior service airport?

We want to develop a hypothesis test for which the decision to reject $H0$ will lead to the conclusion that the population mean rating for the Heathrow Airport is *greater* than 7. Thus, an upper tail test with $H_a = \mu > 7$ is required. The null and alternative hypotheses for this upper tail test are as follows:

$$H_0: \mu \leq 7$$

$$H_1: \mu > 7$$

We will use α _ .05 as the level of significance for the test. Using the formula with$\bar{x} = 7.25$, $\mu_0 = 7$, $s = 1.052.$, and $n = 60,$ the value of the test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$t = \frac{7.25 - 7}{1.052/\sqrt{60}}$$

$$= 1.84$$

The sampling distribution of *t* has $n - 1 = 60 - 1 = 59$ degrees of freedom. Because the test is an upper tail test, the *p*-value is the area under the curve of the *t* distribution to the right of $t = 1.84$
The *t* distribution table provided in most textbooks will not contain sufficient detail to determine the exact *p*-value, such as the *p*-value corresponding to $t = 1.84$. For instance using the *t* distribution with 59 degrees of freedom provides the following information

Area in upper tail	**0.2**	0.1	0.05	0.025	0.01	0.005
T –value (59 df)	0.848	1.296	1.671	2.001	2.391	2.662

We see that $t = 1.84$ is between 1.671 and 2.001. Although the table does not provide the exact *p*-value, the values in the "Area in Upper Tail" row show that the *p*-value must be less than .05 and greater than .025. With a level of significance of α _ .05, this placement is all we need to know to make the decision to reject the null hypothesis and conclude that Heathrow should be classified as a superior service airport.

With *t* = 1.84 provides the upper tail *p*-value of .0354 for the Heathrow Airport hypothesis test.
With .0354 < 0.05, we reject the null hypothesis and conclude that Heathrow should be classified as a superior service airport.

1.4.3 Two-Tailed Test

To illustrate how to conduct a two-tailed test about a population mean for the σ unknown case, let us consider the hypothesis testing situation facing Holiday Toys. The company manufactures and distributes its products through more than 1000 retail outlets. In planning production levels for the coming winter season, Holiday must decide how many units of each product to produce prior to knowing the actual demand at the retail level. For this year's most important new toy, Holiday's marketing director is expecting demand to average 40 units

per retail outlet. Prior to making the final production decision based upon this estimate, Holiday decided to survey a sample of 25 retailers in order to develop more information about the demand for the new product. Each retailer was provided with information about the features of the new toy along with the cost and the suggested selling price. Then each retailer was asked to specify an anticipated order quantity.

With μ denoting the population mean order quantity per retail outlet, the sample data will be used to conduct the following two-tailed hypothesis test:

$$H_0: \mu = 40$$

$$H_1: \mu \neq 40$$

If H_0 cannot be rejected, Holiday will continue its production planning based on the marketing

director's estimate that the population mean order quantity per retail outlet will be $\mu = 40\, units$ units. However, if H_0 is rejected, Holiday will immediately reevaluate its production plan for the product. A two-tailed hypothesis test is used because Holiday wants to re-evaluate the production plan if the population mean quantity per retail outlet is less than anticipated or greater than anticipated. Because no historical data are available (it's a new product), the population mean μ and the population standard deviation must both be estimated using $\bar{x}$ and s from the sample data

The sample of 25 retailers provided a mean of $\bar{x} = 37.4$ and a standard deviation of $s = 11.79$ units. Using the formula, with $\bar{x} = 37.4$, $\mu_0 = 40$, $s = 11.79$, and $n = 25$, the value of the test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$t = \frac{37.4 - 40}{11.79/\sqrt{25}}$$

$$= -1.10$$

Because we have a two-tailed test, the $p - value$ is two times the area under the curve of the t distribution for $t \leq -1.10$. Using Table, the t distribution table for 24 degrees of freedom provides the following information.

Area in upper tail	**0.2**	0.1	0.05	0.025	0.01	0.005
T –value (24 df)	0.857	1.318	1.711	2.064	2.492	2.797

t=1.10

The *t* distribution table only contains positive *t* values. Because the *t* distribution is symmetric, however, the area under the curve to the right of *t* =1.10 is the same as the area under the curve to the left of *t* = -1.10. We see that *t* =1.10 is between 0.857 and 1.318.

From the "Area in Upper Tail" row, we see that the area in the tail to the right of *t* = 1.10 is between .20 and .10. When we double these amounts, we see that the *p*-value must be between .40 and .20. With a level of significance of α = .05, we now know that the *p*-value is greater than α. Therefore, H_0 cannot be rejected. Sufficient evidence is not available to conclude that Holiday should change its production plan for the coming season.

The *p*-value obtained is .2822. With a level of significance of α =.05, we cannot reject H_0because .2822 > .05. The test statistic can also be compared to the critical value to make the two-tailed hypothesis testing decision. With α = .05 and the *t* distribution with 24 degrees of freedom, $-t_{0.25} = -2.064$ and $t_{0.25} = 2.064$are the critical values for the two-tailed test. The rejection rule using the test statistic is:

Reject H_0 if $t_{0.25} \leq -2.064$ or $t_{0.25} \geq 2.0644$

Based on the test statistic *t* = -1.10, H_0 cannot be rejected. This result indicates that Holiday should continue its production planning for the coming season based on the expectation that $\mu = 40$.

1.5 Solved Examples

Example: Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 11.8kg and standard deviation is 0.15kg. Does the sample mean differ significantly from the intended weight of 12kg, α=0.05

Hint: You are given that for *d.f* =9, $t_{0.05}$ = 2.26

Solution: n= 10, = 11.8kg, s = 0.15kg

Null hypothesis, H_O : μ = 12 kg (i.e the sample mean of =11.8 kg does not differ significantly from the population mean μ = 12 kg

*Alternative Hypothesis.*Ho: ≠ 12kg (Two tailed)

Solution

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

$$t = \frac{11.8 - 12}{0.15/\sqrt{10}}$$

$$t = \frac{-0.2}{0.15/\sqrt{3.1623}}$$

$$t = \frac{-0.2}{0.0474}$$

$$t = \frac{-0.2}{0.0474}$$

$$t = 4.219$$

The tabulated value of t for 9 d.f at 5% level of significance is 2.26. Since the the calculated t is much greater than the tabulated t, it is highly significant. Hence, null hypothesis is rejected at 5%

level of significance and we conclude that the sample mean differ significantly.

Example 2

The mean life of a battery used in a digital clock is 305 days. The lives of the batteries follow the normal distribution. The battery was recently modified with the objective of making it last longer. A sample of 20 of the modified batteries had a mean life of 311 days with a standard deviation of 12 days. Did the modification increase the mean life of the battery?

(a) State the null hypothesis and the alternate hypothesis.

(b) Compute the value of t. What is your decision regarding the null hypothesis? Briefly summarize your results.

Solution

(a)

$$H_0: \mu \leq 305$$
$$H_1: \mu > 305$$

(b)

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$
$$t = \frac{311 - 305}{12/\sqrt{20}}$$
$$t = \frac{6}{12/4.4721}$$
$$t = \frac{6}{2.6833}$$
$$t = 2.236$$

Reject Ho because 2.236> 1.729. The modification increased the mean battery life to more than 305 days.

Self-Assessment Exercise 1 (SAE 1)

1. The mean length of a small counterbalance bar is 43 millimeters. The production supervisor is concerned that the adjustments of the machine producing the bars have changed. He asks the Engineering Department to investigate. Engineering selects a random sample of 12 bars and measures each. The results are reported below in millimeters.

42	39	42	45	43	40	39	41	40	42	43	42

(a) Formulate the hypothesis

(b) Is it reasonable to conclude that there has been a change in the mean length of the bars? Use the .05 significance level.

1.6 SUMMARY

In summary, learners would have learnt how to apply t-test in solving statistical problems such as test to confirm if mean is a certain value, to test significance of the difference between two mean among others.

1.7 REFERENCES/FURTHER READING/WEB RESOURCES

Spiegel, M. R. and Stephens L.J. (2008).*Statistics*. (4th ed.). New York: McGraw Hill Press. Gupta S.C. (2011). *Fundamentals of Statistics*. (6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House.

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance.* London: Macmillan.

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

1.8 Possible Answers to Self-Assessment Exercise(s) within the content

1 (a)

$H_0: \mu = 43$

$H_1: \mu \neq 43$

(b)

$t = -2.913$

UNIT 2: F Distribution

UNIT STRUCTURE

2.1 Introduction
2.2 Learning Outcomes
2.3 F distribution
2.3.1 Characteristics of F distribution
2.4 Applications of the F-distribution
2.5 Comparing Two Population Variances
2.5.1 Assumption for F distribution for equality of variances
2.5.2 Solved Examples

2.6 Summary
2.7 References/Further Reading/ Web Resources
2.8 Possible Answers to Self-Assessment Exercises (SAEs) within the content

2.1 INTRODUCTION

The *F* distribution was named to honor Sir Ronald Fisher, one of the founders of modern-day statistics. This probability distribution is used as the distribution of the test statistic for several situations. It is used to test whether two samples are from populations having equal variances, and it is also applied when we want to compare several population means simultaneously.

2.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- List the characteristics of the F distribution
- Conduct a test of hypothesis to determine whether the variances of two populations are equal.

2.3 F distribution

The *F* family of distributions resembles the χ^2 distribution in shape: it is always non-negative and is skewed to the right. It has two sets of degrees of freedom (these are its parameters, labelled ν_1 and ν_2) and these determine its precise shape.

2.3.1 Characteristics of F distribution

1. There is a "family" of *F* distributions
2. The F distribution is continuous
3. The F distribution cannot be negative
4. It is positively skewed
5. It is asymptotic

2.4 Applications of the F-distribution

F-distribution has a number of applications in the field of statistics. This includes but not limited to the following:

(1) To test for equality of population variances

(2) To the equality of several populations means *i.e.* for testing Ho: $\mu_1 = \mu_2 = ... = \mu_k$.
This is by far the most important application of F-statistic and is done through the technique of Analysis of Variance (ANOVA). This shall be treated as a separate unit later.

(3) For testing the significance of an observed sample multiple correlation

(4) For testing the significance of an observed sample correlation ratio

2.5 **Comparing Two Population Variances**

The *F* distribution is used to test the hypothesis that the variance of one normal population equals the variance of another normal population. The following examples will show the use of the test:

• Two Barth shearing machines are set to produce steel bars of the same length. The bars, therefore, should have the same mean length. We want to ensure that in addition to having the same mean length they also have similar variation.

• The mean rate of return on two types of common stock may be the same, but there may be more variation in the rate of return in one than the other. A sample of 10 Internet stocks and 10 utility socks shows the same mean rate of return, but there is likely more variation in the Internet stocks.

• A study by the marketing department for a large newspaper found that men and women spent about the same amount of time per day reading the paper. However, the same report indicated
there was nearly twice as much variation in time spent per day among the men than the women.

The *F* distribution is also used to test assumptions for some statistical tests. Recall that. in the previous chapter when small samples were assumed, we used the t test to investigate whether the means of two independent populations differed. To employ that test, we assume that the variances of two normal populations are the same. The *F* distribution provides a means for conducting a test regarding the variances of two normal populations.

Regardless of whether we want to determine if one population has more variation than another population or validate an assumption for a statistical test, we first state the null hypothesis. The null hypothesis could be that the variance of one normal population, σ_1^2, equals the variance of the other normal population, σ_2^2. The alternate hypothesis is that the variances differ. In this instance the null hypothesis and the alternate hypothesis are:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_0 : \sigma_1^2 \neq \sigma_2^2$$

To conduct the test, we select a random sample of n_1 observations from one population, and a sample of n_2 observations from the second population. The test statistic is defined as follows:

$$F = \frac{s_1^2}{s_2^2}$$

Where the terms s_1^2 and s_2^2 are the respective sample variances
The test statistic follows the F distribution with $n_1 - 1$ and $n_2 - 1$degrees of freedom. In order to reduce the size of the table of critical values, the larger sample variance is placed in the numerator; hence, the tabled F ratio is always larger than 1.00. Thus, the right-tail critical value is the only one required. The critical value

of *F* for a two-tailed test is found by dividing the significance level in half $(\alpha/2)$ and then referring to the appropriate degrees of freedom.

Example 2

2.5.1 Assumption for F-test for equality of variances

1. The samples are simple random samples
2. The samples are independent of each other
3. The parent populations from which the samples are drawn are normal

N.B (1) Since the most available tables of the significant values of F are for the right-tail test, i.e against the alternative $H_o : \zeta_1^2 > \zeta_2^2$, in numerical problems we will take greater of the variances or as the numerator and adjust for the degree of freedom accordingly. Thus, in $F \sim (v_1, v_2)$, v_1 refers to the degree of freedom of the larger variance, which must be taken as the numerator while computing *F*.

If Ho is true i.e $\zeta_1^2 = \zeta_2^2 = \zeta^2$ the value of F should be around 1, otherwise, it should be greater

than 1. If the value of *F* is far greater than 1 the Ho should be rejected. Finally, if we take larger of or as the numerator, all the tests based on the F-statistic become right tailed tests.

- All one tailed tests for Ho at level of significance –∝‖ will be right tailed tests only with area –∝‖ in the right.
- For two-tailed tests, the critical value is located in the right tail of F-distribution with area ($\alpha/2$) in the right tail.

Formula for F distribution

2.5.2 Solved Examples

Example 1: The time taken (in minutes) by drivers to drive from Town A to Town B driving two different types of cars X and Y is given below

Car Type X:	20	16	26	27	23	22	
Car Type Y:	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population from which the samples are drawn do not differ significantly?

Solution:

X	***d = x – 22***	d^2	***Y***	***d = y -35***	D^2
20	-2	4	27	-8	64
16	-6	36	33	-2	4
26	4	16	42	7	49
25	5	9	35	0	0
23	1	1	32	-3	9
22	0	0	34	-1	1
			38	3	9
Total	**2**	d^2 **= 82**		**-4**	ΣD^2 **=136**

The formula is:

$$F = \frac{s_1^2}{s_2^2} = \frac{82}{136}$$
$$= 0.603$$

Tabulated $F_{0.05(6,5)} = 4.95$

Since the calculated F is less than tabulated F, it is not significant. Hence Ho may be accepted at 5% level of significance or risk level. We may therefore conclude that variability of the time distribution in the two populations is same.

Example 2: Lammers Limos offers limousine service from the city hall in Toledo, Ohio, to Metro Airport in Detroit. Sean Lammers, president of the company, is considering two routes. One is via U.S. 25 and the other via Interstate-75. He wants to study the time it takes to drive to the airport using each route and then compare the results. He collected the following sample data, which is reported in minutes. Using the .10 significance level, is there a difference in the variation in the driving times for the two routes?

US Route 25	**Interstate 75**
52	59
67	60
56	61
45	51
70	56
54	63
64	57
	65

U.S. Route 25

$$\bar{X} = \frac{\sum X}{n}$$

$$X = \frac{408}{7} = \ 58.29$$

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

$$s = \sqrt{\frac{485.43}{7-1}}$$

$$s = \sqrt{\frac{485.43}{6}}$$

$$s = 8.9947$$

Interstate 75

$$\bar{X} = \frac{\sum X}{n}$$

$$\bar{X} = \frac{472}{8} = 59.00$$

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

$$s = \sqrt{\frac{134}{8 - 1}}$$

$$s = \sqrt{\frac{134}{7}}$$

$$s = 4.3753$$

There is more variation, as measured by the standard deviation, in the U.S. 25 route than in the 1-75 route. This is somewhat consistent with his knowledge of the two routes; the U.S. 25 route contains more stoplights, whereas 1-75 is a limited-access interstate highway. However, the 1-75 route is several miles longer. It is important that the service offered be both timely and consistent, so he decides to conduct a statistical test to determine whether there really is a difference in the variation of the two routes.

The usual five-step hypothesis-testing procedure will be employed.

Step 1: We begin by stating the null hypothesis and the alternate hypothesis. The test is two-tailed because we are looking for a difference in the variation of the two routes. We are *not* trying to show that one route has more variation than the other.

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_0 : \sigma_1^2 \neq \sigma_2^2$$

Step 2: We selected the .10 significance level.

Step 3: The appropriate test statistic follows the F distribution.

Step 4: The critical value is reproduced as Table 2-1. Because we are conducting a two-tailed test, the tabled significance level is .05, found by $(\alpha/2)$ = .10/2 = .05. There are $n_1 - 1 = 7 - 1 = 6$ degrees of freedom in the numerator, and $n_1 - 1 = 8 - 1 = 7$ degrees of freedom in the denominator. To find the critical value, move horizontally across the top portion of the F table (Table 2-1 for the .05 significance level to 6 degrees of freedom in the numerator. Then move down that column to the critical value opposite 7 degrees of freedom in the denominator. The critical value is 3.87.

Thus, the decision rule is: Reject the null hypothesis if the ratio of the sample variances exceeds 3.87.

TABLE 2-1 Critical Values of the F Distribution, $\alpha = 0.05$

Degree of freedom for denominator	Degree of freedom for numerator			
	5	6	7	8
1	230	234	237	239
2	19.3	19.3	19.4	19.4
3	9.01	8.94	8,89	8.85
4	6.26	6.16	6.09	6.04
5	5.05	4.95	4.88	4.82
6	4.39	4.28	4.21	4.15
7	3.97	3.97	3.79	3.73
8	3.69	3.58	3.50	3.44
9	3.48	3.37	3.29	3.23
10	3.33	3.22	3.14	3.07

Step 5: The final step is to take the ratio of the two sample variances, determine the value of the test statistic, and make a decision regarding the null hypothesis. Note that formula (12-1) refers to the sample *variances* but we calculated the sample *standard deviations.* We need to square the standard deviations to determine the variances.

$$F = \frac{s_1^2}{s_2^2}$$

$$F = \frac{(8.9947)^2}{(4.3753)^2}$$

$$F = 4.23$$

The decision is to reject the null hypothesis, because the computed F value (4.23) is larger than the critical value (3.87). We conclude that there is a difference in the variation of the travel times along the two routes.

Self-Assessment Exercise 1 (SAE 1)

1. State the characteristics
2. NOUN logistics offers car service from Yaba to MMA Airport. The company, is considering two routes. One is route A and the other is route B. the company wants to study the time it takes to drive to the airport using each route and then compare the results. It collected the following sample data, which is reported in minutes. Using the .10 significance level, is there a difference in the variation in the driving times for the two routes?

Route A	52	67	56	45	70	54	64	
Route B	59	60	61	51	56	63	57	65

2.6. SUMMARY

In summary learners would have learnt the theories and application of the F-test. Such knowledge would definitely enhance learners‘ ability to solve more challenging statistical problems related to F-test.

2.7 REFERENCE/ FURTHER READING/WEB RESOURCES

Spiegel, M. R. and Stephens L.J. (2008).*Statistics*. (4th ed.). New York: McGraw Hill Press

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance*. London: Macmillan.

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

2.8. Possible Answers to Self-Assessment Exercise(s) within the content

SAE 1

1. (i)There is a "family" of *F* distributions
   (ii) The F distribution is continuous
   (iii) The F distribution cannot be negative
   (iv) It is positively skewed
   (v) It is asymptotic

2. $F = 4.23$

UNIT 3: CHI-SQUARE TEST

UNIT STRUCTURE

3.1 Introduction
3.2 Learning Outcomes
3.3 Chi-Square

3.3.1 Characteristics of Chi-square
3.3.2 Application of Chi-Square Distribution
3.3.3 Observed and Theoretical Frequencies
3.4 Chi-squared test of goodness of fit
3.4.1 Steps for computing χ^2 and drawing conclusions
3.5 Chi-Square test for independence of attributes
3.5.1 Solved Examples
3.6 Summary
3.7 References/Further Reading/ Web Resources
3.8 Possible Answers to Self-Assessment Exercise(s) within the content

3.1 INTRODUCTION

A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.

A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.

3.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- List the characteristics of the chi-square distribution.
- Conduct a test of hypothesis comparing an observed set of frequencies to an expected distribution
- Compute a Chi-square test

3.3 Chi-Square

Chi-square is a measure of discrepancy existing between the observed and expected frequencies is supplied by the statistics χ^2 given by:

$$\chi^2 = \frac{(f_0 - f_e)^2}{f_e}$$

3.3.1 Characteristics of Chi-square

Chi-square values are never negative
There is a family of chi-square distributions.
The chi-square distribution is positively skewed

3.3.2 Applications of the χ^2-Distribution

Chi-square distribution has a number of applications, some of which are enumerated below: (i)

Chi-square test of goodness of fit.

(ii) χ^2-test for independence of attributes

(iii) To test if the population has a specified value of variance ζ^2. (iv) To test the equality of several population proportions

3.3.3 Observed and Theoretical Frequencies

Suppose that in a particular sample a set of possible events $E_1, E_2, E_3, \ldots\ldots\ldots\ldots E_k$ are observed to occur with frequencies $O_1, O_2, O_3, \ldots\ldots\ldots O_k$, called observed frequencies, and that according to probability rules they are expected to occur with frequencies $e_1, e_2, e_3, \ldots\ldots e_k$, called expected or theoretical frequencies. Often we wish to know whether the observed frequencies differ significantly from expected frequencies.

3.4 Chi-Square test of goodness of fit

The chi-square test can be used to determine how well theoretical distributions 9such as the normal and binomial distributions) fit empirical distributions (i.e. those obtained from sample data). Suppose we are given a set of observed frequencies obtained under some experiment and we want to test if the experimental results support a particular hypothesis or theory. Karl Pearson in 1900, developed a test for testing the significance of the discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis. This test is known as χ^2-test of goodness of fit and is used to test if the deviation between observation (experiment) and theory may be attributed to chance (fluctuations of sampling) or if it is really due to the inadequacy of the theory to fit the observed data.

Under the null hypothesis that there is no significant difference between the observed (experimental and the theoretical or hypothetical values i.e there is good compatibility between theory and experiment.

Karl Pearson proved that the statistic

$$\chi^2 = \sum \quad \text{———}$$

$$= \text{———} \qquad \text{———} \qquad\qquad \text{———}$$

Follows χ^2-distribution with $v = n-1$, *d.f* where $O_1, O_2, \ldots\ldots\ldots\ldots O_n$ are the observed frequencies and $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ are the corresponding expected or theoretical frequencies obtained under some theory or hypothesis.

3.4.1 Steps for computing χ^2 and drawing conclusions

(i) Compute the expected frequencies $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ corresponding to the observed frequencies $O_1, O_2, \ldots\ldots\ldots\ldots O_n$ under some theory or hypothesis

(ii) Compute the deviations *(O-E)* for each frequency and then square them to obtain $(O- E)^2$.

(iii) Divide the square of the deviations $(O-E)^2$ by the corresponding expected frequency to obtain $(O-E)^2/E$.

(iv) Add values obtained in step (iii) to compute $\chi^2 = \sum$ ———

(v) Under the null hypothesis that the theory fits the data well, the statistic follows χ^2-distribution with v = n-1 *d.f*

(vi) Look for the tabulated (critical) values of χ^2 for (n-1) d.f at certain level of significance, usually 5% or 1%, from any Chi-square distribution table.

If calculated value of χ^2 obtained in step (iv) is less than the corresponding tabulated value obtained in step (vi), then it is said to be non-significant at the required level of significance.

This implies that the discrepancy between observed values (experiment) and the expected values (theory) may be attributed to chance, i.e fluctuations of sampling. In other words, data do not provide us any evidence against the null hypothesis [given in step (v)] which may, therefore, be accepted at the required level of significance and we may conclude that there is good correspondence (fit) between theory and experiment.

(vii) On the other hand, if calculated value of χ^2 is greater than the tabulated value, it is said to be significant. In other words, discrepancy between observed and expected frequencies cannot be attributed to chance and we reject the null hypothesis. Thus, we conclude that the experiment does not support the theory.

3.5. Chi-Square test for independence of attributes

Consider a given population consisting of *N* items divided into *r* mutually disjoint (exclusive) and exhaustive classes A_1, A_2, ...*Ar* with respect to (*w.r.t*) the attribute A, so that randomly selected item belongs to one and only one of the attributes A_1, A_2, ...*Ar*. Similarly, let us suppose that the same population is divided into *s* mutually disjoint and exhaustive classes B_1, B_2, ...B_s*w.r.t* another attribute B_s so that an item selected at random possesses one and only one of the attributes B_1, B_2, ...B_s can be represented in the following *r x s* manifold contingency e.g like below:

B	B_1	B_2	.....	B_j	.....	B_s	*Total*
A_1	$(A_1 B_1)$	$(A_1 B_2)$		(A_1B_j)	.....	(A_1B_s)	(A_1)
A_2	$(A_2 B_1)$	$(A_2 B_2)$	.....	(A_2B_j)	.....	(A_2B_s)	(A_2)
:	:	:		.....	.....	:	:
A_i	$(A_i B_1)$	$(A_i B_2)$	.....	(A_iB_j)	.....	(A_iB_s)	(A_i)
:	:	:		.....	.....	:	:
A_r	$(A_r B_1)$	$(A_r B_2)$	.....	A_rB_j	.....	(A_rB_s)	(A_r)
Total	(B_1)	(B_2)	.....	(B_j)	.....	(B_s)	$\sum$ $\sum$

Where (A_i) is the frequency of the *ith* attribute A_i,i.e, it is, number of persons possessing the attribute A_i , *i*=1,2, ...*r*; (B_j) is the number of persons possessing the attribute B_j, j=1,2.....s; and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j ; (*i*: 1, 2, ...r; j: 1, 2, ..., s)

Under the hypothesis that the two attributes A and B are independent, the expected frequency for (A_i, B_i) is given by

$$E[(A_iB_j)] = N.P[A_iB_j] = N.P[A_i \cap B_j] = N.P[A_i].P[B_j]$$

[By compound probability theorem, since attributes are independent]

$$= N \text{ —} \quad \text{—} \quad \text{———}$$

If $(A_iB_j)_O$ denotes the expected frequency of (A_iB_j) then

$(A_iB_j)_O$ = ——— ; *(i = 1, 2, ...,r; j=1,2, ...s)*

Thus, under the null hypothesis of independence of attributes, the expected frequencies for each of the cell frequencies of the above table can be obtained on using this last equation. The rule in the last can be stated in the words as follows:

–Under the hypothesis of independence of attributes the expected frequency for any of the cell frequencies can be obtained by multiplying the row totals and the column totals in which the frequency occurs and dividing the product by the total frequency N".

Here, we have a set of *r* x *s* observed frequencies (A_iB_j) and the corresponding expected frequencies $(A_iB_j)_O$. Applying χ^2–test of goodness of fit , the statistic

$$\chi^2 = \Sigma\,\Sigma\,[\frac{[\qquad\qquad]}{\qquad\qquad}]$$

follows χ^2-distribution with (r-1)X(s-1) degrees of freedom.

Comparing this calculated value of χ^2 with the tabulated value for (r-1)X(s-1) *d.f* and at certain level of significance, we reject or retain the null hypothesis of independence of attributes at that level of significance.

Note: For the contingency table data, the null hypothesis is always set up that the attributes under consideration are independent. It is only under this hypothesis that formula $(A_iB_j)_O$ = ———; *(i = 1, 2, ...,r; j=1,2, ...s)* can be used for computing expected frequencies.

3.5.1 Solved Examples

Example 1: A pair of dice is rolled 500 times with the sums in the table below

Sum (x)	Observed Frequency
2	15
3	35
4	49
5	58
6	65
7	76
8	72
9	60
10	35
11	29
12	6

Take α = 5%

It should be noted that the expected sums if the dice are fair, are determined from the distribution of *x* as in the table below:

Sum (x)	P(x)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

To obtain the expected frequencies, the *P(x)* is multiplied by the total number of trials

Sum (x)	Observed Frequency (0)	P(x)	Expected Frequency (P(x).500)
2	15	1/36	13.9
3	35	2/36	27.8
4	49	3/36	41.7
5	58	4/36	55.6
6	65	5/36	69.5
7	76	6/36	83.4
8	72	5/36	69.5
9	60	4/36	55.6
10	35	3/36	41.7
11	29	2/36	27.8
12	6	1/36	13.9

Recall that $\chi^2_i = (O_i - E_i)^2/E_i$

Therefore $\chi_1^2 = (O_1 - E_1)^2/E_1 = (15 - 13.9)^2/13.9 = 0.09$ $\chi_2^2 = (O_2 - E_2)^2/E_2 = (35 - 27.8)^2/27.8 = 1.86$ $\chi_3^2 = (O_3 - E_3)^2/E_3 = (49 - 41.7)^2/41.7 = 1.28$ $\chi_4^2 = (O_4 - E_4)^2/E_4 = (58 - 55.6)^2/55.6 = 0.10$ $\chi_5^2 = (O_5 - E_5)^2/E_5 = (65 - 69.5)^2/69.5 = 0.29$

$\chi_6^2 = (O_6 - E_6)^2/E_6 = (76 - 83.4)^2/83.4 = 0.66$

$\chi_7^2 = (O_7 - E_7)^2/E_7 = (72 - 69.5)^2/69.5 = 0.09$ $\chi_8^2 = (O_8 - E_8)^2/E_8 = (60 - 55.6)^2/55.6 = 0.35$ $\chi_9^2 = (O_9 - E_9)^2/E_9 = (35 - 41.7)^2/41.7 = 1.08$

$\chi_{10}^2 = (O_{10} - E_{10})^2/E_{10} = (29 - 27.8)^2/27.8 = 0.05$

$\chi_{11}^2 = (O_{11} - E_{11})^2/E_{11} = (6 - 13.9)^2/13.9 = 4.49$

To calculate the overall Chi-squared value, recall that $\chi^2 = \sum$ i,e we add the individual χ^2 value.

Therefore, $\chi^2 = 0.09 + 1.86 + 1.28 + 0.10 + 0.29 + 0.66 + 0.09 + 0.35 + 1.08 + 0.05 + 4.49$

$\chi^2 = 10.34$

For the critical value, since n=11, $d.f$ = 10

Therefore, table value = 18.3

Decision: since the calculated value which is 10.34 is less than table (critical) value the null hypothesis is accepted.

Conclusion: There is no significant difference between observed and expected frequencies. The slight observed differences occurred due to chance.

Example 2: A movie producer is bringing out a new movie. In order to map out her advertising, she wants to determine whether the movie will appeal most to a particular age group or whether it will appeal equally to all age groups. The producer takes a random sample from persons attending a pre-reviewing show of the new movie and obtained the result in the table below. Use Chi-square (χ^2) test to arrive at the conclusion (α=0.05).

	Age-groups (in years)				
Persons	***Under 20***	***20-39***	***40 – 59***	***60 and over***	***Total***
Liked the movie	320	80	110	200	**710**
Disliked the movie	50	15	70	60	**195**
Indifferent	30	5	20	40	**95**
Total	**400**	**100**	**200**	**300**	**1,000**

Solution:
It should be noted that the two attributes being considered here are the age groups of the people and their level of likeness of the new movie. Our concern here is to determine whether the two attributes are independent or not.

Null hypothesis (H_0): Likeness of the of the movie is independent of age group (i.e. the movie appeals the same way to different age group)

Alternative hypothesis (H_a): Likeness of the of the movie depends on age group (i.e. the movie appeals differently across age group)

As earlier explained, to calculate the expected value in the cell of row 1 column 1, we divide the product of row 1 total and column 1 total by the grand total (N) *i.e.*

$E_{ij} = (A_iB_j)/N$

Therefore, $E_{11} = E_{12} = E_{13} = E_{14} = E_{21} = E_{22} = E_{23} = E_{24} = E_{31} = E_{32} = E_{33} = E_{34} =$

We can get a table of expected values from the above computations

Table of expected values

	Under 20	***20-39***	***40-59***	***60 & above***
Like	284	71	142	213
Dislike	78	19.5	39	58.5
Indifferent	38	9.5	19	28.5

χ^2 value = $\sum\sum$ [ ] = $\chi^2 = \sum\sum$ [ ]

where O_{ij} are the observed frequencies while the E_{ij} are the expected values.

$\chi^2_{calculated} = \sum \sum$ ~~=4.56+~~1.14+7.12+0.79+10.05+1.04+24.64+0.04+1.68+2.13+0.05

+4.64 = **57.97**

Recall, that the *d.f* is (number of row minus one) X (number of column minus one)

$\chi^2_{(r-1)(s-1)} = 12.59$ (critical value)

Decision: Since the calculated χ^2 value is greater than the table (critical value) we shall reject the null hypothesis and accept the alternative.

Conclusion: It can be concluded that the movie appealed differently to different age groups (i.e likeness of the movie is dependent on age).

Self-Assessment Exercises 1 (SAE 1)

1. State the characteristics of Chi-square

2. The sample below is from the marketing manager for a manufacturer of sports cards. The number of cards sold for each player is shown below:

Players	TL	TM	TN	TO	TP	TQ
Card sold	13	33	14	7	36	17

(a) Can the manager conclude the sales are not the same for each player?
(b) What is the decision regarding the null hypothesis?

3.6 SUMMARY

In this unit, we have examined the concept of chi-square and its scope. We also look at its methodology and applications. It has been emphasized that it is not just an ordinary statistical exercise but a practical tool for solving day-to-day business and economic problems.

3.7 REFERENCES/FURTHER READING/WEB RESOURCES

Spiegel, M. R. and Stephens L. J. (2008).*Statistics*. (4th ed.). New York: McGraw Hill Press. Gupta S.C. (2011). *Fundamentals of Statistics*. (6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House.

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance*. London: Macmillan.
Lind, A. D., Marchal, W. G. & Wathen, S. A. (2006) *Basic Statistics for Business and Economics* (5th ed.). New York: McGraw-Hill.
Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.
Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.
McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.
https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test

3.8 Possible Answers to Self-Assessment Exercise(s) within the content

SAE 1

1. (i) Chi-square values are never negative
   (ii)There is a family of chi-square distributions.
   (iii)The chi-square distribution is positively skewed

2.(a) $\chi^2 = \frac{(f_0 - f_e)^2}{f_e}$= 34.40

(b) The computed χ^2 of 34.40 is in the rejection region beyond the critical value of 11.070. The decision, therefore,is to reject H_0 at the .05 level and to accept H_1• The difference between the observed and the expected frequencies is not due to chance. Rather, the differences between f_0 and f_e are large enough to be considered significant. The chance these differences are due to sampling error is very small. So we conclude that it is unlikely that card sales are the same among the six players.

UNIT 4: ANALYSIS OF VARIANCE (ANOVA)

UNIT STRUCTURE

4.1 Introduction

4.2 Learning Outcomes

4.3 Analysis of Variance (ANOVA)

4.3.1 Assumption for ANOVA test

4.4 The one-way classification

4.5 Steps for testing hypothesis for more than two means (ANOVA)
4.5.1 Solved Examples

4.6 Summary

4.7 References/Further Reading/Web Resources
4.8 Possible Answers to Self-Assessment Exercise(s) within the content

4.1 INTRODUCTION

In day-to-day business management and in sciences, instances may arise where we need to compare means. If there are only two means e.g. average recharge card expenditure between male and female students in a faculty of a University, the typical t-test for the difference of two means becomes handy to solve this type of problem. However in real life situation man is always confronted with situation where we need to compare more than two means at the same time. The typical t-test for the difference of two means is not capable of handling this type of problem; otherwise, the obvious method is to compare two means at a time by using the t-test earlier treated. This process is very time consuming, since as few as 4 sample means would require ${}^{4}C_{2} =$ 6, different tests to compare 6 possible pairs of sample means. Therefore, there must be a

procedure that can compare all means simultaneously. One such procedure is the analysis of variance (ANOVA). For instance, we may be interested in the mean telephone recharge expenditures of various groups of students in the university such as student in the faculty of Science, Arts, Social Sciences, Medicine, and Engineering. We may be interested in testing if the average monthly expenditure of students in the five faculties are equal or not or whether they are drawn from the same normal population. The answer to this problem is provided by the technique of analysis of variance. It should be noted that the basic purpose of the analysis of variance is to test the homogeneity of several means.

The term Analysis of Variance was introduced by Prof. R.A Fisher in 1920s to deal with problem s in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes and (ii) chance causes

The variation due to assignable causes can be detected and measured whereas the variation due to chances is beyond the control of human and cannot be traced separately.

4.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Discuss the general idea of analysis of variance.
- Organize data into a one-way ANOVA table.
- Conduct a test of hypothesis among three or more treatment means.
- Compute total sum of square

4.3 ANOVA

4.3.1 Assumption for ANOVA test

ANOVA test is based on the test statistic *F* (or variance ratio). For the validity of the *F*-test in

ANOVA, the following assumption are made:

(i) The observations are independent.

(ii) Parent population from which observation are taken are normal. (iii) Various treatment and environmental effects are additive in nature.

ANOVA as a tool has different dimensions and complexities. ANOVA can be (a) One-way classification or (b) two-way classification. However, the one-way ANOVA will dealt with in this course material

Note

(i) ANOVA technique enables us to compare several population means simultaneously and thus results in lot of saving in terms of time and money as compared to several experiments required for comparing two populations means at a time.

(ii) The origin of the ANOVA technique lies in agricultural experiments and as such its language is loaded with such terms as treatments, blocks, plots etc. However, ANOVA technique is so versatile that it finds applications in almost all types of design of experiments in various diverse fields such as industry, education, psychology, business, economics etc.

(iii) It should be clearly understood that ANOVA technique is not designed to test equality of several population variances. Rather, its objective is to test the equality of several population means or the homogeneity of several independent sample means.

(iv) In addition to testing the homogeneity of several sample means, the ANOVA technique is now frequently applied in testing the linearity of the fitted regression line or the significance of the correlation ratio.

4.4 The one-way classification

Assuming *n* sample observations of random variable *X* are divided into *k* classes on the basis of some criterion or factor of classification. Let the *i*thclass consist of $n_i$observations and let:

X_{ij} = *jth* member of the *i*th class; {j=1,2,.....n_i; *i*= 1,2, ...*k*}

$n = n_1 + n_2 + \ldots\ldots\ldots\ldots + n_k = \sum$

The *n* sample observations can be expressed as in the table below:

Class	***Sample observation***	***Total***	***Mean***
1	$X_{11}, X_{12}, \ldots\ldots\ldots\ldots X_{1n}$	T_1	Mean X_1
2	$X_{21}, X_{22}, \ldots\ldots\ldots\ldots X_{2n}$	T_2	Mean X_2
:	**: : : :**	**:**	**:**
:	**: : : :**	**:**	**:**
I	$X_{i1}, X_{i2}, \ldots\ldots\ldots\ldots X_{in}$	$T_i = \sum$	Mean X_i
:	**: : : :**	**:**	**:**
:	**: : : :**	**:**	**:**
K	$X_{k1}, X_{k2}, \ldots\ldots\ldots\ldots X_{kn}$	T_k	Mean X_k

Such scheme of classification according to a single criterion is called one-way classification and its analysis of variance is known as one-way analysis of variance.

The total variation in the observations X_{ij} can be split into the following two components:

(i) The variation between the classes or the variation due to different bases of classification (commonly known as treatments in pure sciences, medicine and agriculture). This type of variation is due to assignable causes which can be detected and controlled by human endeavour.

(ii) The variation within the classes, i.e. the inherent variation of the random variable within the observations of a class. This type of variation is due to chance causes which are beyond the control of man.

The main objective of the analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability within the separate classes.

4.5 Steps for testing hypothesis for more than two means (ANOVA): Here, we adopt the rejection region method and the steps are as follows:

Step1: Set up the hypothesis:

Null Hypothesis: Ho: $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots\ldots\ldots = \mu_k$ i.e, all means are equal
Alternative hypothesis: H_1 : At least two means are different.

Step 2: Compute the means and standard deviations for each of the by the formular:

$= \quad \sum \qquad\qquad \sum$

Also, compute the mean of all the data observations in the k-classes by the formula:

$= \quad \sum \quad \sum \qquad \frac{\sum}{\sum}$

Step 3: Obtain the Between ClassesSum of Squares (BSS) by the formula:

BSS = (

Step 4: Obtain the Between Classes Mean Sum of Squares (MBSS)

Step 5: Obtain the Within Classes Sum of Squares (WSS) by the formula:

$\sum \sum \qquad\qquad \sum$

Step 6: Obtain the Within Classes Mean Sum of Squares (MWSS)

Step 7: Obtain the test statistic *F* or Variance Ratio (V.R)

Which follows *F*-distribution with (v_1 = k-1, v_2 = n-k)*d.f* (This implies that the degrees of freedom are two in number. The first one is the number of classes (treatment) less one, while the second *d.f* is number of observations less number of classes)

Step 8: Find the critical value of the test statistic *F* for the degree of freedom and at desired level of significance in any standard statistical table.

If computed value of test-statistic F is greater than the critical (tabulated) value, reject (H_O, otherwise H_O may be regarded as true.

Step 9: Write the conclusion in simple language.

4.5.1 Solved examples

Example 1: To test the hypothesis that the average number of days a patient is kept in the three local hospitals A, B and C is the same, a random check on the number of days that seven patients stayed in each hospital reveals the following:

Hospital A:	8	5	9	2	7	8	2
Hospital A:	4	3	8	7	7	1	5
Hospital A:	1	4	9	8	7	2	3

Test the hypothesis at 5 percent level of significance.

Solution: Let X_{1j}, X_{2j}, X_{3j} denote the number of days the *j*th patient stays in the hospitals A, B and C respectively

Calculations for various Sum of Squares

X_{1j}	X_{2j}	X_{3j}			
8	4	1	4.5796	1	14.8996
5	3	4	0.7396	4	0.7396
9	8	9	9.8596	9	17.1396
2	7	8	14.8996	4	9.8596
7	7	7	1.2996	4	4.5796
8	1	2	4.5796	16	8.1796
2	5	3	14.8996	0	3.4596
Total=ΣX_{1j} = T_1 = 41	**ΣX_{2j} = T_2= 35**	**ΣX_{3j} = T_3 = 41**	**$\sum$ =50.8572**	**$\sum$ =38**	**=58.8572**

= Σ ___ _ ;

= Σ ___ _

= Σ ___ _

_____ __

= _____

Within Sample Sum of Square: To find the variation within the sample, we compute the sum of the square of the deviations of the observations in each sample from the mean values of the respective samples (see the table above)

Sum of Squares within Samples = Σ Σ Σ

= 50.8572 + 38 + 58.8572 = 147.7144 ~ 147.71

Between Sample sum of Squares: $\sum$

To obtain the variation between samples, we compute the sum of the squares of the deviations of the various sample means from the overall (grand) mean.

= 0.3844;

= 0.0576;

= 0.1444;

Sum of square Between Samples (hospitals):

$\sum$ = (

= 7(0.3844) + 7(0.0576) + 7(0.1444)

= 2.6908 + 0.4032 + 1.0108 = 4.1048 = 4.10

Total Sum of Squares: = $\sum$ $\sum$

The total variation in the sample data is obtained on calculating the sum of the squares of the deviations of each sample observation from the grand mean, for all the samples as in the table below:

X_{1j} = –	X_{2j} = –	X_{3j} = –
8	7.6176	4	1.5376	1	17.9776
5	0.0576	3	5.0176	4	1.5376
9	14.1376	8	7.6176	9	14.1376
2	10.4976	7	3.0976	8	7.6176
7	3.0976	7	3.0976	7	3.0976
8	7.6176	1	17.9776	2	10.4976
2	10.4976	5	0.0576	3	5.0176
Total = 41	**53.5232**	**35**	**38.4032**	**34**	**59.8832**

Total sum of squares (TSS) = ∑ ∑ ∑

= 53.5232 + 38.4032 + 59.8832 = 151.81

Note: Sum of Squares Within Samples + S.S Between Samples = 147.71 + 4.10 =151.81

= Total Sum of Squares

Ordinarily, there is no need to find the sum of squares within the samples (i.e, the error sum of squares), the calculations of which are quite tedious and time consuming. In practice, we find the total sum of squares and between samples sum of squares which are relatively simple to calculate. Finally within samples sum of squares is obtained by subtracting Between Samples Sum of Squares from the Total Sum of Squares:

W.S.S.S = T.S.S – B.S.S.S

Therefore, Within Sample (Error) Sum of Square = 151.8096 – 4.1048 = 147.7044

Degrees of freedom for:

Between classes (hospitals) Sum of Squares = k-1 = 3-1=2

Total Sum of Squares = n-1 = 21-1 = 20

Within Classes (or Error) Sum of Squares = n-k = 21 – 3= 18

ANOVA TABLE

Sources of variation **(1)**	***d.f*** **(2)**	**Sum of Squares (S.S)** **(3)**	**Mean Sum of Squares** — **(4) =**	**Variance Ratio (F)**
Between Samples (Hospitals)	3-1 =2	4.10	———	———
Within Sample (Error)	20-2=18	147.71	———	
Total	21-1=20	151.81		

Critical Value: The tabulated (critical) value of *F* for *d.f* (v_1=2, v_2=18) *d.f* at 5% level of significance is 3.55

Since the calculated *F* = 0.25 is less than the critical value 3.55, it is not significant. Hence we fail to accept Ho.

However, in cases like this when MSS between classes is less than the MSS within classes, we need not calculate *F* and we may conclude that the means,, and do not differ significantly. Hence, H_O may be regarded as true.

Conclusion: H_O : $\mu_1 = \mu_2 = \mu_3$, may be regarded as true and we may conclude that there is no significant difference in the average stay at each of the three hospitals.

Critical Difference: If the classes (called treatments in pure sciences) show significant effect then we would be interested to find out which pair(s) of treatment differs significantly. Instead of calculating Student's *t* for different pairs of classes (treatments) means, we calculate the Least Significant Difference (LSD) at the given level of significance. This LSD is also known as Critical Difference (CD).

The LSD between any two classes (treatments) means, say and at level of significance ‗α' is given by:
LSD (-) = [The critical value of t at level of significance α and error d.f] X [S.E (-)]

Note: S.E means Standard Error. Therefore, the S.E (-) above mean the standard error of the difference between the two means being considered.

$= t_{n-k}\ (\alpha/2)\ X \sqrt{\quad\quad}$

MSSE means sum of squares due to Error
If the difference | | between any two classes (treatments) means is greater than the LSD or CD, it is said to be significant.

Another Method for the computation of various sums of squares
Step 1: Compute: $G = \sum \sum$
Step 2: Compute Correction Factor (CF) = , where n = $n_1+n_2=.....n_k$, is the total number of observations.
Step 3: Compute Raw Sum of Square (RSS) = $\sum \sum$

= Sum of squares of all observations
Step 4: Total Sum of Square = $\sum \sum$

Step 5: Compute $\sum$

Step 6: Between Classes (or Treatment) Sum of Squares = $\sum$

=
Step 7: Within Classes or Error Sum of Squares = Total S.S – Between Classes S.S The calculations here are much simpler and shorter than in the first method
Application: Let us now apply this alternative method to solve the same problem treated earlier. n= Total number of observation = 7 + 7 + 7 = 21

Grand Total (G) = $\sum \sum$

Correction Factor = (CF) = — ——

Raw Sum of Square (RSS) = $\sum \sum$

$= (8^2 + 5^2 + 9^2 + 2^2 + 7^2 + 8^2 + 2^2) + (4^2 + 3^2 + 8^2 + 7^2 + 7^2 + 1^2 + 5^2)$

$+ (1^2 + 4^2 + 9^2 + 8^2 + 7^2 + 2^2 + 3^2)$

$= 291 + 213 + 224 = 728$

Total Sum of Square (TSS) = RSS – CF = 728 – 576.1905 = 151.8095

Between Classes (hospitals) Sum of Squares = — — —

But $\sum$ $\sum$ $\sum$

Therefore, BCSS = —— —— ——

= ————————

Therefore, Within Classes (hospitals) Sum of Squares or Error S.S = TSS – BCSS

= 151.8095 – 4.0957 = 147.7138

Having arrived at the same Sums of Squares figures, computations can proceed as done earlier.

Example 2: The table below gives the retail prices of a commodity in some shops selected at random in four cities of Lagos, Calabar, Kano and Abuja. Carry out the Analysis of Variance (ANOVA) to test the significance of the differences between the mean prices of the commodity in the four cities.

City	***Price per unit of the commodity in different shops***			
Lagos	9	7	10	8
Calabar	5	4	5	6
Kano	10	8	9	9
Abuja	7	8	9	8

If significant difference is established, calculate the Least Significant Difference (LSD) and use it to compare all the possible combinations of two means (α=0.05).

Solution:

Using the alternative method of obtaining the sum of square

City	*Price per unit of the commodity in different shops*				*Total*	*Means*
Lagos	9	7	10	8	34	8.5
Calabar	5	4	5	6	20	5
Kano	10	8	9	9	36	9
Abuja	7	8	9	8	32	8

Grand Total (G) = $\sum \sum$ = (9+7+10+8) + (5+4+5+6) + (10+8+9+9) + (7+8+9+8)

= 34 + 20 + 36 +32=
122

Correction Factor (CF) =

=

= 930.25

Raw Sum of Square (RSS) = $\sum \sum$

= $(9^2 + 7^2 + 10^2 + 8^2) + (5^2 + 4^2 + 5^2 + 6^2) + (10^2 + 8^2 + 9^2 + 9^2) +$
$(7^2 + 8^2 + 9^2 + 8^2)$
= 294 + 102 + 326 + 258

RSS = 980

Total Sum of Square (TSS) = RSS – CF

= 980 – 930.5

TSS= 49.75

Between Classes (cities) Sum of Squares =

BCSS = 289 + 100 + 324 + 256 – 930.25

= 969 – 930.25

BCSS = 38.75

Within Class (cities) or Error Sum of Squares = TSS – BCSS

TSS –BCSS

= 49.75 – 38.75, **WSS**= 11

Between Class Mean Sum of Square Error = ; where k is the number of classes

= 12.92

Within Class Mean Sum of Square Error (WCMSSE) =

= 0.92

Variance Ratio (*F* calculated) =

***F* calculated** =

***F* calculated** = 14.04

F-table (critical value) = $F_{(v1,\ v2,\ \alpha)}$ = $F_{(3,\ 12,\ 0.05)}$ = 3.49

Decision: Sincethe computed *F* is greater than the table value $F_{(v1,\ v2,\ \alpha)}$, the null hypothesis is rejected and the alternative is accepted.

Conclusion: At least one of the means is significantly different from others.

LSD = $t_{n-k(\alpha/2)}$. *S.E*

But the standard error of $= \sqrt{\quad}$

Therefore, **LSD** = 2.18 X $\sqrt{\quad}$

= 2.18 X $\sqrt{\quad}$

= 2.18 X 0.678

LSD = 1.48

Comparison between different means

Cities	Absolute Difference	Comparison	Conclusion
Lagos and Calabar	| |	> LSD	Significant
Lagos and Kano	| |	< LSD	Not Significant
Lagos and Abuja	| |	< LSD	Not Significant
Calabar and Kano	| |	> LSD	Significant
Calabar and Abuja	| |	> LSD	Significant
Kano and Abuja	| |	< LSD	Not Significant

Self-Assessment Exercise (SAE 1)

1. Professor Ade had the 22 students in his 10 A.M. Introduction to Statistics rate his performance as Excellent, Good, Fair, or Poor. A graduate student collected the ratings and assured the students that Professor Ade would not receive them until after course grades had been sent to the Deans' office. The rating (i.e., the treatment) a student gave the professor was matched with his or her course grade, which could range from 0 to 100. The sample information is reported below.

Course grades			
Excellent	Good	Fair	Poor
94	75	70	68
90	68	73	70
85	77	76	72
80	83	78	65
	88	80	74
		68	65
		65	

(a) Is there a difference in the mean score of the students in each of the four rating categories? Use the .01 significance level.

(b) Interpret your answer

2. MAH Clean is a new all-purpose cleaner being test marketed by placing displays in three different locations within various supermarkets. The number of 12-ounce bottles sold from each location within the supermarket is reported below.

MZ	18	14	19	17
MM	12	18	10	16
Other cleaners	26	28	30	32

At the .05 significance level, is there a difference in the mean number of bottles sold at the three locations?

(a) State the null hypothesis and the alternate hypothesis.

(b) Compute the values of SS total, SST, and SSE.

4.6 SUMMARY

In summary, ANOVA is very useful in the multiple comparison of mean among other important uses in both social and applied sciences.

4.7 REFERENCES/FURTHER READINGS/WEB RESOURCES

Spiegel, M. R. and Stephens L. J. (2008).*Statistics*. (4th ed.). New York: McGraw Hill Press. Gupta S.C.

(2011). *Fundamentals of Statistics.* (6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House.

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance*. London: Macmillan
Lind, A. D., Marchal, W. G. & Wathen, S. A. (2006) *Basic Statistics for Business and Economics* (5th ed.). New York: McGraw-Hill.
Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.
Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.
McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

4.8 Possible Answers to Self-Assessment Exercise(s) within the content

1.

Source of variation	Sum of square	Degree of freedom	Mean square	F
Treatment	890.68	3	296.89	8.99
Error	594.41	18	33.02	
Total	1485.09	21		

(b) The computed value of *F* is 8.99, which is greater than the critical value of 5.09, so the null hypothesis is rejected. We conclude the population means are not all equal. The mean scores are not the same in each of the four ratings groups. It is likely that the grades students earned in the course are related to the opinion they have of the overall competency and classroom performance of Prof Ade, the instructor.

(a)

$$H_0 = \mu_1 = \mu_2 = \mu_3$$

$H_1 =$ At least one treatment mean is different

(b)

SS Total = 578

SSE = 74

SST = 504

c.

Source of variation	Sum of square	Degree of freedom	Mean square	F
Treatment	504	2	252	30.65
Error	74	9	8.22	
Total	578	11		

(d)

H_0 is rejected. There is a difference in the mean number of bottles sold at the various locations.

UNIT 5: N ON-PARAMETRIC TEST METHODS

UNIT STRUCTURE

5.1 Introduction

5.2 Learning Outcomes

5.3 Non-parametric Test Method
- 5.3.1 Meaning of Non-parametric
- 5.4 Tests used for Non-Parametric Statistics
- 5.4.1 The H-Test or the Kruskal-Wallis Test
- 5.4.2 The Sign Test
- 5.4.3 Solved Example 1
- 5.5 Test Based on Runs
- 5.5.1 Solved Example 2

5.6 Summary
5.7 References/Further Reading/Web Resources
5.8 Possible Answers to Self-Assessment Exercise(s) within the content

5.1 INTRODUCTION

In statistics, the term non-parametric statistics refers to statistics that do not assume the data or population have any characteristic structure or parameters. For example, non-parametric statistics are suitable for examining the order in which runners complete a race, while parametric statistics would be more appropriate for looking at the actual race times (which may possess parameters such as a mean and standard deviation). In other words, the order (or "rank") of the values is used rather than the actual values themselves.

5.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Explain the meaning of non-parametric statistics
- State the tests used for non-parametric statistics
- Compute the Sign test
- Compute the Kruskal-Wallis test

5.3 NON-PARAMETRIC METHODS

5.3.1 Meaning of Non-parametric Statistics

In statistics, the term non-parametric statistics has at least two different meanings:

(1) The first meaning of *non-parametric* covers techniques that do not rely on data belonging to any particular distribution. These include, among others:

(a) distribution free methods, which do not rely on assumptions that the data are drawn from a given probability distribution. As such it is the opposite of parametric statistics. It includes non-parametric descriptive statistics, statistical models, inference and statistical tests.

(b) non-parametric statistics (in the sense of a statistic over data, which is defined to be a function on a

sample that has no dependency on a parameter), whose interpretation does not depend on the population fitting any parameterised distributions. Order statistics, which are based on the ranks of observations, are one example of such statistics and these play a central role in many non-parametric approaches.

(2) The second meaning of *non-parametric* covers techniques that do not assume that the *structure* of a model is fixed. Typically, the model grows in size to accommodate the complexity of the data. In these techniques, individual variables *are* typically assumed to belong to parametric distributions, and assumptions about the types of connections among variables are also made. These techniques include, among others:

(a) *non-parametric regression*, which refers to modeling where the structure of the relationship between variables is treated non-parametrically, but where nevertheless there may be parametric assumptions about the distribution of model residuals.

(b) non-parametric hierarchical Bayesian models, such as models based on the Dirichlet process, which allow the number of latent variables to grow as necessary to fit the data, but here individual variables still follow parametric distributions and even the process controlling the rate of growth of latent variables follows a parametric distribution.

Non-parametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of non-parametric methods may be necessary when data have a ranking but no clear numerical interpretation, such as when assessing preferences. In terms of levels of measurement, non-parametric methods result in "ordinal" data.

As non-parametric methods make fewer assumptions, their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Also, due to the reliance on fewer assumptions, non- parametric methods are more robust.

Another justification for the use of non-parametric methods is simplicity. In certain cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due both to this simplicity and to their greater robustness, non-parametric methods are seen by some statisticians as leaving less room for improper use and misunderstanding.

The wider applicability and increased robustness of non-parametric tests comes at a cost: in cases where a parametric test would be appropriate, non-parametric tests have less power. In other words, a larger sample size can be required to draw conclusions with the same degree of confidence.

Non-parametric models differ from parametric models in that the model structure is not specified *a priori* but is instead determined from data. The term **non-parametric**is not meant to imply that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance.

Non-parametric (or **distribution-free**) **inferential statistical methods** are mathematical procedures for statistical hypothesis testing which, unlike parametric statistics, make no assumptions about the probability distributions of the variables being assessed.

5.4 Tests used for Non-Parametric Statistics

1. Kruskal-Wallis one-way analysis of variance by ranks: tests whether >2 independent samples are drawn from the same distribution
2. Spearman's rank correlation coefficient: measures statistical dependence between two variables using a monotonic function
3. Sign test: tests whether matched pair samples are drawn from distributions with equal medians

4. Wilcoxon signed-rank test: tests whether matched pair samples are drawn from populations with different mean ranks
5. Mann–Whitney U or Wilcoxon rank sum test: tests whether two samples are drawn from the same distribution, as compared to a given alternative hypothesis
6. Anderson–Darling test: tests whether a sample is drawn from a given distribution
7. Statistical Bootstrap Methods: estimates the accuracy/sampling distribution of a statistic

8. Cochran's Q: tests whether *k* treatments in randomized block designs with 0/1 outcomes have identical effects
9. Cohen's kappa: measures inter-rater agreement for categorical items

10. Friedman two-way analysis of variance by ranks: tests whether *k* treatments in randomized block designs have identical effects
11. Kaplan–Meier: estimates the survival function from lifetime data, modelling censoring

12. Kendall's tau: measures statistical dependence between two variables

13. Kendall's W: a measure between 0 and 1 of inter-rater agreement

14. Kolmogorov–Smirnov test: tests whether a sample is drawn from a given distribution, or whether two samples are drawn from the same distribution
15. Kuiper's test: tests whether a sample is drawn from a given distribution, sensitive to cyclic variations such as day of the week
16. Logrank Test: compares survival distributions of two right-skewed, censored samples

17. McNemar's test: tests whether, in 2 × 2 contingency tables with a dichotomous trait and matched pairs of subjects, row and column marginal frequencies are equal
18. Median test: tests whether two samples are drawn from distributions with equal medians

19. Pitman's permutation test: a statistical significance test that yields exact *p* values by examining all possible rearrangements of labels
20. Rank products: detects differentially expressed genes in replicated microarray experiments

21. Siegel–Tukey test: tests for differences in scale between two groups

22. Squared ranks test: tests equality of variances in two or more samples

23. Wald–Wolfowitz runs test: tests whether the elements of a sequence are mutually independent/random

5.4.1The H-Test or the Kruskal-Wallis Test

The nonparametric Kruskal-Wallis test is based on the analysis of independent random samples from each of k populations. This procedure can be used with either ordinal data or quantitative data and does not require the assumption that the populations have normal distributions.

This is a non-parametric test alternative to the one-way analysis of variance. The data are ranked jointly from the smallest to the highest as though they constitute on sample. Then, letting R_i be the sum of the ranks of the values in the *ith* sample, the test is based on the statistic:

$$\Sigma$$

Where $n = n_1 + n_2 + \ldots\ldots\ldots\ldots\ldots + n_k$ and k is the number of populations sampled.

The test is usually based on large-sample theory that the sampling distribution of H can be closely

approximated with a chi-square distribution with k-1 degree of freedom

5.4.2 **The Sign Test**

The **sign test** is a versatile nonparametric method for hypothesis testing that uses the binomial distribution with p =50 as the sampling distribution. It does not require an assumption about the distribution of the population.

Suppose we are interested in testing

Ho: P(+) = P (-) H1:
P(+) ≠ P(-)
We shall require the following:

(i) +'s and –'s

(ii) n = number of +'s and –'s

(iii) T number of +'s

If T ≥ n-t, we shall reject Ho

Note always that P = ½. To get *t*, we look at value close to our α *e.g.* let α = 0.05, we look at value closer to this value, the value on the left hand side of the table corresponding to this is our *t. i.e.* the value under the column of P=0.5 in the Binomial Distribution table.

5.4.3 Solved Examples 1
Example 1:

From the following information, test: Ho:
P(+) = P (-)
H1: P(+) ≠ P(-)

Number of +'s = 8

Number of –'s =1

Number of ties =1

Solution:

n = number of +'s and –'s =9
T = number of +'s = 8

We now go into the Binomial distribution table with P= ½ (*i.e.* under the column of 0.50), n =9 and we look for value close to 0.05 but not more than. In this case, what we have is 0.0195. This corresponds to 1. *i.e.* t=1.

Therefore, we reject Ho if

T ≥ n-t

T = 8, n=9, t =1

Therefore, 8 ≥ 9-1

8 = 8

Hence, we reject the null hypothesis

Example 2: The following are measurements of the households' weekly demand for water in litres: 163, 165, 160, 189, 161, 171, 158, 151, 169, 162, 163, 139, 172, 165, 148, 166, 172, 163, 187, 173.

Test the null hypothesis

$\mu = 160$ against the alternative $\mu > 160$ at $\alpha = 0.05$

Solution:

Ho: $\mu = 160$

H_1: $\mu > 160$

Critical region: Reject Ho if T $\geq$ n-t Where T = number of +'s

Computations:

Replace each value exceeding 160with a plus sign, each value less than 160 with a negative sign and discarding those actually equal to 160, we have the following:

+ + + + + - - + + + - + + - + + + + +

n = 19 (+'s and –s added) T = 15 (+'s signs)

From the Binomial table, with n=19, P= 0.5, look for the value very close to $\alpha = 0.05$, (say α_1)

Therefore, $\alpha_1 = 0.0318$, it corresponds t= 5

Therefore, n-t = 19-5 = 14

Since T > n-t

i.e. 15 > 14 we therefore reject the null hypothesis

5.5 Tests Based on Runs

A run is a succession of identical letters (or other kinds of symbol) which is preceded and followed by different letters or no letters at all. *E.g*. consider the following:

M F M MM F FF M F M F M MM F F M MM

In the above example, there are 11 runs and they are represented by u *i.e.*

u = 11

n_1 = 12 (for m's)

n_2 = 8 (for F's)

When n_1 and n_2 are small, tests of the null hypothesis of randomness are usually based on specially constructed tables in the any statistical tables. However, when n_1 and n_2 are either 10 or more, the sampling

distribution of u (the total number of Runs) can be approximated with a normal distribution. For this, we require the following results:

E(u) =

Var (u) =

Use (- ½) when u> E(u)

and (+ ½) when u < E(u)

5.5.1 Solved Examples 2

Example: Consider the following: nnnnnddddnnnnnnnnnnddnndddd n ddnn. Test the null hypothesis of randomness at α = 1%.

Solution:

H_0: arrangement is random

H_1: arrangement is not random

Critical region is -2.58 to 2.58

Where

Computation:

n_1 = 20 (for n's)

n_2 = 12 (for d's)

U = 9 (total number of runs)

E(u) =

=

= 16

E(9) < E(u) (16), hence we use + ½ Var

(u) =

= 6.77

= -2.50

Decision: Since Z = -2.50 falls between -2.58 and 2.58, then we cannot reject the null hypothesis of randomness.

Example 2: Consider the following sample observations taken from three different populations

Population I	**Population II**	**Population III**
94	85	89
88	82	67
91	79	72
74	84	76
87	61	69
97	72	
	80	

Use the H-test at $\alpha = 0.05$ to test H_o: $\mu_1 = \mu_2 = \mu_3$

Solution:

H_o: $\mu_1 = \mu_2 = \mu_3$

H_1: The three means are not equal

Rank all the observations together from the smallest to the highestas if they are from one sample.

Population I	R_1	Population II	R_{1I}	Population III	R_{1I} I
94	17	85	12	89	15
88	14	82	10	67	2
91	16	79	8	72	4.5
74	6	84	11	76	7
87	13	61	1	69	3
97	18	72	4.5		
		80	9		
	84		**55.5**		**31.5**

Therefore, $R_1 = 84$, $R_2 = 55.5$, $R_3 = 31.5$

$$\Sigma$$

$$= 61.67$$

$$\chi^2_{0.05,\ 2} = 5.991$$

Since $H > 5.991$

The null hypothesis is rejected

5.6 SUMMARY

The unit has explored the concept of non-parametric test viz: the definition, types, applications (including hypothesis setting and testing)and interpretation of various tests. It emphasized that non-parametric test as distribution free tests do not make assumptions about the probability

distributions of the variables being assessed. This also contributes to its flexibility and wide applicability.
Self-Assessment Exercises (SAE 1)

1. Explain the meaning of non-parametric method
2. State 5 tests that can be used for non-parametric statistics

5.7 REFERENCES/FURTHER READINGS/WEB RESOURCES

Gupta S.C (2011). *Fundamentals of Statistics*,(6th Rev. and Enlarged ed.). Mumbai, India:Himalayan Publishing House

FisherR. A. (1925).*Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd, p.43.

5.8 Possible Answers to SAE 1

1. Non-parametric covers techniques that do not rely on data belonging to any particular distribution.
2. (i).Kruskal-Wallis one-way analysis of variance by ranks: tests whether >2 independent samples are drawn from the same distribution

   (ii). Spearman's rank correlation coefficient: measures statistical dependence between two variables using a monotonic function

(iii). Sign test: tests whether matched pair samples are drawn from distributions with equal medians

(iv) Wilcoxon signed-rank test: tests whether matched pair samples are drawn from populations with different mean ranks

(v) Mann–Whitney U or Wilcoxon rank sum test: tests whether two samples are drawn from the same distribution, as compared to a given alternative hypothesis

MODULE 3: CORRELATION AND REGRESSION ANALYSIS

This module focuses on explaining the statistical relationship and interdependence among economic variables. The two technique considered are correlation and regression analysis. The two techniques are applied to measure strength of relationship between 2 or more economic variables and the level of significance`

Correlation provides an estimate of the relationship between two measurements, without any assumption of whether one comes before the other. For example, muscle mass and fat mass are correlated, both depends on body size. Correlation coefficients have a value between -1 and +1. A positive coefficient means that x and y values increases and decrease in the same direction. A negative correlation means that as x and y move in opposite directions, one increases as the other decreases. Coefficient of 0 means x and y are associated randomly.

The correlation measures only the degree of linear association between two variables while regression analysis is a statistical process for estimating the relationships among variables. In this module the under listed topics will be considered:

Unit 1: Pearson's Correlation Coefficient

Unit 2: Spearman's Rank Correlation Coefficient

Unit 3: Methods of Curve and Eye Fitting of Scattered Plot

Unit 4: The Least Square Regression Line

Unit 5: Forecasting in Regression

UNIT 1: PEARSON'S CORRELATION CONTENTS

1.1 Introduction

1.2 Learning Outcomes

1.3 Pearson's Correlation
1.3.1 Types of correlation
1.3.2 Reason for high correlation between two variables
1.3.3 Coeffient of rank correlation
1.3.4
1.4

1.7 References/Further Reading/ Web Resources
1.8 Possible Answers to Self-Assessment Exercise(s) within the content1.8

1.1 INTRODUCTION

For Pearson's correlation coefficient is based on pairs of measurement (x,y) and the data is entered in 2 columns, each pair in a row. The coefficients, and whether it significantly differs from null (0), are usually presented. More recently, the 95% confidence interval of the coefficient is presented, and correlation can be considered statistically significant if the 95% confidence interval does not overlap the zero (0) value. Sample size calculations or tables can be used for estimating sample size requirements or power of the results in correlation.

1.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Explain the types of correlation
- Calculate the coefficient of correlation
- Interpret the coefficient of correlation
- Explain methods of studying correlation

1.3 PEARSON CORRELATION

Correlation measures the strength of the relationship between two (or more) ratio scale variables. Correlation is a technique used by economists and forecasters.

(i) They can be used to answer such questions as

(ii) Is there a link between the money supply and the price level?

(iii) Do bigger firms produce at lower cost than smaller firms?

(iv) Does spending more on advertising increase sales?

Each of these questions is about economics or business as much as about statistics. The statistical analysis is part of a wider investigation into the problem; it cannot provide a complete answer to the problem but, used sensibly, is a vital input. Correlation technique may be applied to time-series or cross-section data.

The graphs are helpful, but it would be useful to have a simple numerical summary measure of each relationship. For this purpose, we use the **correlation coefficient** between any pair of variables. The correlation coefficient is a number which summarizes the relationship between two variables.

The different types of possible relationship between any two variables, *X* and *Y*, may be summarised as follows:

-High values of *X* tend to be associated with low values of *Y* and vice versa. This is termed **negative correlation** and appears to be the case for *B* and *G*.

-High (low) values of *X* tend to be associated with high (low) values of *Y*. This is **positive correlation** and reflects (rather weakly) the relationship between *B* and the income ratio (*IR*).

- No relationship between *X* and *Y* exists. High (low) values of *X* are associated

about equally with high and low values of *Y*. This is **zero**, or the absence of, **correlation**.

There appears to be little correlation between the birth rate and per capita GNP. It should be noted that positive correlation does not mean that high values of *X* are *always* associated with high values of *Y*, but usually they are. It is also the case that correlation only measures a *linear* relationship between the two variables. As a counter-example, consider the backward-bending labour supply curve, as suggested by economic theory (higher wages initially encourage extra work effort, but above a certain point the benefit of higher wage rates is taken in the form of more leisure). The relationship is non-linear and the measured degree of correlation between wages and hours of work is likely to be low, even though the former obviously influences the latter.

The sample correlation coefficient, *r*, is a numerical statistic which distinguishes between the types of cases shown in Figure 7.1. It has the following properties:

- It always lies between -1 and +1. This makes it relatively easy to judge the strength of an association.

-A positive value of *r* indicates positive correlation, a higher value indicating a stronger correlation between *X* and *Y* (i.e. the observations lie closer to a straight line). A value of $r = 1$ indicates perfect positive correlation and means that all the observations lie precisely on a straight line with positive slope.

-A negative value of *r* indicates negative correlation. Similar to the above, a larger negative value indicates stronger negative correlation and $r = -1$ signifies perfect negative correlation.

- A value of $r = 0$ (or close to it) indicates a lack of correlation between *X* and *Y*.

- The relationship is symmetric, i.e. the correlation between *X* and *Y* is the same as between *Y* and *X*. It does not matter which variable is labelled *Y* and which is labelled *X*.

1.3.1Types of correlation

(a) **Positive Correlation:** Situations may arise when the values of two variables deviate in the same direction i.e. if the increase in the values of one variables results, on an average, in a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable. Correlation is said to be positive. Some examples of possible positive correlations are:

- Price and supply of a commodity
- Household income and demand for luxury items
- Height and Weight
- Rainfall and Farm yield

(b) **Negative Correlation:** Correlation is said to be negative or inverse if the variables deviate in the opposite direction *i.e.*; if the increase (or decrease) in the values of one variable results, on the average, in a corresponding decrease (or increase) in the value of the other variable. Example of negative correlation are:

- Quantity demanded and price
- Tax rate and consumption demand

(c) Linear Correlation: This describes a situation where for a unit change in one variable there is a constant corresponding change in the other variable over the entire range of values. E.g.

x: 1 2 3 4 5

y:2 5 8 11 14

As seen above, for a unit change in *x*, there is a constant change (*i.e.* 3) in the corresponding value of *y*. This can be expressed as $y = 2+3x$

I general two variables are said to be linearly related if they have a relationship of the form

$y = a+bx$

(c) **Non-linear or curvilinear correlation:** This describes situations if corresponding to a unit change one variable; the other variable does not change at a constant rate but at a fluctuating rate.

1.3.2 Reasons for high correlation between two variables

(a) Mutual dependence: This is the situation when the phenomena under study inter-influence each other. Such instances are usually observed in data relating to economic and business situations.

(b) Both variables being influenced by the same external factor(s): A high degree of correlation between the two variables may be due to the effect or interaction of a third variable or a number of variables on each of these two variables.

(c) Chance: It may happen that a small randomly selected sample from a bivariate distribution may show a fairly high degree of correlation though, actually, the variables may not be correlated in the population. Such correlation may be due to chance fluctuation. For example, one may observe a high degree of correlation between the height and intelligence in a group of people. Such correlation is called spurious or non-sense correlation.

1.3.3 The coefficient of rank correlation

On occasion it is inappropriate or impossible to calculate the correlation coefficient as described above and an alternative approach is helpful. Sometimes the original data are unavailable but the ranks are. For example, schools may be ranked in terms of their exam results, but the actual pass rates are not available. Similarly, they may be ranked in terms of spending per pupil, with actual spending levels unavailable. Although the original data are missing, one can still test for an association between spending and exam success by calculating the correlation between the ranks. If extra spending improves exam performance, schools ranked higher on spending should also be ranked higher on exam success, leading to a positive correlation.

Second, even if the raw data are available, they may be highly skewed and hence the correlation coefficient may be influenced heavily by a few outliers. In this case the hypothesis test for correlation may be misleading as it is based on the assumption of underlying Normal distributions for the data. In this case we could transform the values to ranks, and calculate the correlation of the ranks. In a similar manner to the median, described in Chapter 1, this can effectively deal with heavily skewed distributions.

Note the difference between the two cases. In the first, we would prefer to have the actual school pass rates and expenditures because our analysis would be better. We could actually see how much extra we have to spend in order to get better results. In the second case we actually prefer to use the ranks because the original data might mislead us, through the presence of outliers for example. Non- parametric statistics are those which are robust to the distribution of the data, such as the calculation of the median, rather than the mean which is a parametric measure. The rank correlation coefficient is one of the parametric measures.

1.3.4 Some Methods of Studying Correlation

1. Scatter Diagram method
2. Karl Pearson's coefficient of correlation
3. Edward Spearman Rank correlation method etc.

1.3.5 Karl Pearson's Correlation Method

Karl Pearson (1867 – 1936), was a British Biometrician and statistician suggested a mathematical method for measuring the magnitude of linear relationship between two variables. Karl Pearson's method, also known as Pearsonian correlation between two variables (series) X and Y, usually denoted by $r(X,Y)$ or r_{xy} or ρ and it is given as:

$$\rho = \frac{}{\sqrt{[\ \Sigma\ \ \Sigma\ \ \Sigma\ \ \Sigma\ \ \Sigma\ ][\ \Sigma\ \ \Sigma\ ]}}$$

Alternative formula that relies on deviation of each individual observation from the mean is also frequently used where the deviation from the mean $x = X-$ and $y = Y-$. Here and are the sample means of the set of data X_i and Y_i respectively. This formula is given as

$$\frac{\Sigma}{\sqrt{\Sigma\ \ \Sigma}}$$

1.3.6 Interpretation of the value of r

Given two variables X and Y:

If r = +1, there is a perfect direct relationship between Y and X.

If r = -1, there is a perfect inverse or negative relationship between Y and X. If r = 0, there is no relationship between Y and X.

1.3.7 Factors which could limit a product-moment correlation coefficient

1. Homogenous group (the subjects are very similar on the variables)
2. Unreliable measurement instrument (your measurements can't be trusted and bounce all over the place)
3. Nonlinear relationship (Pearson's r is based on linear relationships...other formulas can be used in this case)
4. Ceiling or Floor with measurement (lots of scores clumped at the top or bottom...therefore no spread which creates a problem similar to the homogeneous group)

1.3.8 Assumptions one must meet in order to use the Pearson product-moment correlation

1. The measures are approximately normally distributed
2. The variance of the two measures is similar (**homoscedasticity**)
3. The relationship is linear
4. The sample represents the population
5 The variables are measured on an interval or ratio scale

1.3.9 Correlation and causality

It is important to test the significance of any result because almost every pair of variables will have a non-zero correlation coefficient, even if they are totally unconnected (the chance of the sample correlation coefficient being exactly zero is very, very small). Therefore, it is important to distinguish between correlation coefficients which are significant and those which are not, using the t test just outlined. But even when the result is significant one should beware of the danger

of 'spurious' correlation. Many variables which clearly cannot be related turn out to be 'significantly' correlated with each other. One now famous example is between the price level and cumulative rainfall. Since they both rise year after year, it is easy to see why they are correlated, yet it is hard to think of a plausible reason why they should be causally related to each other.

Apart from spurious correlation, there are four possible reasons for a non-zero

value of *r*:

(1) *X* influences *Y*.

(2) *Y* influences *X*.

(3) *X* and *Y* jointly influence each other.

(4) Another variable, *Z*, influences both *X* and *Y*.

Correlation alone does not allow us to distinguish between these alternatives. For example, wages (*X*) and prices (*Y*) are highly correlated. Some people believe this is due to cost–push inflation, i.e. that wage rises lead to price rises. This is case (1) above. Others believe that wages rise to keep up with the cost of living (i.e. rising prices), which is (2). Perhaps a more convincing explanation is (3), a wage– price spiral where each feeds upon the other. Others would suggest that it is the growth of the money supply, *Z*, which allows both wages and prices to rise. To
distinguish between these alternatives is important for the control of inflation, but correlation alone does not allow that distinction to be made.

Correlation is best used therefore as a suggestive and descriptive piece of analysis, rather than a technique which gives definitive answers. It is often a preparatory piece of analysis, which gives some clues to what the data might yield, to be followed by more sophisticated techniques such as regression.

1.5 Solved Examples

Example 1: Calculate Karl Pearson's correlation coefficient between expenditure on advertising and sales from the data given below:

Advertising Expenses (in '000 Naira)	39	65	62	90	82	75	25	98	36	78
Sales (in '000, 000 Naira)	47	53	58	86	62	68	60	91	51	84

Solution: Let the advertising expenses (in ₦000 Naira) be denoted by the variable *x* and the sales (in ₦000,000) be denoted by the variable *x* and the sale (in ₦000,000) be denoted by the variable *y*.

X	***Y***	***x=X-***	***y = Y-***	x^2	y^2	***Xy***
39	47	-26	-19	676	361	494
65	53	0	-13	0	169	0
62	58	-3	-8	9	64	24
90	86	25	20	625	400	500
82	62	17	-4	289	16	-68
75	68	10	2	100	4	20
25	60	-40	-6	1600	36	240
98	91	33	25	1089	625	825
36	51	-29	-15	841	225	435
78	84	13	18	169	324	234
ΣX=650	**ΣY = 660**	**Σx= 0**	**Σy= 0**	**Σx^2= 5398**	**Σy^2=2224**	**Σxy=2704**

Σ ___ Σ ___ ___

Therefore, $x = X - = = X-65;$ $y = Y -$

Using the deviation from the mean formula: $\frac{\Sigma}{\sqrt{\Sigma \quad \Sigma}}$

Example 2: The following table shows the marks obtained in Mathematics (X) and English (Y) by ten students chosen randomly from a group of final year students in a Senior Secondary School.

Mathematics (X)	**English (Y)**
75	82
80	78
93	86
65	72
87	91
71	80
98	95
68	75
84	89
77	74

Calculate the product moment correlation coefficient between the two subjects and interpret your result.

Solution: Using the direct observation data method given by the formula:

$$\rho = \frac{\Sigma \quad \Sigma \quad \Sigma}{\sqrt{[\ \Sigma \quad \Sigma\ ][\ \Sigma \quad \Sigma\ ]}}$$

X	Y	X^2	Y^2	XY
75	82	5625	6724	6150
80	78	6400	6084	6240
93	86	8649	7396	7998
65	72	4225	5184	4680
87	91	7569	8281	7917
71	80	5041	6400	5680
98	95	9604	9025	9310
68	75	4624	5625	5100
84	89	7056	7921	7476
77	74	5929	5476	5698
798	**822**	**64,722**	**68,116**	**66,249**

$$\rho = \frac{\bar{\quad}}{\sqrt{[\quad[\qquad\qquad][\qquad\qquad]}}$$

$$\rho = \frac{\quad}{\sqrt{\quad}}$$

It can be said that there is a strong positive relationship between the marks obtained in English and Mathematics by the 10 ten students.

Self-Assessment Exercise 1 (SAE 1)

1. Explain different types of correlation
2. Highlight the reasons for high correlation between variables

1.6 SUMMARY

In summary, learners should have been able to find out here that Pearson's Correlation is a measure of relationship between two (or more) variables that change together.

1.7 REFERENCES/FURTHER READING/WEB RESOURCES

Spiegel M. R. and Stephens L. J. (2008).*Statistics*. (4th ed.). New York: McGraw Hill Press.

Gupta S.C. (2011). *Fundamentals of Statistics.* (6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House.

Swift L. (1997). *Mathematics and Statistics for Business, Management and Finance*. London: Macmillan.

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

1.8 Possible Answers to Self-Assessment Exercise(s) within the content

1. (i) Positive correlation
   (ii) Negative correlation
   (iii. Linear correlation
   (iv) Non-linear correlation
2. (i). Mutual dependence
   (ii. Chance
   (iii. Both variables being influenced by the same external factor(s)

UNIT 2: SPEARMAN'S RANK CORRELATION METHOD UNIT STRUCTURE

2.1 Introduction
2.2 Learning Outcomes
2.3 Spearman's Rank Correlation Method
2.3.1 Meaning of Spearman's Rank Correlation
2.4 Solved Examples
2.5 Challenges using Spearman's Rank
2.6 Summary

2.7 References/Further Reading/Web Resources
2.8 Possible Answers to Self-Assessment Exercise(s) within the content

2.1 INTRODUCTION

In certain instances, we come across statistical series in the variables under consideration cannot be measured quantitatively but can only be arranged in serial order. This is always the situation when we are dealing with qualitative attributes such as intelligence, preference, honesty, morality etc. In such case, Karl Pearson's coefficient of correlation cannot be used. A British Psychologist Charles Edward Spearman developed a formula in 1904 which can be used to obtain the correlation coefficient between the ranks of *n* individuals in the two variables or attributes being study.

For example, assuming we are interested in determining the correlation between fluency in English Language (A) and Beauty (B) among a group young ladies numbering *n*. These are variables which cannot be measured but we can arrange the group of *n* individuals in order of merit (ranks) with respect to their proficiency in the two attributes. Let the random variables X

and Y denote the rank of the individuals in the characteristics A and B respectively. Also, if it is assumed that there is no tie, i.e no two individuals get the same rank in a characteristic, then, X and Y assume numerical values ranging from 1 to *n.*

Spearman's Rank Correlation Coefficient, usually denoted by ρ (Rho) is given by the formula:

$$\Sigma$$

Where *d* is the difference between the pair of ranks of the same individual in the two characteristics and *n* is the number of pairs.

2.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Explain the types of correlation
- Calculate the coefficient of correlation
- Interpret the coefficient of correlation
- Explain methods of studying correlation

2.3 SPEARMAN'S RANK CORRELATION METHOD

2.3.1 **Meaning of Spearman's Rank Correlation**

Spearman's correlation coefficient measures correlation when the data is non-parametric, when either x or y is not a continuous and normally distributed measurement.

2.4 Solved Examples

Example 1: Fifteen members of staff of the administrative unit of an organization were studied to determine the correlation between their punctuality at work (X) and the compliance of their dresses with organizational dress code (Y) and the following ranks as given in the table below were observed:

Rank in (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Rank in (Y)	10	7	2	6	4	8	3	1	11	15	9	5	14	12	13

Calculate the Spearman's rank correlation coefficient between the two characteristics.

Solution:

Spearman's Rank Correlation Coefficient is given by the formula:

$$\Sigma$$

Rank X	**Rank Y**	d **= X-Y**	d^2
1	10	-9	81
2	7	-5	25
3	2	1	1
4	6	-2	4
5	4	1	1
6	8	-2	4
7	3	4	16
8	1	7	49
9	11	-2	4
10	15	-5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	2	4
15	13	2	4
		$\Sigma d = 0$	$\Sigma d^2 = 272$

Example 2: Calculate the Spearman's rank correlation coefficient between advert expenditure and sales revenue recorded by some randomly selected companies in an industrial estate as given below:

Advert (₦ '000)	24	29	19	14	30	19	27	30	20	28	11
Sales (₦ '000)	37	35	16	26	45	27	28	33	16	41	21

Solution:

X(advert)	***Y (sales)***	***Rank (Rx)***	***Rank (Ry)***	***d = Rx-Ry***	d^2
24	37	6	3	3	9
29	35	3	4	-1	1
19	16	8.5	10.5	-2	4
14	26	10	8	2	4
30	45	1.5	1	0.5	0.25
19	27	8.5	7	1.5	2.25
27	28	5	6	-1	1
30	33	1.5	5	-3.5	12.25
20	16	7	10.5	-3.5	12.25
28	41	4	2	2	4
11	21	11	9	2	4
				$\Sigma d = 0$	**$\Sigma d^2 = 54$**

2.5 Some challenging cases may arise when using Spearman's rank correlation method: These include:

Case I: When ranks are not given: The Edward Spearman's rank correlation formula can be used even when dealing with variables which are measured quantitatively, i.e. when the actual data but not the ranks relating to two variables are given. In such case, we shall have to convert the data into ranks. The highest (or smallest) observation is given rank 1. The next highest (or next lowest) observation is given rank 2and so on. It does not matter in which way (ascending or descending) the ranks are assigned. However, the same approach should be followed for all the variables under consideration.

Case II: Repeated ranks: In case of attributes, if there is a tie, *i.e.* if any two or more individuals are

placed together in any classification with respect to an attributes or if in any case of variable data there are more than one items with the same value in either or both the series, then the Spearman's formula for calculating the rank correlation coefficient breaks down, since in this case the variable X (the rank of individuals in the first characteristic series) and Y (the rank of individuals in the second characteristics series) do not take the values from 1 to *n* and consequently *as* assumed in the derivation of the formula. In such instance, common
ranks are assigned to the repeated items. The common rank assigned is the arithmetic mean of the ranks which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common ran. For example, suppose an item is repeated at rank 6, then the common rank to be assigned to each of the two items is (6+7)/2 i.e. 6.5which is the average of 6 and 7, the ranks which the two observations would have assumed if they were different. Therefore, the next item will be assigned the rank 8. Meanwhile, if an item is repeated thrice at 9 for instance, then the common rank to be assigned to each of the three will be (9+10+11)/3 i.e. 10which is the arithmetic mean of the three ranks. The next rank to be assigned will be 12.

Self-Assessment Exercise (SAE 1)

1. What is Spearman's rank
2. Highlight challenges in using Spearman's rank

2.6 SUMMARY

This unit by now has been able to help learners to calculate and interpret the simple correlation between two variables and to determine whether the correlation is significant

2.7 REFERENCES/FURTHER READING/WEB RESOURCES

Armitage P. (1980) *Statistical Methods in Medical Research.*Oxford:Blackwell Scientific

Publications, ISBN 0-632-05430-1 p. 147-166.

Siegel S. and Castellan Jr. N. J. (1988).*Nonparametric Statistics for the Behavioral Sciences,*(2nd. ed.).Boston Massachusetts:McGraw Hill, Inc. ISBN 0-07-057357-3 p235-244.

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

2.8 Possible Answers to SAE 1

1. Spearman's correlation coefficient measures correlation when the data is non-parametric, when either x or y is not a continuous and normally distributed measurement.
2. (i)When ranks are not given. (ii) Repeated ranks

UNIT 3: LEAST SQUARE REGRESSION ANALYSIS

UNIT STRUCTURE

3.1 Introduction
3.2 Learning Outcomes
3.3 Least Square Regression Analysis
3.4 Meaning of Regression Analysis
3.5 Solved Examples
3.6 Summary
3.7 References/Further Reading/Web Resources
3.8 Possible Answers to Self-Assessment Exercise(s) within the content

3.1 INTRODUCTION

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as _Regression Analysis'.

The literal or dictionary meaning of the word ―regression‖ is ‗stepping back or returning to the average value‖. The term was first used by the British Biometrician Sir Francis Galton in late 19th century in connection with some studies he conducted on estimating the extent to which the stature of the sons of tall parents reverts or regresses back to the mean stature of the population.

He studied the relationship between the heights of about one thousand fathers and sons and published the results in a paper ‗*Regression towards Mediocrity in Hereditary Stature*‖.

3.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Identify the dependent and independent variables
- Calculate the least square regression line
-

3.3 REGRESSION ANALYSIS

3.4 Meaning of Regression Analysis

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which are extensively used in almost all sciences – Natural, Social and Physical. It is specially used in business and economics to study the relationship betweentwo or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Prediction or estimation is one of the major problems in almost all the spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc. are of very great importance to business professionals. Similarly, population estimates and population projections, GNP, Revenue and Expenditure etc. areindispensable for economists and efficient planning of an economy.

Regression analysis was explained by M. M. Blair as follows: ―*Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*‖

Regression analysis is a more sophisticated way of examining the relationship between two (or more) variables than is correlation. The major differences between correlation and regression are the following:

-Regression can investigate the relationships between two *or more* variables.

- A *direction* of causality is asserted, from the explanatory variable (or variables) to the dependent variable.

- The *influence* of each explanatory variable upon the dependent variable is measured.

- The *significance* of each explanatory variable's influence can be ascertained.

Thus regression permits answers to such questions as:

- Does the growth rate influence a country's birth rate?

- If the growth rate increases, by how much might a country's birth rate be expected to fall?

-Are other variables important in determining the birth rate?

3.4.1 Simple Regression: This a type of regression in which more than two variables are studied. This is always the case in our day-to-day life because more often than not, a particular phenomenon is affected by multiplicity of factors. For example, demand for a particular product is not only determined by its market price but also by prices of substitutes, income of buyers, population and taste and fashion among others.In regression analysis there are two types of variables and these are:

Dependent Variable: This is the variable whose value is influenced or is to be predicted. For example, elementary economic theory states that ―the higher the price the lower the quantity demanded‖ In this, it is clear that quantity demanded is influenced by price of the commodity. Therefore, quantity demanded of the commodity is described as the ―Dependent variable‖

Independent Variable: This is the variable which influences the value of the dependent variable or which is used for prediction. In our example involving the law of demand, price of the commodity determines or influences the quantity demanded. Therefore, price is described as the ―independent variable‖.

In regression analysis, the dependent variable is also known as *regressand, regressed or explained variable*. On the other hand, the independent variable is also known as the *regressor, predictor or explanatory variable*.

3.4.2 Line of Regression: This is the line which gives the best estimate of one variable for any given value of the other variable. Therefore, the line of regression of *y* on *x* is the line which gives the best estimates for the value of *y* for any specified value of *x*.

The term best fit is interpreted in accordance with the principle of least squares which involves minimising the sum of the squares of the residuals or the errors of the estimates i.e, the deviations between the given observed values of the variable and their corresponding estimated values as given by the line of best fit.

It should be noted that several lines can be drawn from the same set of pairs of observations plotted in the form of a scattered diagram, but the best fit line gives the best estimate of the dependent variable for any given level of independent variable.

Typical regression model is specified in form of $Y = a + bX + e.$

Meanwhile the best fit line can be given as $y = a + bx$

The term –*a*‖ represents the intercept of the model and it is the value of Y when X is equal to zero. It is represented by the formula‖

$$\frac{\Sigma \quad \Sigma \quad \Sigma \quad \Sigma}{\Sigma \quad \Sigma}$$

Furthermore, the term –*b*‖ represents the slope off the regression model and it is the amount of change in the dependent variable *Y* as a result of a unit change in the value of the independent variable *X*. It is represented by the formula:

$$\frac{\Sigma \quad \Sigma \quad \Sigma \quad \Sigma \quad \Sigma}{\Sigma \quad \Sigma}$$

ECO 253 STATISTICS FOR ECONOMIST 1

Where is the sample mean of the dependent variable Y and is the sample mean of the independent variable X. From the foregoing –*a*‖ can be obtained having obtained –*b*‖ by:

b can be obtained using deviation from mean approach where:

$x = X -$; $y = Y -$

$$\frac{\Sigma}{\Sigma}$$

3.5 Solved Examples

Example 1: Ten households were randomly selected in Abeokuta and data were collected on household monthly income and demand for beef as follows:

Income (X) in (₦ '000)	25	24	43	23	30	50	15	34	21	45
Demand for beef (Y) in Kg	10	8	12	11	13	16	5	13	7	15

Estimate the regression equation of the form Y= a + bX

Solution:

X	***Y***	X^2	***XY***	***x***	***y***	***xy***	x^2	y^2
25	10	625	250	-6	-1	6	36	1
24	8	576	192	-7	-3	21	49	9

43	12	1849	516	12	1	12	144	1
23	11	529	253	-8	0	0	64	0
30	13	900	390	-1	2	-2	1	4
50	16	2500	800	19	5	95	361	25
15	5	225	75	-16	-6	96	256	36
34	13	1156	442	3	2	6	9	4
21	7	441	147	-10	-4	40	100	16
45	15	2025	675	14	4	56	196	16
ΣX=310	**ΣY=110**	**ΣX^2=10826**	**3740**			**Σxy=330**	**1216**	**112**

$$\frac{\Sigma \quad \Sigma \qquad \Sigma \qquad \Sigma}{\Sigma \qquad \Sigma}$$

$$\frac{}{}$$

$$\Sigma \quad \Sigma$$
$$\Sigma \qquad \Sigma$$

Alternatively, having obtained estimate value for the slope parameter *b*, the intercept term *a* can be obtained using the formula:

$$\frac{}{}$$

In the above problem $\frac{}{} \qquad \Sigma\frac{}{}$

Example 2: Suppose data were collected from a sample of 10 Foodco restaurants located near NOUN campuses are shown in the table below:

Foodco restaurant	**Student population (1000s)**	**Sales per quarter (N1000s)**
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

(a) Estimate the regression equation

The estimated regression equation is:

$$y_t = b_0 + b_1 x$$

Where

y_t= denoting the observed quarterly sales for restaurant

b_0= intercept of the estimated regression line

b_1= slope of the estimated regression line

x= size of the student population (1000s)

$$\bar{x} = \frac{\sum X}{n}$$

$$\bar{x} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum Y}{n}$$

$$\bar{y} = \frac{1300}{10} = 130$$

$\bar{x}$= mean of the independent variable

$\bar{y}$= mean of the dependent variable

Foodco resta	**x**	**y**	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	-7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Total	$\sum X = 140$	$\sum Y = 1300$			2840	568

The formula for regression is:

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

$$b_1 = \frac{2840}{568}$$

$$b_1 = 5$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = 130 - 3\,(14)$$

$$b_0 = 130 - 70$$

$$b_0 = 60$$

The estimated equation is:

$$y_t = 60 + 5x$$

The slope of the estimated regression equation $(b_1 = 5)$ is positive, implying that as student population increases, sales increase. In fact, we can conclude (based on sales measured in N1000s and student population in 1000s) that an increase in the student population of 1000 is associated with an increase of N5000 in expected sales; that is, quarterly sales are expected to increase by N5 per student.

If we believe the least squares estimated regression equation adequately describes the relationship between *x* and *y*, it would seem reasonable to use the estimated regression equation
to predict the value of *y* for a given value of *x*. For example, if we wanted to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute:

$$y_t = 60 + 5(16)$$

$$y_t = 60 + 80$$

$$y_t = 140$$

Hence, we would predict quarterly sales of N140,000 for this restaurant.

Self-Assessment Exercises (SAE 1)

1. Explain the meaning of regression
2. Differentiate between dependent and independent variables

3.6 SUMMARY

With reference to the explanations and illustrations demonstrated above, learners can now apply least square method to study the nature of the relation between two variables.

3.7 REFERENCES / FURTHER READING/WEB RESOURCES

Spiegel, M. R. and Stephens L. J. (2008).*Statistics*. (4th ed.). New York: McGraw Hill Press.

Gupta S.C. (2011). *Fundamentals of Statistics.* (6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House.

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance*. London: Macmillan.

Lucey T. (2002). *Quantitative Techniques*. 6th ed. BookPower

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

Anderson, D. R., Sweeney, D. J., Camm, J. D. & Cochran, J. J. (2014). *Statistics for Business and Economics* (12th ed.).South-Western.

Barrow, M. (2017). *Statistics for Economics, Accounting and Business Studies*, (7th ed.). Pearson, United Kingdom.

McClave, J. T., Benson, P. G. & Sincich. T. (2017*) Statistics for Business and Economics*, (13th ed.), Pearson, UK.

3.8 Possible Answers to Self-Assessment Exercise(s) within the content

1. Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable.

2. Dependent variable is the variable whose value is influenced or is to be predicted while independent variable is the variable which influences the value of the dependent variable or which is used for prediction

UNIT 4: FORECASTING IN REGRESSION UNIT STRUCTURE

4.1 Introduction
4.2 Learning Outcomes
4.3 Forecasting in Regression
4.4 Model Evaluation
4.5 Solved Examples
4.6 Summary
4.7 References/Further Reading/Web Resources
4.8 Possible Answers to Self-Assessment Exercise(s) within the content

4.1 INTRODUCTION

In general, we are interested in *point forecasts*, which predict a single number in each forecast period. Needless to say, the information provided by the forecasts can be very useful. For instance, it can help not only in policy and decision making, but also in validating the model from which the forecasts are made. In the forecasting process, we are usually given the values of the independent variables and our goal is to predict the value of the dependent variable. This raises the question of whether the values of the independent variables are known with certainty. If so, then we are making an *unconditional forecast*. Otherwise, we are making a *conditional forecast*. To see the difference between the two, consider the following settings:

1. Suppose that we use the following linear regression model to describe the relationship between the demand for beef and household income Y = a + bX + u as used in the previous unit. Once we obtain estimators of the regression parameters *a; b*, we can use the resulting regression line to make forecasts. Specifically, the forecasted household demand for beef will be given by . Sometimes, the value X which is the household income may

   depend on some unpredictable factors that are not known with certainty at the time of forecast. Thus, our forecast for *Y* will be conditional on our forecast for *X*.
2. Suppose that we use the linear regression model to describe the relationship betweenthe monthly auto sales *st*and the production capacity *Ct*:
   $S_t = a + bC_{t-2} + \epsilon t:$

In other words, sales in the *t*-th month depends linearly on the production capacity of the

(*t*-2)nd month. If we are currently at the *T*-th month and we want to forecast the auto sales in the (*T* +1)-st month, then we need the production capacity of the (*T*-1)st month, which can be determined with certainty. Thus, in this case, the forecast will be unconditional.

4.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Explain unconditional forecast
- Explain conditional forecast
- State the methods of evaluating reliability of a linear regression

4.3 FORECASTING IN REGRESSION

Forecasting methods can be classified as qualitative or quantitative. Qualitative methods generally involve the use of expert judgment to develop forecasts. Such methods are appropriate when historical data on the variable being forecast are either not applicable or unavailable. Quantitative forecasting methods can be used when (1) past information about the variable being forecast is available, (2) the information can be quantified, and (3) it is reasonable to assume that the pattern of the past will continue into the future. In such cases, a forecast can be developed using a time series method or a causal method.

If the historical data are restricted to past values of the variable to be forecast, the forecasting procedure is called a *time series method* and the historical data are referred to as a time series. The objective of time series analysis is to discover a pattern in the historical data or time series and then extrapolate the pattern into the future; the forecast is based solely on past values of the variable and/or on past forecast errors.

Causal forecasting methods are based on the assumption that the variable we are forecasting has a cause-effect relationship with one or more other variables. In the discussion of regression analysis in, we showed how one or more independent variables could be used to predict the value of a single dependent variable. Looking at regression analysis as a forecasting tool, we can view the time series value that we want to forecast as the dependent variable. Hence, if we can identify a good set of related independent, or explanatory, variables, we may be able to develop an estimated regression equation for predicting or forecasting the time series. For instance, the sales for many products are influenced by advertising expenditures, so regression analysis may be used to develop an equation showing how sales and advertising are related. Once the advertising budget for the next period is determined, we could substitute this value into the equation to develop a prediction or forecast of the sales volume for that period. Note that if a time series method were used to develop the forecast, advertising expenditures would not be considered; that is, a time series method would base the forecast solely on past sales.

By treating time as the independent variable and the time series as a dependent variable, regression analysis can also be used as a time series method. To help differentiate the application of regression analysis in these two cases, we use the terms *cross-sectional regression* and *time series regression*. Thus, time series regression refers to the use of regression analysis when the independent variable is time.

4.4 Model Evaluation

The various ways to evaluate the reliability of a linear regression model include:

_ the *t* and *F*, which test the explanatory power of the independent variables;

_ the R^2 which measures the goodness of fit; and

_ the forecast confidence interval.

It should be noted that these are quite different measures of model reliability, and they need not subsume each other. For instance, a regression model can have significant *t*{statistics and a high R^2 value, and yet it still forecasts poorly. This could happen if there is a structural change in the system during the forecasting period, which occurs after the estimation period (i.e., the period during which we collect data and estimate the coefficients of the regression model). On the other hand, one may be able to obtain good forecasts from regression models that have insignificant regression coefficients or relatively low R^2 values. This could happen when there is very little variation in the dependent variable, so that although it is not well explained by the regression model, it can still be forecast easily

4.5 Solved Examples

Example 1: Let us use our example on the household demand for beef in Abeokuta which was states thus: Ten households were randomly selected in Abeokuta and data were collected on household monthly income and demand for beef as follows:

Income (X) in (₦ '000)	25	24	43	23	30	50	15	34	21	45
Demand for beef (Y) in Kg	10	8	12	11	13	16	5	13	7	15

Forecast or predict the demand for beef by households whose incomes are ₦35,000,₦40,000 and ₦45,000.

Solution: Following the regression formulas for obtaining the intercept term *a* and the slope estimate *b* discussed in the last unit one can easily obtain these values by substituting the values of income (X) into the estimated regression equation.

Recall that intercept term is given as:

$$\frac{\Sigma \quad \Sigma \quad \Sigma}{\Sigma}$$

$$\Sigma$$

$$\Sigma$$

and the slope term b is given as:

$$\frac{\Sigma \quad \Sigma}{\Sigma}$$

$$\Sigma$$

$$\Sigma$$

However, one may obtain the value of the estimate of the slope parameter *b* and use the formula:

to estimate the value of the intercept term a. For the above problem, the estimated regression equation is given as:

$= 2.6 + 0.27X$

For values of explanatory variables 35, 40 and 45 (remember that _000 was factored out of the calculations)

(i) For household income X = ₦35,000

Therefore, $= 2.6 + 0.27(35)$

(ii) For household income X = ₦40,000

Therefore, $= 2.6 + 0.27(40)$

(iii) For household income X = ₦45,000

Therefore, $= 2.6 + 0.27(45)$

Self-Assessment Exercises (SAEs 1)

1. Highlight the methods of evaluating reliability of a linear regression
2. Discuss the meaning of conditional forecasting

4.6 SUMMARY

In summary, the unit has explored the rudiments of Least Square Regression, the condition under which it could be used and the procedure for estimating the relevant parameter estimates and their interpretations

4.7 REFERENCES / FURTHER READING/WEB RESOURCES

Spiegel, M. R. and Stephens L. J. (2008).*Statistics*. (4th ed.). New York: McGraw Hill Press. Gupta

S.C. (2011). *Fundamentals of Statistics*. (6th Rev. and Enlarged ed.). Mumbai, India: Himalayan Publishing House.

Swift L. (1997).*Mathematics and Statistics for Business, Management and Finance*. London: Macmillan.

Lucey T. (2002). *Quantitative Techniques*.6th ed. Book Power.

4.8 Possible answers to SAE 1

(i) the *t* and *F*, which test the explanatory power of the independent variables;
(ii) The R^2 which measures the goodness of fit; and
(iii) the forecast confidence interval.

MODULE 4 INTRODUCTION TO THE CENTRAL LIMIT THEORY (CLT)

Central limit theorem: it states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all the samples will follow an approximate normal distribution pattern, with a variance being approximately equal to the variance of the population divided by the sample size. This Module is divided into five units. These are:

Unit 1: Central Limit Theorems for Independent Sequences Unit
2: Central Limit Theorems for dependent Processes Unit 3:
Relation to the law of large numbers
Unit 4: Extensions to the theorem of CLT and Beyond the classical framework

UNIT 1: THE CENTRAL LIMIT THEOREM FOR INDEPENDENT SEQUENCE

UNIT STRUCTURE

1.1 Introduction
1.2 Learning Outcomes
1.3 Central Limit Theorems for Independent Sequences
1.4 CLT Rules
1.5 Classical CLTs
1.6 Summary
1.7 References/Further Reading/Web Resources
1.8. Possible Answers to Self-Assessment Exercise(s) within the content

1.1 INTRODUCTION

If a random sample of N cases is drawn from a population with a mean μ and standard deviation s, then the sampling distribution of the mean has:

1. a mean equal to the population mean μ_X
2. a standard deviation (standard error) equal to the standard deviation divided by the square root of the sample size N: $\frac{}{\sqrt{}}$
3. the shape of the sampling distribution of the mean approaches normal as N increases.

1.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- State the CLT rules
- State the classical CLT
- State the various theorem of classical CLT

1.3 THE CENTRAL LIMIT THEOREM FOR INDEPENDENT SEQUENCE

1.4 CLT Rules

Let's denote the mean of the observations in a random sample of size n from a population having a mean μ and standard deviation σ. Denote the mean of the *Y* distribution by μ_Y and the standard deviation of the *Y* distribution by $\sigma_{\bar{y}}$.

Then the following rules hold:

Rule 1. $\mu_{\bar{y}} = \mu$

Rule 2. $\frac{}{\sqrt{}}$ This rule is approximately correct as long as no more than 5% of the population is included in the sample.

Rule 3. When the population distribution is normal, the sampling distribution of *Y* is also normal for any sample size *n*.

Rule 4. When *n* is sufficiently large, the sampling distribution of *Y* is well approximated by a normal curve, even when the population distribution is not itself normal.

Suppose that a sample of size n is selected from a population that has mean μ and standard deviation ζ. Let $X_1; X_2; \_ \_ \_ X_n$ be the n observations that are independent and identically distributed (i.i.d.). Define now the sample mean and the total of these n observations as follows:

$$\Sigma$$

$$\frac{\Sigma}{}$$

The central limit theorem states that the sample mean follows approximately the normal distribution with mean μ and standard deviation $\frac{}{\sqrt{}}$, where μ and ζ are the mean and standard deviation of the population from where the sample was selected. The sample size n has to be large (usually $n>30$) if the population from where the sample is taken is non-normal. If the population follows the normal distribution then the sample size n can either be small or large. The sample mean of a large random sample of random variables with mean μ and finite variance ζ^2 has approximately the normal distribution with mean μ and variance ζ^2/n. This result helps to justify the use of the normal distribution as a model for many random variables that can be thought of as being made up of many independent parts. Another version of the central limit theorem is given that applies to independent random variables that are not identically distributed.

To summarize: $\frac{}{\sqrt{}}$

To transform into z we use: $\frac{}{\sqrt{}}$

Let us suppose that Y_1, Y_2, Yn ,..., are independent and identically distributed with mean = μand

finite variance σ^2. We now prove these two theorems about the mean and variance of the sample mean.

Theorem 1

Theorem 2:

Pro

In probability theory central limit theorem states that given a certain conditions the mean of a sufficiently large number of iterates.

The CLT can tell us about the distribution of large sums of random variables even if the distribution of the random variables is almost unknown. With this result we are able to approximate how likely it is that the arithmetic mean deviates from its expected value.

Using the CLT we can verify hypotheses by making statistical decisions, because we are able to

determine the asymptotic distribution of certain test statistics.

$$\frac{}{\sqrt{\quad}} \Rightarrow$$

i.e. a centred and normalized sum of independent and identically distributed (i.i.d.) random

variables is distributed standard normally as n goes to infinity.

Example: Let X be a random variable with μ = 10 and ζ = 4. A sample of size 100 is taken from this population. Find the probability that the sample mean of these 100 observations is less than 9. We write

$$(\quad)$$

$$\frac{\sqrt{\quad}}{=}$$

(from the standard normal probabilities table).

Similarly the central limit theorem states that sum T follows approximately the normal distribution, $\sqrt{}$

from where the sample ~~was~~ selected.

, where μ and ζ are the mean and standard deviation of the population

To transform T into z we use: $\frac{}{\sqrt{}}$

1.5 Classical CLT

Let $\{X_1, ..., X_n\}$ be a random sample of size n — that is, a sequence of independent and identically distributed random variables drawn from distributions of expected values given by μ and finite variances given by ζ^2. Suppose we are interested in the sample average

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

of these random variables. By the law of large numbers, the sample averages converge in probability and almost surely to the expected value μ as $n \to \infty$. The classical central limit theorem describes the size and the distributional form of the stochastic fluctuations around the deterministic number μ during this convergence. More precisely, it states that as *n* gets larger, the distribution of the difference between the sample average S_n and its limit μ, when multiplied by the factor $\sqrt{n}$ (that is $\sqrt{n}(S_n - \mu)$), approximates the normal distribution with mean 0 and variance ζ^2. For large enough *n*, the distribution of S_n is close to the normal distribution with mean μ and variance $\zeta^2 n$. The usefulness of the theorem is that the distribution of $\sqrt{n}\ (S_n - \mu)$ approaches normality regardless of the shape of the distribution of the individual X_i's. Formally, the theorem can be stated as follows:

1. Lindeberg–Lévy CLT.

Suppose $\{X_1, X_2, ...\}$ is a sequence of independent and identically distributed (i.i.d) random variables with $E[X_i] = \mu$ and $Var[X_i] = \zeta^2 < \infty$. Then as *n* approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0, \zeta^2)$:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \xrightarrow{d} N(0,\ \sigma^2).$$

In the case $\zeta > 0$, convergence in distribution means that the cumulative distribution functions (cdf) of $\sqrt{n}(S_n - \mu)$ converge point wise to the cdf of the $N(0, \zeta^2)$ distribution: for every real number *z*,

$$\lim_{n\to\infty} \Pr[\sqrt{n}(S_n - \mu) \le z] = \Phi(z/\sigma),$$

where $\Phi(x)$ is the standard normal cdf evaluated at *x*. Note that the convergence is uniform in *z* in the sense that

$$\lim_{n\to\infty} \sup_{z\in\mathbf{R}} \left| \Pr[\sqrt{n}(S_n - \mu) \le z] - \Phi(z/\sigma) \right| = 0,$$

where sup denotes the least upper bound (or supremum) of the set.

2. Lyapunov CLT

The theorem is named after Russian mathematician Aleksandr Lyapunov. In this variant of the central limit theorem the random variables X_i have to be independent, but not necessarily identically distributed. The theorem also requires that random variables $|X_i|$ have moments of some order (2 + δ), and that the rate of growth of these moments is limited by the Lyapunov condition given below.

Suppose $\{X_1, X_2, ...\}$ is a sequence of independent random variables, each with finite expected value μ_i and variance σ^2_i . Define

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

If for some δ > 0, the *Lyapunov's condition*

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathrm{E}\left[\,|X_i - \mu_i|^{2+\delta}\right] = 0$$

is satisfied, then a sum of $(X_i - \mu_i)/s_n$ converges in distribution to a standard normal random variable, as *n* goes to infinity:

$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,\ 1).$$

In practice it is usually easiest to check the Lyapunov's condition for δ = 1. If a sequence of random variables satisfies Lyapunov's condition, then it also satisfies Lindeberg's condition. The converse implication, however, does not hold.

3. Lindeberg CLT

In the same setting and with the same notation as above, the Lyapunov condition can be replaced with the following weaker one called Lindeberg's condition, for every ε > 0

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathrm{E}\left[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\{|X_i - \mu_i| > \varepsilon s_n\}}\right] = 0$$

where $\mathbf{1}_{\{\ldots\}}$ is the indicator function. Then the distribution of the standardized sums $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{X_i} - E(X_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{X_i} - \mu] = \sqrt{n}\left(\overline{\mathbf{X}}_n - \mu\right)$ $\frac{1}{s_n}\sum_{i=1}^{n}(X_i - \mu_i)$ converges towards the standard normal distribution N(0,1).

4. Multidimensional CLT

Proofs that the used characteristic functions can be extended to cases where each individual X_1, ..., X_n is an independent and identically distributed random vector in $\mathbf{R}^k$, with mean vector μ = E(X_i) and covariance matrix**Σ** (amongst the individual components of the vector). Now, if we take the summations of these vectors as being done component wise, then the multidimensional central limit theorem states that when scaled, these converge to a multivariate normal distribution.

Let

$$\mathbf{X_i} = \begin{bmatrix} X_{i(1)} \\ \vdots \\ X_{i(k)} \end{bmatrix}$$

be the *i*-vector. The bold in $\mathbf{X}_i$ means that it is a random vector, not a random (univariate) variable. Then the sum of the random vectors will be

$$\begin{bmatrix} X_{1(1)} \\ \vdots \\ X_{1(k)} \end{bmatrix} + \begin{bmatrix} X_{2(1)} \\ \vdots \\ X_{2(k)} \end{bmatrix} + \cdots + \begin{bmatrix} X_{n(1)} \\ \vdots \\ X_{n(k)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}\left[X_{i(1)}\right] \\ \vdots \\ \sum_{i=1}^{n}\left[X_{i(k)}\right] \end{bmatrix} = \sum_{i=1}^{n}[\mathbf{X_i}]$$

and the average will be

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n}[\mathbf{X_i}] = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n}\left[X_{i(1)}\right] \\ \vdots \\ \sum_{i=1}^{n}\left[X_{i(k)}\right] \end{bmatrix} = \begin{bmatrix} \bar{X}_{i(1)} \\ \vdots \\ \bar{X}_{i(k)} \end{bmatrix} = \bar{\mathbf{X}}_\mathbf{n}$$

and therefore

The multivariate central limit theorem states that

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mu\right) \xrightarrow{D} \mathcal{N}_k(0, \Sigma)$$

Self-Assessment Exercises (SAEs 1)

1. State any 2 rules of CLT rule
2. Mention 4 types of classical CLT rule

1.6 SUMMARY

Some level of independence sequence in CLT is highlighted here to make learners have an understanding of inferential statistics and hypothesis testing. Your ability to attempt the assignments below will go a long way in improving on the knowledge already acquired above.

1.7 REFERENCES/FURTHER READING/WEB RESOURCES

Billingsley, P. (1995), *Probability and Measure* (3rded.). John Wiley &Sons Publishers, ISBN0-

471-00710-2
Bradley, R. (2007), *Introduction to Strong Mixing Conditions* (1st ed.), Heber City, UT: Kendrick

Press, ISBN0-9740427-9-X

Bradley, R. (2005), *Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions*, *Probability Surveys***2**:107–144, arXiv:math/0511078v1,doi:10.1214/154957805100000104, http://arxiv.org/pdf/math/0511078.pdf

Dinov, I., Christou, N. and Sanchez, J. (2008), *Central Limit Theorem: New SOCR Applet and Demonstration Activity, Journal of Statistics Education* (ASA) **16** (2), http://www.amstat.org/publications/jse/v16n2/dinov.html, website: www. Wikipedia.com

1.8 Possible Answers to SAE 1

1.

(i) .$\mu_{\bar{y}} = \mu$

(ii) When the population distribution is normal, the sampling distribution of Y is also normal for any sample size n

2.. (i) Lindeberg–Lévy CLT
(ii) Lyapunov CLT
(iii.) Lindeberg CLT
(iv.) Multidimensional CLT

UNIT 2 CENTRAL LIMIT THEOREM FOR DEPENDENT PROCESSES

CONTENTS UNIT STRUCTURE

2.1 Introduction
2.2 Learning Outcomes
2.3 Central Limit Theorem for Dependent Processes
2.4 Theorems
2.5 Proof of Classical CLT
2.6 Summary
2.6 References/Further Reading/Web Resources
2.8 Possible Answers to Self-Assessment Exercise(s) within the content

2.1 INTRODUCTION

A useful generalization of a sequence of independent identically distributed random variables is a mixing random process in discrete time; "mixing" means, roughly, that random variables temporally far apart from one another are nearly independent. Several kinds of mixing are used in ergodic theory and probability theory. Strong mixing (also called α-mixing) is defined by $\alpha(n) \to 0$ where $\alpha(n)$ is so-called strong mixing coefficient.

2.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Explain the meaning of central limit theorem

2.3 CENTRAL LIMIT THEOREM FOR DEPENDENT PROCESSES

A simplified formulation of the central limit theorem under strong mixing is provided for in the following:

- CLT under weak dependence
- Martingale difference CLT

2.4 Theorems

1. **Theorem*i*.** Suppose that $X_1, X_2, \ldots$ is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $\mathrm{E}(X_n) = 0$ and $\mathrm{E}(X_n^{12}) < \infty$. Denote $S_n = X_1 + \ldots + X_n$, then the limit

exists, and if $\zeta \neq 0$ then $\sqrt{\quad}$ converges in distribution to $N(0, 1)$.

In fact, $\sum$

where the series converges absolutely.

The assumption $\zeta \neq 0$ cannot be omitted, since the asymptotic normality fails for $X_n = Y_n - Y_{n-1}$ where Y_n are another stationary sequence.

There is a stronger version of the theorem: the assumption $\mathrm{E}(X_n^{12}) < \infty$ is replaced with $\mathrm{E}(|X_n|^{2+\delta}) < \infty$, and the assumption $\alpha_n = O(n^{-5})$ is replaced with $\sum$

Existence of such $\delta > 0$ ensures the conclusion.

2. Theorem*ii*. Let a martingaleM_n satisfy

$\sum$ in probability as n tends to infinity,

-for every $\varepsilon > 0$, $\sum$ | | $\sqrt{}$

Then $\sqrt{}$ converges in distribution to $N(0,1)$ as $n \to \infty$.

2.5 Proof of classical CLT

For a theorem of such fundamental importance to statistics and applied probability, the central limit theorem has a remarkably simple proof using characteristic functions. It is similar to the proof of a (weak) law of large numbers. For any random variable, Y, with zero mean and a unit variance (var(Y) = 1), the characteristic function of Y is, by Taylor's theorem,

where$o$ (t^2) is "little o notation" for some function of t that goes to zero more rapidly than t^2. Letting Y_i be $(X_i - \mu)/\zeta$, the standardized value of X_i, it is easy to see that the standardized mean of the observations X_1, X_2, ..., X_n is

$\sqrt{}$ Σ $=$ $\sqrt{}$

By simple properties of characteristic functions, the characteristic function of Z_n is

[( )] * ( )+ /

But this limit is just the characteristic function of a standard normal distribution N(0, 1), and the central limit theorem follows from the Lévy continuity theorem, which confirms that the convergence of characteristic functions implies convergence in distribution.

2.6 SUMMARY

The central limit theorem applies in particular to sums of independent and identically distributed discrete random variables. A sum of discrete random variables is still a discrete random variable, so that we are confronted with a sequence of discrete random variables whose cumulative probability distribution function converges towards a cumulative probability distribution function

corresponding to a continuous variable (namely that of the normal distribution). This means that if we build a histogram of the realisations of the sum of n independent identical discrete variables, the curve that joins the centres of the upper faces of the rectangles forming the histogram converges toward a Gaussian curve as n approaches infinity, this relation is known as de Moivre– Laplace theorem. The binomial distribution article details such an application of the central limit theorem in the simple case of a discrete variable taking only two possible values.

2.7 REFERENCES/FURTHER READING/WEB RESOURCES

Artstein, S.; Ball, K.; Barthe, F. andNaor, A. (2004), *"Solution of Shannon's Problem on the Monotonicity of Entropy",Journal of the American Mathematical Society***17** (4): 975–982, doi:10.1090/S0894-0347-04-00459-X

Rosenthal, J. S. (2000) *A first look at rigorous probability theory*, World Scientific, ISBN 981-02-

4322-7.(Theorem 5.3.4, p. 47)

Website: www.wikipedia.com

 **2.8** Possible Answers to Self-Assessment Exercise(s) within the content

UNIT 3:THE LAW OF LARGE NUMBERS UNIT STRUCTURE

3.1 Introduction
3.2 Learning Outcomes
3.3 The Law of Large Numbers
3.4 Solved Examples
3.5 Summary
3.6 References/Further Reading/Web Resources
3.7 Possible Answers to Self-Assessment Exercise(s) within the content

3.1 INTRODUCTION

It is a rule that assumes that as the number of samples increases, the average of the samples is likely to reach the mean of the population. The law of large numbers says that the sample mean of a random sample converges in probability to the mean μ of the individual random variables, if the variance exists. This means that the sample mean will be close to μ if the size of the random sample is sufficiently large.
Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the mean is μ and for which the variance is finite. Let X_n denote the sample mean. Then

Proof: Let the variance of each X_i be σ^2. It then follows from the Chebyshev inequality that for every number $\varepsilon > 0$,

| | | —

Hence,

3.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Explain law of large numbers

3.3 THE LAW OF LARGE NUMBERS

The law of large numbers as well as the central limit theorem is partial solutions to a general problem of example; "What is the limiting behavior of S_n as n approaches infinity?" In statistical analysis, asymptotic series are one of the most popular tools employed to approach such questions.

Suppose we have an asymptotic expansion of $f(n)$:

$$f(n) = a_1\varphi_1(n) + a_2\varphi_2(n) + O(\varphi_3(n)) \qquad (n \to \infty).$$

Dividing both parts by $\varphi_1(n)$ and taking the limit will produce a_1, the coefficient of the highest- order term in the expansion, which represents the rate at which $f(n)$ changes in its leading term.

$$\lim_{n\to\infty} \frac{f(n)}{\varphi_1(n)} = a_1.$$

Informally, one can say: "$f(n)$ grows approximately as $a_1\ \varphi_1(n)$". Taking the difference between $f(n)$ and its approximation and then dividing by the next term in the expansion, we arrive at a more refined statement about $f(n)$:

$$\lim_{n\to\infty} \frac{f(n) - a_1\varphi_1(n)}{\varphi_2(n)} = a_2.$$

Here one can say that the difference between the function and its approximation grows approximately as $a_2\varphi_2(n)$. The idea is that dividing the function by appropriate normalizing functions, and looking at the limiting behavior of the result, can tell us much about the limiting behavior of the original function itself.

Informally, something along these lines is happening when the sum, S_n, of independent identically distributed random variables, X_1, ...,X_n, is studied in classical probability theory. If each X_i has finite mean μ, then by the law of large numbers, $S_n/n \to \mu$. If in addition each X_i has finite variance ζ^2, then by the central limit theorem,

$$\frac{S_n - n\mu}{\sqrt{n}} \to \xi,$$

where ξ is distributed as N(0, ζ^2). This provides values of the first two constants in the informal expansion

$$S_n \approx \mu n + \xi\sqrt{n}.$$

In the case where the X_i's do not have finite mean or variance, convergence of the shifted and rescaled sum can also occur with different centering and scaling factors:

$$\frac{S_n - a_n}{b_n} \to \Xi,$$

or informally

$$S_n \approx a_n + \Xi b_n.$$

Distributions Ξ which can arise in this way are called **stable**. Clearly, the normal distribution is stable, but there are also other stable distributions, such as the Cauchy distribution, for which the mean or variance are not defined. The scaling factor b_n may be proportional to n^c, for any $c \geq 1/2$; it may also be multiplied by a slowly varying function of n.

3.4 Solved Examples

Examples: Suppose that a random sample is to be taken from a distribution for which the value of the mean μ is not known, but for which it is known that the standard deviation σ is 2 units or less. We shall determine how large the sample size must be in order to make the probability at least 0.99 that $|Xn - \mu|$ will be less than 1 unit. Since $\sigma^2 \leq 2^2 = 4$, it follows from the relation that for every sample size n,

| |

Since n must be chosen so that $\Pr(|Xn - \mu| < 1) \geq 0.99$, it follows that n must be chosen so that $4/n \leq 0.01$. Hence, it is required that $n \geq 400$.

For example a single roll of a six-sided die produces one of the numbers 1, 2,3,4,5 or 6 each with equal probability. Therefore the expected value of a single die roll is

According to the law of large numbers if a large number of six-sided dice are rolled the average of their values sometimes called the sample mean is likely to be close to 3.5 with the accuracy increasing as more dice are rolled.

3.5 SUMMARY

In summary, the explanations and illustrations presented above would have provided clear understanding of this unit. In case learners encounter some difficulties in understanding any area, it is suggested that they make reference to further reading list at the end of this unit. Such reference is expected to enhance learners‘ knowledge to be able to solve problems arising from large numbers.

3.6 REFERENCES/FURTHER READING/WEB RESOURCES

Artstein, S.; Ball, K.; Barthe, F. andNaor, A. (2004), *"Solution of Shannon's Problem on the Monotonicity of Entropy"*, *Journal of the American Mathematical Society***17** (4): 975–982, doi:10.1090/S0894-0347-04-00459-X

Rosenthal, J. S**.** (2000) *A first look at rigorous probability theory*, World Scientific, ISBN 981-02-

4322-7.(Theorem 5.3.4, p. 47)

Johnson, O. T**.** (2004) *Information theory and the central limit theorem*, Imperial College Press,

2004, ISBN 1-86094-473-6. (p. 88)

Vladimir V. U. and Zolotarev V. M. (1999) *Chance and stability: stable distributions and their applications*, VSP. ISBN 90-6764-301-7.(pp. 61–62)

3.7 **Possible Answers to Self-Assessment Exercise(s) within the content**

UNIT 4 EXTENSION TO THE CLT AND BEYOND THE CLASSICAL FRAMEFORK

CONTENTS UNIT STRUCTURE

4.1 Introduction
4.2 Learning Outcomes
4.3 Extension to the CLT and beyond the Classical Framework
4.4 Lacunary trigonometric series
4.5 Linear functions of orthogonal matrices
4.6 Summary
4.7 References/Further Reading/Web Resources
4.8 Possible Answers to Self-Assessment Exercise(s) within the content

4.1 INTRODUCTION

Convex body

Theorem. There exists a sequence $\varepsilon_n \downarrow 0$ for which the following holds. Let $n \geq 1$, and let random variables $X_1, \ldots, X_n$ have a log-concave joint density f such that $f(x_1, \ldots, x_n) = f(|x_1|, \ldots, |x_n|)$ for all $x_1, \ldots, x_n$, and $\mathrm{E}(X_k^2) = 1$ for all $k = 1, \ldots, n$. Then the distribution of

$$\frac{}{\sqrt{\quad}}$$

is ε_n-close to $N(0, 1)$ in the total variation distance.

These two ε_n-close distributions have densities (in fact, log-concave densities), thus, the total variance distance between them is the integral of the absolute value of the difference between the densities. Convergence in total variation is stronger than weak convergence.

An important example of a log-concave density is a function constant inside a given convex body and vanishing outside; it corresponds to the uniform distribution on the convex body, which explains the term "central limit theorem for convex bodies".

Another example: $f(x_1, \ldots, x_n) = \text{const} \cdot \exp(-(|x_1|^{\alpha} + \ldots + |x_n|^{\alpha})^{\beta})$ where $\alpha > 1$ and $\alpha\beta > 1$. If $\beta = 1$ then $f(x_1, \ldots, x_n)$ factorizes into const $\cdot \exp(-|x_1|^{\alpha}) \ldots \exp(-|x_n|^{\alpha})$, which means independence of $X_1, \ldots, X_n$. In general, however, they are dependent.

The condition $f(x_1, \ldots, x_n) = f(|x_1|, \ldots, |x_n|)$ ensures that $X_1, \ldots, X_n$ are of zero mean and uncorrelated; still, they need not be independent, nor even pairwise independent. By the way, pairwise independence cannot replace independence in the classical central limit theorem.
Below is a Berry-Esseen type result.

Theorem. Let $X_1, \ldots, X_n$ satisfy the assumptions of the previous theorem, then

$$\left| \left(\frac{}{} \right) \sqrt{\int} \; / \; \right| \; -$$

for all $a<b$; here C is a universal (absolute) constant. Moreover, for every $c_1, \ldots, c_n \in \mathbf{R}$ such that $c_1^2 + \ldots + c_n^2 = 1$,

The distribution of need not be approximately normal (in fact, it can be uniform). However, the distribution of $c_1X_1 + \ldots + c_nX_n$ is close to $N(0, 1)$ (in the total variation distance) for most of vectors $(c_1, \ldots, c_n)$ according to the uniform distribution on the sphere $c_1^2 + \ldots + c_n^2 = 1$.

Products of positive random variables

The logarithm of a product is simply the sum of the logarithms of the factors. Therefore when the logarithm of a product of random variables that take only positive values approaches a normal distribution, the product itself approaches a log-normal distribution. Many physical quantities (especially mass or length, which are a matter of scale and cannot be negative) are the products of different random factors, so they follow a log-normal distribution.

Whereas the central limit theorem for sums of random variables requires the condition of finite variance, the corresponding theorem for products requires the corresponding condition that the density function be square-integrable.

Beyond the classical framework

Asymptotic normality, that is, convergence to the normal distribution after appropriate shift and rescaling, is a phenomenon much more general than the classical framework treated in the precious units, namely, sums of independent random variables (or vectors). New frameworks are revealed from time to time; no single unifying framework is available for now.

4.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Prove convex body theorem
- Explain products of positive random variables
- Linear functions of orthogonal matrices

4.3 MAIN CONTENT EXTENSION TO THE CLT AND BEYOND THE CLASSICAL FRAMEFORK

4.4 Lacunary trigonometric series

Theorem (Salem-Zygmund) Let U be a random variable distributed uniformly on $(0, 2\pi)$, and $X_k = r_k\cos(n_kU + a_k)$, where

- n_k satisfy the lacunarity condition: there exists $q > 1$ such that $n_{k+1} \geq qn_k$ for all k,
- r_k are such that
- $0 \leq a_k < 2\pi$.

Then

√

converges in distribution to $N(0, 1/2)$.

Gaussian polytopes

Theorem Let $A_1, ...,A_n$ be independent random points on the plane $\mathbf{R}^2$ each having the two-dimensional standard normal distribution. Let K_n be the convex hull of these points, and X_n the

area of K_n. Then; $\frac{\quad}{\sqrt{\quad}}$

converges in distribution to $N(0, 1)$ as n tends to infinity.The same holds in all dimensions (2, 3,

olds for the number of vertices (of the Gaussian polytope), the number of edges, and in fact, faces of all dimensions.

4.5 Linear functions of orthogonal matrices

A linear function of a matrix M is a linear combination of its elements (with given coefficients), $M \mapsto \operatorname{tr}(AM)$ where A is the matrix of the coefficients; see Trace_(linear_algebra)#Inner product.

A random orthogonal matrix is said to be distributed uniformly, if its distribution is the

normalized Haar measure on the orthogonal group $O(n, \mathbf{R})$; see Rotation matrix#Uniform random rotation matrices.

Theorem. Let M be a random orthogonal $n \times n$ matrix distributed uniformly, and Aa fixed $n \times n$ matrix such that $\operatorname{tr}(AA^*) = n$, and let $X = \operatorname{tr}(AM)$. Then the distribution of X is close to N(0, 1) in the total variation metric up to $2\sqrt{3}/(n-1)$.

Implications

Theorem. Let random variables $X_1, X_2, \ldots \in L_2(\Omega)$ be such that $X_n \to 0$ weakly in $L_2(\Omega)$ and $X_n^2 \to$

1 weakly in $L_1(\Omega)$. Then there exist integers $n_1 < n_2 < \ldots$ such that
√converges in distribution to $N(0, 1)$ as k tends to infinity.

Q-analogues

A generalized q-analog of the classical central limit theorem has been described by Umarov, Tsallis and Steinberg in which the independence constraint for the i.i.d. variables is relaxed to an extent defined by the q parameter, with independence being recovered as q->1. In analogy to the classical central limit theorem, such random variables with fixed mean and variance tend towards the q-Gaussian distribution, which maximizes the Tsallis entropy under these constraints. Umarov, Tsallis, Gell-Mann and Steinberg have defined q-analogs of all symmetric alpha-stable distributions, and have formulated a number of conjectures regarding their relevance to an even
more general Central limit theorem.

4.6 SUMMARY

In this unit, we have been able to treat the issues which are addition to the Classical Limit Theorem beyond the classical framework and this includes: Products of positive random variables, the theorem around convex body, the Lacunary trigonometric series, the Linear Functions of Orthogonal Matrices and its implication among others. Students are expected to be proficient in the use of the theorem in order to be able to apply it to solving practical day-to-day problems.

4.7 REFERENCES/FURTHER READING/WEB RESOURCES

Rempala, G. and Wesolowski**,** J.(2002) *"Asymptotics of products of sums and U-statistics", Electronic Communications in Probability*, 7, 47–54.

Zygmund, A.(1959), *Trigonometric series, Volume II,* Cambridge. (2003 combined volume I,II: ISBN 0-521-89053-5) (Sect. XVI.5, Theorem 5-5)

Meckes, E. (2008), *"Linear functions on the classical matrix groups", Transactions of the*

*American Mathematical Society***360** (10): 5355–5366, arXiv:math/0509441, doi:10.1090/S0002-

9947-08-04444-9

4.8 Possible Answers to Self-Assessment Exercise(s) within the content

MODULE 5: Index Numbers and Introduction to Research Methods in Social Sciences

CONTENTS

Unit 1: Index Number
Unit 2: Statistical Data
Unit 3: Sample and Sampling Techniques

UNIT 1: INDEX NUMBER

UINT STRUTURE

1.1 Introduction
1.2 Learning Outcomes
1.3 Index Number
1.3.1 Uses of index numbers
1.3.2 Types of index number
1.4 Methods of constructing index numbers
1.4.1 Solved Examples
1.5 Problems encountered in the construction of index numbers
1.6 Summary
1.7 References / Further Reading/Web Resources
1.8 Possible Answers to Self-Assessment Exercise(s) within the content

1.1 INTRODUCTION

Index numbers are indicators which reflect the relative changes in the level of certain phenomenon in any given period (or over a specified period of time) called the current period with respect to its value in some fixed period called the base period selected for comparison. The phenomenon or variable under consideration may be price, volume of trade, factory production, agricultural production, imports or exports, shares, sales, national income, wage structure, bank deposits, foreign exchange reserves, cost of living of people of a particular community etc.

1.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Highlight the uses of index number
- Explain types of index number
- Explain problems in the construction of index number

1.3 INDEX NUMBER

1.3.1 Uses of Index Number

1. Index numbers are used to measure the pulse of the economy.
2. It is used to study trend and tendencies
3. Index numbers are used for deflation
4. Index numbers help in the formulation of decisions and policies
5. It measures the purchasing power of money

1.3.2 Types of Index Numbers

Index number may be classified in terms of the variables they measure. They are generally classified into three categories:

1. **Price Index Number:** The most common index numbers are the price index numbers which study changes in price level of commodities over a period of time. They are of two types:
   (a) **Wholesale price index number** – They depict changes in the general price level of the economy.
   (b) **Retail Price Index Number –** The reflect changes in the retail prices of different commodities. The are normally constructed for different classes of consumers.
2. **Quantity Index Number –** They reflect changes in the volume of goods produced or consumed
3. **Value Index Number** – They study changes in the total value (price X quantity) e.g index number of profit or sales.

.4 Methods of constructing index numbers

(1) **Simple (unweight) Aggregate Method** – Aggregate of prices (of all the selected commodities) in the current year as a percentage of the aggregate of prices in the base year.

P_{01} → Price index number in the current year with respect to the base year

$$\frac{\Sigma}{\Sigma}$$

Exercise: From the following data calculate Index Number by Simple Aggregate method.

Commodity	A	B	C	D
Price 2011	81	128	127	66
Price 2012	85	82	95	73

$$\frac{\Sigma}{\Sigma}$$

(2) Weighted Aggregate Method - In this method, appropriate weights are assigned to various commodities to reflect their relative importance in the group. The weights can be production figures, consumption figure or distribution figure

Limitations of the Simple Aggregate Method

(i. The prices of various commodities may be quoted in different units

(ii Commodities are weighted according to the magnitude of their price. Therefore, highly priced commodity exerts a greater influence than lowly priced commodity. Therefore, the method is dominated by commodities with higher prices.

(iii. The relative importance of various commodities is not taken into consideration

Based on this method quantity index is given by the formula:

$$\frac{\sum}{\sum}$$

By using different systems of weighting, we obtain a number of formulae, some of which nclude:

(i) **Laspeyre's Price Index or Base year method –** Taking the base year quantity as weights i.e w = q_0 in the equation above, the Laspeyre's Price Index is given as:

$$\frac{\sum}{\sum}$$

This formula was invented by French economist Laspeyre in 1817.

(ii) **Paasche's Price Index –** Here, the current year quantities are taken as weights and we obtain:

$$\frac{\sum}{\sum}$$

This formula was introduced by German statistician Paasche, in 1874.

(iii) Dorbish-Bowley Price Index – This index is given by the arithmetic mean of Laspeyre's and Paasche's price index numbers.It is also sometimes known as L-P formula:

$$\frac{1}{2}\left[\frac{\sum}{\sum} + \frac{\sum}{\sum}\right]$$

(iv) Fisher's Price Index – Irving Fisher advocated the geometric cross of Laspeyre's and Paasche's Price index numbers and is given as:

$$\sqrt{\frac{\sum}{\sum} \times \frac{\sum}{\sum}}$$

Fisher's Index is termed as an ideal index since it satisfies time reversal and factor reversal test for the consistency of index numbers.

1.3.3 Deflating a Series by Price Indexes

Many business and economic series reported over time, such as company sales, industry sales, and inventories, are measured in dollar amounts. These time series often show an increasing growth pattern over time, which is

generally interpreted as indicating an increase in the physical volume associated with the activities. For example, a total dollar amount of inventory up by 10% might be interpreted to mean that the physical inventory is 10% larger. Such interpretations can be misleading if a time series is measured in terms of Naira or dollars, and the total Naira or dollar amount is a combination of both price and quantity changes. Hence, in periods when price changes are significant, the changes in the Naira or dollar amounts may not be indicative of quantity changes unless we are able to adjust the time series to eliminate the price change effect.

For example, let us assume that from 2006 to 2020, the total amount of spending in the construction industry increased approximately by 75%. That figure suggests excellent growth in construction activity. However, construction prices were increasing just as fast as—or sometimes even faster than—the 75% rate. In fact, while total construction spending was increasing, construction activity was staying relatively constant or, as in the case of new housing starts, decreasing. To interpret construction activity correctly for the 2006–2020 period, we must adjust the total spending series by a price index to remove the price increase effect. Whenever we remove the price increase effect from a time series, we say we are **deflating the series**.

In relation to personal income and wages, we often hear discussions about issues such as "real wages" or the "purchasing power" of wages. These concepts are based on the notion of deflating an hourly wage index. For example, if salaries increase from N30,000 per month to N60,000 per month from 2022 - 2023. Should workers be pleased with this growth in monthly salaries? The answer depends on what happened to the purchasing power of their salaries. If we can compare the purchasing power of the N30,000 monthly salaries in 2022 with the purchasing power of the N60,000 monthly salaries hourly wage in 2023, we will be better able to judge the relative improvement in wages.

1.4.1 Solved Examples

Example 1: Consider the table below which gives the details of price and consumption of four commodities for 2010 and 2012. Using an appropriate formula calculate an index number for 2012 prices with 2010 as base year.

Commodities	**Price per unit 2010 (₦)**	**Price per unit 2012 (₦)**	**Consumption value 2010**
Yam flour	70	85	1400
Vegetable oil	45	50	720

Beans	90	110	900
Beef	100	125	600

Solution: In the above problem, we are given the base year (2010) consumption values (p_oq_o) and current year quantities (q_1) are not given, the appropriate formula for index number here is the

Laspeyre's Price Index.

Commodities	**Price per unit 2010 (₦) p_0 (1)**	**Consumption value 2010 (₦) p_oq_o(2)**	**Price per unit 2012 (₦) p_1 (3)**	**2010 quantities q_0= —**	
Yam flour	70	1400	85	20	1700
Vegetable oil	45	720	50	16	800
Beans	90	900	110	10	1100
Beef	100	600	125	6	750
		∑ **3620**			∑ **4 350**

Therefore, the Laspeyre's Price Index for 2012 with respect to (w.r.t) base 2010 is given by:

$$= \frac{\sum}{\sum} =$$

Example 2: From the following data calculate price index for 2012 with 2007 as the base year by
(i) Laspeyre's method (ii) Paasche's method (iii) Fisher's method and
(iv) Dowbish-Bowley price index methods

Commodities	***2007***		***2012***	
	Price	**Quantity**	**Price**	**Quantity**
Gaari	20	8	40	6
Rice	50	10	60	5
Fish	40	15	50	15
Palm-oil	20	20	20	25

Solution:

Commoditis	***2007***		***2012***					
	Price (p_o)	**Quantity (q_o)**	**Price (p_1)**	**Quantity (q_1)**	p_oq_o	p_oq_1	p_1q_o	p_1q_1

Gaari	20	8	40	6	160	120	320	240
Rice	50	10	60	5	500	250	600	300
Fish	40	15	50	15	600	600	750	750
Palm-oil	20	20	20	25	400	500	400	500
Total					***poqo=*** ***1660***	***poq1=*** ***1470***	***p1q0=*** ***2070***	***p1q1=*** ***1790***

(i) Laspeyre's Price Index

$$= \frac{\Sigma}{\Sigma} \quad \text{——} \quad \text{——}$$

= 1.24699 X 100

= 124.7

$$\Sigma$$
$$\Sigma$$

$$\text{——} \quad \text{—} \quad \text{——}$$

= 1.2177 X 100 = 121.77

(iii) Fisher's Price Index

$$\sqrt{\frac{\Sigma}{\Sigma} \quad \frac{\Sigma}{\Sigma} \quad \text{——}}$$

$$\sqrt{\text{——} \quad \text{——}}$$

(iv) Dorbish-Bowley Price Index

= 123.23

$$\Sigma - \quad \Sigma$$

$$\frac{\Sigma}{\Sigma} \quad \frac{\Sigma}{\Sigma} \quad - \text{——} \quad \text{——} \quad \text{——} \quad \text{——} \quad \text{——}$$

= ½ [1.247 + 1.2177] X100

= 1.23235 X 100

= 123.24

(ii) Paasche's Price Index

1.5 Problems in the construction of Index Numbers

1. The purpose of index number – This must be carefully defined as there is no general purpose index number.

2. Selection of base period – The base period is the previous period with which comparison of some later period is made. The index of the base period is taken to be 100. The following points should be borne in mind while selecting a base period:
   (a) Base period should be a normal period devoid of natural disaster, economic boom, depression, political instability, famine etc.
   (b) The base period should not be too distant from the given period. This is because circumstances such as tastes customs, habits and fashion keep changing.
   (c) One must determine whether to use fixed-base or chain-base method

3. Selection of commodities – Commodities to be selected must be relevant to the study; must not be too large nor too small and must be of the same quality in different periods.
4. Data for the index number- Data to be used must be reliable.

5. Type of average to be used – ie, arithmetic, geometric, harmonic etc.

6. Choice of formula – There are different types of formulas and the choice is mostly dependent on available data.
7. System of weighting – Different weights should be assigned to different commodities according to their relative importance in the group.

1.6 SUMMARY

Price and quantity indexes are important measures of changes in price and quantity levels within the business and economic environment. Price relatives are simply the ratio of the current unit price of an item to a base-period unit price multiplied by 100, with a value of 100 indicating no difference in the current and base-period prices. Aggregate price indexes are created as a composite measure of the overall change in prices for a given group of items or products. Usually the items in an aggregate price index are weighted by their quantity of usage. A weighted aggregate price index can also be computed by weighting the price relatives by the usage quantities for the items in the index.

In this unit, we have been able to introduce students to the concept of index numbers, its uses and methods of calculation. Students are now expected to be proficient in the calculation, use and interpretation of index numbers. This is useful in the study and interpretation of inflation, cost of living, trends of economic variables among others.

Self-Assessment Exercises (SAEs 1)

1. Highlight the uses of index number
2. Explain types of index number
3. Enumerate the problems faced in constructing index number

1.7 REFERENCES / FURTHER READING/WEB RESOURCES

Gupta S.C. (2011). *Fundamentals of Statistics.*(6th Rev. and Enlarged ed.).Mumbai India: Himalayan Publishing House.

Lucey T. (2002). *Quantitative Techniques*.(6thed.).BookPower

1.8 Possible Answers to Self-Assessment Exercise(s) within the content

SAEs 1

1. (i).Index numbers are used to measure the pulse of the economy.

(ii). It is used to study trend and tendencies

(iii). Index numbers are used for deflation

(iv). Index numbers help in the formulation of decisions and policies

2 It measures the purchasing power of money

2. (i) **Price Index Number:** The most common index numbers are the price index numbers which study changes in price level of commodities over a period of time. They are of two types:

(a) **Wholesale price index number** – They depict changes in the general price level of the economy.

(b) **Retail Price Index Number –** The reflect changes in the retail prices of different commodities. The are normally constructed for different classes of consumers.

(ii). **Quantity Index Number –** They reflect changes in the volume of goods produced or consumed

(iii). **Value Index Number**

3. (i) Selection of commodities – Commodities to be selected must be relevant to the study; must not be too large nor too small and must be of the same quality in different periods.

(ii) Data for the index number- Data to be used must be reliable.

(iii). Type of average to be used – ie, arithmetic, geometric, harmonic etc.

(iv). Choice of formula – There are different types of formulas and the choice is mostly dependent on available data.

(v). System of weighting – Different weights should be assigned to different commodities according to their relative importance in the group.

UNIT STRUCTURE

2.1 Introduction
2.2 Learning Outcomes
2.3. Statistical Data
2.3.1 Types of data
2.4 Classification based on form of the data
2.5 Sources of Data
2.6 Summary
2.7 References/Further Reading/Web Resources
2.8 Possible Answers to Self-Assessment Exercise(s) within the content

2.1 INTRODUCTION

Statistics is a branch of mathematics that deals with the collection, organization, and analysis of numerical data and with such problems as experiment design and decision making. Simple forms of statistics have been used since the beginning of civilization, when pictorial representations or other symbols were used to record numbers of people, animals, and inanimate objects on skins, slabs, or sticks of wood and the walls of caves. Before 3000BC the Babylonians used small clay tablets to record tabulations of agricultural yields and of commodities bartered or sold. The Egyptians analyzed the population and material wealth of their country before beginning to build the pyramids in the 31st century BC. The biblical books of Numbers and first Chronicles are primarily statistical works, the former containing two separate censuses of the Israelites and the latter describing the material wealth of various Jewish tribes. Similar numerical records existed in China before 2000BC. The ancient Greeks held censuses to be used as bases for taxation as early as 594BC. The Roman Empire was the first government to gather extensive data about the population, area, and wealth of the territories that it controlled.

At present, however, statistics is a reliable means of describing accurately the values of economic, political, social, psychological, biological, and physical data and serves as a tool to correlate and analyze such data. The work of the statistician is no longer confined to gathering and tabulating data, but is chiefly a process of interpreting the information. The development of the theory of probability increased the scope of statistical applications. Much data can be approximated accurately by certain probability distributions, and the results of probability distributions can be used in analyzing statistical data. Probability can be used to test the reliability of statistical inferences and to indicate the kind and amount of data required for a particular problem.

2.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Define Statistics
- Explain types of data
- Explain sources of data
- Highlight the classification of data

2.3 STATISTICAL DATA

2.3.1 Types of Data

Data can be classified into types based on different criteria viz:

1. Based on sources – Data can be classified base on the sources from which they are obtained. In this regards, we have:

(a) **Primary data** – These are data collected directly from the field of enquiries by the user(s) or researcher(s) themselves.

Advantages

– They are always relevant to the subject under study because they are collected primarily for the purpose.

- They are more accurate and reliable
- Provide opportunity for the researcher to interact with study population.
- Information on other relevant issues can be obtained

Disadvantages

– Always costly to collect

- Inadequate cooperation from the study population
- Wastes a lot of time and energy

(b) **Secondary Data:** These are data which have been collected by someone else or some organization either in published or unpublished forms.

Advantages

- It is easier to get
- It is less expensive

Disadvantages

- May not completely meet the need of the research athand because it was not collected primarily for particular purpose
- There is always a problem of missing periods

2.4 **Classification based on form of the data:**

Sometimes, data are classified based on the form of the data at hand and may be classified as:

(a) **Cross-sectional data –** These are data collected for cross-section of subjects (population under study) at a time. For example, data collected on a cross-section of household on demand for recharge card for the month of August 2013; . data collected on stocks of 25 banks on the floor of Nigerian Stock Exchange (NSE) on August 22, 2023.

(b) **Time-series data –** These are data collected on a particular variable or set of variables over time e.g a set Nigeria‘s Gross Domestic Product (GDP) values form 1970 to2012.

(c) **Panel Data –** These combine the features of cross-sectional and time-series data. They are type of data collected from the same subjects over time. For example, a set of data collected on monthly recharge card expenditure from about 100 households in Lagos from January to

December 2013 will form a panel data.

Note that Social and Economic data of national importance are collected routinely as by-product of governmental activities e.g. information on trade, wages, prices, education, health, crime, aids and grants etc.

2.5 Sources of Data

Data can be obtained from existing sources or from surveys and experimental studies designed to collect new data.

2.5.1 Existing Sources

In some cases, data needed for a particular application already exist. Companies maintain a variety of databases about their employees, customers, and business operations. Data on employee salaries, ages, and years of experience can usually be obtained from internal personnel records. Other internal records contain data on sales, advertising expenditures, distribution costs, inventory levels, and production quantities. Most companies also maintain detailed data about their customers.

Organizations that specialize in collecting and maintaining data make available substantial amounts of business and economic data. Companies access these external data sources through leasing arrangements or by purchase. Dun & Bradstreet, Bloomberg, and Dow Jones & Company are three firms that provide extensive business database services to clients. ACNielsen and Information Resources, Inc., built successful businesses collecting and processing data that they sell to advertisers and product manufacturers. Data are also available from a variety of industry associations and special interest organizations.

The Travel Industry Association maintains travel-related information such as the number of tourists and travel expenditures by states. Such data would be of interest to firms and individuals in the travel industry. The Post-graduate Admission Council maintains data on test scores, student characteristics, and graduate management education programs. Most of the data from these types of sources are available to qualified users at a modest cost. The Internet continues to grow as an important source of data and statistical information. Almost all companies maintain websites that provide general information about the company as well as data on sales, number of employees, number of products, product prices, and product specifications. In addition, a number of companies now specialize in making information available over the Internet. As a result, one can obtain access to stock quotes, meal prices at restaurants, salary data, and an almost infinite variety of information. Government agencies are another important source of existing data. For instance, the National Bureau of Statistics maintains considerable data on employment rates. Most government agencies that collect and process data also make the results available through a website.

2.5.2 Statistical Studies

Sometimes the data needed for a particular application are not available through existing sources. In such cases, the data can often be obtained by conducting a statistical study. Statistical studies can be classified as either *experimental* or *observational.* In an experimental study, a variable of interest is first identified. Then one or more other variables are identified and controlled so that data can be obtained about how they influence the variable of interest. For example, a pharmaceutical firm might be interested in conducting an experiment to learn about how a new drug affects blood pressure. Blood pressure is the variable of interest in the study. The dosage level of the new drug is another variable that is hoped to have a causal effect on blood pressure. To obtain data about the effect of the new drug, researchers select a sample of individuals. The dosage level of the new drug is controlled, as different groups of individuals are given different dosage levels. Before and after data on blood pressure are collected for each group. Statistical analysis of the experimental data can help determine how the new drug affects blood pressure.

Non-experimental, or observational, statistical studies make no attempt to control the variables of interest. A survey is perhaps the most common type of observational study. For instance, in a personal interview survey, research questions are first identified. Then a questionnaire is designed and administered to a sample of individuals. Some restaurants use observational studies to obtain data about customer opinions on the quality of food, quality of service, atmosphere, and so on. A customer opinion questionnaire is used to gather information about the restaurant. The customers are asked to fill out the questionnaire by providing ratings, including overall experience, greeting by hostess, manager (table visit), overall service, and so on. The response categories of excellent, good, average, fair, and poor provide categorical data enable the restaurant management to maintain high standards for the restaurant's food and service.

In summary, data sources can be classified into

1. **Source of Primary data:**
   - (i) Census
   - (ii) Surveys
2. **Sources of Secondary data:**
   - (i) Publications of the Federal Bureau of statistics
   - (ii) Publications of Central Bank of Nigeria
   - (iii) Publications of National population commission
   - (iv) Nigerian Custom Service
   - (v) Nigeria Immigration Service
   - (vi) Nigerian Port Authority
   - (vi) Federal and State Ministries, Departments and Agencies

Some of the publications referred to above are: (i) Annual Digest of statistics (by NBS)

(ii) Annual Abstract of statistics (by NBS)

(ii) Economic and Financial Review (by CBN) (iv)
Population of Nigeria (by NPC)

2.5.3 Scales of Measurement

Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.

When the data for a variable consist of labels or names used to identify an attribute of the element, the scale of measurement is considered a **nominal scale**. In cases where the scale of measurement is nominal, a numeric code as well as nonnumeric labels may be used. For example, to facilitate data collection and to prepare the data for entry into a computer database, we might use a numeric code by letting 1 denote primary education, 2 denote secondary education, and 3 denotes tertiary education. In this case the numeric values 1, 2, and 3 identify the level of education. The scale of measurement is nominal even though the data appear as numeric values.

The scale of measurement for a variable is called an **ordinal scale** if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. For example, Abuja Automotive sends customers a questionnaire designed to obtain data on the quality of its automotive repair service. Each customer provides a repair service rating of excellent, good, or poor. Because the data obtained are the labels—excellent, good, or poor—the data have the properties of nominal data. In addition, the data can be ranked, or ordered, with respect to the service quality. Data recorded as excellent indicate the best service, followed by good and then poor. Thus, the scale of measurement is ordinal. Ordinal data can also be provided using a numeric code, for example, your class rank in school.

The scale of measurement for a variable is an **interval scale** if the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric. Scholastic Aptitude Test (SAT) scores are an example of interval-scaled data. For example, three students with SAT math scores of 620, 550, and 470 can be ranked or ordered in terms of best performance to poorest performance. In addition, the differences between the scores are meaningful. For instance, student 1 scored 620 - 550 =70 points more than student 2, while student 2 scored 550 – 470 = 80 points more than student 3.

The scale of measurement for a variable is a **ratio scale** if the data have all the properties of interval data and the ratio of two values is meaningful. Variables such as distance, height, weight, and time use the ratio scale of measurement. This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point For example, consider the cost of an automobile. A zero value for the cost would indicate that the automobile has no cost and is free. In addition, if we compare the cost of N300,000 for one automobile to

the cost of N150,000 for a second automobile, the ratio property shows that the first automobile is N300,000/N15,000 = 2 times, or twice, the cost of the second automobile..

2.6 SUMMARY

This unit has acquainted you with the transformation of the processed data into statistics and steps in the statistical cycle. The transformation involves analysis and interpretation of data to identify important characteristics of a population and provide insights into the topic being investigated.

Self-Assessment Exercise 1 (SAE 1)

1. Define Statistics
2. Explain various types of data
3. Discuss the 2 main sources of data

2.7 REFERENCES / FURTHER READINGS/WEB RESOURCES

Frankfort-Nachmias C. and Nachmias D. (2009).*Research in the Social Sociences.(5th ed.)*.Hodder Education.

Gupta S.C. (2011). *Fundamentals of Statistics*.(6th Rev. and Enlarged ed.). Mumbai India: Himalayan Publishing House

Microsoft ® Encarta ® 2009. © 1993-2008 Microsoft Corporation.

2.8 Possible Answers to Self-Assessment Exercise(s) within the content

SAE 1

1. Statistics is a branch of mathematics that deals with the collection, organization, and analysis of numerical data and with such problems as experiment design and decision making.
2. (i) Primary data
   (ii) Secondary data
3. (i). Primary source
   (ii) Secondary source

UNIT 3: SAMPLE AND SAMPLING TECHNIQUES

CONTENTS UNIT STRUCTURE

3.1 Introduction
3.2 Learning Outcomes
3.3 Sample and Sampling Techniques
3.3.1 Population
3.3.2The Sampling Unit
3.3.3 Sampling Frame
3.3.4 Sample Design
3.4 Probability and Non-probability Sampling
3.4.1 Non-probability Sample Designs
3.4.2 Probability Sample Designs
3.5 Sample size
3.5.1 Standard Error
3.6 Summary
3.7 References/Further Reading/Web Resources
3.8 Possible Answers to Self-Assessment Exercise(s) within the content

3.1 INTRODUCTION

Researchers collect data in order to test hypotheses and to provide empirical support for explanations and predictions. Once investigators have constructed their measuring instrument in order to collect sufficient data pertinent to the research problem, the subsequent explanations and predictions must be capable of being generalized to be of scientific value. Generalizations are important not only for testing hypotheses but also for descriptive purposes. Typically, generalizations are not based on data collected from all the observations, all the respondents, or the events that are defined by the research problem as this is always not possible or where possible too expensive to undertake. Instead, researchers use a relatively small number of cases (a sample) as the bases for making inferences for all the cases (a population).

3.2 LEARNING OUTCOMES

By the end of this unit, you will be able to:

- Define population
- Explain different types of sampling technique
- Explain methods of data collection
- State advantages of each method of data collection
- State disadvantages of each method of data collection

3.3 SAMPLE AND SAMPLING TECHNIQUES

Empirically supported generalizations are usually based on partial information because it is often

impossible, impractical, or extremely expensive to collect data from all the potential units of analysis covered by the research problem. Researchers can draw precise inferences on all the units (a set) based on relatively small number of units (a subset) when the subsets accurately represent
the relevant attributes of the whole sets. For example, in a study of patronage of campus photographer among students in a university, it may be very expensive and time consuming to reach out to all students (some universities have as high as 40,000 students). A careful selection of relatively small number of students across faculties, departments and levels will possibly give a representation of the entire student population.

The entire set of relevant units of analysis, or data is called the population. When the data serving as the basis for generalizations is comprised of a subset of the population, that subset is called a **sample**. A particular value of the population, such as the mean income or the level of formal education, is called **a parameter**; its counterpart in the sample is termed the **statistic**. The major objective of sampling theory is to provide accurate estimates of unknown values of the parameters from sample statistics that can be easily calculated. To accurately estimate unknown parameters from known statistics, researchers have to effectively deal with three major problems:

(1) the definition of the population, (2)
the sample design, and
(3) the size of the sample.

3.3.1 Population

Methodologically, a population is the ―aggregate of all cases that conform to some designated set of specifications‖. For example, a population may be composed of all the residents in a specific neighbourhood, legislators, houses, records, and so on. The specific nature of the population depends on the research problem. If you are investigating consumer behaviour in a particular city, you might define the population as all the households in that city. Therefore, one of the first problems facing a researcher who wishes to estimate a population value from a sample value is how to determine the population involved.

3.3.2The Sampling Unit

A single member of a sampling population (e.g. a household) is referred to as a sampling unit. Usually sampling units have numerous attributes, one or more of which are relevant to the research problem. The major attribute is that it must possess the typical characteristics of the study population. A sampling unit is not necessarily an individual. It can be an event, a university, a city

or a nation.

3.3.3 Sampling Frame

Once researchers have defined the population, they draw a sample that adequately represents that population. The actual procedures involve in selecting a sample from a sample frame comprised of a complete listing of sampling units. Ideally, the sampling frame should include all the sampling units in the population. In practice, a physical list rarely exists; researchers usually compile a substitute list and they should ensure that there is a high degree of correspondence between a sampling frame and the sampling population. The accuracy of a sample depends, first and foremost, on the sampling frame. Indeed, every aspect of the sample design – the population covered, the stages of sampling, and the actual selection process – is influenced by the sampling frame. Prior to selecting a sample, the researcher has to evaluate the sampling frame for potential problems.

3.3.4 Sample Design

The essential requirement of any sample is that it be as representative as possible of the population from which it is drawn. A sample is considered to be representative if the analyses made using the researcher‘s sampling units produce results similar to those that would be obtained had the researcher analyzed the entire population.

3.4 Probability and Non-probability Sampling

In modern sampling theory, a basic distinction is made between probability and non-probability sampling. The distinguishing characteristic of probability sampling is that for each sampling unit of the population, you can specify the probability that the unit will be included in the sample. In the simplest case, all the units have the same probability of being included in the sample. In non- probability sampling, there is no assurance that every unit has some chance of being included.

A well – designed sample ensures that if a study were to be repeated on a number of different samples drawn from a given population, the findings from each sample would not differ from the population parameters by more than a specified amount. A probability sample design makes it possible for researchers to estimate the extent to which the findings based on one sample are likely to differ from what they would have found by studying the entire population. When a researcher is using a probability sample design, it is possible for him or her to estimate the
population‘s parameters on the basis of the sample statistics calculated.

3.4.1 Non-probability Sample Designs

Three major designs utilizing non-probability samples have been employed by social scientists: convenience samples, purposive samples, and quota samples.

(a) Convenience sampling: Researchers obtain a convenience sample by selecting whatever sampling units are conveniently available. Thus a University professor may select students in a class; or a researcher may take the first 200 people encountered on the street who are willing to be interviewed. The researcher has no way of estimating the representativeness of convenience sample, and therefore cannot estimate the population's parameters.

(b) Purposive sampling: With purposive samples (occasionally referred to as judgment samples), researchers select sampling units subjectively in an attempt to obtain a sample that appears to be representative of the population. In order words, the chance that a particular sampling unit will selected for the sample depends on the subjective judgment of the researcher. At times, the main reason for selecting a unit in purposive sampling is the possession of pre-determined characteristic(s) which may be different from that the main population. For example, in a study of demand preference for cigarette brands in a city, researcher will need to select smokers purposively.

(c.) Quota sampling: The chief aim of quota sample is to select a sample that is as similar as possible to the sampling population. For example, if it is known that the population has equal numbers of males and females, the researcher selects an equal numbers of males and females in the sample. In quota sampling, interviewers are assigned quota groups characterized by specific variables such as gender, age, place of residence, and ethnicity.

3.4\.2 Probability Sample Designs

Four common designs of probability samples are simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

(a) Simple random sampling – is the basic probability sampling design, and it is incorporated into all the more elaborate probability sampling designs. Simple random sampling is a procedure that gives each of the total sampling units of the population an equal and known nonzero probability of being selected. For example, when you toss a perfect coin, the probability that you will get a head or a tail is equal and known (50 percent), and each subsequent outcome is independent of thprevious outcomes.

Random selection procedures ensure that every sampling unit of the population has an equal and known probability of being included in the sample; this probability is n/N, where n stands for the size of the sample and N for the size of the population. For example if we are interested in selection 60 household

from a population of 300 households using simple random sampling, the probability of a particular household being selected is 60/300 = 1/5.

(b) Systematic Sampling: It consists of selecting every k^{th} sampling unit of the population after the first sampling unit is selected at random from the total of sampling units. Thus if you wish to select a sample of 100 persons from total population of 10,000, you would take every hundredth individual (K=N/n = 10,000/100 = 100). Suppose that the fourteenth person were selected; the sample would then consist of individuals numbered 14,114, 214, 314, 414, and so on. Systematic sampling is more convenient than simple random sampling. Systematic samples are also more amenable for use with very large populations or when large samples are to be selected.

(c) Stratified Sampling:Researchers use this method, primarily to ensure that different groups of population are adequately represented in the sample. This is to increase their level of accuracy when estimating parameters. Furthermore, all other things being equal, stratified sampling considerably reduces the cost of execution. The underlying idea in stratified sampling is to use available information on the population –t0 divide it into groups such that the elements within each group are more alike than are the elements in the population as a whole. That is, you create a set of homogeneous samples based on the variables you are interested in studying. If a series of homogenous groups can be sampled in such a way when the samples are combined they constitute a sample of a more heterogeneous population, you will increase the accuracy of your parameter estimates.

(d) Cluster sampling: it is frequently used in large-scale studies because it is the least expensive sample design. Cluster sampling involves first selecting large groupings, called clusters, and then selecting the sampling units from the clusters. The clusters are selected by a simple random sample or a stratified sample. Depending on the research problem, researchers can include all the sampling units in these clusters in the sample or make a selection within the clusters using simple or stratified sampling procedures.

3.5 Sample size

A sample is any subset of sampling units from a population. A subset is any combination of sampling units that does not include the entire set of sampling units that has been defined as the population. A sample may include only one sampling unit, or any number in between.

There are several misconceptions about the necessary size of a sample. One is that the sample size must be certain proportion (often set as 5 percent) of the population; another is that the sample should total about 2000; still another is that any increase in the sample size will increase the precision of the sample results.

These are faulty notions because they do not derive from the ***sampling theory***. To estimate the adequate size of the sample properly, researchers need to determine what level of accuracy is expected of their estimates; that is, how large a standard error is acceptable.

3.5.1 Standard error

Some people called it *error margin*or*sampling error*. The concept of standard error is central to sampling theory and to determining the size of a sample. It is one of the statistical measures that indicate how closely the sample results reflect the true value of a parameter.

Self-Assessment Exercise 1 (SAE 1)

1. Discuss the main types on non-probability sample design.
2. Highlight 4 types of probability sample design

3.5 Methods of data collection

There are three methods of data collection with survey and these are mail questionnaires, personal interviews, and telephone interviews.

Mail questionnaire: It is an impersonal survey method. Here, survey instrument (the questionnaire) is mailed to the selected respondents and the questionnaires are mailed back to the researcher after the respondents must have filled it up. This is very common in developed countries where the citizens appreciate the relevance of data and research. Under certain conditions and for a number of research purposes, an impersonal method of data collection can be
useful.

Advantages and disadvantages of mail questionnaires

Advantages

- The cost is low compared to others
- Biasing error is reduced because respondents are not influenced by interviewed characteristics or techniques.
- Questionnaires provide a high degree of anonymity for respondents. This is especially important when sensitive issues are involved.
- Respondents have time to think about their answers and /or consult other sources.
- Questionnaires provide wide access to geographically dispersed samples at low cost

Disadvantages

- Questionnaires require simple, easily understood questions and instructions
- Mail questionnaires do not offer researchers the opportunity to probe for additional information or to clarify answers.
- Researchers cannot control who fills out the questionnaire.
- Response rate are low

Factors affecting the response rate of mail questionnaires

Researchers use various strategies to overcome the difficulty of securing an acceptable response rate to mail questionnaires and to increase the response rate.

- Sponsorship: The sponsorship of a questionnaire motivates the respondents to fill the questionnaires and return them. Therefore, investigators must include information on sponsorship, usually in the cover letter accompanying the questionnaire.
- Inducement to response: Researchers who use mail surveys must appeal to the respondents and persuade them that they should participate by filling out the questionnaires and mailing them back. For example, a student conducting a survey for a class project may mention that his or her grade may be affected by the response to the questionnaire.
- Questionnaire format and methods of mailing- Designing a mail questionnaire involves

  several considerations: typography, colour, and length and type of cover letter.

Personal interview

The personal interview is a face-to-face, interpersonal role situation in which an interviewer asks respondents question designed to elicits answers pertinent to the research hypotheses. The questions, their wording, and their sequence define the structure of the interview.

Advantages of personal interview

- Flexibility: The interview allows great flexibility in the questioning process, and the greater the flexibility, the less structure the interview. Some interviews allow the interviewer to determine the wording of the questions, to clarify terms that are unclear, to control the order in which the question are presented, and to probe for additional information and details.
- Control of the interview situation: An interviewer can ensure that the respondents answer the questions in the appropriate sequence or that they answer certain questions before they ask subsequent questions.

- High response rate: The personal interview results in a higher response rate than the mail questionnaire.
- Fuller information: An interviewer can collect supplementary information about respondents. This may include background information, personal characteristics and their environment that can aid the researcher in interpreting the results.

Disadvantages of the personal interview

- Higher cost: The cost of interview studies is significantly higher than that of mail survey. Costs are involved in selecting, training, and supervising interviewers; in paying them; and in the travel and time required to conduct interviews.
- Interviewer bias: The very flexibility that is the chief advantage of interviews leaves room for the interviewer's personal influence and bias.
- Lack of anonymity: The interview lacks the anonymity of the mail questionnaire. Often the interviewer knows all or many of the potential respondents (their names, addresses, and telephone numbers). Thus respondents may feel threatened or intimidated by the interviewer, especially if a respondent is sensitive to the topic or some of the questions.

Telephone interview

It is also called telephone survey, and can be characterised as a semi-personal method of collecting information. In comparison, the telephone is convenient, and it produces a very significant cost saving.

Advantages of Telephone interview

- Moderate cost
- Speed: Telephone interviews can reach a large of respondents in a short time. Interviewers can code data directly into computers, which can later compile the data.
- High response rate: Telephone interviews provide access to people who might be unlikely to reply to a mail questionnaire or refuse a personal interview.
- Quality: High quality data can be collected when interviewers are centrally located and supervisors can ensure that questions are being asked correctly and answers are recorded properly.

Disadvantages of Telephone interview

- Reluctant to discuss sensitive topics: Respondents may be resistant to discuss some issues over the phone.
- The ―broken off‖ interview: Respondents can terminate the interview before it is

completed.

- Less information Interviewers cannot provide supplemental information about the respondents‘ characteristics or environment.

SAE 2

1. Explain the 3 methods of data collection

3.6 SUMMARY

You now would be able to discern that a sample is a subset of a population selected to meet specific objectives. And also familiar with the guiding principle and sampling techniques in selecting a sample, is that it must, as far as possible have the essential characteristics of the target population

3.7 REFERENCES/FURTHER READINGS/WEB RESOURCES

Frankfort-Nachmias C. and Nachmias D. (2009).*Research in the Social Sociences*.(5th ed.).Hodder Education.

Gupta S.C (2011). *Fundamentals of Statistics*.(6th Rev. and Enlarged ed.). Mumbai India: Himalayan Publishing House

Esan E. O. and Okafor R. O. (1995) *Basic Statistical Methods*,(1st ed.). Lagos Nigeria: JAS Publishers, pages 72-89

3.8 Possible Answers to Self-Assessment Exercise(s) within the content

SAE 1

1. (i) Convenience samples
   (ii) Purposive samples,
   (iii) Quota samples
2. (i) Simple random sampling
   (ii) Systematic sampling
   (iii) Stratified sampling
   (iv) Cluster sampling.

SAE 2

1. (i) mail questionnaires
   (ii) Personal interviews
   (iii) Telephone interviews.