NATIONAL OPEN UNI VERSITY OF NIGERIA

SCHOOL OF MANAGEMENT SCIENCE

COURSE CODE: ECO 203

COURSE TITLE: STATISTICS FOR ECONOMICS

COURSE GUIDE

ECO 203
STATISTICS FOR ECONOMICS

Course Team Okojie, Daniel Esene (Course Developer) - UNILAG

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island, Lagos

Abuja Office
5 Dar es Salaam Street
Off Aminu Kano Crescent
Wuse II, Abuja

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

Published by
National Open University of Nigeria

Printed 2014

ISBN: 978-058- 203-9

CONTENTS **PAGE**

Introduction……... iv
Course Outline…….. iv
Course Aim……... iv
Course Objectives……. v
Working through the Course……. v
Course Materials…… vi
Study Units…… vi
Textbooks and Reference Resources…… viii
Assignment Folder…… viii
Presentation Plan……... viii
Assessment…… ix
Tutor-Marked Assignments (TMAs)…… ix
Final Examination and Grading…… xi
Marking Scheme……... xi
Course Overview…….. xi
How to Get the Most from this Course…… xiii
Tutors and Tutorials…….. xiii
Summary……... xiii

INTRODUCTION

This course (**ECO 203**) focuses progressively on elementary understanding of distribution functions and other inferential statistical techniques. It 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 you are required to perform several statistical analyses on your own.

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.

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 economics' 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 OUTLINE

ECO 203 is made up of 25 units spread across twelve lectures weeks and covering areas such as Normal, Binomial and Poisson Distributions, Correlation and Regression Analysis, Estimation Theory which will introduce you to the different discrete nature and models. Also to be outlined is test of statistical hypothesis using t, f and chi-square. 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 AIM

The aims of this course are to give you thorough understanding and an appreciative importance of statistics in the study of economics as you are becoming an Economist. It is almost impossible to find a problem which does not require a general use of statistical data. Important phenomena in all branches of economics can be described, compared and correlated with the help of Statistics with examples to demonstrate its applications in business and economics. There will be a strong

emphasis on the concepts and application of tests analysis methods, random variables, distributions, sampling theory, statistical inference, correlation and regression.

Others are statistical inference techniques such as estimation and significance testing are important in the fitting and interpretation of econometric models. Correlation and regression analysis are essential tools for measuring relationships between variables and for prediction.

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 is 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
- perform basic data manipulation and hypothesis testing
- state statistical estimation of economic relationships
- apply correlation and regression analyses models to data
- discuss 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 (SAEs). 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 12 weeks to complete and some components of the course are outlined below:

COURSE MATERIALS

The major components of the course are:

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

STUDY UNIT

This course consist five modules that are subdivided into 20 units which should be studied diligently and with utmost care. They are as follows:

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

Module 3 Correlation and Regression Coefficient Analyses

Unit 1 Pearson's Correlation Coefficient
Unit 2 Spearman's Rank Correlation Coefficient
Unit 3 Methods of Curve and Eye Fitting of Scattered Plot and the Least Square Regression Line
Unit 4 Forecasting in Regression

Module 4 Introduction to the Central Limit Theory (CLT)

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 and Beyond the Classical Framework

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

Unit 1 Index Number
Unit 2 Statistical Data
Unit 3 Sample and Sampling Techniques

Module 1 (units 1- 4) presents you with the common probability distribution functions as a general background of the course, the discreteness of Bernoulli, Binomial and Poisson distributions and the continuous natures of Normal distribution are discussed.

Module 2 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 treated in Module 3. This module explores Pearson's Correlation Coefficient, Spearman's Rank Correlation Coefficient, the Least Square Regression Line and Forecasting in Regression. Module 4 gives a detailed description of an introduction to central limit theory (CLT). CL theorems for independent sequences, dependent processes and the relation to law of large numbers are brought to your knowledge in this module. Also, extensions to the theorem and beyond the classical framework are presented.While basic concepts and notation of elementary Index Numbers and Introduction to Research Methods in Social Sciences are in Units 2 and 3 of Module 5. Module 5 has present in it: Index Number, Statistical Data, and Sample and 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.

TEXTBOOKS AND REFERENCES

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

Adebayo, O. A. (2006). ***Understanding Statistics.*** (5^{th} ed.). Lagos, Nigeria: JAS Publisher.

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

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

Dowling, Edward T. (2001). *Mathematical Economics.* (2^{nd} ed.). Schaum Outline Series.

Esan, E. O. & Okafor, R. O. (n.d.). *Basic Statistical Methods***.** Lagos**,** Nigeria**:** JAS Publishers.

ASSIGNMENT FOLDER

There are assignments on this course and you are expected to do all of them by following the schedule prescribed for them in terms of when to attempt them and submit same for grading by your facilitator. The marks you obtain for these assignments will count toward the final mark you obtain for this course. Further information on assignments will be found in the Assignment File itself and later in this Course Guide in the section on Assessment.

PRESENTATION SCHEDULE

The presentation schedule included in your course materials gives you the important dates 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.

TUTOR-MARKED ASSIGNMENT (TMA)

You are expected to submit all the assignments. You are encouraged to work all the questions thoroughly. The TMAs constitute 30% of the total score.

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

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

Unit	**Unit Title**	**Week's Activity**	**Assessment (end of unit)**
	Course Guide		
	Index Numbers and Introduction to Research Methods in Social Sciences		
18	Index Number	Week 18	
19	Statistical Data	Week 19	
20	Sample and Sampling Techniques	Week 20	Assignment 5
	Total	**20 Weeks**	

			Examination
5	T-test	Week 5	
6	F-test	Week 6	
7	Chi square test	Week 7	
8	ANOVA	Week 8	
9	Parametric and Non-Parametric test methods	Week 9	Assignment 2
	Correlation and Regression Coefficient Analysis		
10	Pearson's Correlation Coefficient	Week 10	
11	Spearman's Rank Correlation Coefficient	Week 11	
12	The Least Square Regression Line	Week12	
13	Forecasting in Regression	Week 13	Assignment 3
	Introduction to the Central Limit Theory (CLT)		
14	Central Limit Theorems for Independent Sequences	Week 14	
15	Central Limit Theorems for dependent Processes	Week 15	
16	Relation to the law of large numbers	Week 16	
17	Extensions to the theorem & Beyond the classical framework	Week 17	Assignment 4
	Index Numbers and Introduction to Research Methods in Social Sciences		
18	Index Number	Week 18	
19	Statistical Data	Week 19	
20	Sample and Sampling Techniques	Week 20	Assignment 5
	Total	**20 Weeks**	
			Examination

FINAL EXAMINATION AND GRADING

At the end of the course, you are expected to sit for a final examination. The final examination grade is 70% while the remaining 30% is taken from your scores in the TMAs. Naturally, the final examination questions will be taken from the materials you have already read and digested in the various study units. So, you need to do a proper revision and preparation to pass your final examination very well.

COURSE OVERVIEW

The table presented below indicates the units, number of weeks and assignments to be taken by you to successfully complete the course, Statistics for Economics (ECO 203).

HOW TO GET THE BEST FROM THIS COURSE

The distance learning system of education is quite different from the traditional or conventional learning system. Here, the prepared study texts replace the lecturers, thus providing you with a unique advantage. For instance, you can read and work through the specially designed study materials at your own pace and at a time and place you find suitable to you.

You should understand from the beginning that the contents of the course are to be worked on carefully and thoroughly understood. Step by step approach is recommended. You can read over a unit quickly y to see the general run of the contents and then return to it the second time more carefully. You should be prepared to spend a little more time on the units that prove more difficult. Always have a paper and pencil by you to make notes later on and this is why the use of pencil (not pen or biro) is recommended.

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. Organise 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, both 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).

TUTORS AND TUTORIALS

There are some hours of tutorials (2-hours 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.

SUMMARY

For a successful conclusion of the course, you would have developed critical thinking skills with the material necessary for efficient understanding of applied statistics. Nonetheless, in order to achieve a lot more from the course please try to apply anything you learn in the course to analysing of data for a better presentation and interpretation of findings in any assignment given both in your academic programme and other spheres of life. We wish you the very best in your studies.

MAIN COURSE

CONTENTS **PAGE**

Module 1	**Probability and Statistic Distribution Functions……...**	**1**
Unit 1	Bernoulli Distribution……...	1
Unit 2	Binomial Distribution……	8
Unit 3	Normal Distribution……..	16
Unit 4	Poisson Distribution…….	23
Module 2	**Statistical Hypothesis Test……....**	**27**
Unit 1	T- test…….	27
Unit 2	F- test…….	33
Unit 3	Chi square test……...	38
Unit 4	ANOVA…….	50
Unit 5	Parametric and Non-Parametric test Methods…….	64
Module 3	**Correlation and Regression Coefficient Analyses……**	**74**
Unit 1	Pearson's Correlation Coefficient……	74
Unit 2	Spearman's Rank Correlation Coefficient	82
Unit 3	Methods of Curve and Eye Fitting of Scattered Plot and the Least Square Regression Line……	88
Unit 4	Forecasting in Regression…….	94
Module 4	**Introduction to the Central Limit Theory (CLT)……...**	**99**
Unit 1	Central Limit Theorems for Independent Sequences……..	99
Unit 2	Central Limit Theorems for dependent Processes…….....**	108
Unit 3	Relation to the law of large numbers……	112
Unit 4	Extensions to the theorem and beyond the Classical Framework……	117

Module 5	**Index Numbers and Introduction to Research Methods in Social Sciences….**	**122**
Unit 1	Index Number……..	122
Unit 2	Statistical Data…….	131
Unit 3	Sample and Sampling Techniques……..	136

MODULE 1 PROBABILITY AND STATISTICAL DISTRIBUTION FUNCTIONS

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

UNIT 1 BERNOULLI DISTRIBUTION

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Bernoulli Process
    3.2 Interpretation of Values
    3.3 Bernoulli Distribution
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

This module intends to provide a broad understanding of the topic Bernoulli Distribution which is preparatory to the more widely used Binomial Distribution. The focus here is to provide learners' with the common probability distribution functions as a general background to the course.

2.0 OBJECTIVES

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

- explain the Bernoulli process
- explain what is meant by Bernoulli scheme
- interprete the Bernoulli process
- solve problems with respect to Bernoulli distribution.

3.0 MAIN CONTENT

3.1 Bernoulli Process

A Bernoulli process is a finite or infinite sequence of binary random variable. 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. All the variables have the same Bernoulli distribution. Much of what can be said about the Bernoulli process can also be generalised to more than two outcomes (such as the process for a six-sided die); this generalisation 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.

3.2 Interpretation of Values

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 as follows:

- 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 formalised 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 $\Omega = 2^N = \{H, T\}^N$ or the two-sided set $\Omega = 2^Z$. 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 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 $\{p, 1-p\}$, then one can define a natural measure on the product space, given by $P = \{p, 1-p\}^N$(or by $P = \{p, 1-p\}^Z$ 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; $P([w_1, w_2, w_3,.....w_n]) = p^k(1-p)^{n-k}$

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 given as:

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

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 $\lim_{n\to\infty} P^n = 0$ for any $0 \leq P \leq 1$. One may say 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, (Ω, F, P) (as defined above).

3.3 Bernoulli Distribution

In probability theory and statistics, the Bernoulli distribution, named after Swiss scientist Jacob Bernoulli, is a discrete probability, which takes the 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:

$$\Pr[X = 1] = 1 - \Pr[X = 0] = 1 - q = p$$

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:

$$f(k;p) = \begin{cases} p & if\ k = 1 \\ 1 - p & if\ k = 0 \end{cases}$$

This can also be expressed as:

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

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-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 $0 \leq P \leq 1$ form an exponential family.
The maximum likelihood estimator based on a random sample is the sample mean.

Further explanation

A Bernoulli random variable is one which has only 0 and 1 as possible value.

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 $X \sim B(1, p)$

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:

$$\mu_X = 0(1-p) + 1p = p$$
$$\sigma_X^2 = 0^2(1-p) + 1^2 p - p^2$$
$$= p - p^2 = p(1-p)$$

Solved Examples

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

Solution:

Then $X \sim B(1, \frac{1}{3})$

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 $p = \frac{1}{2}$. Here, *X* denotes the number of heads obtained in the toss.

=> Probability of success $= \frac{1}{2}$ and the probability of failure$= \frac{1}{2}$.

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 $= \frac{1}{6}$and the probability of failure $= \frac{5}{6}$

4.0 SUMMARY

In this unit, you have learnt the essentials and applications of Bernoulli distribution. Also, by now you would have been able to identify Bernoulli distribution function problems and solve them accordingly.

5.0 CONCLUSION

In conclusion, the Bernoulli, distribution is a discrete distribution whose corresponding random variables assume only integer values; 0 and 1. It does not assume value such as 1.5. The Bernoulli distribution as an example of a discrete probability distribution is an appropriate tool in the analysis of proportions and rates.

6.0 TUTOR-MARKED ASSIGNMENT

1. If a student blindly guesses the answer to a multiple choice test questions taken, the test has 10 questions, each of which has 4 possible answers (only one correct). Do the questions form a sequence of Bernoulli trials? If so, identify the trial outcomes and the parameter *p*.
2. An American roulette wheel has 38 slots; 18 are red, 18 are black, and 2 are green. A gambler plays roulette 15 times, betting on red each time. Do the outcomes form a sequence of Bernoulli trials? If so, identify the trial outcomes and the parameter *p*.

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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. & Nelder, J. (1989). *Generalized Linear Models,* (2nd ed.). Boca Raton: Chapman and Hall/CRC. Web site:http://en.wikipedia.org/wiki/Bernoulli_distribution.

Johnson, N. L., Kotz, S. & Kemp, A. (1993).*Univariate Discrete Distributions* (2nd ed.). Wiley. Website: http://en.wikipedia.org/wiki/Bernoulli_distribution.

UNIT 2 BINOMIAL DISTRIBUTION

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Probability Density Function of Binomial Distribution
    3.2 The Mean of a Binomial Distribution
    3.3 Variance of Binomial Distribution
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

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.

2.0 OBJECTIVES

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

- explain the meaning of Binomial distribution
- apply Binomial distribution to instances when it is applicable.

3.0 MAIN CONTENT

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:

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

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 ... (2)
The probability of r successes and consequently (n-r) failures in 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:

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

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}
… (3)

But in n trials, the total number of possible ways of obtaining r successes and (n-r) failure is $\frac{n!}{r!(n-r)!} = {}^nC_r$,
all of which are mutually disjoint. The probability for each of these ${}^nC_r$ 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:

$$P(X=r) = P^r q^{n-r} + P^r q^{n-r} + ...+ p^r q^{n-r} \ ({}^nC_r \text{ terms})$$
$$= {}^nC_r P^r q^{n-r}; \ r = 0, 1, 2.....n$$

Table 1.2: Binomial Probabilities

R	***P*(r) = *P*(X = r)**
0	${}^nC_0 P^0 q^n = q^n$
1	${}^nC_1 P^0 q^{n-1}$
2	${}^nC_2 P^2 q^{n-2}$
.	.
.	.
.	.
.	.
N	${}^nC_0 P^0 q^n = P^n$

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;

$$\sum_{r=0}^{n} p(r) = p(0) + p(1) + \ldots\ldots\ldots\ldots\ldots\ldots p(n)$$
$$= q^n + {}^nC_1 q^{n-1} p + {}^nC_2 q^{n-1} P^2 + ...P^n$$
$$= (q+p)^n = 1 \quad Therefore, p+q = 1$$

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 ${}^nC_r = {}^nC_{n-r}$, 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 $(1+x)n = {}^nC_0 + {}^nC_1 + {}^nC_2 + \ldots\ldots\ldots + {}^nC_n = 2^n$
Therefore, the sum of binomial coefficients is 2^n

3.2 The Mean of a Binomial Distribution

Mean $= \Sigma rp(r) = {}^nC_1 q^{n-1} p + 2\ {}^nC_2 q^{n-2}.p^2 + 3\ {}^nC_3 q^{n-3}.p^3 + ... + np^n$

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

$= np[q^{n-1}+(n-1)q^{n-2}p+\frac{(n-1)(n-2)}{2!}q^{n-3}p^2+.....+p^{n-1}]$
$= np[q^{n-1}+ {}^{n-1}C_1q^{n-2}p+ {}^{n-1}C_2q^{n-3}p^2+ .....P^{n-1}]$
$= np\ (q+p)^{n-1}$ *(By Binomial expansion for positive integer index),*
Therefore, $p+q = 1$

Therefore, ***Mean = np***

3.3 Variance of a Binomial

Variance = $\Sigma r^2p(r)- [\Sigma rp(r)]^2 = \Sigma r^2p(r) – (mean)^2$ .....(*)

$\Sigma r^2p(r)=1^2X\ {}^nC_1q^{n-1}p + 2^{2n}C_2q^{n-2}.p^2 + 3^{2n}C_3q^{n-3}.p^3+.....+n^2p^n$
$=nq^{n-1}p+\frac{4n(n-1)}{2}q^{n-2}p^2+\frac{9n(n-1)(n-2)}{3!}q^{n-3}p^3+.....+n^2p^n$
$= np[q^{n-1}+2(n-1)q^{n-2}p+{}^3/_2(n-1)(n-2)q^{n-3}p^2+.....+np^{n-1}]$
$= np[\{q^{n-1}+(n-1)q^{n-2}p+\frac{(n-1)(n-2)}{2}q^{n-3}p^2+.....+1p^{n-1}\}$
$+\{(n-1)q^{n-2}p+(n-1)(n-2)q^{n-3}p^2+.....+(n-1)P^{n-1}\}]$
$=np[\{(q+p)^{n-1}+(n-1)p\{q^{n-2}+(n-2)q^{n-3}p+.....p^{n-2}\}]$
$=np[(q+p)^{n-1}+(n-1)p(q+p)^{n-2}]$
$=np[1+(n-1)p]$

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

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

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

(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 ${}^{\mathrm{n}}\mathrm{C}_{\mathrm{r}} = \frac{n!}{r!(n-r)!}$,

Therefore, ${}^{10}\mathrm{C}_6 = \frac{10!}{6!(10-6)!}$

$\frac{7X8X9X10}{1X2X3X4} = 210$

$p(\text{exactly } 6\ heads) = 210 \cdot (½)^6 \cdot (½)^4$

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

$= \frac{210}{1024}$

p (exactly *6 heads*) $= \frac{105}{512}$

(ii) Probability of at least eight heads = $P(X\geq 8) = p(8)+P(9)+p(10)$

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) $= {}^{10}C_8 P^8 q^{10-8}$

$= \frac{10!}{8!(10-8)!} (½)^8 \cdot (½)^2$

$= 45 \text{ X } {}^{1}/_{256} \text{ X } ¼$

$= \frac{45}{1024}$

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

$= \frac{10!}{9!(10-9)!} (½)^9 \cdot (½)^1$

$= 10 \text{ X } \frac{1}{1024}$

$= \frac{10}{1024}$

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

$= \frac{1}{1024}$

Therefore, Probability of at least 8 heads $= \frac{45}{1024} + \frac{10}{1024} + \frac{1}{1024}$

$= \frac{56}{1024}$

P(at least 8 heads) $= \frac{7}{128}$

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

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

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

$= 1 \text{ X } 1 \text{ X } \frac{1}{1024}$

$P(0 \text{ head}) = \frac{1}{1024}$

(iv) Probability of at least one head

$= 1 - P[\text{No head}]$

$= 1 - P(0)$

Recall that $P(0) = \frac{1}{1024}$

$= 1 - \frac{1}{1024}$

$= 1 - \frac{1}{1024}$

$= \frac{1023}{1024}$

(v) Probability of not more than three heads

$= P(X \leq 3) = P(0) + P(1) + P(2) + P(3)$

$= \frac{1}{1024}[{}^{10}C_0 + {}^{10}C_1 + {}^{10}C_2 + {}^{10}C_3] = \frac{1+10+45+120}{1024}$

$= \frac{176}{1024} = \frac{11}{64}$

(vi) Probability (at least 4 heads) $= (X \geq 4) = 1 - P(X \leq 3)$

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

$1 - \frac{11}{64}$

$= \frac{53}{64}$

4.0 SUMMARY

In this unit, you have learnt that a Binomial Distribution is the sum of Independent Bernoulli Random Variables and that the Binomial distribution describes the distribution 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).*

5.0 CONCLUSION

In probability theory and statistics, the Binomial distribution is the discrete probability distribution of the number of successes in a sequence of *n* independent yes/no experiments, each of which yields success with probability *p*. Such a success/failure experiment is also called a Bernoulli experiment or Bernoulli trial; when $n = 1$, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance. The Binomial distribution is frequently used to model the number of successes in a sample of size *n* drawn with replacement from a population of size *N*. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hyper-geometric distribution, not a binomial one. However, for *N* much larger than *n*, the binomial distribution is a good approximation, and widely used.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define Binomial Distribution. What is the probability of guessing correctly at least six of the ten answers in a TRUE-FALSE objective test?
2. A merchant's file of 20 accounts contains 6 delinquent and 14 non-delinquent accounts. An auditor randomly selects 5 of these accounts for examination.
    a. What is the probability that the auditor finds exactly 2 delinquent accounts?
    b. Find the expected number of delinquent accounts in the sample selected?
    c. What is the variance of the distribution?

7.0 REFERENCES/FURTHER READING

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

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

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

UNIT 3 NORMAL DISTRIBUTION

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Properties of Normal Distribution
    3.2 Relationship between Binomial and normal distribution
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

The Normal probability distribution commonly called the normal distribution is one of the most important continuous 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 in Figure 1.1.:

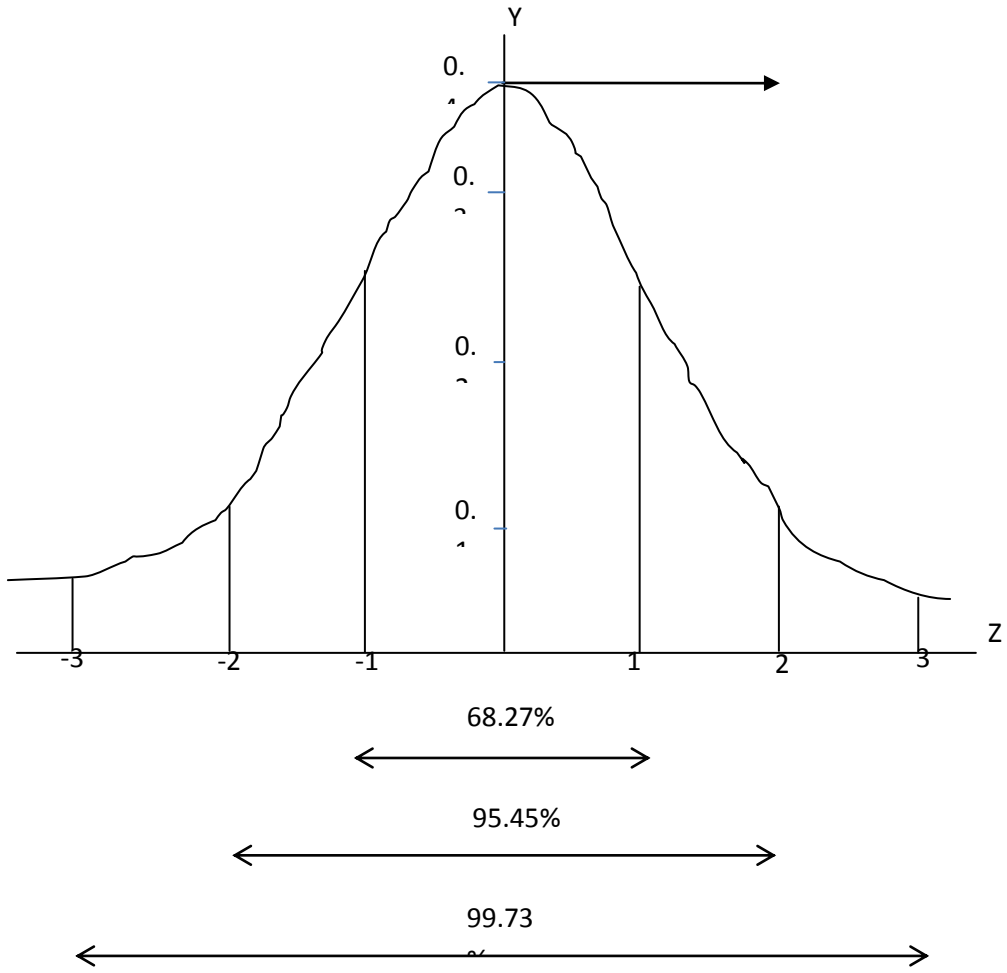

Fig. 1.1: Normal Curve

2.0 OBJECTIVES

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

- explain the meaning of normal distribution
- appreciate the applicability of normal distribution in day-to-day business and scientific live
- use normal distribution as appropriate in practical statistical studies

3.0 MAIN CONTENT

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.

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:

$$n(X,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty < x < \infty \quad or$$

$$n(X,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x-\mu^2}{2\sigma^2}}, -\infty < x < \infty \text{ and } \sigma > 0$$

Where π and e are constants given by $\pi = {}^{22}/_{7}$,$\sqrt{2\pi}$ =2.5066 and e = 2.71828 (which is the base system of Natural Logarithm)

2. The normal distribution with μ = 0 and σ = 1 is referred to as the standard normal distribution.
3. If X has a normal distribution with the mean of μ and the deviation σ, then $Z = \frac{x-\mu}{\sigma}$ is the standard normal distribution

Note:

(i) Definition 3 above 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.

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

Examples

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) $Z = \frac{x-\mu}{\sigma}$
X = 280, μ= 230, σ = 20
Therefore, Z = $\frac{280-230}{20}$
=2.50
Therefore, P(X > 280) = P(z > 2.50)
= 1 – P(z < 2.50)
= 1 – 0.9938
= 0.0062

(b) P(X < 220) $Z = \frac{x-\mu}{\sigma}$
Z = $\frac{220-230}{20}$
= -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

3.2 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* $\infty \longrightarrow$
(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 that is:

$$Z = \frac{X-E(X)}{\sigma_x} = \frac{X-np}{\sqrt{npq}}$$

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, 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 $n \longrightarrow \infty$.

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

4.0 SUMMARY

In summary, you would have understood that the normal distribution can be described completely by the two parameter $\mu and \sigma$s.The mean is the center of the distribution and the standard deviation is the measure of the variation around the mean.

5.0 CONCLUSION

The normal distribution is the most important distribution. It describes well the distribution of random variables that arise in practice, such as the heights or weights of people, the total annual sales of a firm, exam scores etc. Also, it is important for the central limit theorem, the approximation of other distributions such as the binomial, etc.

6.0 TUTOR-MARKED ASSIGNMENT

1. Suppose the monthly salaries of 500 parents of students in a senior school are normally distributed with mean ₦63,000 and standard deviation of ₦7,500. Find:

    a. The probability of a parent's salary lying between ₦59,500 and ₦66,000.
    b. Probability of a parent's salary greater than or equal to ₦65,000.
    c. Find the number of parents with monthly salaries between ₦62,000 and ₦67,000.

2. Time taken by the crew of a company to construct a small bridge is a normal variate with a mean of 400 labour hours and a standard deviation of 100 labour hours.

   a. What is the probability that the bridge is constructed between 350 and 450 labour hours?
   b. If the company promises to construct a bridge in 450 labour hours or less and agrees to pay a penalty of ₦1,000 for each labour hour spent in excess of 450, what is the probability that the company pays a penalty of at least ₦20,000

3. The National Bureau for Economic Data is organising a stakeholder symposium on how to tackle the problem of dearth of reliable economic data in Nigeria. The Director of Accounts states that they can afford to entertain no more than 200 guests at the symposium, and the hotel at which it is to be held will only cater for a minimum of 140 guests. The Public Relation Unit of the Bureau is to send invitation to 240 people. The Statistics, Planning and Research unit of the Bureau estimates that the probability each individual accepts the invitation is about 70%. Using this model, estimate the probability that between 140 and 200 guests accept the invitation. Is 240 a good number of invitations to send?

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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.

UNIT 4 POISSON DISTRIBUTION

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Condition for using Poisson Distribution
    3.2 Application of Poisson Distribution
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 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:

$$P(r) = P(X = r) = \frac{e^{-\mu}.\ \mu^{x}}{x!}\ , r = 0, 1, 2, 3, \ldots\ldots$$

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)

2.0 OBJECTIVES

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

- explain the concept of Poisson distribution
- State the importance of Poisson distribution
- apply Poisson distribution
- link Poisson distribution and Binomial distribution
- state the distinguishing features of both Poisson distribution and Binomial distribution.

3.0 MAIN CONTENT

3.1 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.

3.2 Application of Poisson distribution

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

$\mu = 1.2$

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

(a) Pr (No misprints)

$= Pr(X=0)$

$$= \frac{e^{-1.2}.\ 1.2^{0}}{0!}$$

$= e^{-1.2}$

$= 0.301$

(b) Pr(or more misprint)

$= Pr(X \geq 3)$

$= 1 - [Pr(0) + Pr(1) + Pr(2)]$

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

$$Pr(1) = \frac{e^{-1.2}.\ 1.2^{0}}{1!}$$

$= 0.3612$

$$Pr(2) = \frac{e^{-1.2}.\ 1.2^{2}}{2!}$$

$= 0.21672$

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

Therefore, $Pr(X \geq 3) = 1 - 0.87892$

$= 0.12108$

$= 0.121$

4.0 SUMMARY

In this unit, you have learnt the rudiments and applications of Poisson distribution. You have equally learnt how to solve problems using Poisson distribution.

5.0 CONCLUSION

In conclusion, Poisson distribution is a limiting case of Binomial distribution, it can be applied in cases when the number is very large tending towards infinity and the probability of success is very low.

6.0 TUTOR-MARKED ASSIGNMENT

1. The number of customers asking for a particular expensive commodity each day in a local supermarket has a mean of 2. On a particular day, what is the probability that:
   (a) No customer asked for the commodity.
   (b) Exactly one customer asked for the commodity.
   (c) Exactly two customers asked for the commodity.
   (d) More than two customers asked for the commodity.

2. Between the hours of 10a.m and 12 noon, the average number of phone calls per minute coming into the switch board of a company is 2.35. Find the probability that during one particular minute, there will be at most 2 phone calls.

3. The average number of customers who appear at a counter of a certain bank per minute is two. Find the probability that during a given time:
   (a). No customer appears.
   (b).Three or more customers appear.

4. Suppose the diameter of a certain car component follows the normal distribution with $X \sim N(10; 3)$. Find the proportion of these components that have diameter larger than 13.4 mm. Or, if we randomly select one of these components, find the probability that its diameter will be larger than 13.4 mm.

5. A manufacturing process produces semiconductor chips with a known failure rate 6:3%. Assume that chip failures are independent of one another. You will be producing 2000 chips tomorrow.
   (a) Find the expected number of defective chips produced.
   (b) Find the standard deviation of the number of defective chips.
   (c) Find the probability (approximate) that you will produce less than 135 defects.

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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.

MODULE 2 STATISTICAL HYPOTHESIS TEST

A **statistical hypothesis test** is a method of making decisions using data from a scientific study. In statistics, a result is interpreted as being statistically 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 analysing 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 analysed 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

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
  3.1 Application of *t*-distribution
  3.2 Test for Single Mean
  3.3 Assumptions for Student's Test
  3.4 t-Test for Difference of Means
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

For large sample test for mean $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), asymptotically$

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 σ^2 is given by:

$$S^2 = \frac{1}{n-1}\sum(\bar{x} - x)^2 \rightarrow ns^2 = (n-1)S^2$$

It is quite conventional to replace σ^2 byS^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 $\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$ 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.....x_n is a random sample of size *n* from a normal population with mean μ and variance σ^2 then the Student's *t* statistic is defined as:

$$t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\bar{x}-\mu}{\sqrt{\frac{S^2}{n}}}$$

Where$\bar{x} = \frac{\Sigma x}{n}$ is the sample mean and $S^2 = \frac{1}{n-1}\sum(\bar{x} - x)^2$ is an unbiased estimate of the population variance σ^2

2.0 OBJECTIVES

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

- solve problems using *t*-distribution
- emphasize *t*-distribution application in statistics.

3.0 MAIN CONTENT

3.1 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.

3.2 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 $\bar{x}$ 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 $\bar{x}$ 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. Under H_o the test-statistic is:

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

Where $\bar{x} = \frac{\Sigma x}{n}$ and $S^2 = \frac{1}{n-1}\sum(\bar{x} - x)^2$

And it follows Student's t-distribution with (n-1) degrees of freedom.

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.

3.3 Assumptions of 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 σ is unknown

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, $\overline{x}$= 11.8kg, s = 0.15kg

Null hypothesis, H_o : μ = 12 kg (i.e. the sample mean of $\bar{x}$ =11.8 kg does not differ significantly from the population mean μ = 12 kg.
Alternative Hypothesis. Ho: ≠ 12kg (Two tailed)

$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{S^2}{n}}} = \frac{\bar{x}-\mu}{\sqrt{\frac{S^2}{n-1}}} \sim t_{n-1} = t_9$$

$$t = \frac{11.8-12}{0.15\sqrt{9}} = \frac{-0.2\ X3}{0.15} - 4.0$$

The tabulated value of t for 9 d.f at 5% level of significance is 2.26. Since 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.

3.4 T-Test for Difference of Means

Assume we are interested in testing if two independent samples have been drawn from two normal populations having the same means, the population variances being equal.

Let $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ and $y_1, y_2 \ldots\ldots\ldots\ldots y_n$ be two independent random samples from the given normal populations.

H_o: $\mu_x = \mu_y$ i.e the two samples have been drawn from the normal populations with the same means. Under the hypothesis that the $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e population variances are equal but unknown, the test statistic under H_o is:

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Where $\bar{x} = \frac{1}{n_1}\sum x, \quad \bar{y} = \frac{1}{n_2}\sum y$

And $S^2 = \frac{1}{n_1+n_2-2}[\sum(\bar{x} - x)^2 + \sum(\bar{y} - y)^2]$

is an unbiased estimate of the common population variance σ^2 based on both the samples. By comparing the computed value of t with the tabulated value of t for $n_1 + n_2 - 2$ *d.f.* and at desired level of significance, usually 5% or 1%, we reject the null hypothesis.

Example: The nicotine content in milligram of two samples of tobacco were found to be as follows:

Sample A:	24	27	26	21	25	
Sample B:	27	30	28	31	22	36

Can it be said that the two samples come from the same normal population having the same mean?

Solution

Hints: Applying the above formula and calculating the variance as appropriate, the calculated t-value is -1.92. The tabulated value for 9 d.f at 5% level of significance for two-tailed test is 2.262. Since calculated t is less than the tabulated t, it is not significant and the null hypothesis is accepted.

4.0 SUMMARY

In summary, you 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.

5.0 CONCLUSION

T-test has very wide applications. It can be applied in the tests of single mean, in the comparison of two different means and in the test of significance of other parameter estimates..

6.0 TUTOR-MARKED ASSIGNMENT

1. The mean weekly sale of the chocolate bar in candy stores was 146.3 bars per store. After advertising campaign the mean weekly sales in 22 stores for typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?
2. Prices of shares of a company on the different days in a month were found to be: 66, 65, 69, 70, 69, 71, 70, 63, 64 and 68. Discuss whether the mean price of the price of the shares in the month is 65.
3. Two salesmen A and B are working in certain district. From a Sample Survey Conducted by the Head Office the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen.

	A	B
No. of sales	20	18
Average sales (in '000₦)	170	205
Average sales (in '000₦)	20	25

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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.

UNIT 2 F-TEST

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 F-Test
    3.2 Applications of the F-Distribution
    3.3 For testing Equality of Population Variances
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

F-Test is a test of whether a sample of observations comes from a larger sample with a standard distribution of statistical properties.

2.0 OBJECTIVES

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

- explain F-distribution
- state F-distribution theories
- apply F-distribution theories to day-to-day business and economic problems.

3.0 MAIN CONTENT

3.1 F-Test

In F-TEST, If X is a χ^2-variate with n_1 degree of freedom and Y is an independent χ^2-variate with n_2 degree of freedom, then *F*-statistic is defined as:

$$F = \frac{X/n_1}{Y/n_2}$$

i.e. F-statistic is the ratio of two independent chi-square variates divided by their respective degrees of freedom. The statistic follows G.W Snedecor's *F*-distribution with (n_1, n_2)degree of freedom with probability density function given by:

$$(F) = y_o . \frac{F^{\frac{n_1}{2}-1}}{(1+\frac{n_1}{n_2}F)^{\frac{n_1+n_2}{2}}} ; 0 \leq F < \infty$$

Where y_o is a constant which is so determined that total area under the probability curves is 1 i.e $\int_0^\infty p(F)dF = 1.$ $This\ gives:\ y_o = \frac{(\frac{n_1}{n_2})^{n\ \ 1/2}}{\beta(\frac{n\ 1}{2}, \frac{n\ 2}{2})}$

Note: The sampling distribution of F-statistics does not involve any population parameters and depends only on the degrees of freedom n_1 and n_2. The graph of the function p(F) varies with the degree of freedom n_1 and n_2.

Critical values of F-distribution: The available *F*-tables in most standard statistical table give the critical values of *F* for the right-tailed test, i.e the critical region is determined by the right tail areas. Thus, the significant value Fα (v_1,v_2) at level of significance α and (v_1,v_2) *d.f. 2* is determined by the equation:

$$P\ [F > F\alpha\ (\mathrm{v}_1, \mathrm{v}_2)] = \alpha$$

Significant values of the variance-Ratio $F = \frac{S_1^2}{S_2^2}\ ;\ S_1^2 > S_2^2$

3.2 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 test 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.

3.3 For Testing Equality of Population Variances

Here, we set up the Null hypothesis Ho: $\sigma_1 = \sigma_2 = \sigma$ i.e. population variances are the same. In other words, H_o is that the two independent estimates of the common population variance do not differ significantly.

Under Ho, the test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1),$$

whereS_1^2 and $S_2^2$are unbiased estimates of the common population variance σ^2 and are given by:

$$S_1^2 = \frac{1}{n_1 - 1}\sum(x - \bar{x})^2 \quad and \quad S_2^2 = \frac{1}{n_2 - 1}\sum(y - \bar{y})^2$$

and it follows Snedecor's F-distribution with $v_1 = n_1 - 1$, $v_2 = n_2 - 1$ *d.f.; i.e* $F \sim F(v_1, v_2)$

Since F-test is based on the ratio of two variances, it is also known as variance ratio test.

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 $H_o : \sigma_1^2 > \sigma_2^2$, in numerical problems we will take greater of the variances S_1^2 or S_2^2 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 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 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 S_1^2 or S_2^2 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 "α" will be right tailed tests only with area "α" 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.

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**

$$S_1^2 = \frac{1}{n_1-1}\sum(x - \bar{x})^2 = \frac{1}{n_1-1}\left[\sum d^2 - \frac{(\sum d)^2}{n_1}\right]$$
$$= \frac{1}{5}\left[82 - \frac{4}{6}\right] = \frac{1}{5}[82 - 0.67] = 16.266$$
$$S_2^2 = \frac{1}{n_2-1}\sum(y - \bar{y})^2 = \frac{1}{n_2-1}\left[\sum D^2 - \frac{(\sum d)^2}{n_1}\right]$$
$$= \frac{1}{6}\left[136 - \frac{16}{7}\right] = \frac{1}{6}[136 - 2.286] = 22.286$$

Since $S_2^2 > S_1^2$, under Ho, the test statistic is

$$F = \frac{S_2^2}{S_1^2} \sim F(n_1 - 1, n_2 - 1) = F(6,5)$$
$$F = \frac{22.286}{16.266} = 1.37$$

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.

4.0 SUMMARY

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

5.0 CONCLUSION

F-test can be used to test the equality of several population variances, several population means, and overall significance of a regression model.

6.0 TUTOR-MARKED ASSIGNMENT

Can the following two samples be regarded as coming from the same normal population?

Sample	**Size**	**Sample Mean**	**Sum of squares of deviation from the mean**
1	10	12	120
2	12	15	314

7.0 REFERENCE/ FURTHER READING

Spiegel, M. R.& 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.

UNIT 3 CHI-SQUARE TEST

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 The Chi-Square
    3.2 Application of Chi-Square Distribution
    3.3 Chi-Squared Test of Goodness of Fit
    3.4 Steps for Computing χ^2 and Drawing Conclusions
    3.5 Chi-Square Test for Independence of Attributes
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

The chi-square test is a non-parametric inferential statistical method commonly used in the analysis of frequencies or nominal data. As a non-parametric statistic, it makes no restrictive assumptions about the distribution of scores. Because of this, the method becomes useful in education and other behavioural sciences; particularly in the analysis of data in the form of frequencies or categories.

The chi-square is a two-tailed test. It can only indicate whether or not a set of observed frequencies differ significantly from the corresponding set of expected frequencies and not possibly the direction in which they differ.

2.0 OBJECTIVES

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

- identify the uses of chi-square (X2).statistics
- test the goodness-of-fit of any data
- test the independence of attributes.

3.0 MAIN CONTENT

3.1 The Chi-Square

The square of a standard normal variable is called a Chi-square variate with 1 degree of freedom, abbreviated as *d.f*. Thus if x is a random

variable following normal distribution with mean μ and standard deviation σ, then $(X- \mu)/ \sigma$ is a standard normal variate.

Therefore, $Z = (\frac{x-\mu}{\sigma})^2$ is a chi-square (abbreviated by the letter χ^2 of the Greek alphabet) variate with 1 *d.f.*

If X_1, X_2, X_3, ...X_v are v independent random variables following normal distribution with means μ_1, μ_2, μ_3,..... μ_v, and standard deviations σ_1, σ_2, σ_3,..... σ_v respectively then the variate

$$\chi^2 = (\frac{x_1-\mu_1}{\sigma 1})^2 + (\frac{x_2-\mu_2}{\sigma 2})^2 + \cdots \ldots \ldots (\frac{x_v-\mu_v}{\sigma v})^2 = \sum_{i=1}^{v} (\frac{x_1-\mu_1}{\sigma 1})^2$$

this is the sum of the squares of v independent standard normal variates, follow Chi-square distribution with v d.f.

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 σ^2
(iv) To test the equality of several population proportions.

Observed and Theoretical Frequencies

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

Definition of χ^2

A measure of discrepancy existing between the observed and expected frequencies is supplied by the statistics χ^2 given by

$$\chi^2 = \frac{(o_1-e_1)^2}{e_1} + \frac{(o_2-e_2)^2}{e_2} + \cdots \ldots \ldots \ldots \ldots \ldots \frac{(o_k-ek\ )^2}{e_k}$$

3.3 Chi-Square Test of Goodness of Fit

The chi-square test can be used to determine how well theoretical distributions (such 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

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$
$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots \ldots \ldots \ldots \ldots \ldots \ldots \frac{(O_n - En\ )^2}{E_n}$$

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 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 $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

(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.

Example 1: A pair of dice is rolled 500 times with the sums as given 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 $\chi_i^2 = (O_i - E_i)^2/E_i$

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 $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ 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.

Exercise: The following figures show the distribution of digits in numbers chosen at random from a telephone directory:

Digit	0	1	2	3	4	5	6	7	8	9	Total
Frequency	1,002	1,107	997	966	1,075	933	1,107	972	964	853	10,000

Test whether the digits may be taken to occur equally frequently in the directory. The table value of χ^2 for d.f at 5% level of significance is 16.92.

Hint: Set up the null hypothesis that the digits 0, 1, 2, 3, ...9 in the numbers in the telephone directory are uniformly distributed, i.e all digits occur equally frequently in the directory. Then, under the null hypothesis, the expected frequency for each of the digits 0, 1, 2, 3,.....9 is 10,000/10 = 1,000

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 / A	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)	$\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$

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]

$$= \mathrm{N}\ X\ \frac{(A_i)}{N}\ X\ \frac{(B_i)}{N} = \frac{(A_i)(B_i)}{N}$$

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

$$(A_iB_j)_o = \frac{(A_i)(B_i)}{N};\ (i = 1, 2, ...,r;\ j=1,2, ...s)$$

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

$$\chi^2 = \sum_i \sum_j \left[\frac{[(A_iB_i) - (AiBj)\mathrm{o}]^2}{(AiBj)\mathrm{o}}\right]$$

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 $(A_iB_j)_o = \frac{(A_i)(B_i)}{N};\ (i = 1, 2, ...,r;\ j=1,2, ...s)$ can be used for computing expected frequencies.

Example: 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_o): 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,
$$E_{11} = \frac{710\ X\ 400}{1000} = 284$$
$$E_{12} = \frac{710\ X\ 100}{1000} = 71$$
$$E_{13} = \frac{710\ X\ 200}{1000} = 142$$
$$E_{14} = \frac{710\ X\ 300}{1000} = 213$$
$$E_{21} = \frac{195\ X\ 400}{1000} = 78$$
$$E_{22} = \frac{195\ X\ 100}{1000} = 19.5$$
$$E_{23} = \frac{195\ X\ 200}{1000} = 39$$
$$E_{24} = \frac{195\ X\ 300}{1000} = 58.5$$
$$E_{31} = \frac{95\ X\ 400}{1000} = 38$$
$$E_{32} = \frac{95\ X100}{1000} = 9.5$$
$$E_{33} = \frac{95\ X200}{1000} = 19$$
$$E_{34} = \frac{95\ X300}{1000} = 28.5$$

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 = $\sum_i \sum_j [\frac{[(A_iB_i)-(AiBj)o]^2}{(AiBj)o}]$ = χ^2 = $\sum_i \sum_j \frac{(Oij-Eij)^2}{Eij}$

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

$$\chi^2_{11} = \frac{(320-284)^2}{284} = 4.56$$

$$\chi^2_{12} = \frac{(80-71)^2}{71} = 1.14$$

$$\chi^2_{13} = \frac{(110-142)^2}{142} = 7.21$$

$$\chi^2_{14} = \frac{(200-213)^2}{213} = 0.79$$

$$\chi^2_{21} = \frac{(50-78)^2}{78} = 10.05$$

$$\chi^2_{22} = \frac{(15-19.5)^2}{19.5} = 1.04$$

$$\chi^2_{23} = \frac{(70-39)^2}{39} = 24.64$$

$$\chi^2_{24} = \frac{(60-58.5)^2}{58.5} = 0.04$$

$$\chi^2_{31} = \frac{(30-38)^2}{38} = 1.68$$

$$\chi^2_{32} = \frac{(5-9.5)^2}{9.5} = 2.13$$

$$\chi^2_{33} = \frac{(20-19)^2}{19} = 0.05$$

$$\chi^2_{34} = \frac{(40-28.5)^2}{28.5} = 4.64$$

$\chi^2_{\text{calculated}} = \sum_i \sum_j \frac{(Oij-Eij)^2}{Eij}$ = 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, the *d.f* is (number of row minus one) X (number of column minus one)

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

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).

4.0 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.

5.0 CONCLUSION

In conclusion, chi-squared analysis has very wide applications which include test of independence of attributes; test of goodness fit; test of equality of population proportion and to test if population has a specified variance among others. This powerful statistical tool is useful in business and economic decision making.

6.0 TUTOR-MARKED ASSIGNMENT

1. A sample of students randomly selected from private high schools and sample of students randomly selected from public high schools were given standardized tests with the following results

Test Scores	**0-275**	**276 - 350**	**351 - 425**	**426 - 500**	**Total**
Private School	6	14	17	9	**46**
Public School	30	32	17	3	**86**
Total	**36**	**46**	**34**	**12**	**128**

H_0: The distribution of test scores is the same for private and public high school students at $\alpha=0.05$

2. A manufacturing company has just introduced a new product into the market. In order to assess consumers' acceptability of the product and make efforts towards improving its quality, a survey was carried out among the three major ethnic groups in Nigeria and the following results were obtained:

	Ethnic groups				
Persons	***Igbo***	***Yoruba***	***Hausa***	***Ijaw***	***Total***
Accept the product	48	76	56	70	**250**
Do not Accept	57	44	74	30	**205**
Total	**105**	**120**	**130**	**100**	**455**

Using the above information, does the acceptability of the product depend on the ethnic group of the respondents? (Take α=1%)

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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.

UNIT 4 ANALYSIS OF VARIANCE (ANOVA)

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 The Analysis of Variance (ANOVA)
    3.2 Assumption for ANOVA Test
    3.3 The One-Way Classification
    3.4 Bernoulli Distribution
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 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 ${}^4C_2 = 6$, 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.

2.0 OBJECTIVES

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

- state the assumptions of ANOVA
- state the theories of ANOVA
- apply the ANOVA to solve business and economic problems.

3.0 MAIN CONTENT

3.1 The Analysis of Variance (ANOVA)

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

3.2 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 assumptions 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 be dealt with in this course.

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.

3.3 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*th class 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*}

$n = n_1 + n_2 + \ldots\ldots\ldots\ldots + n_k = \sum_{i=1}^{k} n_i$

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

Class	***Sample observation***	***Total***	***Mean***
1	$X_{11}, X_{12}, \ldots\ldots\ldots X_{1n}$	T_1	**Mean** X_1
2	$X_{21}, X_{22}, \ldots\ldots\ldots X_{2n}$	T_2	**Mean** X_2
:	**: : : :**	**:**	**:**
:	**: : : :**	**:**	**:**
I	$X_{i1}, X_{i2}, \ldots\ldots\ldots X_{in}$	$T_i = \sum_{i=1}^{n} X_{ij}$	**Mean** $X_i = \frac{T_i}{n_i}$
:	**: : : :**	**:**	**:**
:	**: : : :**	**:**	**:**
K	$X_{k1}, X_{k2}, \ldots\ldots\ldots X_{kn}$	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.

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:

$$\overline{X_1} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \; ; \qquad S_i^2 = \frac{1}{n}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \; ; \; (i = 1,2, \ldots\ldots k)$$

Also, compute the mean $\bar{X}$ of all the data observations in the k-classes by the formula:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{\sum_i n_i \overline{X_i}}{\sum_i n_i}$$

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

$$\textbf{BSS} = n_1(\overline{X_1} - \bar{X})^2 + n_2(\overline{X_2} - \bar{X})^2 + \ldots\ldots\ldots\ldots\ldots + n_{3k}(\overline{X_k} - \bar{X})^2$$

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

$$\boldsymbol{MBSS} = \frac{Between\ classes\ Sum\ of\ Square}{Degrees\ of\ freedom} = \frac{BSS}{k-1}$$

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

$$\boldsymbol{WSS} = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k} n_i\, s_i^2 = n_1 s_1^2 + n_2 s_2^2 \ldots\ldots\ldots + n_k s_k^2$$

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

$$\boldsymbol{MBSS} = \frac{Within\ classes\ Sum\ of\ Square}{Degrees\ of\ freedom} = \frac{WSS}{n-k}$$

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

$$F = \frac{Between\ classes\ Mean\ Sum\ of\ Square}{Within\ classes\ Mean\ Sum\ of\ Square} = \frac{Step\ 4}{Step\ 6} \sim F(k-1, n-k)$$

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.

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 jth patient stays in the hospitals A, B and C respectively

Calculations for various Sum of Squares

X_{1j}	X_{2j}	X_{3j}	$(X_{1j} - \bar{X}_1)^2$	$(X_{2j} - \bar{X}_2)^2$	$(X_{3j} - \bar{X}_3)^2$
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**	**$\sum_{j=1}^{7}(X_{1j} - \bar{X}_1)^2$ =50.8572**	**$\sum_{j=1}^{7}(X_{2j} - \bar{X}_2)^2$ =38**	**=58.8572**

$\bar{X}_1 = \frac{\sum X_{1j}}{n_1} = \frac{41}{7} = 5.86$;

$\bar{X}_2 = \frac{\sum X_{2j}}{n_2} = \frac{35}{7} = 5$

$\bar{X}_3 = \frac{\sum X_{3j}}{n_3} = \frac{34}{7} = 4.86$

$\bar{X} = \frac{Grand\ Total}{Total\ number\ of\ observation} = \frac{41+35+34}{7+7+7} = \frac{110}{21} = 5.24$

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 = $\sum_{j=1}^{7}(X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{7}(X_{2j} - \bar{X}_2)^2 + \sum_{j=1}^{7}(X_{3j} - \bar{X}_3)^2$

= 50.8572 + 38 + 58.8572 = 147.7144 ~ 147.71

Between Sample sum of Squares: $\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$

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.

$(\overline{X_1} - \bar{X})^2 = (5.86 - 5.24)^2 = (0.62)^2 = 0.3844$;
$(\overline{X_2} - \bar{X})^2 = (5 - 5.24)^2 = (-0.24)^2 = 0.0576$;
$(\overline{X_3} - \bar{X})^2 = (4.86 - 5.24)^2 = (-0.38)^2 = 0.1444$;

Sum of square Between Samples (hospitals):

$\sum_{i=1}^{k} n_i(\overline{X_i} - \bar{X})^2 = n_1(\overline{X_1} - \bar{X})^2 + n_2(\overline{X_2} - \bar{X})^2 + n_3(\overline{X_3} - \bar{X})^2$

= 7(0.3844) + 7(0.0576) + 7(0.1444)
= 2.6908 + 0.4032 + 1.0108 = 4.1048 = 4.10

Total Sum of Squares: = $\sum_i \quad \sum_j (X_{ij} - \bar{X}_i)^2$

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}	$(X_{1j} - \bar{X})^2$ = $(X_{1j} - 5.24)^2$	X_{2j}	$(X_{2j} - \bar{X})^2$ = $(X_{2j} - 5.24)^2$	X_{3j}	$(X_{3j} - \bar{X})^2$ = $(X_{3j} - 5.24)^2$
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) = $\sum (X_{1j} - \bar{X})^2 + \sum (X_{2j} - \bar{X})^2 + \sum (X_{3j} - \bar{X})^2$

= 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) = $\frac{(3)}{(2)}$**	**Variance Ratio (F)**
Between Samples (Hospitals)	3-1 =2	4.10	$\frac{4.10}{2} = 2.05$	$\frac{2.05}{8.21} = 0.25$
Within Sample (Error)	20-2=18	147.71	$\frac{147.71}{18} = 8.21$	
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,$\overline{X_1}$, $\overline{X_2}$ and $\overline{X_3}$ 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 $\overline{X}_\iota$ and $\overline{X}_J$ at level of significance 'α' is given by:

LSD ($\overline{X}_\iota$ -$\overline{X}_J$) = [The critical value of t at level of significance α and error d.f] X [S.E ($\overline{X}_\iota$ -$\overline{X}_J$)]

Note: S.E means Standard Error. Therefore, the S.E ($\overline{X}_\iota$ -$\overline{X}_J$) above mean the standard error of the difference between the two means being considered.

$$= t_{\text{n-k }(\alpha/2)} \text{ X } \sqrt{MSSE(\frac{1}{n_i} + \frac{1}{n_j})}$$

MSSE means sum of squares due to Error

If the difference $\left|\overline{X}_\iota - \overline{X}_J\right|$ 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: $G = \sum_i \sum_j X_{ij} = Grand\ Total\ of\ all\ observations$

Step 2: Compute Correction Factor (CF) = $\frac{G^2}{n}$, where n = $n_1+n_2=.....n_k$, is the total number of observations.

Step 3: Compute Raw Sum of Square (RSS) = $\sum_i \sum_j {X_{ij}}^2$
= Sum of squares of all observations

Step 4: Total Sum of Square = $\sum_i \sum_j (X_{ij} - \overline{X})^2 = RSS - CF$

Step 5: Compute
$T_i = \sum_{j=1}^{n_i} X_{ij} =$
$The\ sum\ of\ all\ observations\ in\ the\ ith\ class; (i = 1,2,..k)$

Step 6: Between Classes (or Treatment) Sum of Squares = $\sum_{i=1}^{k} \frac{T_i^2}{n_i} - CF$

$$= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \cdots \ldots \ldots \ldots \ldots \ldots \frac{T_k^2}{n_k} - CF$$

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 $(G) = \sum_i \sum_j X_{ij} = (8 + 5 + 9 + 2 + 7 + 8 + 2) + (4 + 3 + 8 + 7 + 7 + 1 + 5) + (1 + 4 + 9 + 8 + 7 + 2 + 3 = 110$

Correction Factor = (CF) = $\frac{G^2}{n} = \frac{110^2}{21} = 576.1905$

Raw Sum of Square (RSS) = $\sum_i \sum_j {X_{ij}}^2$

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

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

$= 291 + 213 + 224 = 728$

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

Between Classes (hospitals) Sum of Squares = $\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} - CF$

But $T_1 = \sum_j X_{1j} = 41,\ T_2 = \sum_j X_{2j} = 35, T_3 = \sum_j X_{3j} = 34,$

Therefore, BCSS $= \frac{41^2}{7} + \frac{35^2}{7} + \frac{34^2}{7} - CF$

$= \frac{1681+1225+1156}{7} - 576.1905 = 580.2857 - 576.1905 = .0952$

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) = $\sum_i \sum_j X_{ij}$ = (9+7+10+8) + (5+4+5+6) + (10+8+9+9) + (7+8+9+8)

= 34 + 20 + 36 +32= 122

Correction Factor (CF) = $\frac{G^2}{n}$

$= \frac{(122)^2}{16}$

$= \frac{14{,}884}{16}$

= 930.25

Raw Sum of Square (RSS) = $\sum_i \sum_j {X_{ij}}^2$

= ($9^2 + 7^2 + 10^2 + 8^2$) + ($5^2 + 4^2 + 5^2 + 6^2$) + ($10^2 + 8^2 + 9^2 + 9^2$) + ($7^2 + 8^2 + 9^2 + 8^2$)

= 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 = $\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} - CF$

= $\frac{34^2}{4} + \frac{20^2}{4} + \frac{36^2}{4} + \frac{32^2}{4} - CF$

= $\frac{1156}{4} + \frac{400}{4} + \frac{1296}{4} + \frac{1024}{4} - CF$

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 = $\frac{BCSS}{k-1}$; where k is the number of classes

$= \frac{38.75}{4-1}$ $= \frac{38.75}{3}$
= 12.92

Within Class Mean Sum of Square Error (WCMSSE) = $\frac{WSS}{n-k} = \frac{11}{16-4}$
= 0.92

Variance Ratio ($F_{calculated}$) = $\frac{BCMSSE}{WCMSSE}$

$F_{calculated} = \frac{12.92}{0.92}$

$F_{calculated}$ = 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 = $t_{n-k(\alpha/2)}$. $S.E(\bar{X}_i - \bar{X}_j)$

But the standard error of $(\bar{X}_i - \bar{X}_j) = \sqrt{WCMMSE\ X\ \frac{1}{n_i} + \frac{1}{n_j}}$

Therefore, **LSD** = 2.18 X $\sqrt{0.92\ X\ \frac{1}{4} + \frac{1}{4}}$

= 2.18 X $\sqrt{0.46}$

= 2.18 X 0.678

LSD = 1.48

Comparison between different means

Cities	Absolute Difference	Comparison	Conclusion
Lagos and Calabar	$\|8.5 - 5\| = 3.5$	> LSD	Significant
Lagos and Kano	$\|8.5 - 9\| = 0.5$	< LSD	Not Significant
Lagos and Abuja	$\|8.5 - 8\| = 0.5$	< LSD	Not Significant

Calabar and Kano	$\|5-9\|=4$	> LSD	Significant
Calabar and Abuja	$\|5-8\|=3$	> LSD	Significant
Kano and Abuja	$\|9-8\|=1$	< LSD	Not Significant

4.0 SUMMARY

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

5.0 CONCLUSION

This unit has espoused the theory and application of Analysis of Variance in statistics with special emphasis on its application in the comparison of more than two means.

6.0 TUTOR-MARKED ASSIGNMENT

Concord Bus Company just bought four different brands of tyres and wishes to determine if the average lives of the brands of tyres are the same or otherwise in order to make an important management decision. The company uses all the brands of tyres on randomly selected buses. The table below shows the lives (in '000Km) of the tyres:

Brand 1:	10,	12,	9,	9	
Brand 2:	9,	8,	11,	8,	10
Brand 3:	11,	10,	10,	8,	7
Brand 4:	8,	9,	13,	9	

Test the hypothesis that the average life for each of brand of tyres is the same. Take $\alpha = 0.01$

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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.

UNIT 5 PARAMETRIC NON-PARAMETRIC TEST METHODS

CONTENTS

1.0 Introduction
2.0 Objective
3.0 Main Content
    3.1 Parametric Non-Parametric Test Methods
        3.1.1 The Sign Test
        3.1.2 Tests Based on Runs
        3.1.3 The H-Test or the Kruskal-Wallis Test
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

If data is normally distributed, the mean is equal to the median and we use the mean as our measure of centre. However, if data is skewed, then the median is a much better measure of centre. Therefore, just like the Z, t and F tests made inferences about the population mean(s), nonparametric tests make inferences about the population median(s).

2.0 OBJECTIVE

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

explain the meaning of non-parametric data

3.0 MAIN CONTENT

3.1 Non-Parametric Data

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.

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 modelling 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. The most frequently used tests include

Kruskal-Wallis one-way analysis of variance by ranks: tests whether >2 independent samples are drawn from the same distribution.

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

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

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

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.

Anderson–Darling test: tests whether a sample is drawn from a given distribution.

Statistical Bootstrap Methods: estimates the accuracy/sampling distribution of a statistic.

Cochran's Q: tests whether *k* treatments in randomised block designs with 0/1 outcomes have identical effects.

Cohen's kappa: measures inter-rater agreement for categorical items.
Friedman two-way analysis of variance by ranks: tests whether *k* treatments in randomised block designs have identical effects.

Kaplan–Meier: estimates the survival function from lifetime data, modelling censoring.

Kendall's tau: measures statistical dependence between two variables.
Kendall's W: a measure between 0 and 1 of inter-rater agreement.

Kolmogorov–Smirnov test: tests whether a sample is drawn from a given distribution, or whether two samples are drawn from the same distribution.

Kuiper's test: tests whether a sample is drawn from a given distribution, sensitive to cyclic variations such as day of the week.
Logrank test: compares survival distributions of two right-skewed, censored samples.

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.

Median test: tests whether two samples are drawn from distributions with equal medians

Pitman's permutation test: a statistical significance test that yields exact *p* values by examining all possible rearrangements of labels.

Rank products: detects differentially expressed genes in replicated microarray experiments.

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

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

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

3.1.1 The Sign Test

Suppose we are interested in testing
H_o: P(+) = P (-)
H_1: 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 H_o

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.

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 H_o 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 $T \geq n\text{-}t$
Where T = number of +'s

Computations:

Replace each value exceeding 160 with 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

3.1.2 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. For example, 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:

$$E(u) = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$\text{Var}(u) = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

$$Z = \frac{u \pm \frac{1}{2} - E(u)}{\sqrt{Var(u)}}$$

Use (- ½) when u> E(u)
and (+ ½) when u < E(u)

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

Solution:

H_o: arrangement is random
H_1: arrangement is not random
Critical region is -2.58 to 2.58
Where $Z = \frac{u \pm \frac{1}{2} - E(u)}{\sqrt{Var(u)}}$

Computation:

n_1 = 20 (for n's)
n_2 = 12 (for d's)
U = 9 (total number of runs)

$$E(u) = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$= \frac{2\ X\ 20\ X\ 12}{20 + 12} + 1$$

$$= 16$$

E(9) < E(u) (16), hence we use + ½

$$\text{Var}(u) = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

$$= \frac{2\ X\ 20\ X\ 12(2\ X\ 20\ X\ 12-20-12)}{(20+12)^2(20\ \ +12\ \ -1)}$$

$$= 6.77$$

$$Z = \frac{9\pm\frac{1}{2}-16}{\sqrt{6.77}}$$

$$= -2.50$$

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

3.1.3 The H-Test or the Kruskal-Wallis Test

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:

$$H = \frac{12}{n(n+1)} . \frac{\sum R_1^2}{n_i} - 3(n+1)$$

Where $n = n_1 + n_2 + ...+ n_k$ 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

Example: 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 highest as if they are from one sample.

Population I	**R_1**	**Population II**	**R_{II}**	**Population III**	**R_{III}**
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, $R_1 = 84$, $R_2 = 55.5$, $R_3 = 31.5$

$$H = \frac{12}{n(n+1)} . \frac{\sum R_1^2}{n_i} - 3(n+1)$$

$$= \frac{12}{18\,X19} . [\frac{84^2}{6} + \frac{55.5^2}{7} + \frac{31.5^2}{5} - 3\,X\,19$$

$= 61.67$

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

Since H > 5.991

The null hypothesis is rejected.

4.0 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.

5.0 CONCLUSION

Various forms of Non-Parametric Test exist which may fit perfectly well into solving some qualitative problems which ordinarily may be difficult to solve with the conventional parametric test. The diversity of the different methods available reflects the varieties of problems they can solve.

6.0 TUTOR-MARKED ASSIGNMENT

1. Eighteen (18) individuals were asked if the prefer apple juice to orange juice and the following responses were obtained. Y YYY N N Y N Y YYY N Y N Y N Y (Y= Yes, N = No). Use the sign test to test the null hypothesis
   H_o: P(+) = P (-) against
   H_1: P(+) ≠ P(-) given that α = 0.05
2. Given YY N Y NN Y N YYYY N Y N Y N Y NNN YY N YYYY NN. Test the null hypothesis of randomness at α = 1%
3. The weights (in Kilograms) of randomly selected students from four different arms of a class were taken and the following results were obtained:

A	**B**	**C**	**D**
45	30	34	57
39	52	40	60
48	49	41	48
54	46	49	52
36	39	45	51
53	43	32	49
52	44	30	50
47	54	31	59
50	40	35	56
46		51	43
51			34
			38

Use the H-test at α=0.05 to test equality of the four means.

7.0 REFERENCES/FURTHER READING

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

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

MODULE 3 CORRELATION AND REGRESSION ANALYSIS

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 whileregression 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

1.0 Introduction
2.0 Objectives
3.0 Main Content
  3.1 Types of Correlation
  3.2 Reasons for High Correlation between Two Variables
  3.3 Some Methods of Studying Correlation
    3.3.1 Karl Pearson's Correlation Method
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

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.

2.0 OBJECTIVES

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

- discover whether there is a relationship between variables
- find out the direction of the relationship – whether it is positive, negative or zero
- find the strength of the relationship between the two variables.

3.0 MAIN CONTENT

3.1 Types 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
- ax 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 $y = a+bx$

(d) **Non-Linear or Curvilinear Correlation:** This situation arises if corresponding to a unit change in one variable; the other variable does not change at a constant rate but at a fluctuating rate.

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.

3.3 Some Methods of Studying Correlation

1. Scatter Diagram method
2. Karl Pearson's coefficient of correlation
3. Edward Spearman Rank correlation method etc.

In this unit, we shall treat Karl Pearson's correlation method only while the scatter diagram method and the Edward Spearman Rank correlation method will be treated in the subsequent units.

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.

3.3.1 Karl Pearson's Correlation Method

Karl Pearson (1867 – 1936), 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:

$$\rho = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Alternative formula that relies on deviation of each individual observation from the mean is also frequently used where the deviation from the mean $x = X - \overline{X}$ and $y = Y - \bar{Y}$. Here $\overline{X}$ and $\bar{Y}$ are the sample means of the set of data X_i and Y_i respectively. This formula is given as:

$$\rho = \frac{\sum xy}{\sqrt{\sum(x^2)\sum(y^2)}}$$

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 cannot 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).

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.

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	$x=X-\overline{X}$	$y = Y-\overline{Y}$	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***

$$\overline{x} = \frac{\sum x}{n} = \frac{650}{10} = 65; \quad \bar{y} = \frac{\sum y}{n} = \frac{660}{10} = 66$$

Therefore, $x = X - \overline{X} = = X - 65$; $y = Y - \overline{Y} = Y - 66$

Using the deviation from the mean formula: $\rho = \frac{\sum xy}{\sqrt{\sum(x^2)\sum(y^2)}}$

$$\rho = \frac{2704}{\sqrt{5398\ X\ 2224}}$$

$$\rho = \frac{2704}{\sqrt{12005152}} = \frac{2704}{3464.8451} = 0.7804$$

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:

$$\rho = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

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**

$$\rho = \frac{10(66{,}249) - (798)(822)}{\sqrt{[10[64{,}722 - 798^2][10X68{,}116 - 822^2]}}$$

$$\rho = \frac{6534}{\sqrt{57038016}}$$

$$\rho = \frac{6534}{7552.35}$$

$$\rho = 0.8651$$

$$\rho = 0.87$$

It can be said that there is a strong positive relationship between the marks obtained in English and Mathematics by the 10 ten students.

4.0 SUMMARY

You have learnt that Pearson's Correlation is a measure of relationship between two (or more) variables that change together.

5.0 CONCLUSION

The test statistic, called the Pearson's correlation coefficient r, measures the strength of the relationship between variables. This measure varies from 0 (no relationship) to +1and to -1 (perfect relationship).

6.0 TUTOR-MARKED ASSIGNMENT

1. Calculate the Pearson's product moment correlation coefficient between households' monthly income and expenditure on beverage using the set of data given below:

Household Income ('000)	Expenditure on beverage ('000)
50	5
76	4
82	7

67	3
96	9
59	8
61	10
75	6
70	11
85	7

2. Give the practical interpretation of the coefficient estimated.

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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.

UNIT 2 SPEARMAN'S RANK CORRELATION METHOD

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 The Spearman's Rank Correlation Method
    3.2 Some Challenging Cases when using Spearman's Rank Correlation Method
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

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.

The Spearman's rho/Rank Correlation Coefficient is a nonparametric measure of correlation coefficient, the aim here is to explain how Spearman's rho is used when your data does not conform to the assumptions of a parametric test.

2.0 OBJECTIVES

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

- explain how rho is used when a data does not conform to the assumptions of a parametric test
- write the Spearman's Rank Correlation Coefficient formular.

3.0 MAIN CONTENT

3.1 The Spearman's Rank Correlation Coefficient Method

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:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

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.

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.

Example 1: Fifteen (15) members of staff of the administrative unit of an organisation were studied to determine the correlation between their punctuality at work (X) and the compliance of their dresses with organisational 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:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

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
		$\Sigma d = 0$	**$\Sigma d^2 = 272$**

$$\rho = 1 - \frac{6X272}{15(15^2 - 1)}$$

$$\rho = 1 - \frac{6\ X\ 272}{15(225 - 1)}$$

$\rho = 1 - \frac{17}{35}$, $\rho = \frac{18}{35}$, $\rho = 0.51$

3.2 Some Challenging Cases which 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 2 and 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 $\overline{X} \neq \overline{Y}$ *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.

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
				$\Sigma d = 0$	**$\Sigma d^2 = 54$**

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6\,X\,54}{11(11^2 - 1)}$$

$$\rho = 1 - \frac{324}{1320}$$

$$\rho = 1 - 0.245$$

$$\rho = 0.75$$

4.0 SUMMARY

This unit by now has been able to help you to calculate and interpret the simple correlation between two variables and to determine whether the correlation is significant.

5.0 CONCLUSION

Spearman's correlation coefficient approximates Pearson's correlation when the sample size is large. Significant tests become problematic when sample size is less than 20, and specific tables of probability will need to be consulted.

6.0 TUTOR-MARKED ASSIGNMENT

The following are the marks obtained by a group of students in two papers. Calculate the rank coefficient of correlation.

Economics:	*78*	*36*	*98*	*25*	*75*	*82*	*92*	*62*	*65*	*39*
Statistics:	*84*	*51*	*91*	*69*	*68*	*62*	*86*	*58*	*35*	*49*

7.0 REFERENCES/FURTHER READING

Armitage, P. (1980). *Statistical Methods in Medical Research.* Oxford: Blackwell Scientific Publications. Pp. 147-166.

Siegel, S. & Castellan Jr., N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* (2nd. ed.). Boston Massachusetts: McGraw Hill, Inc. pp. 235-244.

UNIT 4 LEAST SQUARE REGRESSION ANALYSIS

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 The Least Square Regression Analysis
    3.2 The Regression Analysis
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

A line of best fit can be roughly determined using an eyeball method by drawing a straight line on a scatter plot so that the number of points above the line and below the line is about equal (and the line passes through as many points as possible). A more accurate way of finding the line of best fit is the **least square method.** This method would be examined with in this unit.

2.0 OBJECTIVES

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

- explain the importance of the basic idea of regression/correlation analysis
- predict the behaviour of one variable (the predict and) based on fluctuations in one or more related variables (the predictors)
- estimate the behaviour of the predict and use multiple predictors (multiple regression
- predict the behaviour of the predict and based on nonlinear relationships with the predictors.

3.0 MAIN CONTENT

3.1 The Least Square Regression Analysis

If two variables are significantly correlated and 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*".

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 between two 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 among others are indispensable 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.*"

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".

3.2 The Regression Analysis

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*.

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"

$$a = \frac{(\sum X^2)(\sum Y) - (\sum X)(\sum XY)}{n\sum X^2 - (\sum X)^2}$$

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:

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

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

$$\bar{Y} = a + b\bar{X}$$

Where $\bar{Y}$ is th the sample is mean of the dependent variable Y and $\bar{X}$ is the sample mean of the independent variable X. From the foregoing "*a*" can be obtained having obtained "*b*" by:

$$a = \bar{Y} - b\bar{X}$$

b can be obtained using deviation from mean approach where:

$x = X - \bar{X}$; $y = Y - \bar{Y}$

$$b = \frac{\sum xy}{\sum x^2}$$

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**

$$a = \frac{(\sum X^2)(\sum Y) - (\sum X\ \ )(\sum XY)}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{(10826)(110) - (310)(3740)}{10((10{,}826) - (310)^2}$$

$$a = \frac{1190860 - 1159400}{108{,}260 - 96{,}100}$$

$$a = \frac{31{,}400}{12{,}160}$$

$$a = 2.59 \cong 2.6$$

$$The\ slope\ parameter\ estimate\ \ b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$b = \frac{10(3740) - (310)(110)}{10(10{,}826) - (310)^2}$$

$$b = \frac{37400 - 34{,}100}{108{,}260 - 96{,}100}$$

$$b = \frac{3300}{12{,}160}$$

$$b = 0.27$$

Alternatively, having obtained estimate value for the slope parameter *b*, the intercept term *a* can be obtained using the formula:

$$a = \bar{Y} - b\bar{X}$$

In the above problem $\bar{X} = \frac{\sum X}{n}$; $\bar{Y} = \frac{\sum Y}{n}$

$$Therefore, \qquad \bar{X} = \frac{310}{10} = 31\ ; \qquad \bar{Y} = \frac{110}{10} = 11$$

$$a = 11 - (0.27)(31)$$

$$a = 2.63 \simeq 2.6$$

4.0 SUMMARY

With reference to the explanations and illustrations demonstrated above, you should by now be able to apply least square method to study the nature of the relation between two variables.

5.0 CONCLUSION

A line of best fit can be roughly determined using an eyeball method by drawing a straight line on a scatter plot so that the number of points above the line and below the line is about equal (and the line passes through as many points as possible). A more accurate way of finding the line of best fit is the **least square method.**

6.0 TUTOR-MARKED ASSIGNMENT

Some fairly used cars were displayed for sales in an automobile sales compound and the data below shows the prices and ages of the cars. Using the observed data calculate the estimates of the slope and intercept parameter of the regression model.

Ages of car (*X*)	**Price Y ('000)**
5	80
7	57
6	58
6	55
5	70
4	88
7	43
6	60
5	69
5	63
2	118

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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*. (6^{th} ed.). BookPower.

UNIT 5 FORECASTING IN REGRESSION

CONTENTS

1.0 Introduction
2.0 Objective
3.0 Main Content
    3.1 Forecasting in Regression
    3.2 Model Evaluation
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

The Least Square Regression is the most widely used statistical tool in the field of social and some applied sciences. It is used in business and economics to study relationship between two or more variables that relate causally. That is, where change in one variable results in changes in other variables.

2.0 OBJECTIVE

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

- use the given value of independent variables to predict the value of dependent variable.

3.0 MAIN CONTENT

3.1 Forecasting in Regression

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 $\hat{Y} = \hat{a} + \hat{b}X$. 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 between the 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.

3.1 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 though it is not well explained by the regression model, it can still be forecast easily.

Example: 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:

$$a = \frac{(\sum X^2)(\sum Y) - (\sum X)(\sum XY)}{n\sum X^2 - (\sum X)^2}$$

And the slope term b is given as:

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

However, one may obtain the value of the estimate of the slope parameter *b* and use the formula: $a = \bar{Y} - b\bar{X}$ to estimate the value of the intercept term *a*.

For the above problem, the estimated regression equation is given as:

$\hat{Y} = 2.6 + 0.27X$

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, $\hat{Y} = 2.6 + 0.27(35)$

$\hat{Y} = 2.6 + 9.45$

$\hat{Y} = 12.05\ kg\ of\ beef$

(ii) For household income X = ₦40,000

Therefore, $\hat{Y} = 2.6 + 0.27(40)$

$\hat{Y} = 2.6 + 10.8$

$$\hat{Y} = 13.4\ kg\ of\ beef$$

(iii) For household income X = ₦45,000

Therefore, $\hat{Y} = 2.6 + 0.27(45)$

$$\hat{Y} = 2.6 + 12.15$$

$$\hat{Y} = 14.75\ kg\ of\ beef$$

4.0 SUMMARY

This 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.

5.0 CONCLUSION

As mentioned earlier, the Least Square Regression is the most widely used statistical tool in the field of social and some applied sciences. It is used in business and economics to study relationship between two or more variables that relate causally. That is, where change in one variable results in changes in other variables.

7.0 TUTOR-MARKED ASSIGNMENT

1. The data below represent sales record of a marketing firm over the last seven years.

Year	1	2	3	4	5	6	7
Sales (in '000)	14	17	15	23	18	22	27

It is required to forecast the sales for the Years 8, 9, and 10.
Hint: This is an example of time series data used for trend analysis where time (year) become an explanatory (independent) variable in a regression model and takes on the values 1, 2, 3.....7

2. A housing consultant believes that the number of houses sold in a region for given year is related to the mortgage rate in that period. He collected the following relevant data.

Year	**Mortgage interest rate (X)**	**Housing Sales Index (Y)**
2002	12	80

2003	10	90
2004	8	105
2005	6	115
2006	7	125
2007	8	120
2008	10	115
2009	12	100
2010	14	85
2011	13	70
2012	11	80

Forecast for the housing sales index for years 2013 to 2015 if the mortgage interests rates are expected to be 15, 12 and 14 respectively.

7.0 REFERENCES/FURTHER READING

Spiegel, M. R. & 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.

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.

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

CONTENTS

1.0 Introduction
2.0 Objective
3.0 Main Content
    3.1 The Central Limit Theorem for Independent Sequence
    3.2 CLT Rules
    3.3 Classical CLTs
        3.3.1 Lindeberg–Lévy CLT
        3.3.2 Lyapunov CLT
        3.3.3 Lindeberg CLT
        3.3.4 Multidimensional CLT
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

The central limit theorem describes the characteristics of the "population of the means" which has been created from the means of an infinite number of random population samples of size (n), all of them drawn from a single "parent population".

2.0 OBJECTIVE

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

- state the importance of CLT for independence sequence.

3.0 MAIN CONTENT

3.1 The Central Limit Theorem for Independent Sequence

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: $\boldsymbol{\sigma_y = \frac{\sigma_x}{\sqrt{N}}}$
3. the shape of the sampling distribution of the mean approaches normal as N increases.

3.2 CLT Rules

Let us 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. $\boldsymbol{\sigma_y = \frac{\sigma}{\sqrt{n}}}$ 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 σ. Let X_1;X_2; _ _ _ Xn 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:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$T = \sum_{i=1}^{n} X_i$$

The central limit theorem states that the sample mean $\bar{X}$ follows approximately the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, where μ and σ 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 σ^2 has approximately the normal distribution with mean μ and variance σ^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 summarise: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

To transform $\bar{X}$into z we use: $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

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: $E(\bar{Y}) = \mu$
Proof:

$$E(\bar{Y}) = E\left[\frac{1}{n}(Y_1 + Y_2 + Y_3 + \cdots + Y_n)\right]$$
$$= \frac{1}{n}E[Y_1 + Y_2 + Y_3 + \cdots + Y_n]$$
$$= \frac{1}{n}[E(Y_1) + E(Y_2) + E(Y_3) \ldots\ldots E(Y_n)]$$
$$= \frac{1}{n}[\mu + \mu + \cdots \ldots\ldots\ldots \mu]$$
$$= \frac{1}{n}[n\mu] = \mu$$

Theorem 2: $V(\bar{Y}) = \frac{\sigma^2}{n}$

Proof:

$$V(\bar{Y}) = V\left[\frac{1}{n}(Y_1 + Y_2 + Y_3 \dots \dots Y_n)\right]$$

$$= \left(\frac{1}{n}\right)^2 V[Y_1 + Y_2 + Y_3 \dots \dots Y_n]$$

$$= \left(\frac{1}{n}\right)^2 [V(Y_1) + V(Y_2) + V(Y_3) \dots \dots V(Y_n)]$$

$$= \left(\frac{1}{n}\right)^2 [\sigma_1^2 + \sigma_2^2 + \sigma_3^2 \dots \dots \sigma_n^2]$$

$$= \left(\frac{1}{n}\right)^2 [n\sigma^2] = \frac{\sigma^2}{n}$$

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.

$$\frac{X_1+X_2+X_3+\dots\dots\dots+X_n - n.\mu}{\sigma.\sqrt{n}} \xrightarrow{d} N(0,1)$$

i.e. a centered and normalised 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 σ = 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:

$$P(\bar{X} < 9) = P\left(z < \frac{9-10}{\frac{4}{\sqrt{100}}}\right) = P(z < -2.5) = 0.0062$$

(from the standard normal probabilities table).

Similarly the central limit theorem states that sum T follows approximately the normal distribution, $\sim N(n\mu, \sqrt{n\sigma})$, where μ and σ are the mean and standard deviation of the population from where the sample was selected.

To transform T into z we use: $z = \frac{T - n\mu}{\sqrt{n\sigma}}$

3.3 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 σ^2. Suppose we are interested in the sample average of these random variables.

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

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 σ^2. For large enough n, the distribution of S_n is close to the normal distribution with mean μ and variance σ^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:

3.3.1 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 $Var[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \xrightarrow{d} N(0,\, \sigma^2).$$

In the case $\sigma > 0$, convergence in distribution means that the cumulative distribution functions (cdf) of $\sqrt{n}(S_n - \mu)$ converge point tends to the cdf of the $N(0, \sigma^2)$ 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:

$$\lim_{n\to\infty} \sup_{z\in\mathbf{R}} \left| \Pr[\sqrt{n}(S_n - \mu) \leq z] - \Phi(z/\sigma) \right| = 0,$$

where sup denotes the least upper bound (or supremum) of the set.

3.3.2 Lyapunov CLT

The theorem is named after Russian mathematician Aleksandra 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 $(2 + \delta)$, 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
If for some $\delta > 0$, 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 $\delta = 1$. If a sequence of random variables satisfies Lyapunov's condition, then it also satisfies Lindeberg's condition. The converse implication, however, does not hold.

3.3.3 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 $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathrm{E}\left[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\{|X_i-\mu_i|>\varepsilon s_n\}}\right] = 0$$

where $\mathbf{1}_{\{\ldots\}}$ is the indicator function. Then the distribution of the standardised sums $\frac{1}{s_n}\sum_{i=1}^{n}(X_i - \mu_i)$ converges towards the standard normal distribution N(0,1).

3.3.4 Multidimensional CLT

The **Multidimensional CLT** Proof 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 $\mu = \mathrm{E}(X_i)$ and covariance matrix $\mathbf{\Sigma}$ (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

$$\begin{bmatrix} X_{1(1)} \\ \vdots \\ X_{1(k)} \end{bmatrix} + \begin{bmatrix} X_{2(1)} \\ \vdots \\ X_{2(k)} \end{bmatrix} + \cdots + \begin{bmatrix} X_{n(1)} \\ \vdots \\ X_{n(k)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}\left[X_{i(1)}\right] \\ \vdots \\ \sum_{i=1}^{n}\left[X_{i(k)}\right] \end{bmatrix} = \sum_{i=1}^{n}\left[\mathbf{X_i}\right]$$

and the average will be

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left[\mathbf{X_i}\right] = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n}\left[X_{i(1)}\right] \\ \vdots \\ \sum_{i=1}^{n}\left[X_{i(k)}\right] \end{bmatrix} = \begin{bmatrix} \bar{X}_{i(1)} \\ \vdots \\ \bar{X}_{i(k)} \end{bmatrix} = \mathbf{\bar{X}_n}$$

and therefore

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathbf{X_i} - E\left(X_i\right)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathbf{X_i} - \mu\right] = \sqrt{n}\left(\overline{\mathbf{X}}_n - \mu\right)$$

.

The multivariate central limit theorem states that:

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mu\right) \xrightarrow{D} \mathcal{N}_k(0, \Sigma)$$

4.0 SUMMARY

Some level of independence sequence in CLT is highlighted here to make you 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.

5.0 CONCLUSION

In more general probability theory, a central limit theorem is any of a set of weak-convergence theories. They all expressed the fact that a sum of many independent and identically distributed random variables, or alternatively, random variables with specific types of dependence, will tend to be distributed according to one of a small set of attractor distributions

6.0 TUTOR-MARKED ASSIGNMENT

1. A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean $\mu = 205$ pounds and standard deviation $\sigma = 15$ pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?
2. From past experience, it is known that the number of tickets purchased by a spectator standing in line at the ticket window for the football match of Nigeria against Ghana follows a distribution that has mean $\mu = 2{:}4$ and standard deviation $\sigma = 2{:}0$.

Suppose that few hours before the start of one of these matches there are 100 eager spectators standing in line to purchase tickets.

If only 250 tickets remain, what is the probability that all 100 students will be able to purchase the tickets they desire?

7.0 REFERENCES/FURTHER READING

Billingsley, P. (1995), *Probability and Measure.* (3rd ed.). John Wiley & Sons Publishers.

Bradley, R. (2007). *Introduction to Strong Mixing Conditions* (1st ed.). Heber City, UT: Kendrick Press.

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. & 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

UNIT 2 CENTRAL LIMIT THEOREM FOR DEPENDENT PROCESSES

CONTENTS

1.0 Introduction
2.0 Objective
3.0 Main Content
    3.1 Formulations of CLT
        3.1.1 Theorem i
        3.1.2 Theorem ii
    3.2 Proof of CLT
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

A useful generalisation 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.

2.0 OBJECTIVES

A main attraction of CLT sequences is to provide examples of processes that are weakly dependent, but not mixing. This way of constructing stationary sequences is very natural.

3.0 MAIN CONTENT

3.1 Formulations of CLT

A simplified formulation of the central limit theorem under strong mixing is provided for in the following:

- CLT under weak dependence
- Martingale difference CLT

3.1.1 Theorem *I (*CLT under weak dependence**)**

Suppose that X_1, X_2, … is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $E(X_n) = 0$ and $E(X_n^{12}) < \infty$. Denote $S_n = X_1 + \ldots + X_n$, then the limit

$$\sigma^2 = \lim_n \frac{E(S_n^2)}{n}$$

exists, and if $\sigma \neq 0$ then $S_n/(\sigma\sqrt{n})$ converges in distribution to $N(0, 1)$.
In fact, $\sigma^2 = E(X_1^2) + 2\sum_{k=1}^{\infty} E(X_1 X_{2+k})$ where the series converges absolutely.

The assumption $\sigma \neq 0$ 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 $E(X_n^{12}) < \infty$ is replaced with

$E(|X_n|^{2+\delta}) < \infty$, and the assumption $\alpha_n = O(n^{-5})$ is replaced with $\sum_n \alpha_n^{\frac{\delta}{2(2+\delta)}} < \infty$
Existence of such $\delta > 0$ ensures the conclusion.

3.1.2 Theorem *ii (*Martingale difference CLT)

Let a martingale M_n satisfy

$\frac{1}{n}\sum_{k=1}^{n} E((M_k - M_{k-1})^2 M_1, \ldots\ldots M_{k-1}) \rightarrow 1$in probability as *n* tends to infinity,

-for every ε >0, $\frac{1}{n}\sum_{k=1}^{n} E((M_k - M_{k-1})^2; |M_k, \ldots\ldots M_{k-1}|) > \varepsilon\sqrt{n}) \rightarrow 0$

Then $M_n/\sqrt{n}$ converges in distribution to $N(0,1)$ as $n \rightarrow \infty$.

3.1 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,

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2), \quad t \rightarrow 0$$

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 $(X_i - \mu)/\sigma$, 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

$$Z_n = \frac{n\overline{X_n} - n\mu}{\sigma\sqrt{n}} = \sum_{i=0}^{n} \frac{Y_i}{\sqrt{n}}$$

By simple properties of characteristic functions, the characteristic function of Z_n is

$$\left[\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \to e^{-t^2/2}, \qquad n \to \infty$$

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.

4.0 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.

5.0 CONCLUSION

The central limit theorem gives only an asymptotic distribution. As an approximation for a finite number of observations, it provides a reasonable approximation only when close to the peak of the normal distribution; it requires a very large number of observations to stretch into the tails.

If the third central moment $E((X_1 - \mu)^3)$ exists and is finite, then the above convergence is uniform and the speed of convergence is at least on the order of $1/n^{1/2}$ can be used not only to prove the central limit theorem, but also to provide bounds on the rates of convergence for selected metrics.

The convergence to the normal distribution is monotonic, in the sense that the entropy of Z_n increases monotonically to that of the normal distribution.

6.0 TUTOR-MARKED ASSIGNMENT

1. Show how the Martingale converges towards normality as n tends towards infinity.
2. Explain how a series will converge towards the limit under the dependence process

7.0 REFERENCES/FURTHER READING

Artstein, S.; Ball, K.; Barthe, F. & Naor, 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. (Theorem 5.3.4, p. 47).

Website: www.wikipedia.com

UNIT 3 THE LAW OF LARGE NUMBERS

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 The Law of Large Numbers (LLN)
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

2.0 OBJECTIVES

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

- apply the importance of the law of large numbers
- state the law of large numbers
- differentiate between stable and normal distributions.

The aim here is to emphasize that the Law of Large Numbers (LLN) is important because it "guarantees" stable long-term results for the averages of random events.

3.0 MAIN CONTENT

3.1 The Law of Large Numbers

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:

$$\bar{X}_n \xrightarrow{P} \mu$$

Proof: Let the variance of each *Xi* be σ^2. It then follows from the Chebyshev inequality that for every number $\varepsilon > 0$,

$$\Pr(|\overline{X_n} - \mu| < e) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

Hence,

$$\bar{X}_n \xrightarrow{P} \mu$$

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 behaviour 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 by dividing the function by appropriate normalising functions, and looking at the limiting behaviour of the result, can tell us much about the limiting behaviour 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 σ^2, then by the central limit theorem:

$$\frac{S_n - n\mu}{\sqrt{n}} \to \xi,$$

Where ξ is distributed as N(0, σ^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.

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,

$$\Pr(|\bar{X}_n - \mu| \geq 1) \leq \frac{\sigma^2}{n} \leq \frac{4}{n}$$

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

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

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.

4.0 SUMMARY

The explanations and illustrations presented above would have provided clear understanding of this unit. In case you encounter some difficulties in understanding any area, it is suggested that you make reference to further reading list at the end of this unit. Such reference is expected to enhance your knowledge to be able to solve problems arising from large numbers.

5.0 CONCLUSION

The law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. LLN states that the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

6.0 TUTOR-MARKED ASSIGNMENT

1. Let X be a random variable for which $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. Construct a probability distribution for X such that
$$\Pr(|X - \mu| \geq 3\sigma) = \frac{1}{9}$$
2. How large a random sample must be taken from a given distribution in order for the probability to be at least 0.99 that the sample mean will be within 2 standard deviations of the mean of the distribution?

7.0 REFERENCES/FURTHER READING

Artstein, S., Ball, K., Barthe, F. & Naor, 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. (Theorem 5.3.4, p. 47)

Johnson, O. T**.** (2004). *Information Theory and the Central Limit Theorem*. Imperial College Press. p. 88.

Vladimir, V. U. & Zolotarev, V. M. (1999). *Chance and Stability: Stable Distributions and Their Applications*. VSP. Pp. 61–62.

UNIT 4 EXTENSION TO THE CLT AND BEYOND THE CLASSICAL FRAMEFORK

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 The Theorem around Convex Body
    3.2 Products of Positive Random Variables
    3.3 Beyond the Classical Framework
    3.4 Lacunary Trigonometric Series
    3.5 Gaussian Polytopes
    3.6 Linear Functions of Orthogonal Matrices
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

In this unit, we shall treat the issues which are addition to the Classical Limit Theorem beyond the classical framework. This shall include: 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.

2.0 OBJECTIVES

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

- highlight the product of positive random variables
- give examples of log-concave density
- explain the Lacunary trigonometric series
- state Linear functions of orthogonal matrices
- apply the theorems to solving practical day-to-day problems.

3.0 MAIN CONTENT

3.1 The Theorem around 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

$$\frac{X_1 + X_2 + \ldots\ldots\ldots + X_n}{\sqrt{n}}$$

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 $\text{const} \cdot \exp(-|x_1|^\alpha)\ldots\exp(-|x_n|^\alpha)$, 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, pair-wise 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

$$\left| P\left(a \le \frac{X_1 + \ldots + X_n}{\sqrt{n}} \le b\right) - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt \right| \le \frac{C}{n}$$

for all $a<b$; here C is a universal (absolute) constant. Moreover, for every $c_1, \ldots, c_n \in \mathbf{R}$ such that $c_1^2 + \ldots + c_n^2 = 1$,

$$\left| P(a \le c_1 X_1 + \ldots + c_n X_n \le b) - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt \right| \le C(c_1^4 + \ldots + c_n^4)$$

The distribution of $(X_1 + \ldots + X_n)/\sqrt{n}$ 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 $c_1^2 + \ldots + c_n^2 = 1$.

3.2 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.

3.3 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.

3.4 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

$$r_1^2 + r_2^2 + \ldots\ldots\ldots = \infty \; and \; \frac{r_k^2}{r_1^2 + \cdots + r_k^2} \to 0$$

- $0 \leq a_k < 2\pi$.

Then $\frac{X_1 + \ldots X_k}{\sqrt{r_1^2 + \cdots + r_k^2}}$ converges in distribution to $N(0, 1/2)$.

3.5 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;

$$\frac{X_n - EX_n}{\sqrt{Var\, X_n}}$$

converges in distribution to $N(0, 1)$ as n tends to infinity. The same holds in all dimensions (2, 3, ...).

The polytope K_n is called Gaussian random polytope.

A similar result holds for the number of vertices (of the Gaussian polytope), the number of edges, and in fact, faces of all dimensions.

3.6 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 normalised 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 $(X_{n1} + \ldots + X_{nk})/\sqrt{k}$converges in distribution to $N(0, 1)$ as k tends to infinity.

Q-analogues

A generalised 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.

4.0 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.

5.0 CONCLUSION

Comparison of probability density functions, $p(k)$ for the sum of n fair 6-sided dice to show their convergence to a normal distribution with increasing n, in accordance to the central limit theorem. In the bottom-right graph, smoothed profiles of the previous graphs are rescaled, superimposed and compared with a normal distribution (black curve).

6.0 TUTOR-MARKED ASSIGNMENT

A simple example of the central limit theorem is rolling a large number of identical, unbiased dice. The distribution of the sum (or average) of the rolled numbers will be well approximated by a normal distribution. Since real-world quantities are often the balanced sum of many unobserved random events, the central limit theorem also provides a partial explanation for the prevalence of the normal probability distribution. It also justifies the approximation of large-sample statistics to the normal distribution in controlled experiments.

7.0 REFERENCES/FURTHER READING

Rempala, G. &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:) (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.

MODULE 5 INDEX NUMBERS AND INTRODUCTION TO RESEARCH METHODS IN SOCIAL SCIENCES

Unit 1 Index Number
Unit 2 Statistical Data
Unit 3 Sample and Sampling Techniques

UNIT 1 INDEX NUMBER

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Index Numbers
    3.2 Uses of Index Numbers
    3.3 Types of Index Number
    3.4 Problems Encountered in the Construction of Index Numbers
    3.5 Methods of Constructing Index Numbers
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

Index number is a number used to indicate the change in a value or quantity such as a price or unemployment, when compared with the level of that value or quantity at an earlier time. The base level is usually arbitrarily set at 100, and the increase or decrease in index numbers over time is often expressed as a percentage change.

2.0 OBJECTIVES

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

- explain the meaning of index numbers
- explain the different types of index numbers
- state the applications of index numbers in economics.

3.0 MAIN CONTENT

3.1 Index Numbers

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.

3.2 Uses of Index Number

- Index numbers are used to measure the pulse of the economy
- It is used to study trend and tendencies
- Index numbers are used for deflation
- Index numbers help in the formulation of decisions and policies
- It measures the purchasing power of money.

3.3 Types of Index Numbers

Index numbers 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 –** They reflect changes in the retail prices of different commodities. They 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.

3.4 Problems in the construction of Index Numbers

- The purpose of index number – This must be carefully defined as there is no general purpose index number.
- 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
- 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.
- Data for the index number- Data to be used must be reliable.
- Type of average to be used – ie, arithmetic, geometric, harmonic etc.
- Choice of formula – There are different types of formulas and the choice is mostly dependent on available data.
- System of weighting – Different weights should be assigned to different commodities according to their relative importance in the group.

3.5 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

$$P_{01} = \frac{\sum P_1}{\sum P_0} \text{ X } \frac{100}{1}$$

Limitations of the Simple Aggregate Method

- The prices of various commodities may be quoted in different units
- 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.

- The relative importance of various commodities is not taken into consideration

Based on this method, quantity index is given by the formula:

$$Q_{01} = \frac{\sum q_1}{\sum q_0} \text{ X } \frac{100}{1}$$

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

$$P_{01} = \frac{\sum P_1}{\sum P_0} \text{ X } \frac{100}{1}$$

$$P_{01} = \frac{335}{402} \text{ X } \frac{100}{1}$$

$$= 83.3\ \%$$

(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

$$P_{01} = \frac{\sum wP_1}{\sum wP_0} \text{ X } \frac{100}{1}$$

By using different systems of weighting, we obtain a number of formulae, some of which include:

i. **Laspeyre's Price Index or Base year method –** Taking the base year quantity as weights i.e w = q_o in the equation above, the Laspeyre's Price Index is given as:

$$P_{01}^{La} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \text{ X } \frac{100}{1}$$

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:

$$P_{01}^{Pa} = \frac{\sum P_1 q_1}{\sum P_0 q_1} X \frac{100}{1}$$

This formula was introduced by German statistician Paasche, in 1874.

i. 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:

$$P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}\right] X \frac{100}{1}$$

ii. Fisher's Price Index – Irving Fisher advocated the geometric cross of Laspeyre's and Paasche's Price index numbers and is given as:

a.

$$\boldsymbol{P}_{\mathbf{01}}^{\boldsymbol{F}} = \sqrt{P_{01}^{La} \; XP_{01}^{Pa}} = \left[\frac{\sum P_1 q_0}{\sum P_0 q_0} X \frac{\sum P_1 q_1}{\sum P_0 q_1}\right]^{\frac{1}{2}} X \frac{100}{1}$$

b. Fisher's Index is termed as an ideal index since it satisfies time reversal and factor reversal test for the consistency of index numbers.

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 ($p_o q_o$) 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 (₦) $p_o q_o$(2)**	**Price per unit 2012 (₦) p_1 (3)**	**2010 quantities q_0= $= \frac{(2)}{(1)}$**	$P_1 q_0$
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
		$\sum P_0 q_0 =$ 3620			**$\sum P_1 q_0$ =4350**

Therefore, the Laspeyre's Price Index for 2012 with respect to (w.r.t) base 2010 is given by:

$$P_{01}^{La} = \frac{\sum P_1 q_0}{\sum P_0 q_0} X \frac{100}{1}$$

$$= = \frac{4350}{3620} X \frac{100}{1}$$

$$= = 120.1657 \cong 120.17$$

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

(iii) Dowbish-Bowley price index methods

Commodities	***2007***		***2012***	
	Price	**Quantity**	**Price**	**Quantity**
Garri	20	8	40	6
Rice	50	10	60	5
Fish	40	15	50	15

Palm-oil	20	20	20	25

Solution:

Commodities	***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
Garri	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					p_oq_o = ***1660***	p_oq_1 = ***1470***	p_1q_o = ***2070***	p_1q_1 = ***1790***

(i) Laspeyre's Price Index

$$P_{01}^{La} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \text{ X } \frac{100}{1}$$

$$P_{01}^{La} = \frac{2070}{1660} \text{ X } \frac{100}{1}$$

$$= 1.24699 \text{ X } 100$$

$$= 124.7$$

(ii) Paasche's Price Index

$$P_{01}^{Pa} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \text{ X } \frac{100}{1}$$

$$P_{01}^{Pa} = \frac{1790}{1470} \text{ X } \frac{100}{1}$$

$$= 1.2177 \text{ X } 100 \qquad = 121.77$$

(iii) Fisher's Price Index

$$P_{01}^{F} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \text{ X } \frac{\sum P_1 q_1}{\sum P_0 q_1}} \text{ X } \frac{100}{1}$$

$$P_{01}^{F} = \sqrt{\frac{2070}{1660} \text{ X } \frac{1790}{1470}} \text{ X } \frac{100}{1}$$

$$P_{01}^{F} = \sqrt{1.24699 \text{ X } 1.2177} \text{ X } 100$$

$$= 123.23$$

(iv) Dorbish-Bowley Price Index

$$P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}\right] X \frac{100}{1}$$

$$= \frac{1}{2}\left[\frac{2070}{1660} + \frac{1790}{1470}\right] X \frac{100}{1}$$

$$= ½ [1.247 + 1.2177] X100$$

$$= 1.23235 X 100$$

$$= 123.24$$

4.0 SUMMARY

In this unit, we have been able to introduce you to the concept of index numbers, its uses and methods of calculation. You 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.

5.0 CONCLUSION

The uses of index numbers are enormous. Its uses and importance goes beyond the field of statistics and economics but also applicable in policy formulation, governance and so on. Methods which can be used to study the statistic is also diverse as different variants have been proposed by statisticians and economics alike.

6.0 TUTOR-MARKED ASSIGNMENT

1. Calculate price index number of the year 2010 with 2000 as the base year from the following data using:

   - Laspeyre's
   - Pasche's
   - Dorbish-Bowley and
   - Fisher's formulae

Commodity		**2000**		**2010**	
	Unit	**Price (₦)**	**Value (₦)**	**Quantity consume**	**Value (₦)**
Rice	Kg	20	3000	320	3520

Garri	Kg	24	2160	200	2600
Cloth	Yards	30	1800	120	1920
Sugar	Packets	18	900	80	960

2. From the following data, construct Fisher's ideal index number

Commodity	**2003**		**2013**	
	Price (₦)	**Value (₦)**	**Price (₦)**	**Value (₦)**
W	15	150	18	216
X	21	252	30	240
Y	30	240	36	288
Z	12	60	15	90

7.0 REFERENCES/FURTHER READING

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

Lucey, T. (2002). *Quantitative Techniques*. (6th ed.). Book Power.

UNIT 2 STATISTICAL DATA

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Statistical Data
    3.2 Types of Data
    3.3 Sources of Data
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

Statistics is a branch of mathematics that deals with the collection, organisation, 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 civilisation, 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. In this unit, we shall examine the statistical data; discuss the types viz a viz the advantages and the disadvantages of each type, and the sources of data.

2.0 OBJECTIVES

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

- appreciate the purpose of statistical tests and data
- determine whether some hypotheses are extremely unlikely given observed data.

3.0 MAIN CONTENT

3.1 Statistical Data

Before 3000BC, the Babylonians used small clay tablets to record tabulations of agricultural yields and of commodities bartered or sold. The Egyptians analysed the population and material wealth of their country before they begin 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 analyse 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 analysing 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.

3.2 Types of Data

Data can be classified into types based on different criteria viz:

4. 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 organisation 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 at hand because it was not collected primarily for particular purpose
- There is always a problem of missing periods.

2. **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.
   (b) **Time-series data:** These are data collected on a particular variable or set of variables over time. For example, a set that contain Nigeria's Gross Domestic Product (GDP) values form 1970 to 2012.
   (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 state 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.

3.3 Sources of Data

1. Sources of Primary data:

(i) Census
(ii) Surveys etc.

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
(iv) 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)
(iii) Economic and Financial Review (by CBN)
(iv) Population of Nigeria (by NPC).

4.0 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.

5.0 CONCLUSION

Here, a further aim of statistical data and testing is shown to you to quantify evidence against a particular hypothesis being true. You were able to think of it as testing to guide research. We believe a certain statement may be true and want to work out whether it is worth investing time investigating it. Therefore, we look at the opposite of this statement. If it is quite likely then further study would seem to not make sense. However, if it is extremely unlikely then further study would make sense.

6.0 TUTOR-MARKED ASSIGNMENT

1. Distinguish between primary and secondary data.
2. What are the advantages of primary data?
3. List 4 source of secondary data you know.
4. Distinguish between cross-sectional and panel data.

7.0 REFERENCES/FURTHER READING

Frankfort-Nachmias, C. & Nachmias, D. (2009). *Research in the Social Sciences. (5th ed.).* Hodder Education.

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

Microsoft ® Encarta ® 2009. © 1993-2008 Microsoft Corporation.

UNIT 3 SAMPLE AND SAMPLING TECHNIQUES

CONTENTS

1.0 Objectives
2.0 Introduction
3.0 Main Content
    3.1 Overview of Sampling and Sampling Techniques
    3.2 Population
    3.3 The Sampling Unit
    3.4 Sampling Frame
    3.5 Sample Design
    3.6 Probability and Non-Probability Sampling
    3.7 Non-Probability Sample Designs
    3.8 Probability Sample Designs
    3.9 Sample Size
4.0 Summary
5.0 Conclusion
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 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 generalised to be of scientific value. Generalisations are important not only for testing hypotheses but also for descriptive purposes. Typically, generalisations 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).

2.0 OBJECTIVES

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

- define population
- explain what is meant by sample size and sample design
- highlight the importance of data collection to aid planning and decision making.

3.0 MAIN CONTENT

3.1 Overview of Sampling and Sampling Techniques

Empirically supported generalisations 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 generalisations 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
(3) the size of the sample.

3.2 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 The 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.

3.4 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.

3.5 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 analysed the entire population.

3.6 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.

3.7 Non-probability Sample Designs

Three major designs utilising non-probability samples have been employed by social scientists: convenience samples, purposive samples, and quota samples.

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.

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.

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 characterised by specific variables such as gender, age, place of residence, and ethnicity.

3.8 Probability Sample Designs

Four common designs of probability samples are simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

1. **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%), and each subsequent outcome is independent of the previous 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.

2. **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.

3. **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 "to 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.

4. **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.

3.9 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.

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.

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.

4.0 SUMMARY

You should by now 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

5.0 CONCLUSION

This unit has relayed to you that a well-chosen sample can usually provide reliable information about the whole of the population to any desired degree of accuracy. In some instances, sampling is an alternative

to a complete census, and may be preferable mainly because of its cheapness and convenience.

6.0 TUTOR-MARKED ASSIGNMENT

1. Explain three non-probability sampling methods.
2. What are the advantages of telephone interview?
3. Is there any disadvantage(s) in personal interview method of data collection?

7.0 REFERENCES/FURTHER READING

Frankfort-Nachmias, C. & Nachmias, D. (2009). *Research in the Social Sciences***.** (5th ed.). Hodder Education.

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

Esan, E. O. & Okafor, R. O. (1995). *Basic Statistical Methods.* (1st ed.). Lagos, Nigeria: JAS Publishers. Pp. 72-89.